Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Computação

CÉSAR OLÍMPIO ANTUNES LIMA

APERFEIÇOAMENTO AUTOMÁTICO DOS CONJUNTOS DE TERMOS UTILIZADOS PELO *YUCCA*: UM COLETOR TEMÁTICO BASEADO EM GÊNERO

Ouro Preto
2018

CÉSAR OLÍMPIO ANTUNES LIMA

APERFEIÇOAMENTO AUTOMÁTICO DOS CONJUNTOS DE TERMOS UTILIZADOS PELO *YUCCA*: UM COLETOR TEMÁTICO BASEADO EM GÊNERO

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

Orientador: Guilherme Tavares de Assis

Ouro Preto
2018

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

FOLHA DE APROVAÇÃO

César Olímpio Antunes Lima

Aperfeiçoamento Automático dos Conjuntos de Termos Utilizados Pelo Yucca: Um Coletor temático Baseado Em Gênero

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 Julho de 2018.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Dayanne Gouveia Coelho (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Amanda Sávio Nascimento e Silva (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 11/07/2018.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/09/2023, às 11:59, 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 **0585394** e o código CRC **FA3454A6**.

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

Resumo

A recente popularização de acesso à Web vem provocando um extraordinário aumento no volume de informações que é produzido e consumido. Nesse contexto, tornam-se fundamentais o desenvolvimento e o aperfeiçoamento de mecanismos que promovam o acesso à informação disponibilizada na Web de maneira fácil, rápida e precisa. Coletores tradicionais não são capazes de identificar sub-espaços relevantes na Web relacionado a um tópico específico de interesse; entretanto, coletores temáticos são ferramentas capazes de resolver, de maneira eficaz e eficiente, o problema mencionado. Geralmente, um processo de coleta temática necessita, como parâmetro de entrada, de um conjunto bem definido de termos que expressam o tópico de interesse desejado; dependendo de tal conjunto de termos, a eficácia de um determinado processo de coleta pode não ser satisfatória. Existem distintas estratégias capazes de promover a expansão automática de termos, que podem, no caso, melhorar a eficácia de coletores. Em tal contexto, esse trabalho possui, como objetivo geral, a proposta, o desenvolvimento e a integração em um coletor temático, denominado *Yucca*, de uma estratégia para o aperfeiçoamento automático de conjuntos de termos necessários para a execução do mesmo. Por meio da análise dos resultados dos experimentos realizados, considerando a estratégia proposta e um *baseline*, foi verificado, de uma forma geral, que a estratégia proposta alcançou um ganho de até 17,24% de F1 (eficácia) em relação ao *baseline* considerado, em um processo de coleta cujo os conjuntos de termos de gênero e conteúdo eram reduzidos.

Palavras-chave: Coleta temática, aperfeiçoamento de termos, expansão de consultas, frequência de termos.

Lista de Figuras

2.1 Arquitetura de funcionamento da abordagem original para coleta temática (Assis et al., 2009) 6
2.2 Arquitetura de funcionamento da geração automática de páginas-semente (Mangaravite et al., 2014) 9
2.3 Arquitetura de funcionamento da estratégia baseada em matriz de associação (Costa et al., 2017) 11
2.4 Arquitetura de funcionamento da estratégia baseada em PLN (Costa et al., 2017) 13
2.5 Um sub-grafo da *Freebase* (Chenyan Xiong, 2015). 16

3.1 Arquitetura de um processo de expansão de termos no *Yucca* 19
3.2 Arquitetura da estratégia baseada na frequência de termos 21

Lista de Tabelas

4.1 Conjuntos de termos de gênero e conteúdo para o tópico "Ementa de disciplinas de Banco de Dados"(Mangaravite et al., 2014). 26
4.2 Precisão e revocação obtidos como *baseline*. 26
4.3 Conjuntos reduzidos de termos de gênero e conteúdo para o tópico "Ementa de disciplinas de Banco de Dados". 27
4.4 Conjuntos de termos de gênero e conteúdo definidos por um não especialista para o tópico "Ementa de disciplinas de Banco de Dados". 28
4.5 Precisão e revocação obtidos como referência para o processo de coleta envolvendo os conjuntos reduzidos de termos. 28
4.6 Precisão e revocação obtidos como referência para o processo de coleta envolvendo os conjuntos de termos definidos por um não especialista. 28
4.7 Melhores resultados relativos à estratégia considerando os conjuntos completos de termos 32
4.8 Melhores resultados relativos à estratégia considerando os conjuntos reduzidos de termos 33
4.9 Melhores resultados relativos à estratégia considerando os conjuntos de termos definidos por um não especialista 35
4.10 Resultados relativos a processos de coleta sem a intervenção de um processo de expansão de termos 36

A.1 Resultados relativos à estratégia considerando os conjuntos completos de termos 42
A.2 Resultados relativos à estratégia considerando os conjuntos reduzidos de termos . 45
A.3 Resultados relativos à estratégia considerando os conjuntos de termos definidos por um não especialista 48

Sumário

1 Introdução **1**

1.1 Motivação 2

1.2 Objetivos Geral e Objetivos Específicos 3

1.3 Método de trabalho 3

1.4 Delineamento da Monografia 4

2 Revisão de Literatura **5**

2.1 Fundamentação Teórica 5

2.1.1 Abordagem original para coleta temática baseada em gênero 5

2.1.1.1 Uso de *Link Context* 8

2.1.1.2 Geração semi-automática de páginas-semente 9

2.1.1.3 Determinação automática de limites de similaridade 10

2.1.1.4 Aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo 11

2.1.2 Expansão de consulta 13

2.2 Trabalhos Relacionados 15

3 Desenvolvimento **18**

3.1 Preparação do *Yucca* para a utilização de um processo de expansão de termos 18

3.2 Estratégia baseada em frequência de termos 21

4 Experimentação prática **23**

4.1 Métricas de avaliação 23

4.2 *Baseline* 24

4.3 Descrição dos experimentos 25

4.4 Resultados 29

4.4.1 Resultados considerando os conjuntos completos de termos 30

4.4.2 Resultados considerando os conjuntos reduzidos de termos 31

4.4.3 Resultados considerando os conjuntos de termos definidos por um não especialista 31

4.4.4 Resultados relativos a processos de coleta sem a intervenção de um processo de expansão de termos 34

5 Conclusão **37**

Referências Bibliográficas **39**

A Todos os resultados dos experimentos relativos à estratégia baseada em frequência de termos **41**

Capítulo 1

Introdução

A popularização do acesso à *Web*, de acordo com Costa et al. (2017), vem provocando um extraordinário aumento no volume de informações que é produzido e consumido. Nesse contexto, tornam-se fundamentais o desenvolvimento e o aperfeiçoamento de mecanismos que promovam o acesso à informação disponibilizada na *Web* de maneira fácil, rápida e precisa. Este acesso é feito, atual e basicamente, por meio de máquinas de busca que exploram a estrutura em grafo da *Web* no intuito de localizar páginas relevantes a uma específica consulta (Menczer et al., 2004). Para permitir isto, uma máquina de busca típica trabalha sobre uma coleção de páginas indexadas, gerada a partir da coleta do maior número possível de páginas da *Web*. Nesse caso, um coletor tradicional coleta páginas da *Web*, começando por uma página-semente e seguindo as ligações contidas nela, visitando assim, outras páginas. Tal processo repete-se com as novas páginas, a partir das quais novas ligações são seguidas, até percorrer um número suficiente de páginas ou alcançar algum objetivo específico.

Entretanto, máquinas de busca de propósito geral, segundo Costa et al. (2017), 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 envolvem pouca informação. Neste contexto, coletores temáticos (Chakrabarti et al., 1999; Menczer et al., 2004; Pant e Srinivasan, 2005; Srinivasan et al., 2005) 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 (Almpanidis et al., 2007; Chakrabarti et al., 1999; Johnson et al., 2003; Pant e Srinivasan, 2005, 2006) 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 níveis de revocação[1] e precisão entre 40% e 70%.

[1]Dentro do contexto, de acordo com Manning et al. (2008), revocação é a proporção de páginas relevantes

Em Costa et al. (2017), foram propostas e validadas estratégias para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo necessários para a execução de processos de coleta temática que seguem a abordagem baseada em gênero. De uma forma geral, segundo os experimentos realizados, a melhor estratégia proposta foi uma estratégia baseada em matriz de associação de termos, que proporcionou um aumento de 6,79% na métrica F1[2]; um aumento não satisfatório.

Este capítulo encontra-se organizado como se segue. A Seção 1.1 apresenta a motivação para a realização desse trabalho. A Seção 1.2 descreve seus objetivos geral e específicos. A Seção 1.3 aborda o método utilizado no desenvolvimento desse trabalho. Por fim, a Seção 1.4 apresenta o delineamento do restante da monografia.

1.1 Motivação

Geralmente, para que um processo de coleta temática ocorra, é necessário especificar de forma detalhada o tópico de interesse desejado. De uma forma geral, tal especificação ocorre por meio da definição de um conjunto de termos que represente o tópico de interesse. A definição desse conjunto de termos influencia diretamente na eficácia de um processo de coleta, uma vez que o coletor temático determina a relevância de uma determinada página da *Web* por meio do cálculo da similaridade entre o conteúdo dessa página e do conjunto de termos utilizado para representar o tópico de interesse. Dessa forma, dependendo do conjunto de termos utilizado, a eficácia de um determinado processo de coleta pode não ser satisfatória e, no caso, um aperfeiçoamento automático desse conjunto de termos pode levar a melhoria de tal eficácia.

Particularmente, Assis et al. (2007, 2008, 2009) propuseram uma abordagem para coleta temática de páginas da Web (vide Subseção 2.1) que se baseia na definição de dois conjuntos de termos: um voltado para termos relativos ao gênero[3] do tópico de interesse desejado e outro relativo ao conteúdo de tal tópico. Em Costa et al. (2017), foram propostas e validadas estratégias para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo necessários para a execução de processos de coleta temática que seguem tal

ao tópico de interesse desejado, dentre um gabarito de páginas relevantes, retornadas por um processo de coleta; precisão é a proporção da quantidade de páginas retornadas por um processo de coleta, que são relevantes ao tópico de interesse desejado.

[2]De acordo com Manning et al. (2008), F1 é a média harmônica entre precisão e revocação.

[3]Por gênero, de acordo com Assis et al. (2009), entende-se o tipo, a categoria ou o estilo de texto de documentos específicos

abordagem baseada em gênero. De uma forma geral, segundo os experimentos realizados, a melhor estratégia proposta foi uma estratégia baseada em matriz de associação de termos que proporcionou um aumento de 6,79% na métrica F1: um aumento não satisfatório. Já que, como já mencionado, a eficácia de um processo de coleta temática depende, inicialmente, de uma boa especificação dos conjuntos de termos de entrada, torna-se importante propor novas estratégias de aperfeiçoamento automático dos termos de coleta.

1.2 Objetivos Geral e Objetivos Específicos

Este trabalho de monografia possui, como objetivo geral, a proposta e o desenvolvimento de uma nova estratégia de expansão dos conjuntos de termos de entrada para a abordagem de coleta temática baseada em gênero, proposta por Assis et al. (2007, 2008, 2009) que supere os resultados obtidos por Costa et al. (2017).

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

- geração da segunda versão do Yucca, coletor temático que segue a abordagem para coleta temática baseada em gênero (Assis et al., 2009, 2008, 2007) desenvolvido por Diniz e Assis (2018), envolvendo a estratégia validada para a expansão automática dos conjuntos de termos de gênero e conteúdo;
- preparação do *Yucca* para a utilização de estratégias de expansão automática dos conjuntos de termos de gênero e conteúdo;
- desenvolvimento e integração ao *Yucca* de uma nova estratégia para a expansão automática dos conjuntos de termos de gênero e conteúdo, baseada na frequência dos termos;
- análise de resultados obtidos por meio de experimentos de validação da estratégia desenvolvida e integrada ao *Yucca* por meio da execução de processos de coleta relativos a distintos tópicos de interesse.

1.3 Método de trabalho

Visando o alcance do objetivo geral deste trabalho, foi integrado ao *Yucca* uma nova estratégia para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo que são fornecidos, como parâmetros de entrada, para processos de coleta temática de páginas da *Web* que seguem a abordagem baseada em gênero (Assis et al., 2009, 2008, 2007); ademais, no intuito de validar o *Yucca* quanto ao seu funcionamento, considerando

a estratégia para o aperfeiçoamento dos conjuntos e termos a ser integrada ao mesmo, experimentos práticos serão realizados envolvendo a coleta de páginas da *Web* referentes a distintos tópicos de interesse, juntamente com a medição da precisão dos resultados obtidos.

1.4 Delineamento da Monografia

O restante desta monografia encontra-se organizado como se segue. O Capítulo 2 apresenta a revisão de literatura para a realização deste trabalho, envolvendo a fundamentação teórica e trabalhos diretamente relacionados. O Capítulo 3 descreve, de maneira detalhada, uma nova estratégia proposta para o aperfeiçoamento automático dos termos a serem utilizados pelo *Yucca* em processos de coleta temática baseada em gênero. O Capítulo 4 descreve os experimentos de validação da estratégia proposta, por meio de processos de coleta realizados via *Yucca*, e apresenta os resultados obtidos. Por fim, o Capítulo 5 apresenta a conclusão deste trabalho e as perspectivas de trabalho futuro.

Capítulo 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 necessária ao bom desenvolvimento deste trabalho e a Seção 2.2 apresenta os trabalhos diretamente relacionados.

2.1 Fundamentação Teórica

Nesta seção, é apresentado o suporte teórico necessário para o entendimento e desenvolvimento deste trabalho. A Subseção 2.1.1 apresenta a abordagem elaborada por De Assis et al. (2009) para a realização de processos de coleta temática baseada em gênero, bem como as melhorias propostas e já desenvolvidas e validadas sobre a abordagem original. Por fim, a Subseção 2.1.2 refere-se a técnicas de expansão de consulta e suas formas de aplicação.

2.1.1 Abordagem original para coleta temática baseada em gênero

A abordagem para coleta temática de páginas da *Web* baseada em gênero, 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 baseada em gênero, 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", gerando erros de revocação.

A Figura 2.1 sumariza os principais passos da abordagem proposta em Assis et al. (2009, 2008, 2007). Percebe-se que o coletor temático, construído de acordo com a abordagem baseada em gênero, 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 e os pesos dos termos de gênero e conteúdo, correspondem a parâmetros de entrada para funcionamento da arquitetura.

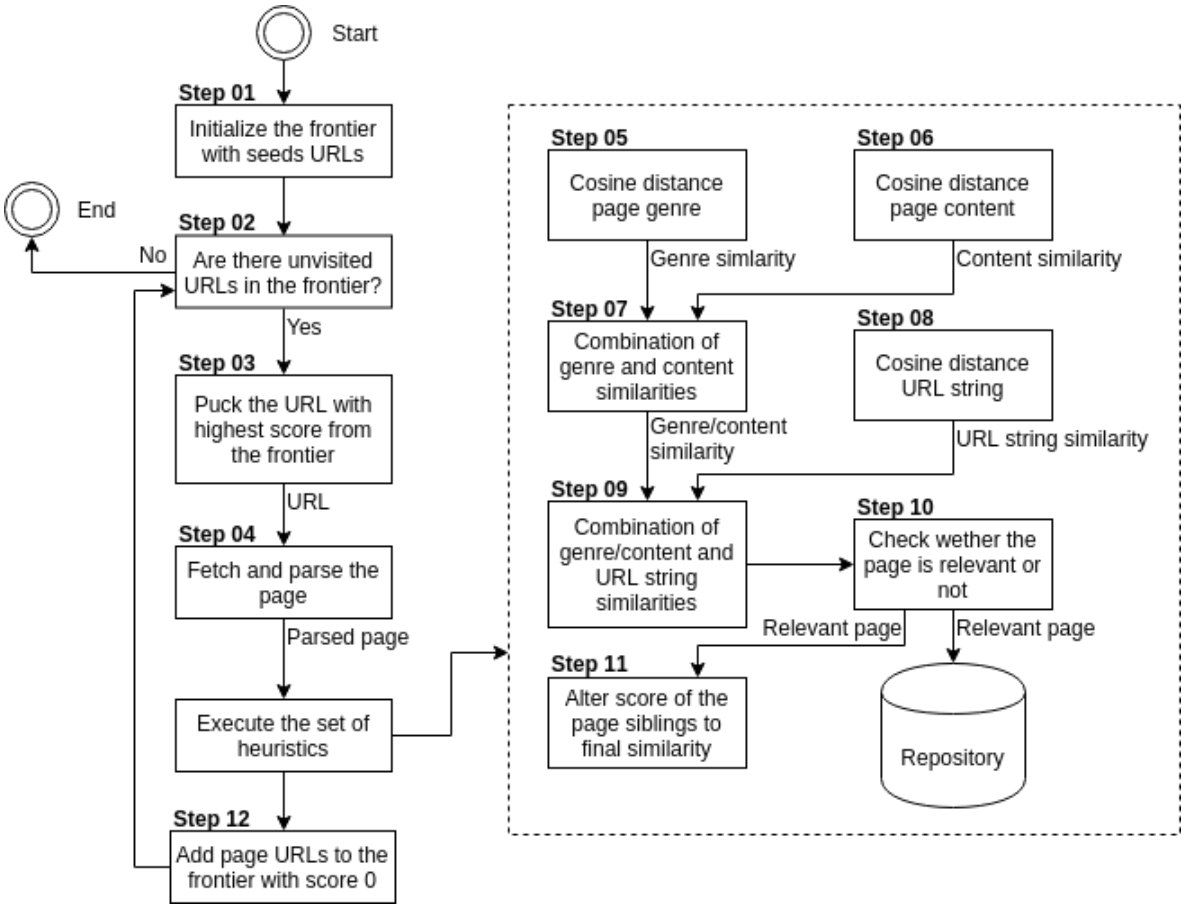


Figura 2.1: Arquitetura de funcionamento da abordagem original para coleta temática (Assis et al., 2009)

De uma forma geral, a arquitetura de funcionamento, ilustrada pela Figura 2.1, 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), 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 (modelo vetorial), 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;
- Passo 08: calcular a similaridade, por meio da distância de cosseno (modelo vetorial), 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;
- Passo11: 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; 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 descrita na Figura 2.1, atingiram níveis de F1 superiores a 88% para todos os tópicos de interesse

considerados.

Nas subseções seguintes, estão descritos trabalhos (Mangaravite et al., 2012, 2014; Siqueira et al., 2016; Costa et al., 2017) que realizaram melhorias referentes à eficiência e à eficácia da abordagem para coleta temática baseada em gênero. A Subseção 2.1.1.1 apresenta a melhoria realizada por Mangaravite et al. (2012), que se baseou na verificação e utilização das características estruturais de *links* (nome da URL, texto de âncora, título do *link*), presentes em uma página visitada pelo coletor. A Subseção 2.1.1.2 aborda o trabalho desenvolvido por Mangaravite et al. (2014), onde o aperfeiçoamento ocorreu por meio da determinação automática das páginas-semente a serem utilizadas em um processo de coleta temática. A Subseção 2.1.1.3 apresenta o trabalho desenvolvido por Siqueira et al. (2016), onde foram propostas e desenvolvidas estratégias para determinação automática do limite de similaridade a ser considerado em processos de coleta temática. Por fim, a Subseção 2.1.1.4 apresenta o trabalho desenvolvido por Costa et al. (2017), onde foram propostas e desenvolvidas estratégias para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo fornecidos, como dados de entrada, para processos de coleta temática.

Assim como tais trabalhos citados, o objetivo geral deste trabalho também está relacionado ao aperfeiçoamento da abordagem para coleta temática proposta por De Assis et al. (2009). No caso, este trabalho preocupa-se com uma melhor especificação dos conjuntos de termos de gênero e conteúdo a serem usados em um processo de coleta temática, além de integrar estratégias automáticas para tal fim ao *Yucca*, gerando uma nova versão do mesmo.

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 em relação ao número de páginas visitadas pelo coletor. Assim, um coletor temático é considerado eficiente se ele localiza uma quantidade significante de páginas relevantes ao visitar uma pequena quantidade de páginas da *Web* em um processo de coleta. Visando a melhoria de eficiência da abordagem de coleta temática baseada em gênero, apresentada na Figura 2.1, sem perda na escalabilidade e na eficácia da mesma, foi proposto em Mangaravite et al. (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 cosseno entre os termos de gênero e de conteúdo, parâmetros de entrada da abordagem apresentada na Figura 2.1, e os textos gerados pela utilização do *Link Context*. Com isso, o passo 12 da arquitetura da abordagem (vide Figura 2.1) é alterado, uma vez que as

pontuações de visita das URLs não visitadas não são mais inicializadas com valor 0 (zero), mas sim com a relevância calculada considerando o *Link Context*.

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

2.1.1.2 Geração semi-automática de páginas-semente

Com o intuito de também melhorar a eficiência da abordagem de coleta temática baseada em gênero, Mangaravite et al. (2014) propuseram uma estratégia para geração semi-automática 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.

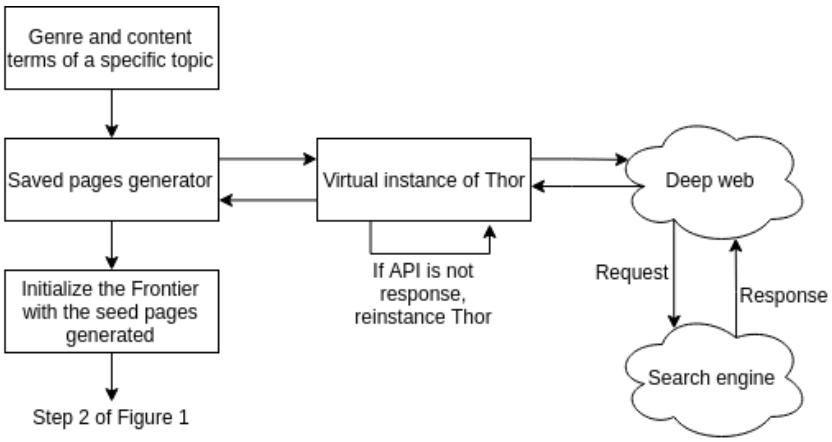


Figura 2.2: Arquitetura de funcionamento da geração automática de páginas-semente (Mangaravite et al., 2014)

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 para coleta temática, apresentada na Figura 2.1, de até 53%.

Uma vez definidas as páginas-semente para um processo de coleta temática, de acordo com a Figura 2.2, as mesmas servem 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 similaridade

A abordagem para coleta temática de páginas *Web* apresentada na Figura 2.1, como já mencionado, utiliza a distância de cosseno para determinar a similaridade entre uma página da *Web* e o conjunto 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, em Siqueira et al. (2016), foram desenvolvidas três estratégias para determinação automática do limite de similaridade utilizado em processos de coleta temática de paginas da *Web* baseada em gênero.

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-semente 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. Com o intuito de validar tais estratégias, foram medidos, para cada coleta, os níveis de precisão, revocação e F1, permitindo observar que a estratégia baseada no método de agrupamento *K-means* foi a que apresentou melhor resultado, chegando a alcançar níveis de precisão e revocação bem próximos daqueles obtidos quando os limites de similaridade foram definidos empiricamente por especialistas.

2.1.1.4 Aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo

Para que um processo de coleta temática baseada em gênero 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 para coleta temática apresentada na Figura 2.1, foram propostas por Costa et al. (2017) duas estratégias para aperfeiçoamento de tais conjuntos.

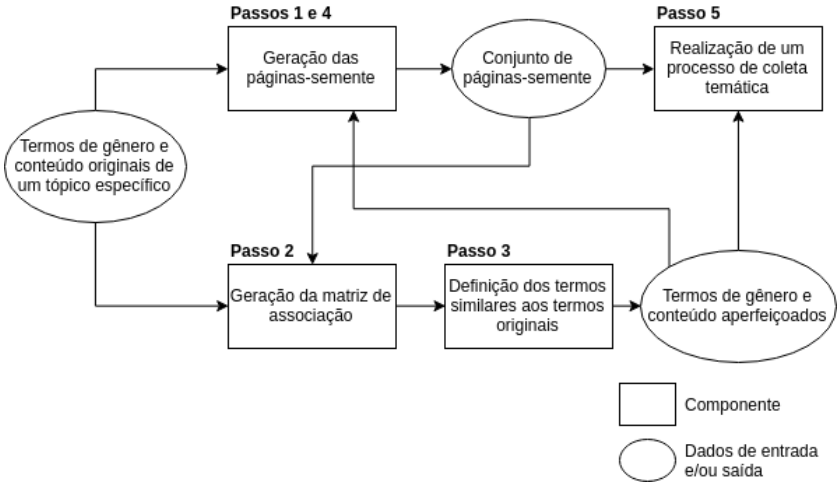


Figura 2.3: Arquitetura de funcionamento da estratégia baseada em matriz de associação (Costa et al., 2017)

Conforme mostrado pela arquitetura da Figura 2.3, a primeira estratégia visa aperfeiçoar os conjuntos de termos de gênero e conteúdo por meio da aplicação de uma técnica de expan-

são de consulta automática baseada no uso de matriz de associação[1]. Inicialmente, no passo 1, a partir dos conjuntos originais de termos de gênero e conteúdo, são geradas automaticamente páginas-semente (vide Subseção 2.1.2) 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 aperfeiçoados de termos de gênero e conteúdo. Esses conjuntos aperfeiçoados juntamente com as páginas sementes, obtidas por meio do passo 4 utilizando os próprios conjuntos aperfeiçoados, são utilizados como dados de entrada para a realização do processo de coleta desejado, conforme apresentado no passo 5.

Para se definir termos similares aos termos dos conjuntos originais (passo 3 da Figura 2.3) foi proposta e implementada a métrica denominada MenorDistância, que 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 os mesmos. A equação da métrica MenorDistância é:

$$s_{i,j} = \frac{\sum_{n}^{p=1} d_{i,j}}{n}$$

$$\begin{cases} d_{i,j} = Distancia(t_i, t_j), & se \quad Distancia(t_i, t_j) \leq k, \\ d_{i,j} = k + 1, & se \quad Distancia(t_i, t_j) > k \end{cases}$$

onde n é a quantidade de páginas que possuem os termos t_i e t_j ; k é a distância máxima pré-estabelecida entre os termos t_i e t_j; $Distancia(t_i, t_j)$ é uma função que retorna a distância,em quantidade de termos, entre os termos t_i e t_j presentes em uma mesma página p; e $d_{i,j}$ é a menor distância entre os termos t_i e t_j presentes em uma mesma página p, calculada de forma que atinja valor máximo igual a $k + 1$.

A segunda estratégia, apresentada na Figura 2.4, visa aperfeiçoar 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 (vide Subseção 2.1.2) 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 aperfeiçoados 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 em Costa et al. (2017), foi possível perceber que a estratégia baseada em matriz de associação de termos, utilizando a métrica MenorDistância, foi a que apresentou melhores resultados quando comparada às demais estratégias propostas por Costa et al. (2017). Contudo, apesar de tal estratégia

[1]Uma matriz de associação, de acordo com (Chartree et al., 2013), é uma estrutura de dados comumente utilizada para mensurar a relação existente entre os termos dos documentos de uma coleção.

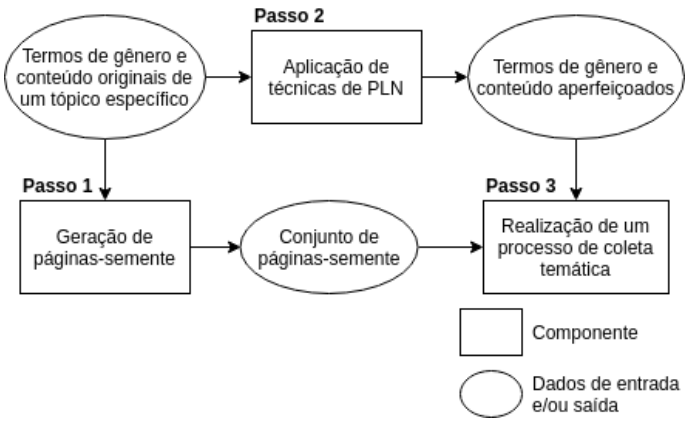


Figura 2.4: Arquitetura de funcionamento da estratégia baseada em PLN (Costa et al., 2017)

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 Expansão de consulta

Sistemas de Recuperação de Informação (RI) recuperam, baseados em consultas geralmente formuladas por usuários, documentos relevantes presentes em uma coleção. Para que um determinado sistema de RI consiga recuperar tais documentos relevantes, o usuário precisa descrever bem sua necessidade de informação em uma consulta, que consiste de um conjunto de termos que resumem a informação desejada. Nesse contexto, de acordo com Aly (2008), a área de expansão de consulta engloba um conjunto de técnicas que fazem com que a consulta original, inicialmente elaborada por um usuário, seja suplementada com termos adicionais no intuito de melhorar, em qualidade, o conjunto de documentos recuperados por um sistema de RI. Dessa forma, em sistemas de RI, segundo Carpineto e Romano (2012), as técnicas de expansão de consulta aplicam-se na fase de formulação da consulta, uma vez que, frequentemente, as consultas elaboradas por usuários são pequenas e podem utilizar uma terminologia diferente da usada para indexar os documentos da coleção.

As técnicas de expansão de consulta podem ser aplicadas de forma manual, automática ou interativa. A expansão manual, segundo Carpineto e Romano (2012), ocorre quando o usuário reformula a consulta adicionando ou removendo termos sem a ajuda do sistema de RI. A seleção desses termos é feita de forma intuitiva, baseado na experiência do usuário ao avaliar resultados de consultas feitas anteriormente, por meio da ajuda de especialistas ou pela consulta a dicionários e *thesaurus*.

A expansão de consulta automática e a interativa são fundamentadas pelo conceito de *feedback* de relevância. Conforme descrito em Manning et al. (2008), a ideia do *feedback* de relevância é envolver o usuário no processo de busca afim de melhor atender a sua necessidade de informação. Para isso, o usuário deve submeter sua consulta inicial ao sistema de RI e classificar os documentos retornados de acordo com a relevância de seu conteúdo. Após a classificação, de acordo com Mitra et al. (1998), os termos que serão utilizados para a expansão são selecionados por meio de uma manipulação automática do conteúdo dos documentos julgados como relevantes.

Na expansão de consulta automática, o usuário provê *feedback* de um conjunto inicial de documentos indicando a relevância com apenas um "sim" ou "não". Já em relação à expansão de consulta iterativa, de acordo com (Greenberg, 2001; Carpineto e Romano, 2012), é dada liberdade ao usuário durante a avaliação do conjunto inicial de resultados: ao invés de apenas classificar um documento como "relevante" ou "não relevante", o usuário pode marcar partes importantes dos documentos, selecionar tópicos de interesse, e aceitar ou recusar sugestões de termos feitas por sistemas de RI. Um dos problemas dessa abordagem é que, além de exigir uma certa habilidade do usuário, o sistema de RI em questão deve oferecer uma interface que permita uma boa interação como o usuário.

Tanto na expansão de consulta automática, quanto na iterativa, é feita uma análise estatística com o intuito de identificar relações semânticas entre os termos da consulta e os demais termos dos documentos marcados como relevantes. Essa análise, segundo Greenberg (2001), mensura a associação entre os termos, ou seja, atribui um valor numérico que representa quão correlacionados os termos são, tornando possível a identificação e extração dos termos que serão utilizados para a expansão.

Nas estratégias propostas neste trabalho para expansão de consulta automática no que diz respeito ao aperfeiçoamento de conjuntos de termos, tal como as estratégias propostas por Costa et al. (2017), apresentadas na Subseção 2.1.1.4, foi considerada a ideia de *pseudo-feedback* de relevância que, de acordo com (Manning et al., 2008), consiste em considerar que os primeiros documentos recuperados por uma máquina de busca são realmente relevantes, podendo, então, ser utilizados para a extração dos termos de expansão.

2.2 Trabalhos Relacionados

Existem diversos trabalhos na literatura que se preocupam em aprimorar o conjunto de termos submetidos por usuários a sistemas de RI, incluindo máquinas de busca. Esse aprimoramento dos termos é motivado pelo fato de que os usuários, na maioria das vezes, fornecem poucos termos para a formulação de suas consultas, não descrevendo apropriadamente o tópico de interesse. Segundo Costa et al. (2017), uma solução bastante utilizada é a aplicação de técnicas relativas à expansão de consulta, que buscam reformular consulta do usuário por meio da adição de um conjunto de novos termos relacionados àqueles que foram originalmente estabelecidos.

Nesse contexto, Chirita et al. (2007) definem estratégias baseadas em frequência de termos como estratégias baratas computacionalmente e capazes de obter resultados satisfatórios no processo de determinação de termos mais relevantes em um determinado documento de uma coleção. No cálculo de relevância dos termos, Chirita et al. (2007) propuseram multiplicar a frequência de cada termo com a posição da primeira ocorrência do mesmo no documento em questão, no intuito de premiar os termos que estão mais próximos ao início do documento. Isso pode ser interessante, especialmente em documentos maiores, pois termos mais importantes tendem a aparecer em seu início. Uma vez encontrando os termos mais relevantes de documentos de uma coleção, os mesmos podem ser usados para aprimorar o conjunto de termos originais de uma determinada consulta. Chirita et al. (2007) obtiveram resultados satisfatórios nas estratégias propostas no trabalho, alcançando um ganho NDCG[2] de até 51,68% em seus experimentos.

Em Chenyan Xiong (2015), foi proposta uma estratégia para expansão de termos de consulta utilizando a Base de Conhecimento colaborativa *Freebase*[3]. A informação armazenada em uma base de conhecimento normalmente é organizada por entidades e relações, onde as entidades correspondem a entidades do mundo-real e conceitos (por exemplo, Ciência da Computação). Cada entidade, por sua vez, possui uma descrição breve, atributos e relações com outras entidades. A informação na *Freebase* é organizada como um grafo, como ilustrado na Figura 2.5. Uma entidade (objeto) é representado como um nodo no grafo com um identificador único. Uma aresta do grafo relaciona uma entidade a outra ou a um atributo. Existem muitos tipos de arestas na *Freebase*, representando diferentes fatos. Na Figura 2.5, por exemplo, *The Brothers Grimm* está conectada ao objeto *Terry Gilliam* por uma aresta do tipo *Directed_by*, mostrando que o filme *The Brothers Grimm* foi dirigido pelo diretor *Terry Gilliam*.

[2] *Normalized Discounted Cumulative Gain* (NDCG) (Kekäläinen, 2000) é uma métrica de qualidade de *ranking*, muito usada para medir a efetividade de estratégias de busca e coleta na *Web*.

[3] `http://www.freebase.com/`

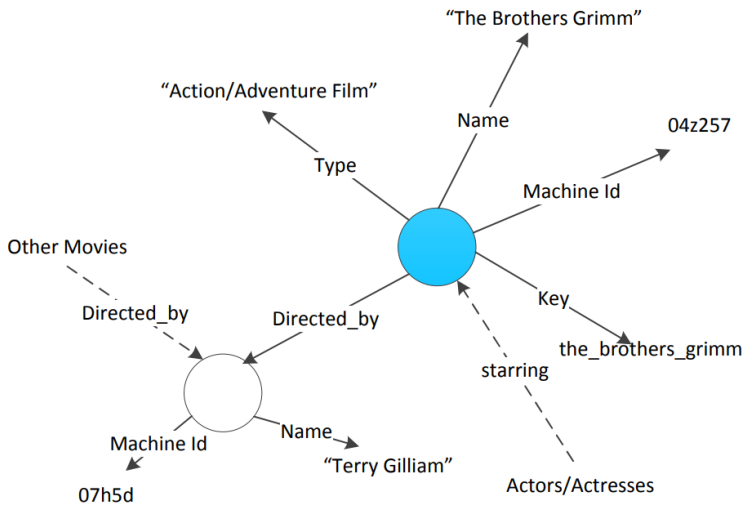


Figura 2.5: Um sub-grafo da *Freebase* (Chenyan Xiong, 2015).

Basicamente, a proposta por Chenyan Xiong (2015) consiste em dois passos: relação de objetos e seleção de termos. Na relação de objetos, a estratégia aplica os termos iniciais, relativos a uma determinada consulta, na máquina de busca da Google, no intuito de recuperar objetos relevantes no contexto dos termos de busca retornados, a partir dos documentos fornecidos pela máquina de busca. Na seleção de termos, a estratégia procura na *Freebase* pelos objetos obtidos no passo anterior, no intuito de recuperar termos relevantes na descrição textual de cada um deles. Esses termos são usados para a expansão dos termos iniciais. A estratégia proposta por Chenyan Xiong (2015) apresentou métodos quase 30% mais eficientes que estratégias clássicas de expansão de termos, segundo seus experimentos.

Em Saar Kuzi (2016), foi proposta uma estratégia de expansão de termos de busca baseada em *word embeddings*[4] usando a abordagem Word2vec[5] *Continuous Bag-of-Words* (CBOW)[6] para representar termos em um espaço vetorial baseado na frequência dos mesmos. Similaridades entre os vetores dos termos representam similaridades semânticas. A estratégia proposta por Saar Kuzi (2016) usa o modelo Word2vec CBOW para treinar uma máquina localmente, no intuito de que ela retorne termos relacionados semanticamente, dado um conjunto de termos. Saar Kuzi (2016) mostrou um aumento da eficiência das estratégias de

[4] *Neural embeddings* ou *word embeddings* (Levy e Goldberg, 2014) é uma proposta de representação de palavras através de vetores, por meio de métodos de treinamento inspirados em modelos de redes neurais aplicados a linguagem natural.

[5] Word2vec (Mikolov et al., 2013) é um grupo de modelos relacionados, usados para produzir *word embeddings*.

[6] *Continuous Bag-of-Words* (Mikolov et al., 2013) é um modelo de representação simplificada, usada em PLN e RI, onde um texto (como uma frase ou um documento) é representado como uma lista das palavras contidas no mesmo.

expansão de consulta abordadas quando se utiliza máquinas previamente treinadas, além de demonstrar a efetividade de sua estratégia integrando seus métodos de expansão de consulta com uma abordagem de expansão de termos de busca baseada em *pseudo-feedback* de relevância (Manning et al., 2008).

Assim como os trabalhos citados, como já mencionado, existem outras estratégias para melhoria dos termos utilizados como dados de entrada em um sistema de RI. É importante notar que nenhuma estratégia abordada nesta seção foi desenvolvida no intuito de ser aplicada em processos de coleta temática. Neste trabalho de monografia, essas abordagens serviram como base para a proposta de uma nova estratégia para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo a serem utilizados na abordagem para coleta temática baseada em gênero (vide Subseção 2.1.1).

Capítulo 3

Desenvolvimento

Este capítulo apresenta, inicialmente, uma arquitetura de preparação do *Yucca* para a utilização de estratégias quaisquer para expansão automática de conjuntos de termos de gênero e conteúdo (Seção 3.1); em seguida, descreve a nova estratégia proposta para a expansão propriamente dita de tais conjuntos de termos, baseada na frequência de termos (Seção 3.2). Para melhor entendimento das seções seguintes, torna-se necessário definir dois conceitos:

- processo de expansão de termos: conjunto de passos a serem seguidos pelo *Yucca* para que os termos expandidos sejam obtidos;
- estratégia de expansão de termos: estratégia utilizada em um processo de expansão de termos, correspondendo a um dos passos de tal processo, responsável pela determinação propriamente dita dos termos expandidos.

3.1 Preparação do *Yucca* para a utilização de um processo de expansão de termos

Esta seção apresenta uma arquitetura de preparação do *Yucca* para que seja possível a utilização de estratégias quaisquer de expansão de termos por tal coletor. Para tanto, foram feitas alterações na arquitetura de funcionamento atual do *Yucca*, de modo a prepará-lo para a utilização de um processo de expansão de termos que seja independente da estratégia escolhida, na intenção de manter a arquitetura do *Yucca* organizada, além de facilitar a utilização de outras estratégias de expansão de termos que possam ser integradas ao *Yucca* futuramente. Basicamente, alterações foram feitas na arquitetura de funcionamento atual do *Yucca* em dois momentos, a saber:

- na inicialização do *Yucca*, onde o coletor utiliza uma estratégia para a expansão dos termos de gênero e conteúdo fornecidos pelo usuário como parâmetros de entrada, e

salva os termos expandidos fornecidos pela estratégia;

- no cálculo de similaridade das páginas durante um processo de coleta, onde o coletor agora considera os termos de gênero e conteúdo expandidos, além dos termos originais especificados.

A Figura 3.1 apresenta a arquitetura definida para um processo de expansão de termos, a ser aplicada apresenta a arquitetura definida para um processo de expansão de termos, a ser aplicada na inicialização do *Yucca*. A ideia, de uma forma geral, é que o *Yucca* estabeleça um conjunto de páginas a ser fornecida para uma estratégia de expansão de termos que, por sua vez, determine e retorne ao *Yucca* termos relacionados às páginas fornecidas. Tal ideia é aplicada, separadamente, ao conjunto de termos de gênero e ao conjunto de termos de conteúdo.

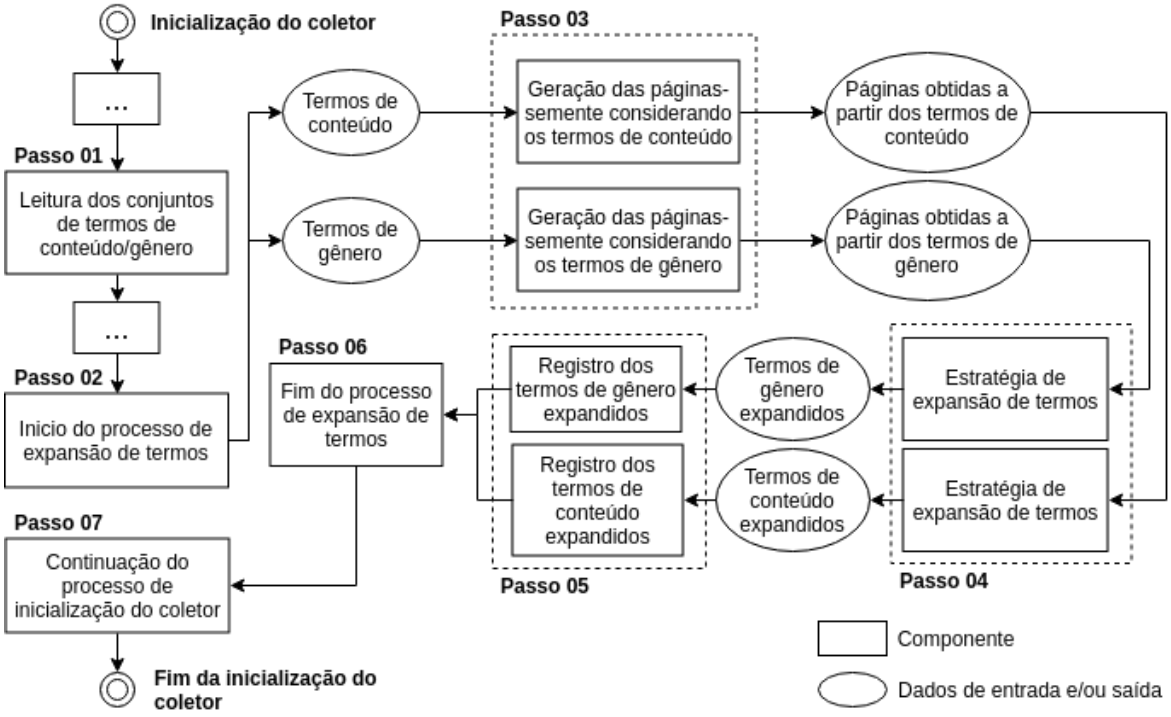


Figura 3.1: Arquitetura de um processo de expansão de termos no *Yucca*

Como é possível observar na Figura 3.1, um determinado processo de expansão de termos na inicialização do *Yucca* segue os seguintes passos:

- Passo 01: de acordo com a arquitetura original do *Yucca* (vide Seção 2.1.1), é feita a leitura dos conjuntos de termos originais de gênero e conteúdo, fornecidos como parâmetros de entrada pelo usuário.
- Passo 02: O processo de expansão de termos é iniciado.

- Passo 03: Para cada conjunto de termos de entrada (gênero ou conteúdo), é gerado um conjunto de páginas, a ser utilizada pela estratégia de expansão de termos.
- Passo 04: a partir do conjunto de páginas gerado no Passo 03, relativas ao gênero ou ao conteúdo do tópico de interesse, uma estratégia para a expansão dos termos é executada.
- Passo 05: a partir do conjunto de termos expandidos gerado no Passo 04, relativos ao gênero ou ao conteúdo do tópico de interesse, o *Yucca* salva tais termos expandidos, para que sejam usados posteriormente pelo cálculo de similaridade das páginas durante um processo de coleta.
- Passo 06: O processo de expansão de termos é finalizado.
- Passo 07: A inicialização do *Yucca* continua seguindo a sua abordagem original (vide Seção 2.1.1).

Em relação ao Passo 03, para a geração do conjunto de páginas para cada conjunto de termos (gênero e conteúdo), foi utilizado o mesmo processo já utilizado pelo coletor para obtenção das páginas-semente, usados no processo de coleta. A diferença, é que ao invés de usar a heurística utilizada no processo de geração das páginas-semente (*unionFirst*), foi utilizada a heurística *justContentOR* para a geração das páginas referentes aos termos de conteúdo, e a heurística *justGenreOR* para a geração das páginas referentes aos termos de gênero. Tais heurísticas são as mesmas propostas por Mangaravite et al. (2014) (vide Subseção 2.2).

Em relação ao Passo 04, a estratégia de expansão de termos a ser utilizada pelo processo de expansão de termos corresponde, neste trabalho, à nova estratégia proposta, que se baseia na frequência de termos (vide Seção 3.2). A expansão de termos, como já mencionado, é feita separadamente para os termos de gênero e conteúdo.

Em relação ao cálculo de similaridade das páginas durante um processo de coleta, com o intuito de calcular o grau de similaridade de uma página em relação a conjuntos de termos (vide Subseção 2.1.1), ao invés de considerar apenas os termos de gênero originais fornecidos pelo usuário, o *Yucca* irá considerar, como termos de gênero, a união dos termos originais com o conjunto de termos de gênero expandidos por uma estratégia de expansão de termos; o mesmo vale para os termos de conteúdo.

3.2 Estratégia baseada em frequência de termos

A estratégia baseada em frequência de termos, como a própria definição diz, visa expandir os conjuntos de termos, sejam de gênero ou de conteúdo separadamente, a serem utilizados em um processo de coleta temática, por meio da aplicação de uma técnica de contagem dos termos. Tal técnica deve extrair os termos de um conjunto de páginas, indexá-los e atribuir um valor de relevância a cada um deles. Este conjunto de páginas é fornecido pelo processo de expansão de termos (vide Figura 3.1). Uma vez estabelecida a relevância de cada termo, tais relevâncias são usadas para escolher os termos a serem adicionados ao conjunto de termos originais (como mostra a Figura 3.2), gerando o conjunto de termos expandidos.

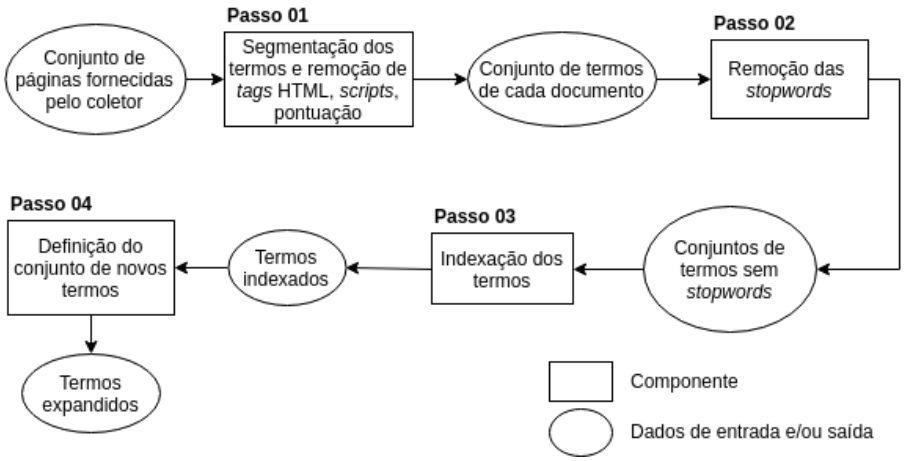


Figura 3.2: Arquitetura da estratégia baseada na frequência de termos

Como é possível observar na Figura 3.2, o funcionamento dessa estratégia segue os seguintes passos para um determinado conjunto de páginas fornecidas pelo *Yucca* (Passo 04 da Figura 3.1):

- Passo 01: para cada página do conjunto de páginas fornecidas pelo *Yucca*, são removidas as cadeias de caracteres indesejadas (*scripts*, *tags* HTML em geral, etc.), sobrando apenas palavras desejadas (termos) entre espaços. Também é feita a segmentação dos termos de cada página, transformando cada página em um conjunto de termos.
- Passo 02: a partir do conjunto de termos gerado no Passo 01, *stopwords* são removidas dos mesmos para não serem consideradas na indexação.
- Passo 03: a partir do conjunto de termos sem *stopwords* fornecido pelo Passo 02, é feita a indexação dos mesmos; nesse processo, a cada termo é associado seu grau de relevância dentro do conjunto, estabelecido por meio de uma função de cálculo de relevância.

- Passo 04: a partir dos termos indexados pelo Passo 03, é selecionado um subconjunto de termos, de acordo com seus graus de relevância, a ser adicionado no conjunto original de termos, formando o conjunto expandido de termos, que é retornado ao *Yucca*, finalizando a execução da estratégia.

Particulamente, em relação ao Passo 03 da Figura 3.2, a indexação é feita para cada termo individual do conjunto de termos sem stopwords e para cada sequência formada por dois ou três termos vizinhos (multi-termos). A estrutura de dados gerada é um dicionário onde cada chave corresponde a um termo (ou multi-termo) e cada valor associado corresponde ao seu grau de relevância no conjunto de termos, definida por uma função denominada função de relevância *R(t)* (vide Equação 3.1), onde t é um termo do dicionário.

$$R(t) = \frac{\sum_{i=1}^{d} \sum_{j=1}^{|d_i|} (T_{i,j} \equiv t) \times \frac{|d_i| - j_0}{|d_i|}}{N} \tag{3.1}$$

De acordo com a Equação 3.1, no contexto desse trabalho, tem-se que:

- t é um termo pertencente ao dicionário;
- *j0* é a posição da primeira ocorrência do termo t_j no documento;
- d é o número de páginas fornecidas como entrada;
- $|d_i|$ é o número de termos da página d_i estabelecidos após o Passo 03;
- $(T_{i,j} \equiv t)$ representa *1* se o termo j da página d_i casa com (equivale a) o termo t, e *0* caso contrário;
- N é a quantidade total de termos de todas as páginas, estabelecidos após o Passo 03.

Na Equação 3.1, o valor $(T_{i,j} \equiv t)$ é ponderado pela posição da primeira aparição do termo j_0 presente na página d_i em questão, valorizando termos que estão no topo de tal página. Segundo Chirita et al. (2007), termos mais importantes tendem a aparecer mais no início do documento. Uma forma de definir a posição de cada termo é dividir a posição do termo pelo número total de termos do documento em questão. Além da indexação dos termos individuais, o Passo 03 faz uma indexação separada para multi-termos. No análise dos multi-termos formados por dois ou três termos vizinhos, o valor de relevância de cada multi-termo é dobrado e triplicado, respectivamente.

Capítulo 4

Experimentação prática

Neste capítulo, são apresentados e analisados os experimentos de validação da estratégia proposta para o aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da *Web* baseada em gênero (De Assis t al. 2009). É importante ressaltar que a estratégia proposta neste trabalho (vide Seção 3.2) envolve a arquitetura de funcionamento de um processo de expansão de termos (vide Figura 3.1), também estabelecida neste trabalho. A Seção 4.1 descreve as métricas que foram utilizadas para avaliar a eficácia da aplicação da estratégia proposta. A Seção 4.2 descreve o *baseline* utilizado para fins comparativos com a estratégia proposta. A Seção 4.3 descreve os experimentos realizados envolvendo a estratégia proposta. Por fim, a Seção 4.4 apresenta e avalia os resultados obtidos por meio dos experimentos realizados, envolvendo a estratégia proposta e o *baseline*.

4.1 Métricas de avaliação

Para avaliação dos experimentos realizados, foram utilizadas a métricas precisão, revocação e F1. De acordo com Manning et al. (2008) e considerando o contexto deste trabalho, precisão (*precision*) é definida como a proporção de páginas relevantes determinadas corretamente pelo *Yucca*, ou seja:

$$precision = \frac{TP}{TP + FP} \tag{4.1}$$

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.

Também conforme definido por Manning et al. (2008) e aplicado ao contexto deste trabalho, revocação (*recall*) é definida como a proporção das páginas relevantes selecionadas pelo coletor dado um gabarito das páginas relevantes ao tópico de interesse; ou seja:

$$recall = \frac{TP}{TP + FN} \tag{4.2}$$

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;
- FN é o número de páginas da *Web* visitadas pelo coletor, que são realmente relevantes ao tópico de interesse, porém classificadas como não relevantes pelo *Yucca*.

Finalmente, também definido por (Manning et al., 2008), F1 corresponde à média harmônica entre as métricas de precisão e revocação, representando a eficácia efetiva de um determinado processo de coleta; ou seja:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.3}$$

4.2 *Baseline*

Para que fosse possível validar a estratégia proposta neste trabalho (vide Seção 3.2) que visa o aperfeiçoamento automático dos conjuntos de termos utilizados em processo de coleta temática de páginas da *Web* baseada em gênero, foi considerado, como *baseline*, a execução de um processo de coleta, que seguiu a abordagem original para a coleta temática sem aperfeiçoamento de termos (vide Figura 2.1), considerando o tópico de interesse proposto por Mangaravite et al. (2014), que também foi usado nos experimentos de Costa et al. (2017). Tal processo de coleta possui as seguintes características:

- Tópico de interesse: Ementas de disciplinas de Banco de Dados.
- Conjuntos de termos de gênero e conteúdo: os mesmos utilizados por Mangaravite et al. (2014), estando apresentados na Tabela 4.1.
- Conjunto de páginas-semente obtido por meio da melhor heurística proposta por Mangaravite et al. (2014) (vide Subseção 2.1.1.2), a saber:
  - https://sites.google.com/site/proftheobaldo/disciplinas/implementacao-de-banco-de-dados/plano-de-ensino

  - http://ulbra-to.br/cursos/Sistemas-de-Informacao/2011/2/turmas/0312/impressao-plano
  - http://slideplayer.com.br/slide/364082
  - http://www.wladmirbrandao.com/course_is-bsc-dbs.html
  - https://www.passeidireto.com/arquivo/1011044/plano_de_ensino
  - http://docplayer.com.br/17920843-Cefet-phb-pi-plano-de-ensino-banco-de-dados-plano-de-ensino-plano-de-ensino-plano-de-ensino-conteudo-plano-de-ensino-conteudo.html
  - https://palmas.ifto.edu.br/index.php/component/phocadownload/category/1-edital%3D573:sistemas-para-internet-administracao-de-banco-de-dados
  - https://sbv.ifsp.edu.br/wiki/index.php/Sistemas_de_Gerenciamento_de_Banco_de_dados_(GBDI3)_(Tecnologia_em_Sistemas_para_Internet)
  - https://www.yumpu.com/pt/document/view/14485746/projeto-de-banco-de-dados-plano-de-ensino

- Número máximo de páginas visitadas: 1000.

Ao longo da execução do processo de coleta mencionado, foi armazenado um *log* contendo as seguintes informações sobre cada página visitada pelo coletor:

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

Ao terminar o processo de coleta relativo ao *baseline*, verificou-se que foram visitadas 865 páginas. Uma vez visitadas tais páginas, para que fosse possível estabelecer as métricas descritas na Seção 4.1, foi gerado um gabarito, contendo todas as páginas realmente relevantes ao tópico de interesse desejado por um especialista. Os valores obtidos para as métricas encontram-se na Tabela 4.2.

4.3 Descrição dos experimentos

Para realizar os experimentos de validação da estratégia proposta neste trabalho (vide Seção 3.2), foram realizados processos de coleta utilizando o *log* relativo à coleta feita para a obtenção do *baseline* (vide Seção 4.2). Em tais processos de coleta, a estratégia apresentada

Tabela 4.1: Conjuntos de termos de gênero e conteúdo para o tópico "Ementa de disciplinas de Banco de Dados"(Mangaravite et al., 2014).

Termos de Gênero	**Termos de Conteúdo**
plano de ensino	banco de dados
disciplina	bancos de dados
creditos	sgbd
professor	sistemas de banco de dados
pre requisitos	gerencia de banco de dados
ementa	modelagem de dados
conteudo programático	modelo de dados
objetivos	entidade relacionamento
descricao	modelo conceitual
planejamento	modelo relacional
aula	modelo logico
notas	integridade referencial
provas	algebra relacional
trabalhos	calculo relacional
referencias bibliograficas	sql
bibliografia	normalizacao
-	dependencias funcionais
-	definicao de dados
-	controle de concorrencia
-	otimizacao de consulta
-	triggers
-	armazem de informacoes
-	silberschatz
-	navathe

Tabela 4.2: Precisão e revocação obtidos como *baseline*.

Precisão	**Revocação**	**F1**
0.698	0.953	0.806

Tabela 4.3: Conjuntos reduzidos de termos de gênero e conteúdo para o tópico "Ementa de disciplinas de Banco de Dados".

Termos de Gênero	**Termos de Conteúdo**
plano de ensino	banco de dados
discilpina	entidade relacionamento
ementa	modelo conceitual
objetivos	algebra relacional
provas	sql
referencias bibliograficas	silberschatz
bibliografia	navathe

no Capitulo 3 foi aplicada utilizando distintos conjuntos de termos de gênero e conteúdo, e considerando o conjunto de páginas-semente do *baseline* e o conjunto de páginas da *Web* visitadas contidos no *log*.

Segundo Costa et al. (2017), ao analisar os conjuntos de termos de gênero e conteúdo estabelecidos por Mangaravite et al. (2014) para o tópico "Ementa de disciplinas de Banco de Dados", percebe-se que tais conjuntos foram adequadamente definidos por um especialista, representando o tópico de interesse de forma detalhada, completa e objetiva. Assim, com o intuito de verificar os resultados obtidos pela aplicação da estratégia proposta sobre conjuntos de termos que definem o tópico de interesse de forma menos detalhada, foram definidos novos conjuntos de termos de gênero e conteúdo para o mesmo tópico de interesse. Os primeiros conjuntos, que estão apresentados na Tabela 4.3, consistem em conjuntos reduzidos de termos de gênero e conteúdo extraídos dos conjuntos apresentados na Tabela 4.1, originalmente estabelecidos por Mangaravite et al. (2014); a redução baseou-se na escolha intuitiva dos 7 termos de gênero e dos 7 termos de conteúdo que melhor descrevem o tópico de interesse. Já para a definição dos segundos conjuntos de termos, considerou-se o caso em que um usuário não especialista fosse responsável pela definição dos termos de gênero e conteúdo para o tópico de interesse "Ementa de disciplinas de Banco de Dados". Logo, a definição desses conjuntos de termos, que se encontram apresentados na Tabela 4.4 foi feita por um estudante do curso de Ciência da Computação arbitrariamente escolhido (Costa et al., 2017).

Para tornar possível a análise comparativa dos resultados relativos à aplicação da estratégia proposta neste trabalho sobre os novos conjuntos de termos de gênero e conteúdo, fez-se necessário estabelecer valores de referência (vide Tabelas 4.5 e 4.6) para as métricas descritas na Seção 4.1 considerando a utilização dos novos conjuntos de termos na determinação da similaridade entre as páginas da *Web* visitadas e o tópico de interesse, seguindo a abordagem original (*baseline*), ou seja, sem aperfeiçoamento de termos. Tais

Tabela 4.4: Conjuntos de termos de gênero e conteúdo definidos por um não especialista para o tópico "Ementa de disciplinas de Banco de Dados".

Termos de Gênero	**Termos de Conteúdo**
ementa	banco de dados
horario	sql
materia	relacionamento
conteudo	consulta
aula	-
cronograma	-

Tabela 4.5: Precisão e revocação obtidos como referência para o processo de coleta envolvendo os conjuntos reduzidos de termos.

Precisão	**Revocação**	**F1**
0.615	0.925	0.739

Tabela 4.6: Precisão e revocação obtidos como referência para o processo de coleta envolvendo os conjuntos de termos definidos por um não especialista.

Precisão	**Revocação**	**F1**
0.607	0.964	0.745

valores foram obtidos por meio de processos de coleta envolvendo os novos conjuntos de termos de gênero e conteúdo e o conjunto de páginas da *Web* contidos no *log*. Uma vez realizados tais processos de coleta, as métricas de avaliação foram obtidas de forma análoga ao que foi feito para a obtenção das métricas relativas ao *baseline* inicial (vide Tabela 4.2). Ou seja, utilizando o gabarito referente às páginas consideradas, para cada caso, foi estabelecido o limite de similaridade ótimo, de forma empírica, visando a maximização da métrica F1.

Uma vez obtidos os valores utilizados como referência, foram realizados experimentos envolvendo processos de coleta temática seguindo a abordagem baseada em gênero e considerando a estratégia proposta neste trabalho para o aperfeiçoamento de termos. Nesses experimentos, além da realização de distintos processos de coleta para os conjuntos de termos originalmente definidos por Mangaravite et al. (2014) (vide Tabela 4.1) e para os novos conjuntos de termos previamente definidos (vide Tabelas 4.3 e 4.4), foram feitas variações dos possíveis parâmetros relacionados à estratégia.

De acordo com a Figura 3.2, a estratégia proposta, que é baseada em frequência de termos, gera um conjunto de termos expandidos a partir de um conjunto distinto de páginas de entrada. Uma vez tendo o conjunto possível de termos expandidos relativos ao gênero (ou ao conteúdo) do tópico desejado, é necessário determinar quantos termos devem ser

realmente selecionados pela mesma para compor cada conjunto de termos expandidos (gênero e conteúdo), definido como valor n. Tal conjunto de termos de gênero (ou conteúdo) expandidos deve ser composto pelos n termos mais relevantes da lista de termos indexados de gênero (ou conteúdo).

Foram realizadas distintas execuções de processos de coleta temática seguindo a abordagem baseada em gênero com a aplicação da estratégia baseada em frequência de termos (gênero e conteúdo). Tais execuções foram feitas de acordo com a combinação dos seguintes parâmetros:

- quantidade n de termos selecionados para compor o conjunto de termos expandidos referentes ao gênero;
- quantidade n de termos selecionados para compor o conjunto de termos expandidos referentes ao conteúdo.

Tais parâmetros sofreram variações de valor em um intervalo de 0 a 10, resultando em 11 valores distintos para cada parâmetro, e uma combinação de 120 casos de processos de coleta. Os resultados referentes aos processos de coleta encontram-se apresentados na Seção 4.4.

Ademais, no intuito de avaliar a estabilidade da eficácia do coletor *Yucca* em seguindo a abordagem original, foram realizadas 10 processos de coleta temática sem a intervenção de um processo de expansão dos termos, usando os conjuntos completos de termos (vide Tabela 4.1).

4.4 Resultados

Nesta seção, são apresentados e analisados os resultados obtidos por meio da experimentação prática realizada. Considerando a estratégia proposta neste trabalho e as possíveis variações de parâmetros, foram realizados 360 processos de coleta temática. Desse total de processos de coleta, 120 consideram os conjuntos completos de termos (vide Tabela 4.1), 120 consideram os conjuntos reduzidos de termos (Tabela 4.3) e 120 consideram os conjuntos de termos definidos por um não especialista (vide Tabela 4.4).

Todos os resultados referentes aos 360 processos de coleta foram organizados de acordo com os conjuntos de termos de entrada utilizados. Dessa forma, os resultados foram estruturados em tabelas onde, para cada processo de coleta realizado (cada caso de teste), foram apresentados os parâmetros utilizados e os valores obtidos para as métricas de avaliação descritas na Seção 4.1. Adicionalmente, com o objetivo de facilitar a análise dos resultados, foi incluído, em cada tabela, o "Caso 0"que apresenta os valores das métricas obtidos como

referência, por meio de um processo de coleta sem o aperfeiçoamento de termos (*baseline*). A partir deste "Caso 0", foi possível incluir, para cada processo de coleta realizado, um valor percentual representando a distância da eficácia obtida (métrica F1) em relação à eficácia apresentada no "Caso 0".

As tabelas contendo todos os 360 processos de coleta temática realizados encontram-se nos Apêndices A e B. Desta forma, as subseções seguintes apresentam e analisam apenas os melhores resultados obtidos (distância em relação ao "Caso 0"maior ou igual a 90%), de acordo com os conjuntos de termos de gênero e conteúdo considerados. Assim, as Subseções 4.4.1, 4.4.2 e 4.4.3 apresentam tais melhores resultados considerando, como termos de entrada, os conjuntos de termos definidos por Mangaravite et al. (2014), os conjuntos reduzidos de termos e os conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista, respectivamente.

4.4.1 Resultados considerando os conjuntos completos de termos

A Tabela 4.7 apresenta os melhores resultados obtidos ao aplicar a estratégia proposta para expansão de termos (vide Figura 3.2) considerando os conjuntos originais e completos de termos de gênero e conteúdo (vide Tabela 4.1).

Para melhor compreensão da Tabela 4.7, analisando por exemplo o caso 2, observa-se que o mesmo gerou valores de precisão, revocação e F1 iguais a 0.782, 0.955 e 0.860, respectivamente, ao considerar nenhum termo expandido de gênero e dois termos expandidos de conteúdo. Em tal caso, a proximidade do valor de F1 obtido pela estratégia proposta em relação ao valor de F1 obtido pelo *baseline* (Caso 0) foi de 106.694, representando uma melhora de 6,694%.

Por meio da análise da Tabela 4.7, houve 15 casos, marcados em negrito, em que o uso da estratégia resultou em uma melhora da métrica F1. Nota-se que, dentre tais casos, existem: (a) casos com poucos termos expandidos de gênero e conteúdo (por exemplo, os Casos 1 e 2); (b) casos com poucos termos expandidos de gênero e muitos termos expandidos de conteúdo (por exemplo, os Casos 16 e 26); (c) casos com muitos termos expandidos de gênero e poucos termos expandidos de conteúdo (por exemplo, os Casos 89 e 90); e (d) casos com muitos termos expandidos de gênero e muitos termos expandidos de conteúdo (por exemplo, os Casos 96 e 97). Desta forma, não é possível encontrar um padrão entre os valores de F1 e as quantidades de termos de gênero e conteúdo expandidos a serem considerados em processos de coleta pelo *Yucca*. Mas, é possível constatar que a estratégia proposta neste trabalho apresentou um resultado bem satisfatório quanto ao processo de expansão de termos, mesmo

considerando conjuntos grandes de termos originais, por ter obtido valores de F1 superiores aos obtidos pelo *baseline* e, em muitos casos, valores de F1 próximos aos obtidos pelo *baseline* (proximidade de F1 acima de 95%). Deve-se considerar, também, que, ao analisar todos os 120 casos (Apêndice A.1), nota-se que alguns casos atingiram um uma proximidade de F1 abaixo de 70% (por exemplo, os Casos 55, 58 e 59), que estão, em sua grande maioria, nos casos em que a quantidade de termos de gênero expandidos é igual ou menor a 5. Ademais, é possível destacar que, ao analisar todos os 120 casos (Apêndice A.1), que a maioria atingiu uma proximidade de F1, em relação ao *baseline*, superior a 80%.

4.4.2 Resultados considerando os conjuntos reduzidos de termos

A Tabela 4.8 apresenta os melhores resultados obtidos ao aplicar a estratégia proposta para expansão de termos (vide Figura 3.2) considerando os conjuntos reduzidos de termos de gênero e conteúdo (vide Tabela 4.3), de forma similar ao que foi feito para a Subseção 4.4.1.

Por meio da análise da Tabela 4.8, houve 26 casos, marcados em negrito, em que o uso da estratégia resultou em uma melhoria da métrica F1, com uma melhora máxima de 17,24% no caso 34. Nota-se que, assim como na análise feita na Subseção 4.4.1, não existe um padrão em relação à melhoria da métrica F1 desses casos e a quantidade de termos de gênero (ou conteúdo) expandidos. Ademais, nota-se que, analisando a tabela onde mostra todos os 120 casos (Apêndice A.2), conforme a quantidade de termos de gênero aumenta, a variação dos valores de proximidade de F1 em relação ao *baseline* diminui, e sua média se mantém próxima de 100%.

4.4.3 Resultados considerando os conjuntos de termos definidos por um não especialista

A Tabela 4.9 apresenta os melhores resultados obtidos ao aplicar a estratégia proposta para a expansão de termos (vide Figura 3.2) considerando os conjuntos de termos definidos por um não especialista (vide Tabela 4.4), de forma similar ao que foi feito para a Subseção 4.4.1.

Por meio da análise da Tabela 4.9, houve 29 casos, marcados em negrito, em que o uso da estratégia resultou em uma melhoria da métrica F1, com uma melhora máxima de 15,99% (Caso 7). Nota-se que, assim como nas análises feitas na Subseção 4.4.1 e Subseção 4.4.2, não existe um padrão entre a relação à melhoria da métrica F1 desses casos e a quantidade de termos de gênero (ou conteúdo) expandidos. Nota-se, também, analisando a tabela onde mostra todos os 120 casos (Apêndice A.3), que houve apenas 7 casos em que o número de

Tabela 4.7: Melhores resultados relativos à estratégia considerando os conjuntos completos de termos

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.698	0.953	0.806	100.000
1	0	1	0.697	0.969	0.811	**100.632**
2	0	2	0.782	0.955	0.860	**106.694**
3	0	3	0.633	0.969	0.766	95.094
5	0	5	0.646	0.967	0.775	96.183
16	1	5	0.743	0.971	0.842	**104.459**
24	2	2	0.660	0.951	0.780	96.766
26	2	4	0.758	0.964	0.849	**105.329**
27	2	5	0.648	0.966	0.776	96.278
34	3	1	0.748	0.962	0.842	**104.494**
36	3	3	0.672	0.931	0.780	96.864
40	3	7	0.696	0.945	0.801	99.449
41	3	8	0.655	0.953	0.776	96.365
44	4	0	0.660	0.951	0.779	96.743
51	4	7	0.674	0.929	0.781	96.972
54	4	10	0.652	0.957	0.775	96.249
57	5	2	0.700	0.958	0.809	**100.453**
66	6	0	0.720	0.958	0.822	**102.064**
72	6	6	0.765	0.957	0.850	**105.516**
75	6	9	0.691	0.959	0.803	99.732
77	7	0	0.654	0.935	0.770	95.522
82	7	5	0.642	0.971	0.773	95.913
86	7	9	0.664	0.955	0.783	97.239
89	8	1	0.758	0.962	0.848	**105.230**
90	8	2	0.764	0.958	0.850	**105.527**
91	8	3	0.663	0.976	0.790	98.014
93	8	5	0.648	0.949	0.770	95.617
94	8	6	0.674	0.971	0.796	98.813
96	8	8	0.714	0.956	0.818	**101.482**
97	8	9	0.774	0.958	0.856	**106.296**
98	8	10	0.658	0.956	0.779	96.753
103	9	4	0.688	0.926	0.789	97.978
104	9	5	0.680	0.957	0.795	98.657
105	9	6	0.693	0.927	0.793	98.426
111	10	1	0.775	0.950	0.854	**105.961**
113	10	3	0.660	0.969	0.785	97.491
114	10	4	0.720	0.956	0.822	**101.989**
118	10	8	0.638	0.969	0.770	95.556

Tabela 4.8: Melhores resultados relativos à estratégia considerando os conjuntos reduzidos de termos

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.615	0.925	0.739	100.000
3	0	3	0.793	0.921	0.852	**115.341**
6	0	6	0.628	0.946	0.755	**102.185**
7	0	7	0.618	0.904	0.734	99.426
10	0	10	0.672	0.895	0.768	**103.987**
18	1	7	0.781	0.944	0.855	**115.713**
19	1	8	0.609	0.937	0.738	99.979
20	1	9	0.575	0.950	0.717	97.012
21	1	10	0.655	0.928	0.768	**104.035**
24	2	2	0.661	0.912	0.767	**103.807**
27	2	5	0.655	0.939	0.772	**104.544**
28	2	6	0.710	0.922	0.802	**108.640**
34	3	1	0.792	0.956	0.866	**117.238**
50	4	6	0.703	0.973	0.816	**110.531**
51	4	7	0.647	0.951	0.770	**104.302**
52	4	8	0.706	0.953	0.811	**109.816**
56	5	1	0.702	0.955	0.809	**109.504**
61	5	6	0.625	0.887	0.733	99.281
70	6	4	0.600	0.917	0.725	98.226
72	6	6	0.703	0.973	0.816	**110.531**
73	6	7	0.637	0.851	0.729	98.684
74	6	8	0.746	0.853	0.796	**107.767**
78	7	1	0.712	0.920	0.802	**108.632**
83	7	6	0.671	0.797	0.729	98.639
87	7	10	0.611	0.959	0.747	**101.109**
88	8	0	0.650	0.963	0.776	**105.041**
89	8	1	0.599	0.956	0.736	99.697
90	8	2	0.583	0.961	0.726	98.264
91	8	3	0.581	0.960	0.724	97.979
92	8	4	0.583	0.951	0.723	97.897
93	8	5	0.766	0.960	0.852	**115.364**
94	8	6	0.677	0.976	0.799	**108.208**
96	8	8	0.623	0.958	0.755	**102.200**
98	8	10	0.612	0.958	0.747	**101.135**
99	9	0	0.708	0.953	0.812	**109.999**
105	9	6	0.607	0.961	0.744	**100.777**
117	10	7	0.638	0.969	0.770	**104.228**
120	10	10	0.628	0.960	0.759	**102.828**

casos em que a proximidade de F1 é inferior a 70%, e que são menos frequentes conforme a quantidade de termos de gênero expandidos aumenta.

4.4.4 Resultados relativos a processos de coleta sem a intervenção de um processo de expansão de termos

A Tabela 4.10 apresenta os resultados obtidos ao executar vários processos de coleta sem a intervenção de um processo de expansão de termos, considerando o mesmo conjunto de termos de entrada (conjuntos completos de termos, vide Tabela 4.1), no intuito de avaliar a estabilidade da eficácia do coletor *Yucca*, e obter respostas para a instabilidade dos resultados obtidos nas Subseções 4.4.1, 4.4.2 e 4.4.3.

Por meio da análise da Tabela 4.10, é possível observar que o próprio processo de coleta possui uma instabilidade em seus resultados. Nota-se que, nos processos de coleta realizados usando o mesmos conjuntos de termos de gênero e conteúdo como entrada, os processos de coleta obtiveram resultados em que a eficácia varia bastante. Nos processos de coleta registrados na Tabela 4.10, por exemplo, o Caso 0 obteve um valor de 0,854 da métrica F1, enquanto o Caso 4 obteve um valor de 0,572; valores bem distantes. Esse experimento mostra que a instabilidade observada nos resultados obtidos nas Subseções 4.4.1, 4.4.2 e 4.4.3 tem influência do próprio processo de coleta do *Yucca*, e não é causado apenas pelo processo de expansão de termos.

Tabela 4.9: Melhores resultados relativos à estratégia considerando os conjuntos de termos definidos por um não especialista

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.607	0.964	0.745	100.000
3	0	3	0.690	0.928	0.792	106.307
4	0	4	0.787	0.954	0.862	115.797
7	0	7	0.790	0.953	0.864	115.995
17	1	6	0.732	0.965	0.833	111.845
18	1	7	0.779	0.954	0.858	115.180
20	1	9	0.709	0.952	0.813	109.127
34	3	1	0.608	0.946	0.741	99.444
49	4	5	0.644	0.942	0.765	102.768
50	4	6	0.590	0.961	0.731	98.178
52	4	8	0.776	0.959	0.858	115.179
54	4	10	0.595	0.956	0.734	98.512
58	5	3	0.632	0.984	0.770	103.371
59	5	4	0.587	0.971	0.732	98.273
61	5	6	0.732	0.969	0.834	111.967
67	6	1	0.674	0.965	0.794	106.609
68	6	2	0.793	0.947	0.863	115.916
69	6	3	0.588	0.959	0.729	97.896
70	6	4	0.641	0.957	0.768	103.109
71	6	5	0.611	0.959	0.747	100.284
72	6	6	0.650	0.963	0.776	104.183
73	6	7	0.599	0.956	0.736	98.883
76	6	10	0.583	0.951	0.723	97.098
77	7	0	0.766	0.960	0.852	114.423
78	7	1	0.677	0.976	0.799	107.324
80	7	3	0.623	0.958	0.755	101.365
82	7	5	0.612	0.958	0.747	100.309
83	7	6	0.708	0.953	0.812	109.101
89	8	1	0.607	0.961	0.744	99.954
103	9	4	0.761	0.954	0.847	113.677
104	9	5	0.616	0.959	0.750	100.750
105	9	6	0.765	0.957	0.850	114.152
108	9	9	0.691	0.959	0.803	107.895
110	10	0	0.654	0.935	0.770	103.340
113	10	3	0.606	0.975	0.747	100.370
114	10	4	0.615	0.954	0.748	100.446
115	10	5	0.642	0.971	0.773	103.763
116	10	6	0.617	0.959	0.751	100.819

Tabela 4.10: Resultados relativos a processos de coleta sem a intervenção de um processo de expansão de termos

Caso	Precisão	Revocação	F1
0	0.770	0.958	0.854
1	0.770	0.956	0.853
2	0.586	0.966	0.729
3	0.639	0.952	0.765
4	0.411	0.940	0.572
5	0.539	0.966	0.692
6	0.770	0.956	0.853
7	0.658	0.967	0.783
8	0.747	0.959	0.840
9	0.610	0.949	0.743
10	0.769	0.957	0.853

Capítulo 5

Conclusão

Neste trabalho, foi proposta e experimentada uma estratégia para o aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da *Web* seguindo a abordagem baseada em gênero proposta (Assis et al. (2007); Assis et al. (2008); Assis et al. (2009)), baseada na frequência dos termos. Como já mencionado, tal aperfeiçoamento faz-se necessário porque, dependendo da forma como tais conjuntos de termos são especificados para um determinado tópico de interesse, a eficácia de um processo de coleta temática pode não ser satisfatória.

Por meio da análise dos resultados dos experimentos descritos no Capítulo 4, foi possível perceber que o melhor resultado resultado foi obtido por meio da aplicação de tal estratégia sobre conjuntos reduzidos de termos de gênero e conteúdo para o tópico de interesse "Ementas de disciplinas de Banco de Dados". Percebe-se, também, que os resultado obtido por meio da aplicação de tal estratégia sobre conjuntos de termos definidos por um não especialista apresentou a maior quantidade de casos em que houve uma melhoria da métrica F1 em relação ao *baseline*, sugerindo que um aperfeiçoamento automático de termos é mais significativo quando os conjuntos de termos de gênero e conteúdo necessários para descrever um tópico de interesse não foram adequadamente definidos.

Nota-se, também, que, em todos os experimentos realizados, dentre os casos em que houve uma melhoria da métrica F1, existem: (a) casos com poucos termos expandidos de gênero e conteúdo; (b) casos com poucos termos expandidos de gênero e muitos termos expandidos de conteúdo; (c) casos com muitos termos expandidos de gênero e poucos termos expandidos de conteúdo; e (d) casos com muitos termos expandidos de gênero e conteúdo. Tal observação torna difícil a definição de uma quantidade de termos expandidos de gênero ótimo para o tópico de interesse definido. Entretanto, foi possível observar que essa instabilidade é causada, em grande parte, pelo próprio processo de coleta do *Yucca*, e não apenas pelo

processo de expansão de termos. Contudo, foi possível observar que, conforme a quantidade dos termos de gênero aumenta, a proximidade da métrica F1 em relação ao *baseline* tende a se estabilizar entre 90% e 105%.

Ademais, tendo como foco a acurácia fornecida pela métrica F1, é importante mencionar que o melhor resultado obtido nos experimentos promoveu um aumento na métrica F1 de 17,24%, superando o melhor resultado da melhor estratégia proposta por Costa et al. (2017) (aumento na métrica F1 de 6,79%), atingindo, então, o objetivo geral deste trabalho. Vale lembrar que os experimentos feitos por Costa et al. (2017) foram obtidos por processos simulados de coleta, onde as páginas eram previamente definidas.

Como perspectivas de trabalho futuro, pretende-se: (1) realizar novos experimentos de validação da estratégia já proposta, por meio da execução de processos de coleta temática envolvendo outros tópicos de interesse; (2) fazer experimentos envolvendo a melhor estratégia proposta por (Costa et al., 2017) utilizando processos reais de coleta, para fins comparativos; (3) estabelecer quantidades ótimas de termos de gênero e conteúdo a serem expandidos, ou um processo que estabeleça tais valores automaticamente, na intenção de maximizar a acurácia do processo de coleta.

Referências Bibliográficas

Assis, G. T.; Laender, A. H. F.; Gonçalves, M. A. e Silva, A. S. (2007). Exploiting genre in focused crawling. In *Proceedings of the 14th Symposium on String Processing and Information Retrieval*, pp. 62–73, Santiago, Chile.

Assis, G. T.; Laender, A. H. F.; Gonçalves, M. A. e Silva, A. S. (2008). The impact of term selection in genre-aware focused crawling. In *Proceedings of the 23th ACM Symposium on Applied Computing*, pp. 1158–1163, Fortaleza, Brazil.

Assis, G. T.; Laender, A. H. F.; Gonçalves, M. A. e Silva, A. S. (2009). A genre-aware approach to focused crawling. *World Wide Web*, 12(3):285–319.

Chenyan Xiong, J. C. (2015). Query expansion with freebase.

Chirita, P. A.; Firan, C. S. e Nejdl, W. (2007). Personalized query expansion for the web. *ACM SIGIR Conf. on Research and development in information retrieval*.

Costa, G. F.; Assis, G. T. e Souza, M. V. O. (2017). Automatic improvement of terms used in focused crawling processes on web page. In *Proceedings of the 16th International Conference WWW/Internet*.

Kekäläinen, K. J. . J. (2000). Ir evaluation methods for retrieving highly relevant documents.

Levy, O. e Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. pp. 2177–2185.

Mangaravite, V.; Assis, G. T. e Ferreira, A. A. (2012). Improving the efficiency of a genre-aware approach to focused crawling based on link context. In *Proceedings of the 8th Latin American Web Congress*, pp. 17–23, Cartagena de Indias, Colômbia.

Mangaravite, V.; Assis, G. T.; Ferreira, A. A. e Pádua, F. L. C. (2014). Semi-automatic generation of seed pages in genre-aware focused crawling. In *Proceedings of the 13th International Conference WWW/Internet*, pp. 51–58, Porto, Portugal.

Manning, C. D.; Raghavan, P.; Schütze, H. et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Mikolov, T.; Chen, K.; Corrado, G. e Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Saar Kuzi, Anna Shtok, O. K. (2016). Query expansion using word embeddings. *CIKM '16 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1929–1932.

Siqueira, G. O.; Assis, G. T.; Ferreira, A. A.; Silva, A. S. N. E.; Mangaravite, V. e Pádua, F. L. C. (2016). Automatic determination of similarity threshold for focused crawling processes on web pages. In *Proceedings of the 15th International Conference WWW/Internet*, Mannheim, Germany.

Apêndice A

Todos os resultados dos experimentos relativos à estratégia baseada em frequência de termos

Tabela A.1: Resultados relativos à estratégia considerando os conjuntos completos de termos

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.698	0.953	0.806	100.000
1	0	1	0.697	0.969	0.811	100.632
2	0	2	0.782	0.955	0.860	106.694
3	0	3	0.633	0.969	0.766	95.094
4	0	4	0.437	0.952	0.599	74.362
5	0	5	0.646	0.967	0.775	96.183
6	0	6	0.409	0.933	0.569	70.569
7	0	7	0.418	0.870	0.565	70.105
8	0	8	0.612	0.977	0.753	93.409
9	0	9	0.463	0.948	0.622	77.170
10	0	10	0.423	0.880	0.571	70.882
11	1	0	0.457	0.934	0.614	76.180
12	1	1	0.281	0.863	0.424	52.603
13	1	2	0.447	0.804	0.575	71.323
14	1	3	0.494	0.916	0.642	79.659
15	1	4	0.441	0.948	0.602	74.728
16	1	5	0.743	0.971	0.842	104.459
17	1	6	0.367	0.810	0.505	62.677
18	1	7	0.433	0.549	0.484	60.063
19	1	8	0.344	0.836	0.487	60.446
20	1	9	0.412	0.894	0.564	70.016
21	1	10	0.467	0.923	0.620	76.985
22	2	0	0.526	0.903	0.665	82.517
23	2	1	0.622	0.939	0.748	92.902
24	2	2	0.660	0.951	0.780	96.766
25	2	3	0.441	0.939	0.601	74.547
26	2	4	0.758	0.964	0.849	105.329
27	2	5	0.648	0.966	0.776	96.278
28	2	6	0.318	0.949	0.476	59.143
29	2	7	0.555	0.966	0.705	87.530
30	2	8	0.617	0.946	0.747	92.687
31	2	9	0.638	0.941	0.760	94.354
32	2	10	0.632	0.970	0.765	94.960
33	3	0	0.552	0.868	0.675	83.758
34	3	1	0.748	0.962	0.842	104.494
35	3	2	0.453	0.820	0.583	72.426
36	3	3	0.672	0.931	0.780	96.864
37	3	4	0.584	0.958	0.726	90.083
38	3	5	0.389	0.941	0.551	68.379

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
39	3	6	0.414	0.949	0.576	71.517
40	3	7	0.696	0.945	0.801	99.449
41	3	8	0.655	0.953	0.776	96.365
42	3	9	0.399	0.941	0.560	69.559
43	3	10	0.418	0.904	0.572	70.976
44	4	0	0.660	0.951	0.779	96.743
45	4	1	0.313	0.900	0.465	57.694
46	4	2	0.401	0.902	0.556	68.961
47	4	3	0.585	0.971	0.730	90.602
48	4	4	0.415	0.906	0.569	70.670
49	4	5	0.365	0.898	0.519	64.375
50	4	6	0.707	0.707	0.707	87.707
51	4	7	0.674	0.929	0.781	96.972
52	4	8	0.465	0.863	0.604	75.000
53	4	9	0.421	0.902	0.574	71.301
54	4	10	0.652	0.957	0.775	96.249
55	5	0	0.315	0.940	0.472	58.634
56	5	1	0.380	0.908	0.536	66.569
57	5	2	0.700	0.958	0.809	100.453
58	5	3	0.377	0.879	0.528	65.549
59	5	4	0.369	0.933	0.528	65.597
60	5	5	0.322	0.957	0.482	59.800
61	5	6	0.578	0.947	0.718	89.125
62	5	7	0.410	0.866	0.557	69.099
63	5	8	0.590	0.861	0.700	86.891
64	5	9	0.428	0.881	0.576	71.521
65	5	10	0.314	0.931	0.469	58.234
66	6	0	0.720	0.958	0.822	102.064
67	6	1	0.518	0.955	0.672	83.408
68	6	2	0.459	0.965	0.622	77.161
69	6	3	0.446	0.974	0.612	75.920
70	6	4	0.761	0.954	0.847	105.077
71	6	5	0.616	0.959	0.750	93.128
72	6	6	0.765	0.957	0.850	105.516
73	6	7	0.539	0.949	0.687	85.330
74	6	8	0.390	0.929	0.549	68.149
75	6	9	0.691	0.959	0.803	99.732
76	6	10	0.476	0.970	0.639	79.288
77	7	0	0.654	0.935	0.770	95.522
78	7	1	0.453	0.974	0.619	76.818
79	7	2	0.565	0.952	0.709	88.007

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
80	7	3	0.606	0.975	0.747	92.776
81	7	4	0.615	0.954	0.748	92.847
82	7	5	0.642	0.971	0.773	95.913
83	7	6	0.617	0.959	0.751	93.191
84	7	7	0.562	0.957	0.708	87.896
85	7	8	0.392	0.933	0.552	68.576
86	7	9	0.664	0.955	0.783	97.239
87	7	10	0.580	0.951	0.720	89.435
88	8	0	0.394	0.915	0.551	68.347
89	8	1	0.758	0.962	0.848	105.230
90	8	2	0.764	0.958	0.850	105.527
91	8	3	0.663	0.976	0.790	98.014
92	8	4	0.442	0.918	0.597	74.049
93	8	5	0.648	0.949	0.770	95.617
94	8	6	0.674	0.971	0.796	98.813
95	8	7	0.607	0.966	0.746	92.579
96	8	8	0.714	0.956	0.818	101.482
97	8	9	0.774	0.958	0.856	106.296
98	8	10	0.658	0.956	0.779	96.753
99	9	0	0.408	0.887	0.559	69.342
100	9	1	0.401	0.871	0.549	68.184
101	9	2	0.375	0.925	0.533	66.202
102	9	3	0.416	0.913	0.571	70.889
103	9	4	0.688	0.926	0.789	97.978
104	9	5	0.680	0.957	0.795	98.657
105	9	6	0.693	0.927	0.793	98.426
106	9	7	0.392	0.953	0.556	68.961
107	9	8	0.434	0.888	0.583	72.381
108	9	9	0.434	0.907	0.587	72.844
109	9	10	0.403	0.608	0.484	60.125
110	10	0	0.697	0.820	0.754	93.547
111	10	1	0.775	0.950	0.854	105.961
112	10	2	0.560	0.940	0.702	87.117
113	10	3	0.660	0.969	0.785	97.491
114	10	4	0.720	0.956	0.822	101.989
115	10	5	0.617	0.968	0.754	93.571
116	10	6	0.600	0.946	0.734	91.111
117	10	7	0.566	0.954	0.710	88.179
118	10	8	0.638	0.969	0.770	95.556
119	10	9	0.378	0.933	0.538	66.747
120	10	10	0.558	0.971	0.709	87.957

Tabela A.2: Resultados relativos à estratégia considerando os conjuntos reduzidos de termos

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.615	0.925	0.739	100.000
1	0	1	0.429	0.869	0.575	77.818
2	0	2	0.514	0.837	0.637	86.239
3	0	3	0.793	0.921	0.852	115.341
4	0	4	0.359	0.850	0.505	68.377
5	0	5	0.459	0.866	0.600	81.237
6	0	6	0.628	0.946	0.755	102.185
7	0	7	0.618	0.904	0.734	99.426
8	0	8	0.418	0.829	0.556	75.219
9	0	9	0.343	0.855	0.490	66.340
10	0	10	0.672	0.895	0.768	103.987
11	1	0	0.428	0.856	0.571	77.265
12	1	1	0.433	0.875	0.579	78.453
13	1	2	0.342	0.785	0.476	64.493
14	1	3	0.438	0.844	0.576	78.017
15	1	4	0.338	0.793	0.474	64.196
16	1	5	0.313	0.848	0.457	61.863
17	1	6	0.448	0.882	0.594	80.455
18	1	7	0.781	0.944	0.855	115.713
19	1	8	0.609	0.937	0.738	99.979
20	1	9	0.575	0.950	0.717	97.012
21	1	10	0.655	0.928	0.768	104.035
22	2	0	0.367	0.924	0.526	71.164
23	2	1	0.602	0.868	0.711	96.220
24	2	2	0.661	0.912	0.767	103.807
25	2	3	0.378	0.839	0.521	70.590
26	2	4	0.435	0.823	0.569	77.039
27	2	5	0.655	0.939	0.772	104.544
28	2	6	0.710	0.922	0.802	108.640
29	2	7	0.309	0.832	0.451	61.078
30	2	8	0.446	0.837	0.582	78.826
31	2	9	0.425	0.907	0.579	78.365
32	2	10	0.462	0.871	0.604	81.745
33	3	0	0.347	0.857	0.494	66.897
34	3	1	0.792	0.956	0.866	117.238
35	3	2	0.448	0.833	0.583	78.940
36	3	3	0.480	0.829	0.608	82.284
37	3	4	0.324	0.847	0.469	63.519
38	3	5	0.341	0.865	0.489	66.167

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
39	3	6	0.425	0.887	0.575	77.809
40	3	7	0.367	0.855	0.514	69.527
41	3	8	0.455	0.861	0.596	80.680
42	3	9	0.453	0.894	0.601	81.398
43	3	10	0.415	0.901	0.568	76.950
44	4	0	0.374	0.915	0.531	71.886
45	4	1	0.435	0.866	0.579	78.408
46	4	2	0.350	0.975	0.515	69.754
47	4	3	0.430	0.809	0.562	76.054
48	4	4	0.560	0.987	0.715	96.766
49	4	5	0.513	0.997	0.678	91.735
50	4	6	0.703	0.973	0.816	110.531
51	4	7	0.647	0.951	0.770	104.302
52	4	8	0.706	0.953	0.811	109.816
53	4	9	0.573	0.961	0.718	97.249
54	4	10	0.496	0.944	0.650	88.031
55	5	0	0.351	0.867	0.500	67.671
56	5	1	0.702	0.955	0.809	109.504
57	5	2	0.548	0.833	0.661	89.559
58	5	3	0.480	0.821	0.606	81.987
59	5	4	0.524	0.947	0.675	91.386
60	5	5	0.441	0.445	0.443	59.948
61	5	6	0.625	0.887	0.733	99.281
62	5	7	0.367	0.805	0.504	68.254
63	5	8	0.615	0.861	0.718	97.210
64	5	9	0.453	0.854	0.592	80.135
65	5	10	0.435	0.901	0.587	79.451
66	6	0	0.374	0.821	0.514	69.573
67	6	1	0.435	0.866	0.579	78.408
68	6	2	0.450	0.875	0.594	80.481
69	6	3	0.450	0.818	0.581	78.635
70	6	4	0.600	0.917	0.725	98.226
71	6	5	0.553	0.947	0.698	94.556
72	6	6	0.703	0.973	0.816	110.531
73	6	7	0.637	0.851	0.729	98.684
74	6	8	0.746	0.853	0.796	107.767
75	6	9	0.572	0.961	0.717	97.145
76	6	10	0.496	0.984	0.659	89.281
77	7	0	0.431	0.857	0.574	77.663
78	7	1	0.712	0.920	0.802	108.632
79	7	2	0.535	0.843	0.655	88.678

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
80	7	3	0.428	0.833	0.565	76.520
81	7	4	0.565	0.927	0.702	95.090
82	7	5	0.441	0.445	0.443	59.948
83	7	6	0.671	0.797	0.729	98.639
84	7	7	0.517	0.845	0.641	86.832
85	7	8	0.620	0.821	0.707	95.717
86	7	9	0.453	0.854	0.592	80.135
87	7	10	0.611	0.959	0.747	101.109
88	8	0	0.650	0.963	0.776	105.041
89	8	1	0.599	0.956	0.736	99.697
90	8	2	0.583	0.961	0.726	98.264
91	8	3	0.581	0.960	0.724	97.979
92	8	4	0.583	0.951	0.723	97.897
93	8	5	0.766	0.960	0.852	115.364
94	8	6	0.677	0.976	0.799	108.208
95	8	7	0.508	0.883	0.645	87.382
96	8	8	0.623	0.958	0.755	102.200
97	8	9	0.517	0.910	0.660	89.337
98	8	10	0.612	0.958	0.747	101.135
99	9	0	0.708	0.953	0.812	109.999
100	9	1	0.461	0.909	0.612	82.814
101	9	2	0.466	0.884	0.610	82.633
102	9	3	0.485	0.914	0.634	85.795
103	9	4	0.414	0.914	0.570	77.139
104	9	5	0.524	0.854	0.650	88.007
105	9	6	0.607	0.961	0.744	100.777
106	9	7	0.509	0.836	0.633	85.671
107	9	8	0.505	0.838	0.630	85.348
108	9	9	0.421	0.963	0.586	79.311
109	9	10	0.417	0.967	0.583	78.885
110	10	0	0.400	0.969	0.566	76.669
111	10	1	0.384	0.876	0.534	72.270
112	10	2	0.511	0.980	0.671	90.887
113	10	3	0.322	0.970	0.484	65.503
114	10	4	0.396	0.967	0.562	76.077
115	10	5	0.447	0.961	0.610	82.635
116	10	6	0.378	0.986	0.546	73.951
117	10	7	0.638	0.969	0.770	104.228
118	10	8	0.378	0.933	0.538	72.805
119	10	9	0.558	0.971	0.709	95.940
120	10	10	0.628	0.960	0.759	102.828

Tabela A.3: Resultados relativos à estratégia considerando os conjuntos de termos definidos por um não especialista

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
0	0	0	0.615	0.925	0.739	100.000
1	0	1	0.429	0.869	0.575	77.818
2	0	2	0.514	0.837	0.637	86.239
3	0	3	0.793	0.921	0.852	115.341
4	0	4	0.359	0.850	0.505	68.377
5	0	5	0.459	0.866	0.600	81.237
6	0	6	0.628	0.946	0.755	102.185
7	0	7	0.618	0.904	0.734	99.426
8	0	8	0.418	0.829	0.556	75.219
9	0	9	0.343	0.855	0.490	66.340
10	0	10	0.672	0.895	0.768	103.987
11	1	0	0.428	0.856	0.571	77.265
12	1	1	0.433	0.875	0.579	78.453
13	1	2	0.342	0.785	0.476	64.493
14	1	3	0.438	0.844	0.576	78.017
15	1	4	0.338	0.793	0.474	64.196
16	1	5	0.313	0.848	0.457	61.863
17	1	6	0.448	0.882	0.594	80.455
18	1	7	0.781	0.944	0.855	115.713
19	1	8	0.609	0.937	0.738	99.979
20	1	9	0.575	0.950	0.717	97.012
21	1	10	0.655	0.928	0.768	104.035
22	2	0	0.367	0.924	0.526	71.164
23	2	1	0.602	0.868	0.711	96.220
24	2	2	0.661	0.912	0.767	103.807
25	2	3	0.378	0.839	0.521	70.590
26	2	4	0.435	0.823	0.569	77.039
27	2	5	0.655	0.939	0.772	104.544
28	2	6	0.710	0.922	0.802	108.640
29	2	7	0.309	0.832	0.451	61.078
30	2	8	0.446	0.837	0.582	78.826
31	2	9	0.425	0.907	0.579	78.365
32	2	10	0.462	0.871	0.604	81.745
33	3	0	0.347	0.857	0.494	66.897
34	3	1	0.792	0.956	0.866	117.238
35	3	2	0.448	0.833	0.583	78.940
36	3	3	0.480	0.829	0.608	82.284
37	3	4	0.324	0.847	0.469	63.519

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
38	3	5	0.341	0.865	0.489	66.167
39	3	6	0.425	0.887	0.575	77.809
40	3	7	0.367	0.855	0.514	69.527
41	3	8	0.455	0.861	0.596	80.680
42	3	9	0.453	0.894	0.601	81.398
43	3	10	0.415	0.901	0.568	76.950
44	4	0	0.374	0.915	0.531	71.886
45	4	1	0.435	0.866	0.579	78.408
46	4	2	0.350	0.975	0.515	69.754
47	4	3	0.430	0.809	0.562	76.054
48	4	4	0.560	0.987	0.715	96.766
49	4	5	0.513	0.997	0.678	91.735
50	4	6	0.703	0.973	0.816	110.531
51	4	7	0.647	0.951	0.770	104.302
52	4	8	0.706	0.953	0.811	109.816
53	4	9	0.573	0.961	0.718	97.249
54	4	10	0.496	0.944	0.650	88.031
55	5	0	0.351	0.867	0.500	67.671
56	5	1	0.702	0.955	0.809	109.504
57	5	2	0.548	0.833	0.661	89.559
58	5	3	0.480	0.821	0.606	81.987
59	5	4	0.524	0.947	0.675	91.386
60	5	5	0.441	0.445	0.443	59.948
61	5	6	0.625	0.887	0.733	99.281
62	5	7	0.367	0.805	0.504	68.254
63	5	8	0.615	0.861	0.718	97.210
64	5	9	0.453	0.854	0.592	80.135
65	5	10	0.435	0.901	0.587	79.451
66	6	0	0.374	0.821	0.514	69.573
67	6	1	0.435	0.866	0.579	78.408
68	6	2	0.450	0.875	0.594	80.481
69	6	3	0.450	0.818	0.581	78.635
70	6	4	0.600	0.917	0.725	98.226
71	6	5	0.553	0.947	0.698	94.556
72	6	6	0.703	0.973	0.816	110.531
73	6	7	0.637	0.851	0.729	98.684
74	6	8	0.746	0.853	0.796	107.767
75	6	9	0.572	0.961	0.717	97.145
76	6	10	0.496	0.984	0.659	89.281
77	7	0	0.431	0.857	0.574	77.663
78	7	1	0.712	0.920	0.802	108.632

Caso	Quantidade de termos de gênero expandidos	Quantidade de termos de conteúdo expandidos	Precisão	Revocação	F1	Proximidade de F1 em relação ao *baseline* (%)
79	7	2	0.535	0.843	0.655	88.678
80	7	3	0.428	0.833	0.565	76.520
81	7	4	0.565	0.927	0.702	95.090
82	7	5	0.441	0.445	0.443	59.948
83	7	6	0.671	0.797	0.729	98.639
84	7	7	0.517	0.845	0.641	86.832
85	7	8	0.620	0.821	0.707	95.717
86	7	9	0.453	0.854	0.592	80.135
87	7	10	0.611	0.959	0.747	101.109
88	8	0	0.650	0.963	0.776	105.041
89	8	1	0.599	0.956	0.736	99.697
90	8	2	0.583	0.961	0.726	98.264
91	8	3	0.581	0.960	0.724	97.979
92	8	4	0.583	0.951	0.723	97.897
93	8	5	0.766	0.960	0.852	115.364
94	8	6	0.677	0.976	0.799	108.208
95	8	7	0.508	0.883	0.645	87.382
96	8	8	0.623	0.958	0.755	102.200
97	8	9	0.517	0.910	0.660	89.337
98	8	10	0.612	0.958	0.747	101.135
99	9	0	0.708	0.953	0.812	109.999
100	9	1	0.461	0.909	0.612	82.814
101	9	2	0.466	0.884	0.610	82.633
102	9	3	0.485	0.914	0.634	85.795
103	9	4	0.414	0.914	0.570	77.139
104	9	5	0.524	0.854	0.650	88.007
105	9	6	0.607	0.961	0.744	100.777
106	9	7	0.509	0.836	0.633	85.671
107	9	8	0.505	0.838	0.630	85.348
108	9	9	0.421	0.963	0.586	79.311
109	9	10	0.417	0.967	0.583	78.885
110	10	0	0.400	0.969	0.566	76.669
111	10	1	0.384	0.876	0.534	72.270
112	10	2	0.511	0.980	0.671	90.887
113	10	3	0.322	0.970	0.484	65.503
114	10	4	0.396	0.967	0.562	76.077
115	10	5	0.447	0.961	0.610	82.635
116	10	6	0.378	0.986	0.546	73.951
117	10	7	0.638	0.969	0.770	104.228
118	10	8	0.378	0.933	0.538	72.805
119	10	9	0.558	0.971	0.709	95.940
120	10	10	0.628	0.960	0.759	102.828