GILSON FARIA COSTA

Orientador: Guilherme Tavares de Assis

APERFEIÇOAMENTO AUTOMÁTICO DOS CONJUNTOS DE TERMOS UTILIZADOS EM PROCESSOS DE COLETA TEMÁTICA DE PÁGINAS DA *WEB* BASEADA EM GÊNERO

Ouro Preto

Março de 2017

Universidade Federal de Ouro Preto
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

APERFEIÇOAMENTO AUTOMÁTICO DOS CONJUNTOS DE TERMOS UTILIZADOS EM PROCESSOS DE COLETA TEMÁTICA DE PÁGINAS DA *WEB* BASEADA 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.

GILSON FARIA COSTA

Ouro Preto
Março de 2017

UNIVERSIDADE FEDERAL DE OURO PRETO

FOLHA DE APROVAÇÃO

Aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da *Web* baseada em gênero

GILSON FARIA COSTA

Monografia defendida e aprovada pela banca examinadora constituída por:

Dr. Guilherme Tavares de Assis – Orientador
Universidade Federal de Ouro Preto

Dra. Amanda Sávio Nascimento e Silva
Universidade Federal de Ouro Preto

Dr. Anderson Almeida Ferreira
Universidade Federal de Ouro Preto

Ouro Preto, Março de 2017

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. Logo, com o objetivo de aperfeiçoar automaticamente os conjuntos de termos necessários para a realização de processos de coleta temática relativos a uma abordagem de coleta baseada em gênero, foram propostas duas estratégias neste trabalho. Experimentos de validação das estratégias foram realizados, gerando, como melhor resultado, um aumento no valor da métrica F1 de 6,79% em processos de coleta cujos conjuntos de termos não foram adequadamente estabelecidos, ao aplicar a estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância.

Palavras-chave: Coleta temática, coleta temática baseada em gênero, aperfeiçoamento de termos

Dedico este trabalho aos meus pais, que foram e sempre serão minha inspiração.

Agradecimentos

Em primeiro lugar, agradeço aos meus pais, Aloisio e Cássia, por todo suporte dado, pelo amor incondicional e pela formação do meu caráter.

Às minhas irmãs, Aline, Carine e Marine, pelo incentivo e cumplicidade.

Aos meus sobrinhos, Isabelly e Kauan, por conseguirem despertar, com a pureza de um sorriso ou de um abraço, o melhor sentimento que já experimentei.

À minha família, pelo carinho e afeto nos momentos em que estivemos juntos, e pela compreensão nos diversos momentos em que eu não pude estar.

Aos meus amigos, agradeço por terem proporcionado momentos inesquecíveis e por nunca deixarem a vida cair na monotonia. Particularmente, agradeço a Isabella pela parceria de todos os momentos.

Agradeço àqueles professores que, de forma responsável e engajada, participaram da minha formação. Em especial, agradeço ao meu orientador, Guilherme Tavares de Assis, por todas as oportunidades que me ofereceu, pelas inúmeras horas dedicadas a este trabalho, e por ser exemplo de pessoa e de profissional.

Por fim, agradeço a UFOP e ao CsF por contribuírem não apenas com minha formação profissional, mas também por me proporcionar experiências que me tornaram uma pessoa melhor.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

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. 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 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 atuais 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%.

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.

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 a metodologia utilizada para o desenvolvimento desse trabalho. Por fim, a Seção 1.4 apresenta o delineamento do restante da monografia.

1.1 Motivação

Distintamente de estratégias de coleta temática existentes na literatura (Almpanidis et al., 2007; Chakrabarti et al., 1999; Johnson et al., 2003; Pant e Srinivasan, 2005, 2006) visando melhorar a eficiência e a eficácia de processos de coleta temática, foi proposta e desenvolvida uma abordagem (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 2009) voltada para atender situações específicas. De uma forma geral, tal abordagem consiste em considerar as evidências de gênero[2] e conteúdo presentes em uma determinada página e estabelece um grau de similaridade entre tais evidências e o tópico específico de interesse. Logo, tal trabalho desenvolvido teve, como objetivo principal, estabelecer um arcabouço que permita a construção de coletores temáticos eficazes, eficientes e escaláveis, sem a necessidade de um treinamento a priori ou qualquer tipo de preprocessamento. Especificamente, a abordagem para coleta temática proposta é útil em situações onde um tópico de interesse possa ser expresso por meio de dois conjuntos distintos de termos: o primeiro descrevendo aspectos de gênero das páginas desejadas e o segundo referente ao assunto ou conteúdo descrito nessas páginas. Por meio de experimentos realizados, tal abordagem para coleta temática baseada em gênero, por ser mais específica, apresentou níveis de revocação e precisão entre 85% e 100%: bem mais satisfatórios que várias outras estratégias de coleta temática existentes na literatura.

[1] Dentro do contexto, de acordo com Manning et al. (2008), revocação é a proporção de páginas relevantes 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] Por gênero, de acordo com De Assis et al. (2009), entende-se o tipo, a categoria ou o estilo de texto de documentos específicos.

Entretanto, para a realização de processos eficazes de coleta temática seguindo a abordagem baseada em gênero, é preciso especificar muito bem conjuntos de termos de gênero e conteúdo que expressam o tópico de interesse desejado. Na experimentação feita para validar tal abordagem, para cada tópico de interesse considerado, especialistas definiram os termos de gênero e conteúdo utilizados nos processos de coleta. Nesses experimentos, foi observado que, dependendo dos conjuntos especificados de termos, a eficácia de um determinado processo de coleta, medida pelo nível de F1[3] alcançado pelo mesmo, pode não ser satisfatória. Além disso, foi observada, em tais conjuntos de termos, a presença de termos parecidos, envolvendo o mesmo prefixo ou diferentes apenas na pluralidade. No caso, então, o aperfeiçoamento automático dos conjuntos originais de termos pode levar a novos conjuntos de termos mais precisos e não-redundantes, ocasionando um desempenho melhor de processos de coleta quanto à eficácia, ou seja, a localização e a determinação correta de um maior número de páginas relevantes ao tópico de interesse desejado. A solução para este problema não é trivial, podendo envolver técnicas de expansão de consulta (Carpineto e Romano, 2012; Aly, 2008) ou ainda a aplicação de técnicas de Processamento de Linguagem Natural (PLN) (Gonzalez e Lima, 2003; Indurkhya e Damerau, 2010).

1.2 Objetivos Geral e Específicos

Este trabalho de monografia possui, como objetivo geral, a proposta e o desenvolvimento de uma estratégia para o aperfeiçoamento automático de conjuntos de termos que são fornecidos, como parâmetros de entrada, para processos de coleta temática de páginas da *Web*. Para tanto, foi considerada a abordagem para coleta temática baseada em gênero (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 2009), onde o tópico de interesse do usuário pode ser expresso por termos que descrevem o conteúdo e o gênero das páginas da *Web* desejadas. Tal abordagem, conforme já mencionado, possibilita a construção de coletores temáticos que realizam processos de coleta eficazes, eficientes e escaláveis, caso tais termos de gênero e conteúdo sejam bem especificados pelo usuário e significativos ao tópico de interesse em questão.

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

- definição e desenvolvimento de distintas estratégias para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo, baseadas em técnicas de expansão de consulta e PLN;
- realização de experimentos de validação das estratégias propostas para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo, por meio da execução de processos de coleta relativos a distintos tópicos de interesse, visando um estudo comparativo entre as mesmas;

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

- definição de métricas para avaliação dos resultados experimentais obtidos pelos processos de coleta temática realizados;
- geração de uma nova versão do coletor existente, que segue a abordagem para coleta temática baseada em gênero, envolvendo a melhor estratégia proposta e validada para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo.

1.3 Metodologia

Visando o alcance do objetivo geral deste trabalho, foram propostas, implementadas e validadas distintas estratégias 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 (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 2009). As estratégias propostas baseiam-se em técnicas de expansão de consulta, que utilizam matrizes de associação para auxiliarem no estabelecimento de novos conjuntos de termos, e técnicas de processamento de linguagem natural voltadas para o aperfeiçoamento de conjuntos de termos.

Após a implementação das estratégias propostas, para a validação das mesmas, foram realizados experimentos por meio de processos de coleta envolvendo um tópico de interesse específico e as estratégias propostas. Para cada processo de coleta, foi medida a eficácia do mesmo por meio das métricas precisão, revocação e F1. Dessa forma, foi feita uma avaliação comparativa da eficácia obtida pelos processos de coleta que utilizaram os conjuntos de termos aperfeiçoados pelas estratégias propostas e por aqueles que utilizaram os conjuntos originais de termos de gênero e conteúdo(*baseline*).

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, as estratégias propostas para o aperfeiçoamento automático dos termos utilizados em processos de coleta temática baseada em gênero. O Capítulo 4 descreve os experimentos de validação das estratégias propostas 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. A Subseção 2.1.2 refere-se às técnicas de expansão de consulta e suas formas de aplicação. A Subseção 2.1.3 descreve o uso de matrizes de associação para determinação de similaridade entre termos. A Subseção 2.1.4 define a área PLN e suas aplicações no contexto de sistemas de Recuperação de Informação (RI). Por fim, a Subseção 2.1.5 apresenta como o *framework* Apache Lucene pode ser útil para a manipulação de documentos textuais em sistemas de RI.

2.1.1 Coleta Temática - Abordagem baseada em gênero

A abordagem para coleta temática de páginas Web baseada em gênero, proposta em (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 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, como já mencionado, essa abordagem foi projetada para situações em que um tópico de interesse pode ser descrito por dois conjuntos distintos de termos: o primeiro conjunto expressa o gênero das páginas desejadas e o segundo descreve o assunto ou os aspectos de conteúdo dessas páginas.

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

A Figura 2.1 sumariza os principais passos da abordagem proposta em (De Assis et al., 2009). 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, correspondem a parâmetros de entrada para funcionamento da arquitetura.

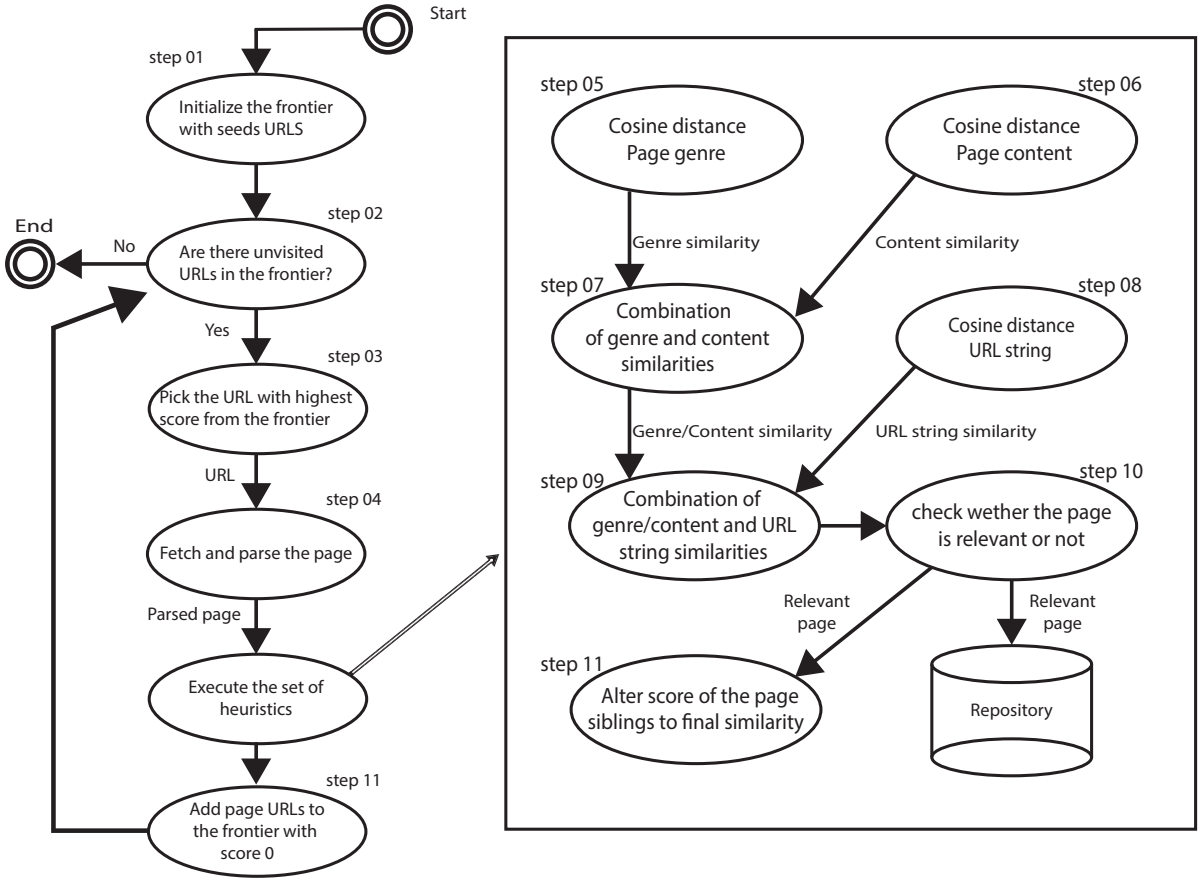

Figura 2.1: Arquitetura de funcionamento da abordagem para coleta temática proposta em (De Assis et al., 2009)

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[1]), entre os termos de gênero do tópico desejado e a página buscada no passo 04;
- Passo 06: calcular a similaridade, por meio da distância de cosseno (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;
- Passo 11: se a página visitada em questão for relevante, alterar a pontuação de prioridade de visita às páginas irmãs, que correspondem a URLs não visitadas no *Frontier*, para a similaridade final obtida no passo 09;
- Passo 12: adicionar as URLs da página visitada pelo coletor no *Frontier* com pontuação de prioridade de visita igual a 0; em seguida, retornar ao passo 02.

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

De acordo com De 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) 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 desenvolvidas estratégias para determinação automática do limite de similaridade a ser considerado em processos de coleta temática.

Assim como os trabalhos citados anteriormente, 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 tem por objetivo realizar uma melhor especificação dos conjuntos de termos de gênero e conteúdo a serem usados em um processo de coleta.

2.1.1.1 Uso de *Link Context* na abordagem baseada em gênero

Segundo (De 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 cossenos (modelo vetorial) entre os termos de gênero e de conteúdo, parâmetros de entrada da abordagem proposta em (De Assis et al., 2009), e os textos gerados pela utilização do *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 na abordagem baseada em gênero

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.

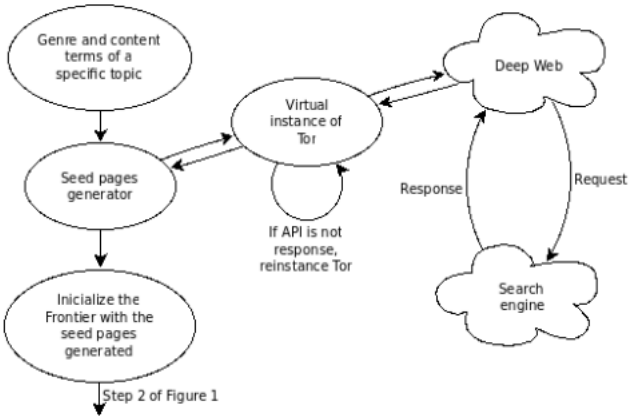

Figura 2.2: Arquitetura de funcionamento da geração de páginas-semente proposta em (Mangaravite et al., 2014)

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 confeccção 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 semi-automática de páginas-semente foi a *unionFirst*, que resultou em uma melhoria de eficiência na abordagem de coleta temática, proposta em (De Assis et al., 2009), de até 53%.

Uma vez definidas as páginas-semente para um processo de coleta temática, de acordo com a Figura 2.2, 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 similaridades na abordagem baseada em gênero

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

A primeira estratégia definida busca determinar o limite de similaridade, para um tópico de interesse específico, por meio da média aritmética ou ponderada das similaridades entre as páginas-sementes e os termos de gênero e conteúdo. A segunda estratégia visa determinar o limite de similaridade mediante a aplicação de métodos de agrupamento sobre os valores de similaridade das páginas-semente; para tanto, foram considerados dois métodos de agrupamento clássicos: K-Means (método de particionamento) e BIRCH (método hierárquico). Por fim, a terceira estratégia objetiva a determinação do valor do limite de similaridade por meio da maximização da métrica coeficiente de silhueta em *clusters* formados por páginas relevantes e não relevantes ao tópico em questão, de acordo com as similaridades das páginas-semente.

Para cada estratégia desenvolvida, foram realizados processos de coleta envolvendo três tópicos de interesse distintos. Por meio dos resultados obtidos, observou-se que os processos de coleta relativos à estratégia baseada no método de agrupamento K-Means, foram os que apresentaram melhores eficácias, chegando a alcançar níveis de F1 bem próximos (dife-

rença de apenas 5,4%) daqueles obtidos quando os limites de similaridade foram definidos por especialistas dos tópicos de interesse considerados.

2.1.2 Expansão de Consulta

Sistemas de 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*[2].

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 interativa, 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

[2]Segundo Kilgarriff e Yallop (2000), *thesaurus* é um recurso linguístico formado por listas de palavras ou expressões agrupadas de acordo com a equivalência de seus significados, geralmente considerando um contexto específico.

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.

Neste trabalho, para o aperfeiçoamento dos conjuntos de termos de gênero e conteúdo, fornecidos como parâmetros de entrada em processos de coleta temática baseada em gênero, uma das estratégias propostas (vide Subseção 3.1) baseia-se em uma técnica de expansão automática de consulta. Logo, como o objetivo é desenvolver um método automático para a melhoria de tais conjuntos de termos, torna-se inviável a aplicação de *feedback* de relevância por parte do usuário. Para tanto, 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.

2.1.3 Matriz de Associação de Termos

Uma das principais fases das técnicas de expansão de consulta, segundo Carpineto e Romano (2012), consiste em gerar os termos candidatos à expansão e ordená-los de acordo com alguma métrica de relevância. Esta ordenação, ou *ranking*, está diretamente relacionada ao desempenho do método de expansão de consulta, uma vez que muitos candidatos são gerados mas apenas alguns deles são efetivamente selecionados para suplementar a consulta inicial. As diversas técnicas existentes para a geração e *ranking* dos candidatos podem ser classificadas de acordo com o tipo de relacionamento entre os termos gerados e os termos da consulta original.

Uma dessas formas de relacionamento é a associação um-para-um entre os termos originais e os de expansão. Nesta forma de relacionamento, ainda de acordo com Carpineto e Romano (2012), estão as técnicas que consideram os termos iniciais isoladamente e, para cada um deles, selecionam termos para expansão. Métodos como *stemming* (vide Subseção 2.1.4) e consulta a *thesaurus* são baseados na ideia de associação um-para-um.

Ademais, as técnicas que utilizam matrizes de associação de termos também seguem a ideia de relacionamento um-para-um. Uma matriz de associação, de acordo com (Chartree et al., 2013), é utilizada para mensurar a relação existente entre os termos dos documentos de uma coleção. A Figura 2.3 ilustra uma matriz de associação, que consiste em uma estrutura bidimensional onde cada entrada s_{ij} representa a similaridade entre os termos t_i e t_j, sendo

que os termos t_1 a t_n são todos os termos contidos na coleção a partir da qual a matriz foi gerada.

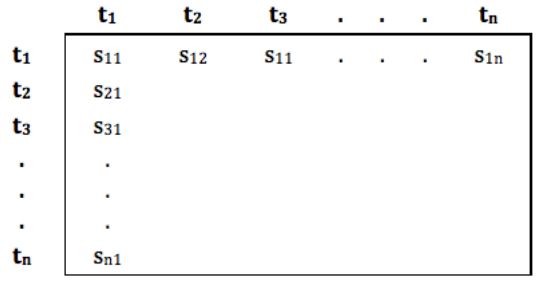

Figura 2.3: Matriz de Associação

Para se determinar a similaridade s_{ij} entre os termos t_i e t_j da matriz de associação apresentada na Figura 2.3, Carpineto e Romano (2012) considera que dois termos são semanticamente relacionados quando eles aparecem juntos em um conjunto de documentos, de forma similar ao fato que dois documentos são similares quando possuem um conjunto de termos em comum. Assim, para o cálculo da similaridade entre dois termos quaisquer de uma matriz de associação, duas métricas comumente usadas são o coeficiente de Sorense-Dice e o de Jaccard, descritos a seguir.

Dados dois termos u e v de uma matriz de associação, o coeficiente de Sorense-Dice é definido como:

$$D = \frac{2 \cdot df_{u \wedge v}}{df_u + df_v}$$

onde $df_{u \wedge v}$ é o número de documentos que contém simultaneamente u e v, e df_u e df_v são os números de documentos que contém u e v, respectivamente.

Já o coeficiente de Jaccard é definido como:

$$J = \frac{df_{u \wedge v}}{df_{u \vee v}}$$

onde $df_{u \wedge v}$ é o número de documentos que contém u e v ao mesmo tempo, e $df_{u \vee v}$ é o número de documentos que contém u ou v.

Considerando a abordagem para coleta temática baseada em gênero (vide Subseção 2.1.1) utilizada como base neste trabalho, dado um conjunto inicial de páginas, obtidas por meio de uma busca na *Web* considerando os termos de gênero e conteúdo relativos aos tópicos de interesse desejados, uma das estratégias propostas nesse trabalho (vide Subseção 3.1) constituiu na geração de matrizes de associação para identificarem a relação entre os termos contidos em tais páginas obtidas: para cada matriz de associação, foi utilizada uma métrica distinta (coeficiente de Jaccard ou MenorDistância (vide Subseção 3.1)) para o cálculo da similaridade entre tais termos. A partir de uma determinada matriz gerada, com o objetivo de melhorar a eficácia de processos de coleta que seguem a abordagem baseada em gênero, foi possível

estabelecer novos termos para suplementar os conjuntos de termos de gênero e conteúdo.

2.1.4 Processamento de Linguagem Natural

Segundo Chowdhury (2003), PLN é uma área de pesquisa que explora como os computadores podem ser utilizados para entender e manipular textos ou sons expressos em linguagem natural. As técnicas de PLN podem ser úteis em muitas áreas, como inteligência artificial, robótica, linguística, interação homem-máquina e RI. Como exemplos de aplicações desenvolvidas usando técnicas de PLN, podem-se citar: sumarização e extração de conhecimento em textos, reconhecimento de fala, geração e interpretação de textos, tradução automática e correção ortográfica.

No contexto da área de RI, particularmente quando o objetivo é recuperar documentos de uma base de dados composta por arquivos de texto, técnicas de PLN podem ser amplamente utilizadas. Um dos usos consiste no processo de construção do índice invertido que, de acordo com Manning et al. (2008), é uma estrutura de dados que mapeia todos os termos de uma coleção de documentos, gerando o vocabulário de termos da mesma e a lista de ocorrências de cada termo, no intuito de promover acesso eficiente aos documentos da coleção mediante consultas. A fase de construção do índice invertido é chamada de indexação e, no caso, técnicas de PLN são aplicadas na tentativa de selecionar os termos que melhor representam a coleção. Os termos presentes no índice devem estar representados de tal forma a promover busca eficiente de documentos da coleção por meio de consultas, que podem ser formuladas de diversas maneiras. Assim, além da seleção de termos mais representativos, PLN também é utilizado para definir o melhor formato de indexação para esses termos. A Figura 2.4 representa a aplicação de técnicas de PLN durante a construção do índice invertido.

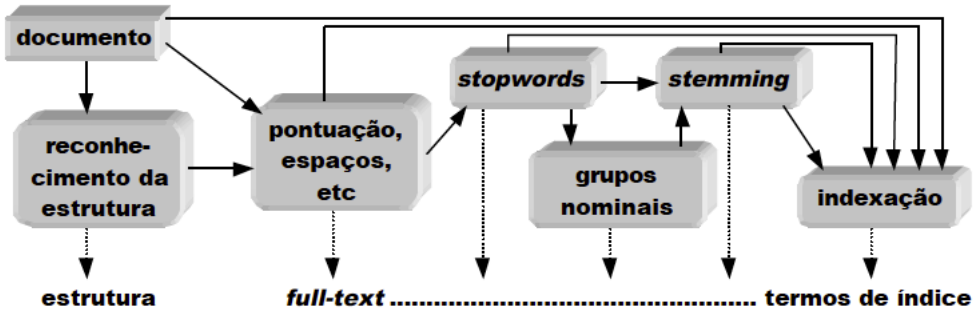

Figura 2.4: Construção de termos de índice (Gonzalez e Lima, 2003)

De acordo com a Figura 2.4, para se gerar os termos de índice relativos a um determinado documento, deve-se fazer, inicialmente, o reconhecimento da estrutura do documento (identificando propriedades como formato de arquivo e idioma) e a análise léxica do mesmo. A análise léxica, em uma definição simples, consiste em fazer uma leitura do texto completo e produzir, como saída, um conjunto de *tokens*. Na maioria das vezes, as palavras que compõem os tex-

tos são delimitadas por espaços, sinais de pontuação ou outros caracteres especiais, tornando possível identificar e extraí-las como um *token*. Apesar de ser uma tarefa simples, segundo Indurkhya e Damerau (2010), a análise léxica ainda é uma área muito pesquisada, uma vez que linguagens como Chinês, Japonês e Tailandês possuem uma forma de representação que não permite a identificação dos *tokens* de maneira trivial.

Uma vez obtidos os *tokens*, seguindo o fluxo denotado na Figura 2.4, é necessário selecionar aqueles que devem compor o índice invertido. Essa seleção, de acordo com Manning et al. (2008), acontece baseada no fato de que todo idioma possui um conjunto de palavras que são extremamente comuns e carregam pouco significado, tais como artigos, preposições e algumas conjunções. Essas palavras são chamadas de *stopwords* e, por não serem relevantes para representar a informação dos documentos, devem ser removidas. Em Manning et al. (2008), são descritas duas estratégias principais para a remoção de *stopwords*. A primeira delas consiste em ordenar todos os termos pela frequência na coleção (somatório do número total de vezes que os termos aparecem em cada documento) e remover os mais frequentes. A segunda consiste em fazer consultas a *stoplists*, que são listas de *stopwords* para cada idioma.

Uma vez removidas as *stopwords*, considerando o fluxo apresentado na Figura 2.4, é aplicada uma técnica denominada *stemming*, amplamente usada em processamento de textos para sistemas de RI. O objetivo da técnica de *stemming*, segundo Orengo e Huyck (2001), é reduzir palavras a uma forma comum de representação chamada de *stem*. Por exemplo, as palavras "apresentação", "apresentado" e "apresentando" podem ser representadas pelo *stem* "apresent". Os *stems* não possuem um significado linguístico relevante, mas capturam o significado da palavra, sem perder muitos detalhes.

Uma das grandes vantagens do uso de técnicas associadas à Figura 2.4, segundo Gonzalez e Lima (2003), é que elas, ao desconsiderar termos sem valor direto de significado, diminuem consideravelmente o tamanho do índice, possibilitando um aumento significativo no desempenho de buscas de documentos por sistemas de RI.

Neste trabalho, foram utilizadas as técnicas de *stemming* e remoção de *stopwors*. Uma das estratégias propostas neste trabalho (vide Subseção 3.2) para o aperfeiçoamento dos conjuntos de termos de gênero e conteúdo, fornecidos como parâmetros de entrada em processos de coleta temática baseada em gênero, envolve a representação de tais termos por seus respectivos *stems*. Ademais, em outra estratégia proposta (vide Subseção 3.1), foi necessária a remoção das *stopwords* para a construção da matriz de associação de termos, explicitada na Subseção 2.1.3.

2.1.5 Apache Lucene

Apache Lucene, de acordo com Sigrist e Higashino (2012), é um *framework* flexível, escalável e de alta performance utilizado para construção de mecanismos de busca. Por ser uma tecnologia gratuita, de código livre, implementada em Java e licenciada sob a liberal

Apache Software Licence, o *framework* permite uma fácil integração dos mecanismos de busca com aplicações Java de diversas naturezas (Sigrist e Higashino, 2012; McCandless et al., 2010). Além da versão original em Java, é oferecido suporte para outras linguagens de programação; dentre elas, Perl, Python, C++ e .NET.

Uma das principais preocupações dos desenvolvedores é esconder a complexidade dos processos de indexação e busca em uma *Application Programming Interface* (API) simples de utilizar, deixando os programadores livres para se ocuparem apenas com as regras de negócio da aplicação final. A Figura 2.5 apresenta os principais componentes do mecanismo de busca do Lucene.

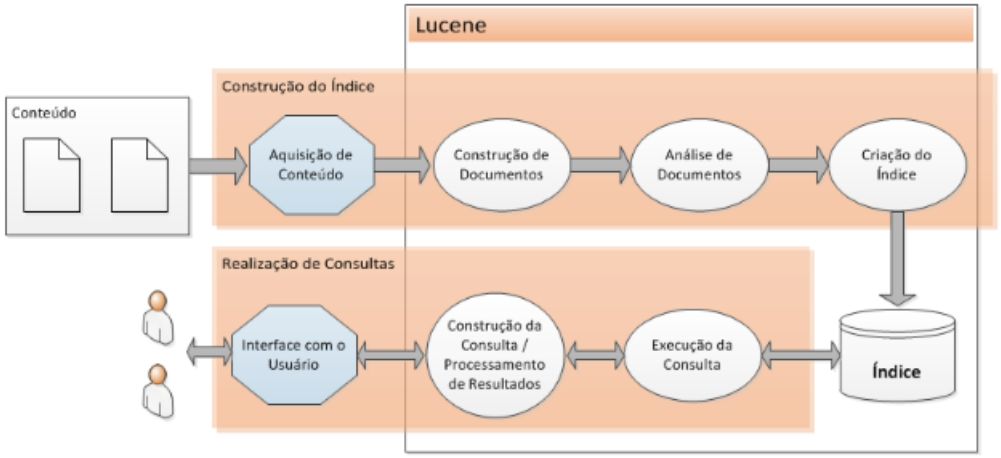

Figura 2.5: Componentes do mecanismo de busca do Lucene - Sigrist e Higashino (2012)

De acordo com a Figura 2.5, percebe-se que, em um sistema de busca simples, o *framework* atua, básica e diretamente, nas seções de indexação e consulta. Para construir o índice, é necessário fornecer algum conteúdo. Considerando, segundo McCandless et al. (2010), que o Lucene é capaz de indexar e tornar pesquisável qualquer tipo de dados que possa ser convertido em um formato textual, a aquisição de conteúdo pode ocorrer de diversas maneiras, dependendo da natureza da aplicação; uma possibilidade seria o uso de um coletor, que navega pela *Web* à procura de páginas a serem indexadas. Uma vez o conteúdo obtido, documentos padronizados são construídos a partir do mesmo, sendo compostos por um conjunto de campos, tais como autor, origem e o próprio conteúdo. Em seguida, com o objetivo de facilitar o processo de busca, esses documentos são analisados e processados. O processamento é simples e, geralmente, envolve a remoção de *stopwords*, conversão de letras maiúsculas em minúsculas e aplicação de técnicas de PLN. Ao fim desse processo, o resultado obtido é um conjunto de termos, que são as palavras que irão ser usadas na criação do índice. O índice, de acordo com McCandless et al. (2010), é a estrutura que irá permitir um eficiente acesso randômico às palavras nele contidas.

Ainda de acordo com a Figura 2.5, uma vez o índice é construído, documentos podem ser

acessados por meio de consultas. Para que a busca ocorra, o Lucene recebe como parâmetro um conjunto de palavras que representam uma dada necessidade de informação. Antes de ser buscado no índice, tal conjunto de palavras precisa passar pelo mesmo processamento que foi aplicado aos documentos no processo de indexação. Feito este processamento, os termos da consulta, oriundos do conjunto inicial de palavras, são buscados no índice afim de encontrar documentos onde eles aparecem. Terminada a busca, os resultados geralmente são ordenados por relevância e retornados ao usuário.

Neste trabalho, o Apache Lucene foi utilizado na construção do índice invertido a partir de um conjunto inicial de páginas da *Web*, de onde serão extraídos termos que, possivelmente, suplementarão os conjuntos originais de termos de gênero e conteúdo para um determinado processo de coleta. Logo, por meio do índice invertido construído e das estatísticas de frequência de ocorrência de termos oferecidas pela API do *framework*, foi possível construir matrizes de associação de termos, utilizadas em uma das estratégias propostas nesse trabalho (vide Subseção 3.1), tornando viável a seleção dos termos mais relevantes.

Existem outras ferramentas que oferecem funcionalidades similares às do Apache Lucene; dentre elas, pode-se citar Xapian[3] e XQEngine[4]. É difícil encontrar, na literatura, trabalhos recentes que avaliem e comparem o desempenho e funcionalidades das versões atuais dessas tecnologias; porém, como o Lucene é utilizado por grandes e conhecidas aplicações, como CiteSeerX, Eclipese IDE, MIT DSpace Federation e Tweeter Trends (Foundation, 2015), pode-se inferir que esse *framework* é eficiente e confiável.

2.2 Trabalhos relacionados

Existem, na literatura, diversos trabalhos que buscam aprimorar o conjunto de termos submetidos por usuários a sistemas de RI, mais especificamente a máquinas de busca. Esse aperfeiçoamento dos termos é motivado pelo fato de que consultas formuladas pelos usuários, na maioria das vezes, são curtas e não descrevem apropriadamente o tópico de interesse. Assim, com o objetivo de fazer com que sistemas de RI consigam recuperar um conjunto de documentos que melhor satisfaça a necessidade de informação do usuário, diversas técnicas são exploradas para a melhoria dos termos de consulta.

Uma solução bastante utilizada é a aplicação de técnicas relativas à expansão de consulta, que buscam reformular a consulta do usuário através da adição de um conjunto de novos termos relacionados com aqueles que foram originalmente estabelecidos. Nesse contexto, o trabalho desenvolvido por Chartree et al. (2013) explora um método automático de expansão de consulta fundamentado no uso de matrizes de associação. O método desenvolvido em tal trabalho foi testado em uma máquina de busca baseada no modelo vetorial da área de RI e

[3] https://xapian.org/
[4] http://xqengine.sourceforge.net/

obteve um aumento de eficácia de 14.3%, quando comparado à mesma máquina de busca sem a aplicação do método proposto de expansão de consulta.

Em Li et al. (2005), foi desenvolvida uma abordagem para expansão automática de consulta que combina o uso de *feedback* automático de relevância e consulta a *thesaurus*. Por considerar que *thesaurus* construídos manualmente são muito gerais ou muito específicos, os autores incorporaram a atualização automática do *thesaurus* ao processo de aperfeiçoamento dos termos de busca. A metodologia proposta em tal trabalho pode ser descrita de acordo com os seguintes passos:

- Passo 1: o usuário submete uma consulta Q_0 ao sistema de busca;
- Passo 2: por meio de consulta ao *thesaurus*, é gerada a consulta Q_1, que possui, além dos termos presentes em Q_0, uma lista de novos termos que estão relacionados aos termos da consulta original;
- Passo 3: a consulta Q_1 é submetida ao sistema de busca e os documentos recuperados são retornados para o usuário, ordenados de acordo com a relevância dos mesmos em relação aos termos de busca da consulta Q_1;
- Passo 4: em *background*, os documentos retornados ao usuário são analisados e as informações extraídas da análise são utilizadas para a atualização automática do *thesaurus*.

Para validar tal abordagem, foi feita uma análise comparativa entre o que foi desenvolvido e abordagens tradicionais de expansão de consulta. O resultado dessa análise mostrou que a abordagem proposta por Li et al. (2005) sobressaiu-se em termos de precisão e velocidade.

Outro trabalho, desenvolvido por Araujo e Pérez-Agüera (2008), propõe um método para expansão automática de consulta baseado em consulta a *thesaurus*, PLN e algoritmo genético. Tal método funciona da seguinte forma: dado um conjunto inicial de termos de consulta, é formado um novo conjunto que inclui os termos originais e outros termos a eles relacionados, derivados de um *thesaurus*; nesse novo conjunto de termos, é aplicada a técnica de *stemming*. Como o resultado das etapas anteriores pode levar a um conjunto grande de termos, é aplicado um algoritmo genético evolucionário para realizar a seleção dos termos que, efetivamente, farão parte da nova consulta. Os experimentos conduzidos para validação do método compararam o mesmo com um método proposto por Porter, baseado em *stemming* e considerado *baseline* na área. Nessa validação, verificou-se que o método proposto por Araujo e Pérez-Agüera (2008) obteve um ganho em precisão de 15.20%, comparado ao *baseline*.

Assim como os trabalhos citados, como já mencionado, existem outras abordagens para melhoria dos termos utilizados como dados entrada em sistemas de RI. Neste trabalho de monografia, tais abordagens foram estudadas e analisadas, tornando possível a proposta e desenvolvimento das estratégias (vide Seções 3.1 e 3.2) para o aperfeiçoamento automático

dos conjuntos de termos de gênero e conteúdo utilizados em processos de coleta temática seguindo a abordagem baseada em gênero.

Capítulo 3

Desenvolvimento

Como já mencionado, este trabalho possui, como objetivo geral, a definição de uma estratégia para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo, a ser utilizada em processos de coleta temática de páginas da *Web* seguindo a abordagem para coleta temática baseada em gênero. Para tanto foram propostas duas estratégias que se encontram descritas nas subseções seguintes.

3.1 Estratégia baseada em matriz de associação de termos

A estratégia baseada em matriz de associação de termos, como a própria definição diz, visa aperfeiçoar os conjuntos de termos de gênero e conteúdo por meio da aplicação de uma técnica de expansão de consulta automática baseada no uso de matriz de associação (vide Subseção 2.1.3). A Figura 3.1 ilustra a arquitetura de funcionamento desta estratégia.

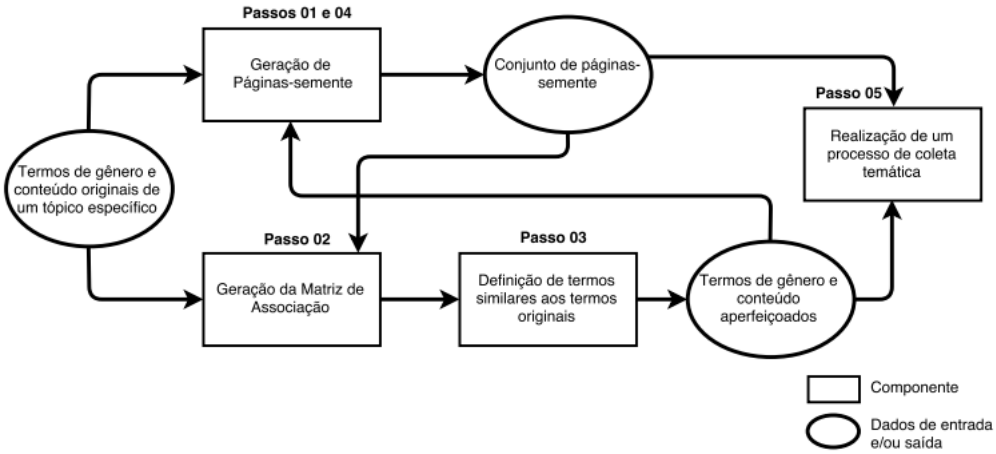

Figura 3.1: Arquitetura da estratégia baseada em matriz de associação.

Como é possível observar na Figura 3.1, o funcionamento dessa estratégia segue os seguintes passos:

- Passo 01: a partir dos conjuntos originais de termos de gênero e conteúdo, é ativado o componente "Geração de Páginas-semente" (vide Subseção 2.1.1.2), pela primeira vez, no intuito de gerar páginas que serão utilizadas pelo componente "Geração da Matriz de Associação".

- Passo 02: a partir do conjunto de páginas-semente gerado pelo Passo 01 e dos conjuntos originais de termos de gênero e conteúdo, é ativado o componente "Geração da Matriz de Associação" no intuito de gerar a matriz de associação de termos;

- Passo 03: a partir da matriz de associação gerada, é ativado o componente "definição de termos similares aos termos originais", responsável em definir os termos que irão suplementar os conjuntos originais no intuito de estabelecer os conjuntos aperfeiçoados de termos de gênero e conteúdo.

- Passo 04: a partir dos conjuntos aperfeiçoados de termos, é ativado o componente "Geração de Páginas-semente" (Subseção 2.1.1.2), pela segunda vez, no intuito de gerar as páginas-semente necessárias a um processo de coleta.

- Passo 05: a partir do conjunto de páginas-semente gerado pelo Passo 04 e dos conjuntos aperfeiçoados de termos definidos no Passo 03, é ativada a abordagem para coleta temática baseada em gênero, ou seja, o processo desejado de coleta a partir do Passo 01 da Figura 2.1.

Em relação ao Passo 2, a matriz de associação gerada pela estratégia, conforme descrito na Subseção 2.1.3, é uma estrutura bidimensional utilizada para mensurar a similaridade s_{ij} entre dois termos t_i e t_j. Além das métricas citadas na Subseção 2.1.3 para calcular s_{ij}, que consideram a frequência de ocorrência dos termos nas páginas fornecidas como entrada, foi definida outra métrica para o cálculo de s_{ij}, que considera um limite de distância entre os termos t_i e t_j. Tal nova métrica, denominada MenorDistância, consiste em calcular a similaridade s_{ij} entre dois termos t_i e t_j por meio das seguintes equações:

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

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

onde:

- s_{ij} é a similaridade entre os termos t_i e t_j, de acordo com a métrica da MenorDistância, dada pela soma normalizada das menores distâncias entre os termos t_i e t_j, considerando todas as páginas p que possuem tais termos;

- 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;
- $d_{i,j}$ 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 $k + 1$.

3.2 Estratégia baseada em PLN

A estratégia baseada em PLN visa aperfeiçoar os conjuntos de termos de gênero e conteúdo por meio da aplicação de técnicas de PLN. A Figura 3.2 ilustra a arquitetura de funcionamento desta estratégia.

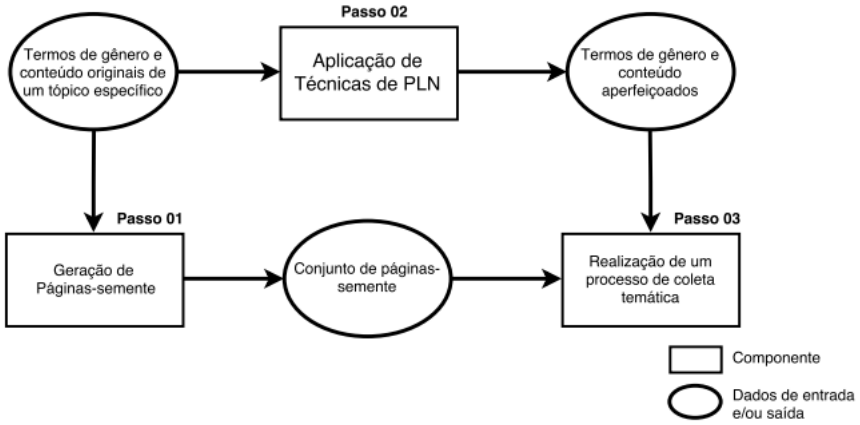

Figura 3.2: Arquitetura da estratégia baseada em PLN.

O funcionamento desta estratégia ocorre, de acordo com a Figura 3.2, por meio dos seguintes passos:

- Passo 01: a partir dos conjuntos originais de termos de gênero e conteúdo, é ativado o componente "Geração de Páginas-semente" (vide Subseção 2.1.1.2), no intuito de gerar o conjunto de páginas-semente necessário para se iniciar um processo de coleta temática.
- Passo 02: a partir dos conjuntos originais de termos de gênero e conteúdo, é ativado o componente "Aplicação de Técnicas de PLN" responsável em aplicar técnicas de PLN sobre os conjuntos originais de termos no intuito de estabelecer os conjuntos aperfeiçoados de termos.
- Passo 03: a partir do conjunto de páginas-semente gerado pelo Passo 01 e dos conjuntos aperfeiçoados de termos definidos no Passo 02, é ativada a abordagem para coleta te-

mática baseada em gênero, ou seja, o processo de coleta a partir do Passo 01 da Figura 2.1.

Em relação ao Passo 02 da Figura 3.2, dentre as possíveis técnicas de PLN (vide Subseção 2.1.4), serão aplicadas a remoção de *stopwords* e a técnica de *stemming*. Em tal contexto, a remoção das *stopwords* consiste em eliminar dos termos originais de gênero e conteúdo as palavras que possuem pouca relevância para descrever o tópico de interesse. Quanto à aplicação de *stemming*, o aperfeiçoamento dá-se por meio da representação dos termos originais de gênero e conteúdo pelos seus respectivos *stems*. Adicionalmente, é aplicada uma técnica de pluralização, que consiste em realizar uma flexão de número nos conjuntos de termos originais, fazendo com que todos os termos sejam representados no plural.

Capítulo 4

Experimentos

Neste capítulo, são apresentados e analisados os experimentos de validação das estratégias propostas para o aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da *Web* baseada em gênero (De Assis et al., 2009). A Seção 4.1 descreve as métricas que foram utilizadas para avaliar a eficácia da aplicação das estratégias propostas no Capítulo 3. A Seção 4.2 descreve o *baseline* utilizado para a avaliação das estratégias propostas. A Seção 4.3 descreve os experimentos realizados que utilizaram as estratégias propostas neste trabalho. Por fim, a Seção 4.4 apresenta e avalia os resultados obtidos por meio dos experimentos realizados.

4.1 Métricas de avaliação

Para avaliação dos experimentos realizados, foram utilizadas as métricas de 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 coletor temático baseado em gênero; ou seja:

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

onde:

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

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:

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

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 coletor.

Finalmente, também definido por Manning et al. (2008), F1 corresponde à média harmônica entre as métricas de precisão e de revocação; ou seja:

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

4.2 *Baseline*

Para que fosse possível validar as estratégias propostas (vide Capítulo 3) que visam o aperfeiçoamento automático dos conjuntos de termos utilizados em processos 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 coleta temática sem aperfeiçoamento de termos, considerando o tópico de interesse proposto por Mangaravite et al. (2014). 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%3Fdownload%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: 5000.

Ao longo da execução do processo de coleta mencionado, foi armazenado um *log* contendo as seguintes informações sobre cada página da *Web* 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, verificou-se que foram visitadas 4026 páginas válidas. Uma vez obtidas as páginas, para que fosse possível estabelecer as métricas descritas na Seção 4.1, foi gerado um gabarito contendo 137 páginas relevantes, dentre as páginas visitadas pelo coletor, e também foi estabelecido o limite de similaridade ótimo para o processo de coleta, de forma empírica, visando a maximização da métrica F1. Os melhores valores obtidos para as métricas encontram-se na Tabela 4.2.

Por meio do limite de similaridade ótimo apresentado na Tabela 4.2, no caso 0.377, foram obtidas um total de 98 páginas que possuem similaridade maior ou igual a tal limite ótimo; ou seja, 98 páginas foram consideradas relevantes para o tópico de interesse desejado, dentre as 4026 páginas visitadas.

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, revocação e F1 obtidos como *baseline*

Limite	**Precisão**	**Revocação**	**F1**
0.377	0.667	0.511	0.579

4.3 Descrição dos experimentos

Para realizar os experimentos de validação das estratégias propostas para o aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da *Web* seguindo a abordagem baseada em gênero, foram simulados processos de coleta utilizando o *log* relativo à coleta feita para a obtenção do *baseline* (vide Seção 4.2). Em tais processos simulados de coleta, as estratégias apresentadas no Capítulo 3 foram aplicadas 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*.

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 das estratégias propostas 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 de 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. Para a definição dos segundos conjuntos, 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 Ciência da Computação arbitrariamente escolhido.

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
disciplina	entidade relacionamento
ementa	modelo conceitual
objetivos	algebra relacional
provas	sql
referencias bibliograficas	silberschatz
bibliografia	navathe

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	-

Para tornar possível a análise comparativa dos resultados relativos a aplicação das estratégias propostas sobre os novos conjuntos de termos de gênero e conteúdo, fez-se necessário

estabelecer valores de referência para as métricas descritas na Seção 4.1 para os casos em que os novos conjuntos de termos foram utilizados para determinar a similaridade entre as páginas da *Web* visitadas e o tópico de interesse, seguindo a abordagem original, ou seja, sem aperfeiçoamento de termos. Tais valores foram obtidos por meio da simulação 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 feitas tais simulações, as métricas de avaliação foram obtidas de forma análoga ao que foi feito para a obtenção das métricas relativas ao *baseline*. 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. Os melhores valores obtidos, utilizados como referência para os experimentos em que as estratégias propostas foram aplicadas sobre os conjuntos apresentados nas tabelas 4.3 e 4.4, estão apresentados, respectivamente, nas tabelas 4.5 e 4.6.

Tabela 4.5: Precisão, revocação e F1 obtidos como referência para o processo simulado de coleta envolvendo os conjuntos reduzidos de termos

Limite	**Precisão**	**Revocação**	**F1**
0.479	0.661	0.526	0.585

Tabela 4.6: Precisão, revocação e F1 obtidos como referência para o processo simulado de coleta envolvendo os conjuntos de termos definidos por um não especialista

Limite	**Precisão**	**Revocação**	**F1**
0.556	0.615	0.409	0.491

Uma vez obtidos os valores utilizados como referência, foram realizados experimentos envolvendo a simulação de processos de coleta temática seguindo a abordagem baseada em gênero e considerando as estratégias propostas para o aperfeiçoamento de termos. Nesses experimentos, para cada estratégia proposta, além da realização de distintas simulações 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 a cada estratégia. Ao fim da execução de cada um desses processos de coleta, utilizando o gabarito referente às páginas da *Web* consideradas, foi estabelecido o limite de similaridade ótimo, de forma empírica, visando a maximização da métrica F1. Os resultados referentes aos processos simulados de coleta encontram-se apresentados na Seção 4.4.

As Subseções 4.3.1 e 4.3.2 descrevem particularidades referentes, respectivamente, aos experimentos relativos à estratégia baseada em matriz de associação de termos e aos experimentos relativos à estratégia baseada em PLN.

4.3.1 Experimentos relativos à estratégia baseada em matriz de associação de termos

A estratégia baseada em matriz de associação de termos, conforme descrito na Seção 3.1, utiliza as páginas-semente obtidas ao início do processo de coleta temática baseada em gênero para gerar a matriz de associação de termos. Para a construção da matriz, é necessário especificar o número de páginas-semente a ser considerado e a métrica a ser utilizada para mensurar a similaridade entre os termos nela contidos. Além disso, uma vez gerada a matriz de associação, é necessário definir quantos termos serão selecionados, em relação a cada termo dos conjuntos originais de termos de gênero e conteúdo, para compor o conjunto aperfeiçoado de termos.

Definidos todos os parâmetros, ao fim do Passo 03 da Figura 3.1, são gerados os conjuntos aperfeiçoados de termos de gênero e conteúdo, tornando possível a simulação de um processo de coleta temática, onde a relevância de cada página da *Web*, contida no *log* da coleta relativa ao *baseline*, é determinada pela similaridade entre tal página e os conjuntos de termos aperfeiçoados.

Assim, por meio da variação de parâmetros, foram realizadas distintas simulações de processos de coleta temática seguindo a abordagem baseada em gênero com a aplicação da estratégia baseada em matriz de associação de termos. Tais simulações foram feitas de acordo com a combinação dos seguintes parâmetros:

- Considerando o coeficiente de Jaccard para o cálculo da similaridade entre termos durante a construção da matriz de associação de termos:
  - conjuntos de termos de gênero e conteúdo originais: conjuntos apresentados nas Tabelas 4.1, 4.3 e 4.4;
  - número de páginas-semente utilizadas na construção da matriz de associação de termos: variando de 1 a 9 páginas-semente;
  - número de termos selecionados para compor os conjuntos de termos aperfeiçoados: variando de 1 a 10 termos de expansão para cada termo original.
- Considerando a métrica MenorDistância para o cálculo da similaridade entre termos durante a construção da matriz de associação de termos:
  - conjuntos de termos de gênero e conteúdo originais: conjuntos apresentados nas tabelas 4.1, 4.3 e 4.4;
  - número de páginas semente utilizadas na construção da matriz de associação de termos: 1, 3, 5 e 9 páginas-semente;
  - número de termos selecionados para compor os conjuntos de termos aperfeiçoados: 1, 2, 5 e 10 termos de expansão para cada termo original;

– k - valor máximo de distância entre termos (vide Seção 3.1): 1, 2, 5 e 10 termos.

4.3.2 Experimentos relativos à estratégia baseada em PLN

A estratégia baseada em PLN, conforme descrito na Seção 3.2, realiza a aplicação de técnicas de PLN sobre os conjuntos originais de termos de gênero e conteúdo para estabelecer os conjuntos aperfeiçoados de termos. Uma vez obtidos os conjuntos aperfeiçoados, é iniciado o Passo 03 da Figura 3.2, ou seja, o processo de coleta temática baseada em gênero. Durante a execução de tal processo, para tornar possível o cálculo da similaridade entre os conjuntos aperfeiçoados de termos de gênero e conteúdo e as páginas da *Web*, as mesmas técnicas de PLN, aplicadas sobre os conjuntos originais de termos, foram aplicadas sobre o conteúdo das páginas da *Web* visitadas pelo coletor.

Assim, por meio da variação dos conjuntos originais de termos de gênero e conteúdo e das técnicas de PLN consideradas, foram realizados distintos processos simulados de coleta temática seguindo a abordagem baseada em gênero com a aplicação da estratégia baseada em técnicas de PLN. Tais simulações foram feitas de acordo com a combinação dos seguintes parâmetros:

- conjuntos de termos de gênero e conteúdo originais: conjuntos apresentados nas Tabelas 4.1, 4.3 e 4.4;
- técnicas de PLN: todas as combinações possíveis entre as técnicas *stemming*, remoção de *stopwords* e pluralização.

4.4 Resultados

Nesta seção, são apresentados e analisados os resultados obtidos por meio da experimentação prática realizada, envolvendo as estratégias propostas para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo, utilizados como dados de entrada em processos de coleta temática de páginas da *Web* baseada em gênero. Considerando as estratégias propostas e as possíveis variações de parâmetros, conforme descritas na Seção 4.3, foram realizados 531 processos simulados de coleta temática. Desse total de simulações, 510 consideram a estratégia baseada em matriz de associação, sendo 270 referentes aos casos em que o coeficiente de Jaccard foi utilizado na construção da matriz de associação e 240 referentes aos casos em que foi considerada a métrica MenorDistância. Os 21 experimentos restantes referem-se å aplicação da estratégia baseada em PLN.

Todos resultados referentes aos 531 processos simulados de coleta foram organizados de acordo com a estratégia proposta de aperfeiçoamento considerada e com os conjuntos de termos de gênero e conteúdo utilizados. Dessa forma, os resultados foram estruturados em tabelas onde, para cada simulação feita, foram apresentados os parâmetros utilizados, as métricas de

avaliação descritas na Seção 4.1 e o limite de similaridade ótimo obtido. 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 da simulação do processo de coleta sem aperfeiçoamento de termos. A partir deste "Caso 0", foi possível incluir, para cada simulação feita, um valor percentual representando a distância da eficácia obtida (métrica F1) em relação à eficácia apresentada no "Caso 0". As tabelas contendo todos os casos de teste considerados (processos simulados de coleta) encontram-se nos Apêndices A, B e C.

As subseções seguintes apresentam e analisam os melhores resultados obtidos (distância em relação ao "Caso 0" maior ou igual a 85%), de acordo com os conjuntos de termos de gênero e conteúdo considerados. Assim, a Seção 4.4.1 apresenta os resultados obtidos considerando como conjuntos originais os termos definidos por Mangaravite et al. (2014) e apresentados na Tabela 4.1. A Seção 4.4.2 apresenta os resultados obtidos ao utilizar os conjuntos reduzidos de termos, apresentados na Tabela 4.3. Por fim, a Seção 4.4.3 apresenta os resultados obtidos considerando os conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista, apresentados na Tabela 4.4.

4.4.1 Resultados considerando os conjuntos completos de termos utilizados no *baseline*

Nesta subseção, são analisados os melhores resultados relativos à aplicação das estratégias propostas neste trabalho utilizando os conjuntos de termos de gênero e conteúdo (vide Tabela 4.1) definidos por Mangaravite et al. (2014). As Tabelas 4.7, 4.8 e 4.9 apresentam tais melhores resultados ao aplicar, para aperfeiçoamento de termos respectivamente, a estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard, a estratégia baseada em matriz de associação utilizando a métrica MenorDistância e a estratégia baseada em PLN.

Por meio da análise da Tabela 4.7, percebe-se que houve 3 casos (Casos 8, 9 e 10) em que o uso da estratégia resultou em uma melhora (aumento máximo de 1,962%) da métrica F1. Nota-se que, nesses casos, foi considerada apenas uma página semente e um número elevado de termos de expansão para cada termo original. Ademais, é possível destacar que, para obtenção dos melhores resultados, de uma forma geral, a quantidade de páginas-semente utilizada para a construção da matriz deve ser pequena (no máximo 4) e que, a medida que se aumenta tal quantidade, o número de termos de expansão deve ser pequeno.

Tabela 4.7: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos completos de termos de gênero e conteúdo

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,377	-	-	0,667	0,511	0,579	100,000
1	0,388	1	1	0,851	0,416	0,559	96,597
2	0,379		2	0,851	0,416	0,559	96,597
3	0,371		3	0,851	0,416	0,559	96,597
4	0,359		4	0,800	0,438	0,566	97,844
5	0,357		5	0,853	0,423	0,566	97,812
6	0,350		6	0,853	0,423	0,566	97,812
7	0,344		7	0,841	0,423	0,563	97,337
8	0,334		8	0,800	0,467	0,590	101,962
9	0,323		9	0,686	0,511	0,586	101,255
10	0,324		10	0,673	0,511	0,581	100,415
11	0,367	2	1	0,811	0,438	0,569	98,307
12	0,351		2	0,831	0,431	0,567	98,063
13	0,337		3	0,779	0,438	0,561	96,929
21	0,350	3	1	0,611	0,482	0,539	93,131
31	0,357	4	1	0,709	0,445	0,547	94,568

Em relação à Tabela 4.8, nota-se que, apesar de não haver melhora da métrica F1 em nenhum dos dos casos, houve um aumento em relação à precisão na maioria deles, chegando a ficar, no Caso 25, 40,8% maior que o *baseline*. Ainda em relação à Tabela 4.8, observa-se que, particularmente no Caso 17, onde foram consideradas 2 páginas-semente, 1 termo de expansão por termo original e distância máxima de 1 termo, a métrica F1 ficou bem próxima ao valor obtido no *baseline*.

Tabela 4.8: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância e os conjuntos completos de termos de gênero e conteúdo

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,377	-	-	-	0,667	0,511	0,579	100,000
2	0,368	1	1	2	0,659	0,394	0,493	85,245
3	0,368			5	0,659	0,394	0,493	85,245
4	0,368			10	0,659	0,394	0,493	85,245
9	0,342		5	1	0,626	0,416	0,500	86,429
17	0,372	2	1	1	0,845	0,438	0,577	99,725
18	0,356			2	0,660	0,467	0,547	94,554
19	0,359			5	0,621	0,467	0,533	92,190
20	0,359			10	0,621	0,467	0,533	92,190
21	0,370		2	1	0,904	0,343	0,497	85,971
22	0,342			2	0,714	0,474	0,570	98,559
23	0,354			5	0,831	0,431	0,567	98,063
24	0,344			10	0,699	0,474	0,565	97,702
25	0,349		5	1	0,939	0,336	0,495	85,499
54	0,327	5	2	2	0,806	0,423	0,555	95,940
55	0,318			5	0,674	0,438	0,531	91,783
56	0,318			10	0,674	0,438	0,531	91,783
57	0,302		5	1	0,545	0,526	0,535	92,533
58	0,285			2	0,508	0,482	0,494	85,457

Por fim, ao analisar a Tabela 4.9, observa-se que a aplicação de técnicas de PLN sobre os conjuntos de termos originais não apresentou resultados significativos em nenhum dos casos.

Tabela 4.9: Melhores resultados relativos à estratégia baseada em PLN utilizando os conjuntos completos de termos de gênero e conteúdo

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,377	-	-	-	0,667	0,511	0,579	100,000
1	0,350			x	0,486	0,518	0,502	86,734
2	0,367		x		0,645	0,438	0,522	90,186
3	0,350		x	x	0,493	0,511	0,502	86,738
5	0,375	x		x	0,667	0,423	0,518	89,515
7	0,377	x	x	x	0,667	0,423	0,518	89,515

4.4.2 Resultados considerando os conjuntos reduzidos de termos derivados do *baseline*

Nesta subseção, de forma similar ao que foi feito para a Seção 4.4.1, serão analisados os melhores resultados obtidos por meio da aplicação das estratégias propostas sobre os conjuntos reduzidos de termos de gênero e conteúdo (vide Tabela 4.3) derivados do *baseline*. As Tabelas 4.10, 4.11 e 4.12 apresentam os melhores resultados obtidos ao aplicar, para aperfeiçoamento de termos respectivamente, a estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard, a estratégia baseada em matriz de associação utilizando a métrica MenorDistância e a estratégia baseada em PLN.

Por meio da análise das Tabelas 4.10 e 4.12, nota-se que, ao aplicar a estratégia baseada em matriz de associação utilizando o coeficiente de Jaccard e a estratégia baseada em PLN, não houve melhorias das métricas F1 e precisão em nenhum dos casos. Em contrapartida, pela análise da Tabela 4.11 é possível perceber que a aplicação da estratégia baseada em matriz de associação, utilizando a métrica MenorDistância, produziu resultados superiores ao apresentado pelo Caso 0 para os casos em que foram consideradas 2 páginas-sementes para a geração da matriz de associação, com distância máxima k igual a 2, 5 ou 10, e 1 ou 2 termos de expansão por termo original. Particularmente, os melhores resultados foram obtidos pelos Casos 19 e 20: métrica F1 5,24% maior que o obtido sem a aplicação da técnica. Novamente, os melhores resultados produzidos por meio da aplicação desta estratégia, de uma forma geral, compreendem os casos em que foram utilizados valores baixos para o número de páginas-semente e termos de expansão considerados.

Tabela 4.10: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos reduzidos de termos de gênero e conteúdo

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,479	-	-	0,661	0,526	0,585	100,000
1	0,451	1	1	0,577	0,547	0,562	95,974
2	0,425		2	0,532	0,547	0,540	92,176
3	0,403		3	0,503	0,547	0,524	89,598
4	0,384		4	0,481	0,569	0,522	89,130
5	0,368		5	0,443	0,569	0,498	85,144
8	0,342		8	0,493	0,547	0,519	88,668
9	0,347		9	0,573	0,460	0,510	87,146
11	0,423	2	1	0,600	0,569	0,584	99,813
12	0,391		2	0,580	0,555	0,567	96,891
61	0,462	7	1	0,636	0,409	0,498	85,037
81	0,423	9	1	0,640	0,416	0,504	86,173
85	0,413		5	0,579	0,482	0,526	89,841

Tabela 4.11: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando a Métrica MenorDistância e os conjuntos reduzidos de termos de gênero e conteúdo

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,479	-	-	-	0,661	0,526	0,585	100,000
1	0,420	1	1	1	0,545	0,489	0,515	88,045
6	0,420		2	2	0,663	0,401	0,500	85,417
9	0,386		5	1	0,667	0,409	0,507	86,576
10	0,371			2	0,659	0,409	0,505	86,186
17	0,416	2	1	1	0,660	0,453	0,537	91,703
18	0,420			2	0,692	0,540	0,607	103,620
19	0,426			5	0,730	0,533	0,616	105,239
20	0,426			10	0,730	0,533	0,616	105,239
22	0,395		2	2	0,716	0,533	0,611	104,358
23	0,394			5	0,705	0,540	0,612	104,477
24	0,394			10	0,705	0,540	0,612	104,477
58	0,316	5	5	2	0,637	0,474	0,544	92,922

Tabela 4.12: Melhores resultados relativos à estratégia baseada em PLN utilizando os conjuntos reduzidos de termos de gênero e conteúdo

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,479	-	-	-	0,661	0,526	0,585	100,000
1	0,440			x	0,551	0,547	0,549	93,864
2	0,447		x		0,600	0,526	0,560	95,720
3	0,442		x	x	0,585	0,526	0,554	94,615
4	0,439	x			0,503	0,533	0,518	88,446
5	0,444	x		x	0,514	0,555	0,533	91,111
6	0,434	x	x		0,473	0,577	0,520	88,788
7	0,445	x	x	x	0,531	0,555	0,543	92,738

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

Nesta subseção, serão apresentados os melhores resultados obtidos por meio da aplicação das técnicas proposta para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo sobre os conjuntos de termos definidos por um usuário não especialista. As Tabelas 4.13, 4.14 e 4.15 apresentam os melhores obtidos ao aplicar, para aperfeiçoamento de termos respectivamente, a estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard, a estratégia baseada em matriz de associação utilizando a métrica MenorDistância e a estratégia baseada em PLN.

Analisando a Tabela 4.13, percebe-se que apenas o Caso 50, onde foram consideradas 5 páginas-semente para a geração da matriz de associação e selecionados 10 termos de expansão para cada termo original, apresentou valor de F1 maior que o obtido no Caso 0: valor de F1 4,15% maior que o valor de referência.

Em relação à Tabela 4.14, os Casos 6, 11, 16, 31 e 32 apresentaram valores de F1 acima do valor considerado como referência. O melhor resultado, obtido no Caso 31, considerou 2 páginas-semente e uma distância máxima de 5 termos, selecionando 10 termos de expansão para cada termo original; em tal caso, o valor de F1 obtido foi 6,79% melhor que o do Caso 0, utilizado como referência.

Em relação a Tabela 4.15, novamente, a estratégia baseada em PLN não apresentou nenhum resultado melhor que o valor de referência apresentado no Caso 0.

Adicionalmente, em relação ao Caso 31 da Tabela 4.14, que foi o melhor resultado obtido dentre os 521 experimentos realizados, a Tabela 4.16 apresenta os conjuntos aperfeiçoados de termos de gênero e conteúdo gerados para tal caso. Por meio da análise desta tabela, é possível perceber que a estratégia retornou alguns termos significativos para o tópico de interesse, como, por exemplo, os termos “relacional”, “conceitual”, “modelagem” e “entidade”

que também foram utilizados por Mangaravite et al. (2014) para descrever o tópico "Ementas de disciplinas de Banco de Dados". Em contrapartida, percebe-se, também, a presença de termos não relacionados ao tópico de interesse, como, por exemplo, “usados”, “wikipedia”, “cada” e “anterior”.

Tabela 4.13: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,556	-	-	0,615	0,409	0,491	100,000
1	0,516	1	1	0,482	0,387	0,429	87,363
8	0,355		8	0,419	0,453	0,435	88,571
9	0,372		9	0,539	0,350	0,425	86,473
11	0,487	2	1	0,602	0,365	0,455	92,532
12	0,438		2	0,521	0,358	0,424	86,364
26	0,441	3	6	0,420	0,423	0,422	85,870
27	0,438		7	0,463	0,409	0,434	88,372
36	0,446	4	6	0,414	0,423	0,419	85,250
41	0,539	5	1	0,509	0,409	0,453	92,308
44	0,510		4	0,645	0,358	0,460	93,662
45	0,495		5	0,491	0,401	0,442	89,931
46	0,491		6	0,529	0,401	0,456	92,916
47	0,489		7	0,645	0,358	0,460	93,662
48	0,472		8	0,529	0,394	0,452	91,990
49	0,471		9	0,646	0,387	0,484	98,532
50	0,457		10	0,705	0,401	0,512	104,153
51	0,540	6	1	0,598	0,358	0,447	91,096
53	0,515		3	0,712	0,307	0,429	87,245
59	0,468		9	0,656	0,307	0,418	85,075
60	0,478		10	0,774	0,299	0,432	87,857
61	0,578	7	1	0,800	0,321	0,458	93,304
62	0,489		2	0,535	0,394	0,454	92,377
68	0,485		8	0,698	0,321	0,440	89,571
69	0,489		9	0,746	0,321	0,449	91,399
70	0,486		10	0,759	0,321	0,451	91,868
71	0,585	8	1	0,880	0,321	0,471	95,798
72	0,518		2	0,750	0,307	0,435	88,601
73	0,483		3	0,485	0,350	0,407	82,809
75	0,480		5	0,588	0,343	0,433	88,183
76	0,481		6	0,605	0,336	0,432	87,928
77	0,457		7	0,529	0,394	0,452	91,990
78	0,497		8	0,804	0,299	0,436	88,792
79	0,494		9	0,851	0,292	0,435	88,509
80	0,488		10	0,788	0,299	0,434	88,322
81	0,549	9	1	0,759	0,321	0,451	91,868
85	0,461		5	0,411	0,438	0,424	86,320

Tabela 4.14: Melhores resultados relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância e os conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,556	-	-	-	0,615	0,409	0,491	100,000
1	0,468	1	1	1	0,432	0,511	0,468	95,318
5	0,425		2	1	0,447	0,518	0,480	97,659
6	0,441			2	0,476	0,511	0,493	100,352
7	0,429			5	0,358	0,533	0,428	87,160
8	0,429			10	0,358	0,533	0,428	87,160
9	0,381		5	1	0,457	0,423	0,439	89,448
10	0,375			2	0,397	0,547	0,460	93,668
11	0,402			5	0,504	0,489	0,496	101,032
12	0,404			10	0,535	0,445	0,486	98,947
15	0,346		10	5	0,442	0,504	0,471	95,880
16	0,370			10	0,629	0,445	0,521	106,136
18	0,493	2	1	2	0,646	0,372	0,472	96,131
19	0,478			5	0,495	0,372	0,425	86,518
20	0,478			10	0,495	0,372	0,425	86,518
21	0,407		2	1	0,435	0,438	0,436	88,831
22	0,450			2	0,492	0,445	0,467	95,156
23	0,437			5	0,509	0,409	0,453	92,308
24	0,428			10	0,500	0,416	0,454	92,459
27	0,385		5	5	0,525	0,460	0,490	99,805
28	0,376			10	0,538	0,409	0,465	94,606
31	0,336		10	5	0,598	0,467	0,525	106,792
32	0,338			10	0,610	0,445	0,515	104,792
54	0,380	5	2	2	0,397	0,453	0,423	86,153
57	0,337		5	1	0,462	0,438	0,449	91,493
77	0,316		10	1	0,470	0,394	0,429	87,245

Tabela 4.15: Melhores resultados relativos à estratégia baseada em PLN utilizando os conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,556	-	-	-	0,615	0,409	0,491	100,000
2	0,541		x		0,568	0,394	0,466	94,766

Tabela 4.16: Conjuntos aperfeiçoados de termos de gênero e conteúdo gerados para o Caso 31 da Tabela 4.14

Termos de gênero		Termos de conteúdo	
whatshotwhatshot	leituras	dml	relacional
indexação	sld	indexação	conceitual
conteudo	violarem	linguagem	insert
slides	explicações	dados	medidas
relacional	modelo	sistema	sql
linguagem	volume	questions	implementação
horario	cada	commit	world
aula	curso	kevin	sido
prova	ementa	implementacao	identificar
usados	york	inserir	custo
wesley	materia	criar	structured
anterior	plano	gerência	modelagem
sql	recursos	álgebra	ensino
world	-	banco	consulta
cronograma	-	enviado	objetos
presença	-	modelo	análise
correspondente	-	sintaxe	editora
visões	-	entidade	modelos
wikipedia	-	relacionamento	ddl
zip	-	curso	home
atividades	-	plano	estendido
índice	-	controlar	sistemas
disponíveis	-	produzidos	york
índices	-	dividivo	relacionais
álgebra	-	luis	pdf
utilizar	-	conceituar	-

Capítulo 5

Conclusão

Neste trabalho, foram propostas e experimentadas estratégias 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 em (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 2009). 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 a estratégia baseada em matriz de associação de termos que utiliza a métrica MenorDistância, proposta neste trabalho, foi a que apresentou melhores resultados quando comparada às outras estratégias experimentadas. Vale ressaltar que o melhor resultado foi obtido por meio da aplicação de tal estratégia sobre conjuntos de termos de gênero e conteúdo definidos por um usuário não especialista para o tópico de interesse "Ementas de disciplinas de Banco de Dados", sugerindo que um aperfeiçoamento automático de termos é mais significativo quando os conjuntos de termos de gênero e conteúdo necessário para descrever um tópico de interesse não foram adequadamente definidos.

Ainda, apesar de tal estratégia ter se sobressaído sobre as demais estratégias que foram propostas e experimentadas neste trabalho, a melhoria apresentada por ela não foi tão satisfatória, uma vez que o melhor resultado promoveu um aumento na métrica F1 de apenas 6,79% ao se comparar com o valor de F1 obtido pelo processo de coleta cujos termos de gênero e conteúdo não foram expandidos. Uma possível explicação para a pequena melhora pode ser concebida por meio da análise da Tabela 4.16, onde se nota que, ao realizar o aperfeiçoamento automático dos conjuntos de termos, os termos de gênero e conteúdo acabam se misturando. Tal fato pode influenciar diretamente na eficácia dos processos de coleta, uma vez que, de acordo com (De Assis et al., 2007; de Assis et al., 2008; De Assis et al., 2009), a abordagem para coleta temática baseada em gênero é útil em situações onde um tópico de interesse possa ser expresso por meio de dois conjuntos distintos de termos: o primeiro descrevendo aspectos

de gênero das páginas desejadas, e o segundo referente ao assunto ou conteúdo descrito nessas páginas.

Como perspectivas de trabalho futuro, pretende-se: (1) realizar novos experimentos de validação das estratégias já propostas, por meio da execução de processos de coleta temática envolvendo outros tópicos de interesse; (2) aperfeiçoar as estratégias já propostas e/ou desenvolver novas estratégias para o aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo; (3) executar novos processos de coleta, considerando a melhor estratégia para aperfeiçoamento de termos, em casos práticos no intuito de se verificar e melhorar a precisão.

Apêndice A

Resultados dos experimentos relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard

Tabela A.1: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos completos de termos de gênero e conteúdo (vide Tabela 4.1)

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,377	-	-	0,667	0,511	0,579	100,000
1	0,388	1	1	0,851	0,416	0,559	96,597
2	0,379		2	0,851	0,416	0,559	96,597
3	0,371		3	0,851	0,416	0,559	96,597
4	0,359		4	0,800	0,438	0,566	97,844
5	0,357		5	0,853	0,423	0,566	97,812
6	0,350		6	0,853	0,423	0,566	97,812
7	0,344		7	0,841	0,423	0,563	97,337
8	0,334		8	0,800	0,467	0,590	101,962
9	0,323		9	0,686	0,511	0,586	101,255
10	0,324		10	0,673	0,511	0,581	100,415
11	0,367	2	1	0,811	0,438	0,569	98,307
12	0,351		2	0,831	0,431	0,567	98,063
13	0,337		3	0,779	0,438	0,561	96,929
14	0,362		4	0,846	0,321	0,466	80,484
15	0,349		5	0,818	0,328	0,469	81,027
16	0,349		6	0,815	0,321	0,461	79,641
17	0,348		7	0,840	0,307	0,449	77,647
18	0,341		8	0,746	0,321	0,449	77,609
19	0,334		9	0,671	0,343	0,454	78,496
20	0,337		10	0,588	0,365	0,450	77,864
21	0,350		1	0,611	0,482	0,539	93,131

22	0,360	3	2	0,750	0,328	0,457	78,970
23	0,352		3	0,582	0,336	0,426	73,624
24	0,337		4	0,348	0,343	0,346	59,737
25	0,346		5	0,357	0,299	0,325	56,247
26	0,344		6	0,568	0,336	0,422	72,949
27	0,340		7	0,701	0,343	0,461	79,650
28	0,320		8	0,373	0,299	0,332	57,386
29	0,312		9	0,325	0,277	0,299	51,721
30	0,308		10	0,291	0,270	0,280	48,452
31	0,357	4	1	0,709	0,445	0,547	94,568
32	0,353		2	0,681	0,343	0,456	78,877
33	0,353		3	0,595	0,321	0,417	72,092
34	0,338		4	0,360	0,328	0,344	59,378
35	0,343		5	0,338	0,350	0,344	59,478
36	0,349		6	0,681	0,343	0,456	78,877
37	0,326		7	0,396	0,292	0,336	58,103
38	0,314		8	0,298	0,307	0,302	52,230
39	0,308		9	0,283	0,285	0,284	49,029
40	0,309		10	0,330	0,277	0,302	52,132
41	0,394	5	1	0,815	0,321	0,461	79,641
42	0,367		2	0,625	0,328	0,431	74,436
43	0,370		3	0,529	0,270	0,357	61,794
44	0,358		4	0,614	0,314	0,415	71,815
45	0,335		5	0,432	0,372	0,400	69,143
46	0,331		6	0,394	0,299	0,340	58,814
47	0,323		7	0,381	0,314	0,344	59,463
48	0,326		8	0,465	0,292	0,359	62,012
49	0,327		9	0,577	0,299	0,394	68,146
50	0,333		10	0,759	0,299	0,429	74,211
51	0,387	6	1	0,703	0,328	0,448	77,399
52	0,376		2	0,606	0,314	0,413	71,470
53	0,370		3	0,683	0,299	0,416	71,951
54	0,374		4	0,864	0,277	0,420	72,581
55	0,364		5	0,804	0,270	0,404	69,899
56	0,354		6	0,702	0,292	0,412	71,281
57	0,359		7	0,870	0,292	0,437	75,566
58	0,354		8	0,870	0,292	0,437	75,566
59	0,347		9	0,837	0,299	0,441	76,206
60	0,351		10	0,889	0,292	0,440	75,981
61	0,383		1	0,719	0,336	0,458	79,119
62	0,360		2	0,561	0,336	0,420	72,616
63	0,374		3	0,800	0,292	0,428	73,950
64	0,370		4	0,816	0,292	0,430	74,347
65	0,367		5	0,813	0,285	0,422	72,880
66	0,354		6	0,737	0,307	0,433	74,845
67	0,361		7	0,833	0,292	0,432	74,749
68	0,342		8	0,776	0,328	0,462	79,780

69	0,343		9	0,789	0,328	0,464	80,191
70	0,334		10	0,738	0,328	0,455	78,571
71	0,403	8	1	0,672	0,314	0,428	73,959
72	0,381		2	0,661	0,285	0,398	68,790
73	0,369		3	0,631	0,299	0,406	70,170
74	0,364		4	0,548	0,292	0,381	65,850
75	0,365		5	0,677	0,307	0,422	72,965
76	0,370		6	0,913	0,307	0,459	79,344
77	0,364		7	0,913	0,307	0,459	79,344
78	0,358		8	0,894	0,307	0,457	78,913
79	0,352		9	0,913	0,307	0,459	79,344
80	0,350		10	0,894	0,307	0,457	78,913
81	0,366	9	1	0,738	0,350	0,475	82,150
82	0,352		2	0,511	0,328	0,400	69,143
83	0,361		3	0,471	0,292	0,360	62,291
84	0,364		4	0,582	0,285	0,382	66,092
85	0,366		5	0,875	0,307	0,454	78,486
86	0,358		6	0,820	0,299	0,439	75,798
87	0,349		7	0,642	0,314	0,422	72,871
88	0,347		8	0,662	0,314	0,426	73,593
89	0,344		9	0,560	0,307	0,396	68,491
90	0,340		10	0,575	0,307	0,400	69,143

Tabela A.2: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos reduzidos de termos de gênero e conteúdo (vide Tabela 4.3)

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,479	-	-	0,661	0,526	0,585	100,000
1	0,451	1	1	0,577	0,547	0,562	95,974
2	0,425		2	0,532	0,547	0,540	92,176
3	0,403		3	0,503	0,547	0,524	89,598
4	0,384		4	0,481	0,569	0,522	89,130
5	0,368		5	0,443	0,569	0,498	85,144
6	0,353		6	0,419	0,569	0,483	82,508
7	0,341		7	0,429	0,569	0,489	83,542
8	0,342		8	0,493	0,547	0,519	88,668
9	0,347		9	0,573	0,460	0,510	87,146
10	0,352		10	0,519	0,409	0,457	78,095
11	0,423		1	0,600	0,569	0,584	99,813
12	0,391		2	0,580	0,555	0,567	96,891
13	0,391		3	0,476	0,358	0,408	69,757

14	0,414	2	4	0,538	0,358	0,430	73,428
15	0,400		5	0,495	0,343	0,405	69,217
16	0,396		6	0,452	0,343	0,390	66,632
17	0,397		7	0,450	0,328	0,380	64,873
18	0,399		8	0,519	0,299	0,380	64,853
19	0,385		9	0,405	0,328	0,363	61,996
20	0,409		10	0,656	0,292	0,404	69,024
21	0,409	3	1	0,344	0,460	0,394	67,266
22	0,425		2	0,521	0,358	0,424	72,475
23	0,413		3	0,352	0,372	0,362	61,791
24	0,397		4	0,240	0,343	0,282	48,223
25	0,403		5	0,230	0,314	0,265	45,345
26	0,416		6	0,646	0,307	0,416	71,040
27	0,412		7	0,764	0,307	0,438	74,740
28	0,380		8	0,369	0,299	0,331	56,485
29	0,367		9	0,376	0,299	0,333	56,944
30	0,356		10	0,308	0,292	0,300	51,186
31	0,409	4	1	0,344	0,460	0,394	67,266
32	0,425		2	0,521	0,358	0,424	72,475
33	0,413		3	0,352	0,372	0,362	61,791
34	0,397		4	0,240	0,343	0,282	48,223
35	0,419		5	0,344	0,321	0,332	56,730
36	0,426		6	0,772	0,321	0,454	77,491
37	0,389		7	0,361	0,285	0,318	54,388
38	0,374		8	0,342	0,299	0,319	54,507
39	0,364		9	0,394	0,299	0,340	58,126
40	0,359		10	0,410	0,299	0,346	59,107
41	0,467	5	1	0,463	0,365	0,408	69,728
42	0,443		2	0,409	0,328	0,364	62,247
43	0,426		3	0,288	0,328	0,307	52,474
44	0,435		4	0,662	0,314	0,426	72,731
45	0,417		5	0,649	0,350	0,455	77,725
46	0,401		6	0,471	0,292	0,360	61,562
47	0,390		7	0,513	0,292	0,372	63,566
48	0,385		8	0,606	0,292	0,394	67,323
49	0,379		9	0,548	0,292	0,381	65,079
50	0,373		10	0,645	0,292	0,402	68,677
51	0,482		1	0,532	0,299	0,383	65,460
52	0,462		2	0,446	0,299	0,358	61,172
53	0,430		3	0,398	0,328	0,360	61,500
54	0,442		4	0,717	0,277	0,400	68,333
55	0,418		5	0,562	0,299	0,390	66,706
56	0,402		6	0,489	0,336	0,398	68,038
57	0,416		7	0,727	0,292	0,417	71,181
58	0,417		8	0,809	0,277	0,413	70,562
59	0,401		9	0,796	0,285	0,419	71,640
60	0,389		10	0,702	0,292	0,412	70,447

61	0,462	6	1	0,636	0,409	0,498	85,037
62	0,426		2	0,410	0,401	0,406	69,342
63	0,455		3	0,783	0,263	0,393	67,213
64	0,440		4	0,795	0,255	0,387	66,068
65	0,409		5	0,385	0,343	0,363	62,001
66	0,426		6	0,661	0,285	0,398	67,985
67	0,411		7	0,625	0,292	0,398	67,993
68	0,409		8	0,732	0,299	0,425	72,582
69	0,407		9	0,774	0,299	0,432	73,728
70	0,398		10	0,745	0,299	0,427	72,960
71	0,492	8	1	0,766	0,358	0,488	83,292
72	0,417		2	0,300	0,453	0,360	61,579
73	0,437		3	0,586	0,299	0,396	67,673
74	0,425		4	0,374	0,336	0,354	60,449
75	0,430		5	0,651	0,299	0,410	70,042
76	0,421		6	0,606	0,314	0,413	70,633
77	0,404		7	0,571	0,321	0,411	70,249
78	0,414		8	0,769	0,292	0,423	72,310
79	0,407		9	0,800	0,292	0,428	73,084
80	0,401		10	0,788	0,299	0,434	74,118
81	0,423	9	1	0,640	0,416	0,504	86,173
82	0,406		2	0,474	0,474	0,474	81,052
83	0,420		3	0,389	0,321	0,352	60,133
84	0,406		4	0,293	0,401	0,338	57,821
85	0,413		5	0,579	0,482	0,526	89,841
86	0,396		6	0,436	0,496	0,464	79,295
87	0,401		7	0,515	0,372	0,432	73,835
88	0,418		8	0,722	0,285	0,408	69,764
89	0,421		9	0,780	0,285	0,417	71,257
90	0,415		10	0,784	0,292	0,426	72,695

Tabela A.3: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando o coeficiente de Jaccard e os conjuntos de termos de gênero e conteúdo definidos por não especialista (vide Tabela 4.4)

	Similaridade	Jaccard - Parâmetros		Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,556	-	-	0,615	0,409	0,491	100,000
1	0,516	1	1	0,482	0,387	0,429	87,363
2	0,490		2	0,495	0,336	0,400	81,429
3	0,457		3	0,479	0,336	0,395	80,380
4	0,428		4	0,459	0,365	0,407	82,753
5	0,405		5	0,395	0,372	0,383	78,061
6	0,385		6	0,369	0,380	0,374	76,156
7	0,370		7	0,389	0,358	0,373	75,856
8	0,355		8	0,419	0,453	0,435	88,571
9	0,372		9	0,539	0,350	0,425	86,473
10	0,364		10	0,495	0,336	0,400	81,429
11	0,487	2	1	0,602	0,365	0,455	92,532
12	0,438		2	0,521	0,358	0,424	86,364
13	0,390		3	0,300	0,416	0,349	70,970
14	0,407		4	0,345	0,431	0,383	77,992
15	0,380		5	0,265	0,445	0,332	67,672
16	0,369		6	0,274	0,401	0,325	66,251
17	0,362		7	0,271	0,438	0,335	68,236
18	0,372		8	0,383	0,358	0,370	75,283
19	0,355		9	0,329	0,380	0,353	71,768
20	0,364		10	0,296	0,387	0,335	68,287
21	0,483	3	1	0,358	0,387	0,372	75,714
22	0,474		2	0,384	0,387	0,385	78,468
23	0,473		3	0,424	0,285	0,341	69,339
24	0,427		4	0,175	0,431	0,248	50,571
25	0,458		5	0,243	0,263	0,253	51,429
26	0,441		6	0,420	0,423	0,422	85,870
27	0,438		7	0,463	0,409	0,434	88,372
28	0,431		8	0,333	0,328	0,331	67,358
29	0,430		9	0,386	0,248	0,302	61,524
30	0,406		10	0,308	0,350	0,328	66,699
31	0,483		1	0,358	0,387	0,372	75,714
32	0,474		2	0,384	0,387	0,385	78,468
33	0,473		3	0,424	0,285	0,341	69,339
34	0,427		4	0,175	0,431	0,248	50,571
35	0,487		5	0,492	0,212	0,296	60,241
36	0,446		6	0,414	0,423	0,419	85,250

37	0,431	4	7	0,298	0,372	0,331	67,417
38	0,442		8	0,453	0,248	0,321	65,296
39	0,411		9	0,283	0,343	0,310	63,154
40	0,410		10	0,367	0,343	0,355	72,210
41	0,539	5	1	0,509	0,409	0,453	92,308
42	0,511		2	0,373	0,387	0,380	77,343
43	0,509		3	0,385	0,328	0,354	72,132
44	0,510		4	0,645	0,358	0,460	93,662
45	0,495		5	0,491	0,401	0,442	89,931
46	0,491		6	0,529	0,401	0,456	92,916
47	0,489		7	0,645	0,358	0,460	93,662
48	0,472		8	0,529	0,394	0,452	91,990
49	0,471		9	0,646	0,387	0,484	98,532
50	0,457		10	0,705	0,401	0,512	104,153
51	0,540	6	1	0,598	0,358	0,447	91,096
52	0,550		2	0,649	0,270	0,381	77,651
53	0,515		3	0,712	0,307	0,429	87,245
54	0,493		4	0,538	0,314	0,396	80,678
55	0,470		5	0,441	0,328	0,377	76,659
56	0,491		6	0,603	0,277	0,380	77,357
57	0,480		7	0,612	0,299	0,402	81,828
58	0,465		8	0,558	0,314	0,402	81,809
59	0,468		9	0,656	0,307	0,418	85,075
60	0,478		10	0,774	0,299	0,432	87,857
61	0,578	7	1	0,800	0,321	0,458	93,304
62	0,489		2	0,535	0,394	0,454	92,377
63	0,521		3	0,740	0,270	0,396	80,558
64	0,494		4	0,507	0,277	0,358	72,978
65	0,486		5	0,440	0,321	0,371	75,588
66	0,482		6	0,556	0,328	0,413	84,043
67	0,478		7	0,542	0,328	0,409	83,279
68	0,485		8	0,698	0,321	0,440	89,571
69	0,489		9	0,746	0,321	0,449	91,399
70	0,486		10	0,759	0,321	0,451	91,868
71	0,585	8	1	0,880	0,321	0,471	95,798
72	0,518		2	0,750	0,307	0,435	88,601
73	0,483		3	0,485	0,350	0,407	82,809
74	0,494		4	0,512	0,321	0,395	80,333
75	0,480		5	0,588	0,343	0,433	88,183
76	0,481		6	0,605	0,336	0,432	87,928
77	0,457		7	0,529	0,394	0,452	91,990
78	0,497		8	0,804	0,299	0,436	88,792
79	0,494		9	0,851	0,292	0,435	88,509
80	0,488		10	0,788	0,299	0,434	88,322
81	0,549	4	1	0,759	0,321	0,451	91,868
82	0,509		2	0,651	0,299	0,410	83,464
83	0,502		3	0,486	0,263	0,341	69,465

84	0,476	9	4	0,440	0,321	0,371	75,588
85	0,461		5	0,411	0,438	0,424	86,320
86	0,445		6	0,371	0,453	0,408	83,036
87	0,434		7	0,351	0,496	0,411	83,643
88	0,481		8	0,750	0,285	0,413	84,014
89	0,462		9	0,532	0,299	0,383	78,004
90	0,466		10	0,731	0,277	0,402	81,859

Apêndice B

Resultados dos experimentos relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância

Tabela B.1: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância e os conjuntos completos de termos de gênero e conteúdo (vide Tabela 4.1)

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,377	-	-	-	0,667	0,511	0,579	100,000
1	0,371	1	1	1	0,663	0,387	0,488	84,437
2	0,368			2	0,659	0,394	0,493	85,245
3	0,368			5	0,659	0,394	0,493	85,245
4	0,368			10	0,659	0,394	0,493	85,245
5	0,375		2	1	0,770	0,343	0,475	82,063
6	0,380			2	0,836	0,336	0,479	82,827
7	0,371			5	0,836	0,336	0,479	82,827
8	0,371			10	0,836	0,336	0,479	82,827
9	0,342		5	1	0,626	0,416	0,500	86,429
10	0,360			2	0,852	0,336	0,482	83,261
11	0,334			5	0,645	0,292	0,402	69,490
12	0,335			10	0,672	0,299	0,414	71,587
13	0,343		10	1	0,758	0,343	0,472	81,651
14	0,330			2	0,536	0,380	0,444	76,825
15	0,309			5	0,390	0,299	0,339	58,571
16	0,307			10	0,442	0,277	0,341	58,911
17	0,372		1	1	0,845	0,438	0,577	99,725
18	0,356			2	0,660	0,467	0,547	94,554
19	0,359			5	0,621	0,467	0,533	92,190
20	0,359			10	0,621	0,467	0,533	92,190
21	0,370			1	0,904	0,343	0,497	85,971

22	0,342	2	2	2	0,714	0,474	0,570	98,559
23	0,354			5	0,831	0,431	0,567	98,063
24	0,344			10	0,699	0,474	0,565	97,702
25	0,349		5	1	0,939	0,336	0,495	85,499
26	0,322			2	0,703	0,328	0,448	77,399
27	0,318			5	0,667	0,307	0,420	72,600
28	0,318			10	0,755	0,292	0,421	72,782
29	0,309		10	1	0,592	0,328	0,423	73,038
30	0,300			2	0,525	0,307	0,387	66,912
31	0,293			5	0,382	0,285	0,326	56,414
32	0,282			10	0,346	0,270	0,303	52,424
33	0,348	3	1	1	0,482	0,489	0,486	83,923
34	0,349			2	0,478	0,474	0,476	82,313
35	0,347			5	0,441	0,467	0,454	78,460
36	0,343			10	0,391	0,496	0,437	75,590
37	0,322		2	1	0,296	0,504	0,373	64,471
38	0,319			2	0,392	0,453	0,420	72,659
39	0,321			5	0,438	0,409	0,423	73,057
40	0,315			10	0,412	0,460	0,434	75,103
41	0,296		5	1	0,342	0,482	0,400	69,143
42	0,294			2	0,588	0,365	0,450	77,864
43	0,281			5	0,389	0,409	0,399	68,897
44	0,279			10	0,314	0,431	0,363	62,760
45	0,283		10	1	0,276	0,307	0,291	50,242
46	0,261			2	0,176	0,350	0,235	40,573
47	0,253			5	0,171	0,358	0,231	39,953
48	0,254			10	0,206	0,314	0,249	42,964
49	0,347	5	1	1	0,479	0,489	0,484	83,620
50	0,346			2	0,464	0,467	0,465	80,457
51	0,353			5	0,598	0,401	0,480	83,032
52	0,353			10	0,598	0,401	0,480	83,032
53	0,332		2	1	0,406	0,474	0,438	75,661
54	0,327			2	0,806	0,423	0,555	95,940
55	0,318			5	0,674	0,438	0,531	91,783
56	0,318			10	0,674	0,438	0,531	91,783
57	0,302		5	1	0,545	0,526	0,535	92,533
58	0,285			2	0,508	0,482	0,494	85,457
59	0,274			5	0,445	0,445	0,445	76,966
60	0,278			10	0,536	0,431	0,478	82,580
61	0,308		10	1	0,764	0,307	0,438	75,625
62	0,275			2	0,506	0,321	0,393	67,908
63	0,258			5	0,420	0,401	0,410	70,949
64	0,256			10	0,426	0,358	0,389	67,222
65	0,342	10	1	1	0,412	0,445	0,428	73,995
66	0,337			2	0,414	0,438	0,426	73,556
67	0,337			5	0,441	0,438	0,440	75,981
68	0,336			10	0,444	0,438	0,441	76,261

69	0,306	9	2	1	0,384	0,482	0,427	73,842
70	0,308			2	0,438	0,460	0,448	77,509
71	0,308			5	0,534	0,401	0,458	79,226
72	0,307			10	0,524	0,401	0,455	78,571
73	0,292		5	1	0,548	0,416	0,473	81,766
74	0,271			2	0,324	0,350	0,337	58,226
75	0,254			5	0,270	0,474	0,344	59,448
76	0,257			10	0,308	0,438	0,361	62,478
77	0,278		10	1	0,645	0,292	0,402	69,490
78	0,255			2	0,742	0,358	0,483	83,448
79	0,229			5	0,190	0,445	0,266	46,045
80	0,229			10	0,226	0,438	0,298	51,471

Tabela B.2: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância e os conjuntos reduzidos de termos de gênero e conteúdo (vide Tabela 4.3)

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,479	-	-	-	0,661	0,526	0,585	100,000
1	0,420	1	1	1	0,545	0,489	0,515	88,045
2	0,415			2	0,512	0,460	0,485	82,788
3	0,415			5	0,512	0,460	0,485	82,788
4	0,415			10	0,512	0,460	0,485	82,788
5	0,419		2	1	0,613	0,416	0,496	84,674
6	0,420			2	0,663	0,401	0,500	85,417
7	0,398			5	0,550	0,445	0,492	84,039
8	0,398			10	0,550	0,445	0,492	84,039
9	0,386		5	1	0,667	0,409	0,507	86,576
10	0,371			2	0,659	0,409	0,505	86,186
11	0,332			5	0,447	0,460	0,453	77,428
12	0,335			10	0,455	0,401	0,426	72,836
13	0,348		10	1	0,452	0,416	0,433	74,049
14	0,343			2	0,609	0,387	0,473	80,841
15	0,302			5	0,293	0,409	0,341	58,333
16	0,319			10	0,557	0,285	0,377	64,372
17	0,416		1	1	0,660	0,453	0,537	91,703
18	0,420			2	0,692	0,540	0,607	103,620
19	0,426			5	0,730	0,533	0,616	105,239
20	0,426			10	0,730	0,533	0,616	105,239
21	0,395		2	1	0,492	0,438	0,463	79,151
22	0,395			2	0,716	0,533	0,611	104,358
23	0,394			5	0,705	0,540	0,612	104,477

		2						
24	0,394	2		10	0,705	0,540	0,612	104,477
25	0,353		5	1	0,513	0,431	0,468	79,993
26	0,321			2	0,485	0,474	0,480	81,950
27	0,340			5	0,657	0,336	0,444	75,926
28	0,332			10	0,551	0,358	0,434	74,078
29	0,311		10	1	0,376	0,387	0,381	65,138
30	0,301			2	0,529	0,394	0,452	77,197
31	0,301			5	0,519	0,292	0,374	63,863
32	0,298			10	0,613	0,277	0,382	65,243
33	0,378	3	1	1	0,278	0,533	0,365	62,354
34	0,375			2	0,291	0,584	0,388	66,343
35	0,375			5	0,291	0,584	0,388	66,343
36	0,375			10	0,291	0,584	0,388	66,343
37	0,338		2	1	0,182	0,533	0,271	46,360
38	0,347			2	0,322	0,533	0,401	68,521
39	0,347			5	0,322	0,533	0,401	68,521
40	0,347			10	0,322	0,533	0,401	68,521
41	0,314		5	1	0,378	0,394	0,386	65,893
42	0,309			2	0,439	0,526	0,478	81,728
43	0,297			5	0,307	0,504	0,381	65,124
44	0,289			10	0,245	0,526	0,334	57,077
45	0,274		10	1	0,128	0,321	0,183	31,254
46	0,280			2	0,236	0,343	0,280	47,793
47	0,275			5	0,246	0,358	0,292	49,826
48	0,261			10	0,166	0,358	0,227	38,754
49	0,372	5	1	1	0,224	0,620	0,329	56,173
50	0,380			2	0,320	0,540	0,402	68,705
51	0,380			5	0,320	0,540	0,402	68,705
52	0,380			10	0,320	0,540	0,402	68,705
53	0,338		2	1	0,200	0,635	0,305	52,058
54	0,345			2	0,455	0,445	0,450	76,907
55	0,345			5	0,455	0,445	0,450	76,907
56	0,345			10	0,455	0,445	0,450	76,907
57	0,309		5	1	0,405	0,453	0,428	73,046
58	0,316			2	0,637	0,474	0,544	92,922
59	0,287			5	0,368	0,489	0,420	71,761
60	0,287			10	0,368	0,489	0,420	71,761
61	0,297		10	1	0,422	0,277	0,335	57,195
62	0,279			2	0,400	0,526	0,454	77,603
63	0,270			5	0,436	0,423	0,430	73,395
64	0,255			10	0,304	0,453	0,364	62,121
65	0,397		1	1	0,414	0,526	0,463	79,100
66	0,374			2	0,376	0,599	0,462	78,920
67	0,371			5	0,361	0,599	0,451	76,969
68	0,368			10	0,368	0,591	0,454	77,521
69	0,343			1	0,281	0,540	0,370	63,208
70	0,334			2	0,341	0,526	0,414	70,690

71	0,325	9	2	5	0,339	0,547	0,419	71,578
72	0,319			10	0,335	0,562	0,420	71,685
73	0,295		5	1	0,292	0,562	0,384	65,607
74	0,291			2	0,374	0,489	0,424	72,442
75	0,271			5	0,347	0,540	0,423	72,238
76	0,267			10	0,354	0,511	0,418	71,393
77	0,269		10	1	0,197	0,314	0,242	41,385
78	0,259			2	0,539	0,453	0,492	84,061
79	0,247			5	0,238	0,365	0,288	49,232
80	0,245			10	0,333	0,336	0,335	57,152

Tabela B.3: Experimentos relativos à estratégia baseada em matriz de associação de termos utilizando a métrica MenorDistância e os conjuntos de termos de gênero e conteúdo definidos por não especialista(vide Tabela 4.4)

	Similaridade	MenorDistância - Parâmetros			Resultados			
Caso	Limite empírico	Número de páginas-semente	Número de termos de expansão por termo original	Distância máxima (k)	Precisão	Revocação	F1	Proximidade de F1 em relação ao baseline (%)
0	0,556	-	-	-	0,615	0,409	0,491	100,000
1	0,468	1	1	1	0,432	0,511	0,468	95,318
2	0,439			2	0,192	0,584	0,289	58,900
3	0,439			5	0,192	0,584	0,289	58,900
4	0,439			10	0,192	0,584	0,289	58,900
5	0,425		2	1	0,447	0,518	0,480	97,659
6	0,441			2	0,476	0,511	0,493	100,352
7	0,429			5	0,358	0,533	0,428	87,160
8	0,429			10	0,358	0,533	0,428	87,160
9	0,381		5	1	0,457	0,423	0,439	89,448
10	0,375			2	0,397	0,547	0,460	93,668
11	0,402			5	0,504	0,489	0,496	101,032
12	0,404			10	0,535	0,445	0,486	98,947
13	0,341		10	1	0,300	0,416	0,349	70,970
14	0,345			2	0,321	0,431	0,368	74,833
15	0,346			5	0,442	0,504	0,471	95,880
16	0,370			10	0,629	0,445	0,521	106,136
17	0,433	2	1	1	0,299	0,467	0,365	74,237
18	0,493			2	0,646	0,372	0,472	96,131
19	0,478			5	0,495	0,372	0,425	86,518
20	0,478			10	0,495	0,372	0,425	86,518
21	0,407		2	1	0,435	0,438	0,436	88,831
22	0,450			2	0,492	0,445	0,467	95,156
23	0,437			5	0,509	0,409	0,453	92,308
24	0,428			10	0,500	0,416	0,454	92,459

25	0,336	2	5	1	0,307	0,540	0,392	79,705
26	0,363			2	0,393	0,387	0,390	79,333
27	0,385			5	0,525	0,460	0,490	99,805
28	0,376			10	0,538	0,409	0,465	94,606
29	0,274		10	1	0,181	0,591	0,277	56,470
30	0,301			2	0,213	0,358	0,267	54,360
31	0,336			5	0,598	0,467	0,525	106,792
32	0,338			10	0,610	0,445	0,515	104,792
33	0,424	3	1	1	0,298	0,518	0,379	77,086
34	0,425			2	0,186	0,474	0,267	54,453
35	0,425			5	0,186	0,474	0,267	54,453
36	0,397			10	0,131	0,635	0,218	44,332
37	0,363		2	1	0,190	0,533	0,280	56,938
38	0,396			2	0,246	0,416	0,309	62,892
39	0,377			5	0,199	0,438	0,274	55,773
40	0,366			10	0,152	0,482	0,232	47,143
41	0,305		5	1	0,196	0,533	0,287	58,392
42	0,311			2	0,133	0,482	0,209	42,451
43	0,319			5	0,136	0,467	0,211	42,928
44	0,309			10	0,108	0,489	0,177	36,035
45	0,297		10	1	0,197	0,387	0,261	53,149
46	0,284			2	0,139	0,474	0,216	43,888
47	0,287			5	0,155	0,467	0,233	47,377
48	0,306			10	0,168	0,307	0,217	44,186
49	0,401	5	1	1	0,235	0,650	0,346	70,361
50	0,412			2	0,262	0,533	0,351	71,446
51	0,412			5	0,262	0,533	0,351	71,446
52	0,412			10	0,262	0,533	0,351	71,446
53	0,379		2	1	0,271	0,496	0,351	71,355
54	0,380			2	0,397	0,453	0,423	86,153
55	0,361			5	0,328	0,445	0,378	76,891
56	0,361			10	0,328	0,445	0,378	76,891
57	0,337		5	1	0,462	0,438	0,449	91,493
58	0,322			2	0,197	0,350	0,252	51,294
59	0,302			5	0,153	0,401	0,222	45,147
60	0,310			10	0,192	0,365	0,251	51,149
61	0,320		10	1	0,370	0,321	0,344	69,978
62	0,303			2	0,219	0,255	0,236	47,980
63	0,289			5	0,281	0,328	0,303	61,688
64	0,310			10	0,432	0,255	0,321	65,367
65	0,400	10	1	1	0,213	0,737	0,330	67,192
66	0,424			2	0,272	0,467	0,344	70,046
67	0,410			5	0,252	0,504	0,336	68,352
68	0,410			10	0,252	0,504	0,336	68,352
69	0,346			1	0,219	0,745	0,339	68,984
70	0,371			2	0,363	0,482	0,414	84,236
71	0,362			5	0,351	0,431	0,387	78,759

72	0,362	9		10	0,351	0,431	0,387	78,759
73	0,304		5	1	0,251	0,540	0,343	69,742
74	0,298			2	0,215	0,511	0,303	61,688
75	0,303			5	0,239	0,380	0,293	59,638
76	0,311			10	0,314	0,321	0,318	64,673
77	0,316		10	1	0,470	0,394	0,429	87,245
78	0,277			2	0,235	0,496	0,319	64,990
79	0,280			5	0,230	0,416	0,296	60,278
80	0,294			10	0,327	0,263	0,291	59,341

Apêndice C

Resultados dos experimentos relativos à estratégia baseada em PLN

Tabela C.1: Experimentos relativos à estratégia baseada em PLN utilizando os conjuntos completos de termos de gênero e conteúdo (vide Tabela 4.1)

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,377	-	-	-	0,667	0,511	0,579	100,000
1	0,350			x	0,486	0,518	0,502	86,734
2	0,367		x		0,645	0,438	0,522	90,186
3	0,350		x	x	0,493	0,511	0,502	86,738
4	0,369	x			0,573	0,431	0,492	84,988
5	0,375	x		x	0,667	0,423	0,518	89,515
6	0,370	x	x		0,562	0,431	0,488	84,286
7	0,377	x	x	x	0,667	0,423	0,518	89,515

Tabela C.2: Experimentos relativos à estratégia baseada em PLN utilizando os conjuntos reduzidos de termos de gênero e conteúdo (vide Tabela 4.3)

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,479	-	-	-	0,661	0,526	0,585	100,000
1	0,440			x	0,551	0,547	0,549	93,864
2	0,447		x		0,600	0,526	0,560	95,720
3	0,442		x	x	0,585	0,526	0,554	94,615
4	0,439	x			0,503	0,533	0,518	88,446
5	0,444	x		x	0,514	0,555	0,533	91,111
6	0,434	x	x		0,473	0,577	0,520	88,788
7	0,445	x	x	x	0,531	0,555	0,543	92,738

Tabela C.3: Experimentos relativos à estratégia baseada em PLN utilizando os conjuntos de termos de gênero e conteúdo definidos por não especialista(vide Tabela 4.4)

	Similaridade	**Técnicas de PLN**			**Resultados**			
Caso	**Limite empírico**	***Stemming***	**Remoção de *stopwords***	**Pluralização**	**Precisão**	**Revocação**	**F1**	**Proximidade de F1 em relação ao *baseline* (%)**
0	0,556	-	-	-	0,615	0,409	0,491	100,000
1	0,505			x	0,216	0,365	0,272	55,318
2	0,541		x		0,568	0,394	0,466	94,766
3	0,505		x	x	0,216	0,365	0,272	55,318
4	0,548	x			0,331	0,365	0,347	70,685
5	0,573	x		x	0,414	0,350	0,379	77,244
6	0,548	x	x		0,331	0,365	0,347	70,685
7	0,573	x	x	x	0,414	0,350	0,379	77,244

Referências Bibliográficas

Almpanidis, G.; Kotropoulos, C. e Pitas, I. (2007). Combining text and link analysis for focused crawling—an application for vertical search engines. *Information Systems*, 32(6):886–908.

Aly, A. A. (2008). Using a query expansion technique to improve document retrieval. In *International Journal "Information Technologies and Knowledge*.

Araujo, L. e Pérez-Agüera, J. R. (2008). Improving query expansion with stemming terms: a new genetic algorithm approach. In *European Conference on Evolutionary Computation in Combinatorial Optimization*, pp. 182–193. Springer.

Cardoso, O. N. P. (2004). Recuperação de informação. *INFOCOMP Journal of Computer Science*, 2(1):33–38.

Carpineto, C. e Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Chakrabarti, S.; Van den Berg, M. e Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11):1623–1640.

Chartree, J.; Cankaya, E. C. e Phithakkitnukoon, S. (2013). Query expansion using association matrix for improved information retrieval performance. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

de Assis, G. T.; Laender, A. H.; da Silva, A. S. e Gonçalves, M. A. (2008). The impact of term selection in genre-aware focused crawling. In *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1158–1163. ACM.

De Assis, G. T.; Laender, A. H.; Gonçalves, M. A. e Da Silva, A. S. (2007). Exploiting genre in focused crawling. In *String Processing and Information Retrieval*, pp. 62–73. Springer.

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

Foundation, A. S. (2015). Lucene-java wiki / poweredby.

Gonzalez, M. e Lima, V. L. S. (2003). Recuperação de informação e processamento da linguagem natural. In *XXIII Congresso da Sociedade Brasileira de Computaçao*, volume 3, pp. 347–395.

Greenberg, J. (2001). Optimal query expansion (qe) processing methods with semantically encoded structured thesauri terminology. *J. Am. Soc. Inf. Sci. Technol.*, 52(6):487–498.

Indurkhya, N. e Damerau, F. J. (2010). *Handbook of natural language processing*, volume 2. CRC Press.

Johnson, J.; Tsioutsiouliklis, K. e Giles, C. L. (2003). Evolving strategies for focused web crawling. In *ICML*, pp. 298–305.

Kilgarriff, A. e Yallop, C. (2000). What's in a thesaurus? In *LREC*.

Li, J.-F.; Guo, M.-Z. e Tian, S.-H. (2005). A new approach to query expansion. In *2005 International Conference on Machine Learning and Cybernetics*, volume 4, pp. 2302–2306. IEEE.

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 *Web Congress (LA-WEB), 2012 Eighth Latin American*, pp. 17–23. IEEE.

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

Manning, C. D.; Raghavan, P. e Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

McCandless, M.; Hatcher, E. e Gospodnetic, O. (2010). *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA.

Menczer, F.; Pant, G. e Srinivasan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, 4(4):378–419.

Mitra, M.; Singhal, A. e Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 206–214, New York, NY, USA. ACM.

Orengo, V. M. e Huyck, C. R. (2001). A stemming algorithmm for the portuguese language. In *spire*, volume 8, pp. 186–193.

Pant, G. e Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems (TOIS)*, 23(4):430–462.

Pant, G. e Srinivasan, P. (2006). Link contexts in classifier-guided topical crawlers. *Knowledge and Data Engineering, IEEE Transactions on*, 18(1):107–122.

Sigrist, P. e Higashino, W. A. (2012). Conhecendo o lucene.

Siqueira, G. O. d.; Assis, G. T. d.; 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 (ICWI)*. WWW.

Srinivasan, P.; Menczer, F. e Pant, G. (2005). A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447.