UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Técnicas de segmentação e agrupamento aplicadas a recursos predeterminados para contextualizar buscas na Internet

João Carlos Prates dos Santos

Orientador

Sean Wolfgand Matsui Siqueira

RIO DE JANEIRO, RJ – BRASIL

Março de 2011

TÉCNICAS DE SEGMENTAÇÃO E AGRUPAMENTO APLICADAS A RECURSOS PREDETERMINADOS PARA CONTEXTUALIZAR BUSCAS NA INTERNET

João Carlos Prates dos Santos

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Sean W. M. Siqueira
Sean Wolfgand Matsui Siqueira, D.Sc. - UNIRIO

Flávia Maria Santoro
Flávia Maria Santoro, D.Sc. - UNIRIO

Neide dos Santos
Neide dos Santos, D.Sc. - UERJ

Geraldo Xexéo
Geraldo Bonorino Xexéo, D.Sc. - UFRJ

RIO DE JANEIRO, RJ – BRASIL

Março de 2011

Santos, João Carlos Prates dos.
S237 Técnicas de segmentação e agrupamento aplicadas a recursos predeterminados para contextualizar buscas na internet / João Carlos Prates dos Santos, 2011.
136 f. + CD ROM

Orientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado em Informática) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2011.

1. Recuperação da informação. 2. Ferramentas de busca na web. 3. Segmentação de tópicos. 4. Agrupamento de tópicos. 5. Extração da informação. I. Siqueira, Sean Wolfgand Matsui. III. Universidade Federal do Estado do Rio de Janeiro (2003-). Centro de Ciências Exatas e Tecnologia. Curso de Mestrado em Informática. III. Título.

CDD – 005.74

A minha esposa, meus pais e meus avós

Agradecimentos

Agradeço aos meus familiares, amigos e todos aqueles que contribuíram de alguma forma para a conclusão desta dissertação.

Agradeço em especial a minha esposa, pelos inúmeros momentos de apoio e compreensão e por ser a minha eterna motivação para alcançar meus objetivos.

Agradeço ao professor Sean por ter sido meu orientador desde antes da minha matrícula na Unirio. Em nenhum momento do desenvolvimento da pesquisa, senti falta de um apoio ou orientação, pois sua presença foi constante. Obrigado pela paciência, pelo profissionalismo e pela amizade.

Aos meus superiores do BNDES por patrocinar esta investida, flexibilizando a minha jornada de trabalho, sendo compreensivos com os períodos mais intensos de trabalho e dando o apoio necessário para que eu seguisse em frente. Agradeço a Paula, Ana Carolina, Alexandre e Oswaldo.

Aos alunos da turma de mestrado (2010) pela participação no estudo de caso piloto, e aos alunos de sistemas de informação (2010/1), pela participação no estudo de caso realizado na Unirio.

Aos companheiros de trabalho Kleber, Ana Lúcia, Carlos Edmar, Lílian, Cláudia e Diego, participantes do estudo de caso realizado no BNDES. Agradeço também as bibliotecárias Margareth, Juceli, Fernanda, Bruna, Tânia e Caroline pelo esforço e dedicação durante o estudo de caso.

A todos os profissionais da Unirio, que fizeram com que o curso de mestrado superasse as minhas expectativas, em especial a Alessandra Nascimento pelas incontáveis ajudas na secretaria; aos professores Angelo, Alexandre e Flávia pela grande contribuição nas disciplinas do mestrado e nos seminários de acompanhamento; e ao Jackson, pelo suporte técnico na infraestrutura utilizada no estudo de caso da Unirio.

Prates, João Carlos. **Técnicas de segmentação e agrupamento aplicadas a recursos predeterminados para contextualizar buscas na Internet**. UNIRIO, 2011. 136 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

A internet é uma fonte incalculável de informações, cujo volume de informações e acessos cresce a cada dia. Ferramentas de busca web ocupam um papel de destaque para a descoberta de informações relevantes, porém esta busca, em geral, é realizada a partir de palavras-chave, o que implica em resultados pouco contextualizados. Este trabalho propõe o uso de técnicas de segmentação e agrupamento (*clustering*) aplicadas a recursos predeterminados, selecionados como conjunto base de documentos do contexto, expandindo as consultas realizadas pelos usuários e as executando em ferramentas de busca web, retornando assim resultados de busca mais úteis. Em estudos de caso, realizados com apoio de uma ferramenta desenvolvida segundo a abordagem proposta, aplicados em ambientes corporativo e educacional, foram obtidos resultados que apontam que este sistema de informação pode ser utilizado apoiando a contextualização de atividades de busca.

Palavras-chave: Recuperação de Informação, Extração de Informação, Contexto, Segmentação, Agrupamento

ABSTRACT

The Internet is a valuable source of information whose volume of data and number of accesses grows every day. Web search engines have a key role in the discovery of relevant information, but this kind of search is usually performed using keywords, which implies on results that do not consider the context. This thesis presents a proposal of using segmentation techniques and clustering applied to previously configured resources, selected as a representative domain context, expanding the queries done by users and executing these queries in web search engines, getting more useful search results. In case studies, conducted with the support from a tool developed according to the proposed approach, executed in corporate and educational environments, the obtained results indicated that this information system could be used to support the contextualization of search activities.

Keywords: Information Retrieval, Information Extraction, Context, Topic Segmentation, Clustering

Sumário

1. Capítulo 1 – Introdução 15

1.1. Motivação 15

1.2. Hipótese de pesquisa 18

1.3. Enfoque de solução 18

1.4. Metodologia de pesquisa 19

1.5. Objetivos da dissertação 21

1.6. Organização da dissertação 22

2. Capítulo 2 – Fundamentação Teórica 23

2.1. Recuperação de informação 23

2.1.1. Abordagens em Recuperação de Informações 24

2.1.2. Cálculo de pesos 27

2.1.3. Expansão de consultas 29

2.1.4. Avaliação 30

2.2. Extração de informação 31

2.2.1. Algumas atividades de extração de informação 33

2.2.2. Técnicas de extração de informação 34

2.2.3. Segmentação de tópicos 35

2.2.4. Agrupamento 36

2.3. Uso de expansão de consultas para tornar a recuperação de informação sensível ao contexto 37

3. Capítulo 3 – Proposta 41

3.1. Arquitetura conceitual 41

3.2. Arquitetura lógica 45

3.2.1. Módulo de configuração da base de conhecimento 46
3.2.2. Módulo de extração de informações 47
3.2.3. Módulo de busca 49
3.2.4. Atualização do contexto 49
3.3. Arquitetura física 50
3.3.1. Módulo de configuração da base de conhecimento 51
3.3.2. Módulo de extração de informações 54
3.3.3. Módulo de busca 61
3.3.4. Interação entre usuário e especialista com a aplicação 62
4. Capítulo 4 – Implementação 65
4.1. O Protótipo 65
4.1.1. Expansão automática 66
4.1.2. Sugestão de termos 68
4.1.3. Avaliações 69
4.1.4. Administração dos contextos 70
5. Capítulo 5 – Estudos de Caso 76
5.1. Metodologia 76
5.2. Estudo de caso 1: Ambiente educacional 78
5.3. Estudo de caso 2: Ambiente corporativo 80
5.4. Os questionários 83
6. Capítulo 6 – Avaliação 85
6.1. Métricas utilizadas na avaliação 85
6.2. Número de avaliações consideradas 88
6.3. Análise quantitativa 89
6.3.1. Justificativas 95

6.3.2. Impacto das variáveis nos julgamentos de relevância 98
6.4. Análise qualitativa 101
6.4.1. Comparação entre as modalidades 102
7. Capítulo 7 – Conclusão 105
7.1. Trabalhos relacionados 108
7.2. Contribuições 109
7.3. Limitações da abordagem e trabalhos futuros 110
8. Referências 113
Anexo I – Questionário para os estudos de caso 118
I.1. Identificação 118
I.2. Perguntas sobre a modalidade expansão automática 118
I.3. Perguntas sobre a modalidade sugestão de termos 121
I.4. Perguntas gerais sobre o experimento 123
Anexo II – Roteiros dos estudos de caso 126
II.1. Roteiro 126

Lista de Figuras

Figura 1 - Organização da dissertação 22

Figura 2 – Motivação da recuperação de informações 23

Figura 3 – Exemplo de extração de informações 32

Figura 4 - Arquitetura conceitual. 43

Figura 5 - Segmentação do texto em tópicos. 44

Figura 6 - Agrupamento de segmentos em assuntos. 45

Figura 7 - Arquitetura lógica. 46

Figura 8 - Arquitetura física. 51

Figura 9 – Diagrama de atividades – Configuração da base de conhecimento. 53

Figura 10 – Diagrama de atividades – Segmentação de tópicos. 56

Figura 11 – Diagrama de atividades – Agrupamento. 58

Figura 12 – Diagrama de atividades – Extração de termos. 59

Figura 13 - Configuração dos contextos em XML. 60

Figura 14 – Caso de uso Configurar Contexto 63

Figura 15 – Caso de uso Executar consulta – Expansão automática. 64

Figura 16 – Caso de uso Executar consulta – Sugestão de termos. 64

Figura 17 - Tela de identificação dos usuários. 65

Figura 18 - Tela de listagem de contextos configurados. 66

Figura 19 - Tela de busca, na modalidade expansão automática. 66

Figura 20 - Exibição de resultados em diferentes abas. 67

Figura 21 - Tela de busca, na modalidade sugestão de termos. 68

Figura 22 - Resultados de busca na modalidade sugestão de termos. 68

Figura 23 - Avaliação de relevância dos resultados de busca. 69
Figura 24 - Registros de consultas e avaliações feitas pelos usuários. 70
Figura 25 - Menu inicial da administração dos contextos. 71
Figura 26 - Formulário de alteração de parâmetros. 71
Figura 27 - Formulário de execução de atividades. 73
Figura 28 – Site de Informações do Coped 81
Figura 29 – Resultados da métrica Precisão total na modalidade Expansão automática 91
Figura 30 – Resultados da métrica Profundidade de busca na modalidade Expansão automática 92
Figura 31 – Resultados da métrica Correlação de ranking na modalidade Expansão automática 93
Figura 32 – Resultados da métrica Precisão total na modalidade Sugestão de termos .. 93
Figura 33 – Resultados da métrica Profundidade de busca na modalidade Sugestão de termos 94
Figura 34 – Resultados da métrica Correlação de ranking na modalidade Sugestão de termos 95
Figura 35 – Resultados da métrica Precisão total na modalidade Expansão automática, limitando o número de termos utilizados na expansão 99
Figura 36 – Resultados da métrica Precisão total na modalidade Expansão automática, com tamanho da expansão maior que o tamanho da consulta original 100

Lista de tabelas

Tabela 1: opções de resposta do questionário 83

Tabela 2: Resultados avaliados 88

Tabela 3: Referências a publicações 101

Tabela 4: Identificação de assuntos 102

Capítulo 1 – Introdução

Neste capítulo serão apresentados a motivação desta pesquisa, o problema abordado e a hipótese assumida. Além disso, serão apresentados os objetivos do trabalho e a organização da dissertação.

1.1. Motivação

É difícil estimar a quantidade de informações disponíveis no mundo atualmente, e qual a taxa de crescimento deste volume. Como exemplos de estimativas, o relatório da Netcraft aponta a existência de mais de 270 milhões de sites no mundo (NETCRAFT 2011) enquanto um estudo realizado com dados entre 2000 e 2008 identificou que a Internet dobra de tamanho a cada 5 anos (ZHANG *et al.*, 2008).

As ferramentas de busca disponíveis na Internet, tais como Google[1], Bing[2] e Yahoo[3], foram projetadas para facilitar a descoberta de informações publicadas em sites de todo o mundo. E o uso destas ferramentas se torna cada vez mais frequente pelos internautas. Enquanto em 2002 o percentual de acessos às ferramentas de busca dentre todos os sites da Internet era de aproximadamente um terço, em 2008 este valor chegou a aproximadamente 50% (PEW RESEARCH CENTER, 2008).

[1] http://www.google.com
[2] http://www.bing.com
[3] http://www.yahoo.com

Na maioria das ferramentas de busca, o usuário informa palavras-chave em consultas, representando a necessidade de informação. Embora esta abordagem apresente como grande vantagem para os usuários a facilidade em expressar a necessidade de informação, ela possui como principal problema a falta de contextualização. Contudo, a recuperação de informações é fortemente dependente do contexto; o que é relevante para um usuário em um local determinado, com experiências e necessidades específicas, pode não ser relevante para outro usuário que possua características e vivências distintas ou que esteja em outro local.

Outros problemas encontrados na abordagem de palavras-chave são: (i) o usuário precisa conhecer as palavras-chave que devem estar presentes no conteúdo dos resultados que atenderão a sua necessidade de informação; e (ii) as palavras-chave utilizadas nas consultas podem ter diferentes significados e a desambiguação destes conceitos deve ser feita pelo próprio usuário, através de realização de novas consultas com a inclusão ou alteração de palavras-chave.

As ferramentas de busca na web procuram incluir em seus mecanismos funcionalidades que aumentem a contextualização dos resultados de pesquisa. Como exemplo destas funcionalidades, podem ser citadas o uso da localização geográfica (obtida a partir do cadastro de preferências ou endereço IP do computador) para priorização dos resultados; definição do idioma; filtragem; sugestões de pesquisas, baseadas em pesquisas históricas feitas pelos usuários; e opções de pesquisa por páginas similares a páginas de resultados inicialmente retornados (GOOGLE, 2011, YAHOO, 2011).

O comportamento dos usuários ao realizar as buscas na Internet potencializa os problemas apontados, ratificando a necessidade da utilização de informações de contexto. Usuários acessam poucas páginas de resultado, frequentemente se limitando

aos cinco (5) primeiros resultados de busca (SPINK e JANSEN, 2004), e embora possa ser lógico que melhores resultados são obtidos com a execução de consultas mais elaboradas, as estatísticas mostram que as expressões de busca normalmente são compostas de poucos termos: 65% das buscas realizadas na Internet contêm de 1 a 3 termos (EXPERIAN, 2010). Logo, é desejado que as ferramentas de busca atendam as necessidades de informação nos primeiros resultados, sendo que para isso sejam fornecidos poucos termos na expressão de busca.

Para tornar os resultados de busca mais relevantes considerando o comportamento dos usuários ao elaborar uma expressão, é utilizada uma técnica de recuperação de informações conhecida como expansão de consultas, que consiste em adicionar termos na consulta original feita pelo usuário em uma tentativa de prover uma maior contextualização, recuperando documentos mais úteis (YATES, NETO, 1999). Os termos adicionados à consulta, extraídos do contexto da necessidade de informação, podem minimizar os problemas relatados. Neste trabalho, considera-se contexto como qualquer informação que possa ser utilizada para caracterizar a situação de uma entidade, sendo esta entidade uma pessoa, lugar ou objeto que possa ser considerada relevante para a interação entre o usuário e a aplicação, incluindo o próprio usuário e a aplicação (DEY e ABOWD, 1999).

Uma aplicação é considerada sensível ao contexto se utiliza informações do contexto para prover informações relevantes ou serviços para o usuário, em que relevância depende da tarefa a ser realizada pelo usuário (DEY e ABOWD, 1999).

Diversos métodos de expansão de consulta já foram propostos na tentativa de tornar as buscas sensíveis ao contexto, porém podem ser diferenciados do presente trabalho em relação à forma de modelar o contexto, e na origem dos dados utilizados

para gerar este modelo. Um resumo sobre estes trabalhos pode ser encontrado na seção 2.3.

1.2. Hipótese de pesquisa

A hipótese desta pesquisa é que a expansão de consultas dos usuários com termos extraídos de recursos pré-existentes, com a aplicação de técnicas de segmentação de tópicos e agrupamento, torna os resultados de busca mais contextualizados e relevantes dos que os obtidos com as consultas originais.

1.3. Enfoque de solução

Para que a busca se torne sensível a um contexto de domínio, o método de expansão de consultas proposto neste trabalho se baseia na seguinte ideia: os termos encontrados com maior frequência em informações representativas de um domínio possuem maior probabilidade de também estarem presentes em outros documentos disponíveis na Internet que sejam relevantes e relacionados com este domínio. Logo, ao utilizar estes termos para expandir as consultas realizadas pelos usuários pode-se obter resultados de pesquisas mais úteis.

Porém, considerando que cada informação representativa de um contexto pode tratar em seu conteúdo de assuntos diferentes, embora todos relacionados ao contexto, a simples extração de termos baseada em ocorrências pode acabar combinando em uma consulta termos de assuntos diferentes, possivelmente reduzindo as probabilidades de obtenção de resultados úteis na busca.

Para minimizar este risco, duas extrações de termos são realizadas: uma em que o cálculo dos pesos dos termos é baseado somente na frequência, para obter os termos gerais do contexto, e outra cujo cálculo leva em consideração a identificação dos

diferentes assuntos tratados no contexto e a frequência dos termos dentro de cada grupo de informações representante de um assunto. Cada extração de termos dará origem a uma diferente expansão de consulta.

Para extrair os termos dos assuntos, é necessário o processamento das informações do contexto com atividades de extração de informação. O conteúdo de cada informação do contexto é segmentado em tópicos, e cada tópico é considerado como um documento independente. Após a segmentação em todas as informações, os tópicos são agrupados em assuntos, para que enfim a extração de termos possa ser realizada.

1.4. Metodologia de pesquisa

Para comprovar a hipótese de pesquisa foi desenvolvida em Java a ferramenta de expansão de consultas chamada *ContextSearch*. Como a solução combina diferentes técnicas de extração de informação, que em seus algoritmos podem utilizar diferentes parametrizações, foi necessária a realização de experimentos em laboratório para definir a influência destes parâmetros nas expansões de consultas realizadas.

Após a definição dos parâmetros a serem utilizados, um estudo de caso piloto foi realizado em uma disciplina de mestrado da UNIRIO (PRATES e SIQUEIRA, 2010) contando com a participação de sete (7) estudantes.

Após alterações na ferramenta e no processo de avaliação, decorrente das observações realizadas no estudo de caso piloto, foram realizados dois estudos de caso em dois ambientes diferentes, um educacional (PRATES e SIQUEIRA, 2011) e outro corporativo. As justificativas para a escolha do estudo de caso como método de pesquisa podem ser encontradas no Capítulo 5.

Nestes estudos de caso foi considerada como variável dependente a relevância dos resultados obtidos nas pesquisas. Como variável independente foi considerado o número

de termos utilizados nas expansões de consulta, e como variáveis intervenientes foram considerados o perfil dos usuários selecionados para cada estudo de caso e os diferentes contextos utilizados.

A aplicação dos estudos de caso em ambientes diferentes teve como objetivo verificar se a motivação das pesquisas e o perfil dos participantes envolvidos traria diferenças significativas nas avaliações da ferramenta. Embora resultados positivos em ambos os ambientes não possam levar a uma conclusão de generalização, um resultado positivo em somente um contexto poderia indicar a utilização em um ambiente específico, com condições específicas.

Em cada estudo de caso, foi proposto aos participantes que elaborassem necessidades de informações, expressões de busca que traduzissem estas necessidades, e utilizassem o ContextSearch para fazer pesquisas. A principal avaliação feita pelos usuários foi a atribuição de relevância dos primeiros resultados encontrados, realizada diretamente na ferramenta. Além da avaliação de relevância, também foram consideradas as condições de uso da ferramenta através de análise do log de atividades (por exemplo, o número de termos selecionado para expansão e digitados na consulta). Também foi aplicado aos participantes dos dois estudos de caso um questionário, que abordava questões qualitativas em relação a aplicação da ferramenta nas atividades propostas.

Os resultados dos julgamentos de relevância foram avaliados de acordo com métricas consolidadas por (TANG e SUN, 2003) para a avaliação de buscadores web: Precisão total dos 10 primeiros resultados, profundidade da busca e correlação de ranking. Já os resultados obtidos com os questionários foram avaliados com a utilização de técnicas estatísticas para as perguntas fechadas, e com análise de conteúdo para as perguntas abertas.

1.5. Objetivos da dissertação

O objetivo principal deste trabalho é propor um método de expansão de consultas que possa ser utilizado em sistemas de informação para apoiar as ferramentas de busca web, tornando-as sensíveis ao contexto de domínio da necessidade de informação dos usuários. Além de tornar os resultados de pesquisa mais úteis, este método deve requerer baixo custo para a definição do contexto, exigindo pouco tempo e esforço de especialistas.

Para atingir este objetivo principal, outros três objetivos foram definidos e alcançados: a definição de uma sistemática para a extração de informações em conteúdo não estruturado que gerem termos para a expansão de consultas; o desenvolvimento de uma ferramenta em Java que implemente o método proposto; e a aplicação desta ferramenta em dois ambientes distintos: um corporativo e um educacional.

1.6. Organização da dissertação

Na figura 1 é apresentado o mapeamento da pesquisa nos demais capítulos da dissertação. A seta tracejada indica relação entre o projeto de pesquisa e os capítulos.

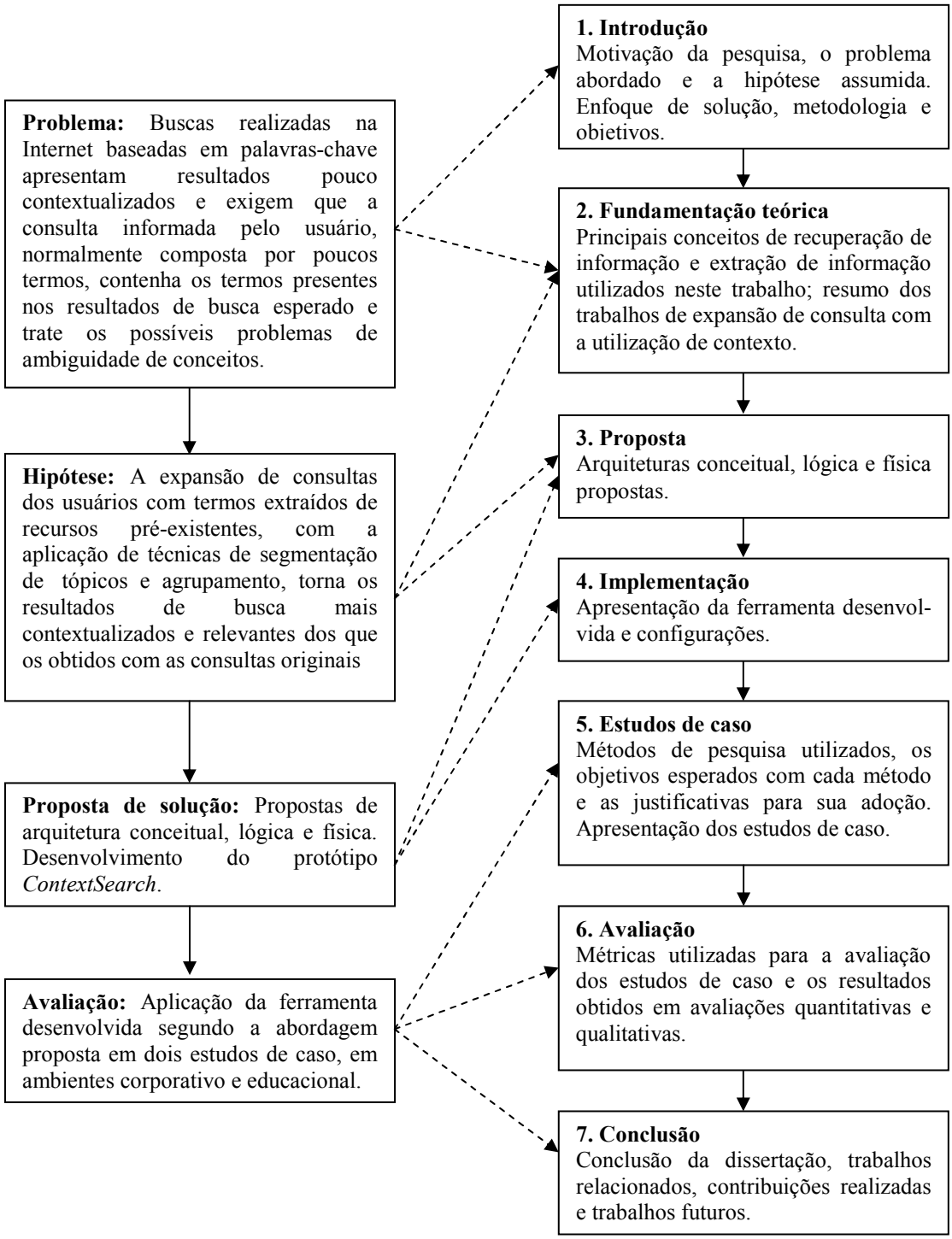


Figura 1 - Organização da dissertação

Capítulo 2 – Fundamentação Teórica

Neste capítulo serão apresentados alguns dos conceitos considerados fundamentais para o entendimento deste trabalho. Os conceitos de recuperação de informação serão necessários para entender o funcionamento básico das ferramentas de busca. A representação dos documentos como vetores de termos e os cálculos de pesos utilizados nos termos serão fundamentais para estabelecer critérios para a seleção dos termos a serem utilizados nas expansões de consultas. Por outro lado, os conceitos de extração de informação serão necessários para entender as atividades aplicadas no método proposto para a preparação dos documentos representativos do contexto. Neste capítulo também são apresentados os trabalhos envolvendo recuperação de informação e expansão de consultas com o uso do contexto.

2.1. Recuperação de informação

Encontrar um documento que contenha a informação desejada em um grande conjunto de documentos que em sua maioria são de natureza não estruturada é a motivação da recuperação de informações (Figura 2).

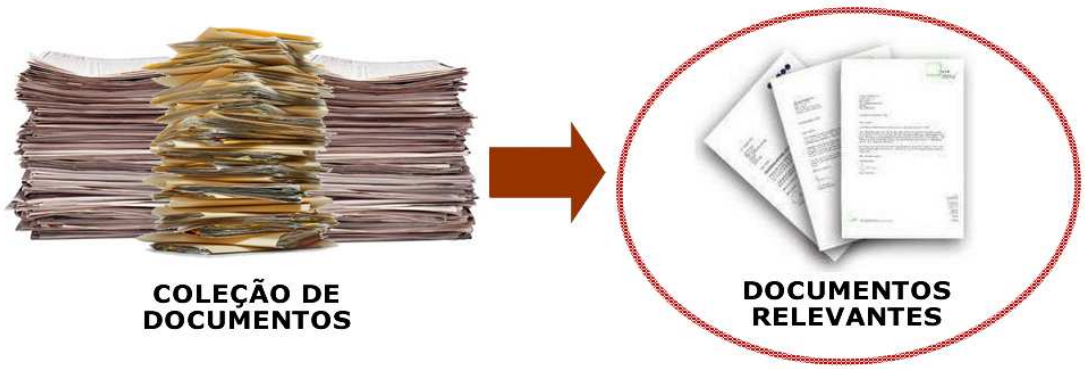


Figura 2 – Motivação da recuperação de informações

Pode-se definir a recuperação de informações (RI) como a área de estudos que trata de buscas por informações de natureza não estruturada que satisfaçam a uma necessidade de informação dentro de grandes coleções (MANNING, RAGHAVAN e SCHÜTZE, 2008). A partir desta definição, três principais conceitos podem ser extraídos:

- Informações de natureza não estruturada: registros que possuam primariamente conteúdo textual em linguagem natural. Referenciados em RI como documentos;
- Necessidade de informação: representada em sistemas de recuperação de informação como uma consulta, geralmente composta por palavras-chave que possam resumir os tipos de documentos que devem ser recuperados;
- Coleções: Conjunto de documentos armazenados em um repositório, também conhecidos em RI como *corpus*.

O foco da recuperação de informações é recuperar documentos com base na comparação de características entre o seu conteúdo textual e o conteúdo de consultas elaboradas pelos usuários, compostas por palavras-chave. Documentos que satisfaçam a uma consulta são ditos relevantes.

2.1.1. Abordagens em Recuperação de Informações

Segundo GREENGRASS (2000), as abordagens em recuperação de informações podem ser divididas (grosseiramente) em semânticas e estatísticas. As abordagens semânticas tentam implementar algum grau de análise sintática e semântica objetivando compreender o texto em linguagem natural da mesma forma que um usuário humano o faria. Na abordagem estatística, os documentos recuperados são aqueles mais semelhantes em relação à consulta do usuário em termos de alguma medida estatística. As abordagens estatísticas dividem-se em várias categorias: booleana, booleana

estendida, espaço vetorial e probabilístico. O princípio básico destas abordagens é segmentar documentos e consultas em termos, que são contados e medidos estatisticamente. Mais comumente, os termos são palavras que ocorrem em uma determinada consulta ou coleção de documentos.

Antes de aplicados os métodos estatísticos, os conteúdos textuais precisam passar por atividades de pré-processamento, como a remoção de palavras normalmente mais encontradas nos textos e que não possuam significância para determinar a relevância ou não dos documentos (conhecida como remoção de *stop words*) e a redução das palavras à sua forma raiz, para que as variações gramaticais de uma mesma palavra sejam eliminadas.

2.1.1.1. Abordagem booleana

Na abordagem booleana, a consulta é formulada como uma combinação booleana de termos, utilizando os operadores booleandos clássicos AND, OR e NOT. Consultas são construídas a partir destes operadores e são avaliadas de acordo com as regras clássicas da álgebra booleana (MANNING, RAGHAVAN e SCHÜTZE, 2008).

Assim como uma consulta booleana é verdadeira ou falsa, um documento é considerado relevante de satisfaz uma consulta ou não relevante se não a satisfaz, sendo que os resultados de consulta não são priorizados de acordo com a relevância.

2.1.1.2. Abordagem booleana estendida

Nesta abordagem se utilizam pesos atribuídos aos termos em cada documento, sendo flexibilizado o retorno das expressões booleanas clássicas, representadas numericamente como 1 caso o documento atenda a expressão ou 0 caso não atenda, para valores que podem variar entre 0 e 1, correspondendo ao grau de correspondência entre a expressão booleana e um dado documento (LEE, 1994).

2.1.1.3. Espaço vetorial (SALTON, 1989)

Uma abordagem comum para a representação e indexação de documentos para fins estatísticos é a representação de cada documento como um conjunto de termos. Este processo é aplicado a uma determinada coleção, gerando um conjunto de termos que representa determinado documento. A união entre todos os conjuntos de termos que representa cada documento gera um conjunto de termos que representa a coleção inteira. Esse conjunto de termos define um "espaço" de tal forma que cada termo representa uma dimensão diferente nesse espaço. Como cada documento é representado como um conjunto de termos, este espaço pode ser visto como um "espaço de documentos".

Nesta representação, são atribuídos pesos numéricos para cada termo em um determinado documento, o que representa uma estimativa de quão útil é um determinado termo para descrever um determinado documento, ou seja, uma estimativa de sua utilidade para distinguir este documento dentre outros documentos na mesma coleção. Os pesos atribuídos aos termos em um determinado documento podem ser interpretados como as coordenadas deste documento no espaço do documento, ou seja, o documento é representado como um vetor no espaço de documentos.

A consulta especificada pelo usuário também pode ser representada como um conjunto de termos com a atribuição de pesos numéricos, recebendo os mesmos tratamentos que os documentos (remoção de *stop words*, redução à forma raiz). Assim, a consulta pode ser interpretada como outro documento no espaço de documentos.

Com documentos e consulta sendo representados no espaço vetorial, são utilizados cálculos matemáticos para determinar o quão semelhantes são a consulta e cada documento da coleção.

2.1.1.4. Abordagem probabilística

Nesta abordagem a probabilidade de que um documento específico seja julgado como relevante para uma consulta específica é baseada na suposição de que os termos são distribuídos de forma diferente em documentos relevantes e não relevantes. A fórmula de probabilidade é normalmente derivada a partir do teorema de Bayes (FUHR, 1992).

2.1.1.5. Abordagens de processamento de linguagem natural

As abordagens que utilizam processamento de linguagem natural são aquelas em que se baseiam no conhecimento da sintaxe e / ou semântica da linguagem natural no qual o documento é escrito ou o conhecimento do domínio da aplicação ao qual os documentos se referem. Desta forma, estas abordagens também podem ser caracterizadas como abordagens semânticas, pois tentam abordar a estrutura e o significado dos documentos textuais ao invés da simples utilização de medidas estatísticas (FALOUTSOS e OARD, 1995).

2.1.2. Cálculo de pesos

Para representar documentos ou consultas em um espaço vetorial, é necessário o cálculo de pesos para cada termo presente, sendo que este valor deve depender do número de ocorrências do termo nos documentos e nas consultas. A abordagem mais simples é atribuir o peso exatamente igual ao número de ocorrências (frequência do termo), sendo denotado por $tf_{t,d}$, *t* representando o termo e *d* o documento.

O cálculo de peso conforme definido acima apresenta o problema de considerar todos os termos igualmente importantes para a determinação da relevância em relação a uma consulta. Porém, certos termos têm pouco ou nenhum poder de discriminação na determinação de relevância e isto deve ser refletido no cálculo do peso de cada termo.

Para atenuar o efeito dos termos que ocorrem com muita frequência em uma coleção, uma idéia inicial seria reduzir o peso dos termos proporcionalmente à frequência deste termo em toda a coleção de documentos, fazendo com que o peso seja reduzido por um fator que cresce com sua frequência na coleção. Porém, é mais comum ser utilizado como fator de redução de peso o cálculo de frequência de documentos, denotado como df_t e representando o número de documentos na coleção que contém o termo *t*.

Como o peso do termo deve ser inversamente proporcional à frequência de documentos, é utilizado o conceito de frequência inversa de documentos em relação ao termo t, definido por:

$$idft = \frac{\log N}{dft}$$

Sendo N o número de documentos na coleção. Desta forma, a frequência inversa de um termo raro na coleção é alta, assim como a de um documento frequente é baixa.

A definição de frequência de termos e a frequência inversa de documentos são combinadas para produzir a regra de cálculo de pesos tf-idf, muito utilizada em recuperação de informações e definida por:

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t$$

Em outras palavras, $tf\text{-}idf_{t,d}$ atribui a um termo t em um documento d um peso:

- Alto quando t ocorre muitas vezes em um pequeno número de documentos;
- Baixo quando o termo t ocorre poucas vezes no documento, ou quando ocorre em vários documentos;
- Menor ainda, quando o termo ocorre em praticamente todos os documentos.

2.1.3. Expansão de consultas

A expansão de consulta é uma técnica aplicada em recuperação de informações que provê uma reformulação da necessidade de informação para melhorar a efetividade da busca (RIJSBERGEN, CRESTANI e LALMAS, 1998). Mais especificamente, no contexto desta dissertação, consiste em adicionar termos na consulta original feita pelo usuário, em uma tentativa de prover uma maior contextualização e, consequentemente, recuperar documentos mais úteis (YATES e NETO, 1999).

Este processo de adição de termos pode ser classificado como manual ou automático (RIJSBERGEN, CRESTANI e LALMAS, 1998). Uma classificação intermediária entre a manual e a automática é a interativa, ou assistida pelo usuário (KACPRZYK, FRANÇA e SOUZA, 2008). A expansão de consulta manual se baseia na experiência do usuário para tomar decisões sobre quais os termos devem ser incluídos na nova consulta, sem o apoio do sistema de busca. No caso de expansão automática de consulta, pesos são calculados para todos os termos e os termos com o maior peso são adicionados à consulta inicial. Funções de cálculo de pesos diferentes produzem resultados diferentes, portanto, o desempenho de recuperação depende de como estes pesos foram calculados. Com a expansão da consulta assistida pelo usuário, ou interativa, o sistema gera possíveis termos a serem incluídos na consulta e o usuário seleciona quais desses deve incluir.

Os novos termos extraídos a partir do método de seleção escolhido deve fornecer informação contextual para a consulta original com o objetivo de melhorar os resultados da busca. A informação contextual pode ser adquirida a partir de realimentação de relevância (*feedback* de relevância), a co-ocorrência entre termos ou derivadas a partir de modelos de conhecimento, tais como ontologias.

2.1.4. Avaliação

Para medir a eficácia de recuperação de informação na forma padrão, é necessária uma coleção de testes composta pelos seguintes elementos:

- Uma coleção de documentos;
- Um conjunto de testes de necessidades de informação, representados como consultas;
- Um conjunto de decisões de relevância, normalmente representada como uma atribuição binária de relevância para cada par documento-consulta. Um par documento-consulta é marcado como relevante se o documento atende a necessidade de informação representada pela consulta, ou não relevante caso contrário.

Como exemplo de coleções de testes existentes podem ser citadas as coleções de Crainfield[4], TREC[5], GOV2[6], Reuters[7], dentre outras.

A efetividade de um sistema de recuperação de informações é medida através do cálculo de métricas específicas. As duas métricas mais básicas e mais utilizadas são a precisão e cobertura.

Precisão é a fração de documentos recuperados por uma consulta e que são relevantes (MANNING, RAGHAVAN e SCHÜTZE, 2008).

$$\text{Precisão} = \frac{\text{total de itens relevantes recuperados}}{\text{total de itens recuperados}}$$

Cobertura é a fração de documentos relevantes dentro da coleção que são recuperados (MANNING, RAGHAVAN e SCHÜTZE, 2008).

[4] https://dspace.lib.cranfield.ac.uk/
[5] http://trec.nist.gov/
[6] http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
[7] http://www.reuters.com/

$$\text{Cobertura} = \frac{\text{total de itens relevantes recuperados}}{\text{total de itens relevantes}}$$

2.2. Extração de informação

Antes de apresentar uma definição formal para a extração de informações, vale a pena descrever um cenário fictício, porém com motivações bastante comuns nos dias atuais: uma empresa de recrutamento e seleção oferece como um de seus serviços a recolocação de profissionais no mercado de trabalho. Para isso, pesquisa todas as oportunidades de trabalho disponíveis no mercado e redireciona para seus clientes, levando em consideração as características das oportunidades e dos perfis profissionais dos seus clientes. A pesquisa por vagas é feita em muitas fontes de informação, tais como jornais online, sites de vagas e sites de grandes empresas. A atualização destas informações é muito rápida, exigindo que essas pesquisas sejam feitas em curtos intervalos de tempo.

O exemplo acima possui algumas características interessantes: (i) a necessidade de informação; (ii) a resposta para esta necessidade está disponível normalmente em conteúdos não estruturados, escritos em linguagem natural; (iii) é extremamente difícil e tedioso para um humano verificar todas as fontes existentes, devido ao fato de serem muitas e, finalmente, (iv) é muito complexo fazer com que um computador consulte e processe diretamente estas informações, pois elas não estão armazenadas de forma estruturada.

Este tipo de problema é tratado com técnicas de extração de informação, que procura capturar em conteúdos não estruturados informações que possam ser armazenadas de forma estruturada, conforme a Figura 3.

Figura 3 – Exemplo de extração de informações

A extração de informações estuda técnicas para a obtenção de informações específicas de um domínio a partir de conteúdos textuais em linguagem natural. As principais características encontradas em uma técnica de extração de informações são: uma necessidade de informação a ser obtida de uma fonte de dados, normalmente composta por conteúdo textual ou imagens; a impossibilidade de obter a resposta por humanos, devido ao grande volume de dados; e a inexistência de estrutura formal nos dados como a obtida em um banco de dados, o que não permite uma consulta direta feita por computadores (MOENS, 2006).

Formalmente, extração de informações pode ser definida como a extração automática de informações estruturadas, tais como entidades, relacionamento entre entidades e atributos descrevendo entidades, a partir de informações não estruturadas,

possibilitando formas mais ricas de consulta a fontes de informação não estruturadas do que utilizando somente busca por palavras-chave (SARAWAGI, 2008).

Explicando a definição, a informação ser classificada como não estruturada significa dizer que seu conteúdo está apresentado de tal forma que dificulta a sua interpretação por computadores. Como exemplos de informação não estruturada estão textos escritos em linguagem natural, áudio, vídeo e imagens.

Informações semânticas podem ser extraídas de informações não estruturadas com a aplicação de técnicas que utilizam padrões de extração, manualmente construídos ou "aprendidos" em coleções de documentos (MOENS, 2006). Estes padrões são obtidos a partir da organização linguística do conteúdo das informações, sendo possível utilizar seus elementos léxicos (palavras e grupos de palavras), construções gramaticais (frases, sentenças, expressões temporais) e estrutura discursiva (parágrafos, capítulos, etc..).

2.2.1. Algumas atividades de extração de informação

Existe um número de tarefas típicas de extração de informação, dentre as quais podem ser citadas (MOENS, 2006):

- Reconhecimento de entidades: reconhecimento e classificação de expressões próprias em texto, tais quais pessoas, empresas ou localidades;
- Correferência de frases: reconhecimento de frases que façam referência a mesma situação ou entidade em um texto;
- Reconhecimento de papel semântico: atribuição de papéis semânticos a elementos sintáticos que constituem uma sentença;
- Relação entre entidades: reconhecimento do relacionamento entre duas ou mais entidades, sendo atribuído a este relacionamento um papel semântico;
- Reconhecimento temporal: detecção e reconhecimento no texto de expressões temporais.

2.2.2. Técnicas de extração de informação

As técnicas de extração de informação podem ser classificadas como técnicas simbólicas e técnicas de reconhecimento de padrão, sendo que esta última também pode ser classificada de acordo com o modo de classificação: supervisionado ou não supervisionado.

As técnicas simbólicas são aquelas que utilizam regras para a extração definidas manualmente com a ajuda de um especialista no domínio da aplicação, também conhecido como engenheiro do conhecimento (MOENS, 2006). Em um cenário típico, o especialista tem acesso a um *corpus* (coleção de documentos) de tamanho moderado que seja representativo do domínio da aplicação. Este *corpus* é inspecionado para a definição das regras de extração.

O alto custo da definição manual de regras e da aquisição do conhecimento fez com que nas décadas de 80 e 90 os pesquisadores investigassem a possibilidade de utilização de abordagens baseadas em aprendizado de máquina.

As técnicas de reconhecimento de padrões (também conhecidas como classificação de padrões) objetivam classificar dados (padrões) baseados tanto em conhecimentos prévios adquiridos por especialistas quanto em conhecimento automaticamente aprendido a partir de dados. A classificação de padrões pode ser vista como uma combinação de recursos (*features*) e seus valores (MOENS, 2006). No caso da extração de informações, esses recursos são características textuais que podem ser identificadas ou medidas, por exemplo termos e pesos associados a estes termos, e que possuam valores dicriminativos quando da organização de padrões em classes semânticas.

A maioria das aplicações em extração de informações utiliza técnicas supervisionadas de reconhecimento de padrões. Nestas técnicas, o classificador treina a

partir de uma coleção de exemplos previamente classificados. A idéia geral é que um especialista possa anotar os fragmentos que deveriam ser extraídos em um pequeno subconjunto de documentos para que então o sistema possa generalizar a partir destes exemplos e produzir uma função ou regra que possa ser aplicada ao resto dos documentos. A aplicabilidade destas técnicas reside no fato de que é mais fácil anotar documentos do que escrever regras de extração.

Embora muito utilizadas em extração de informação, as técnicas supervisionadas tem como principal crítica o fato de funcionam bem somente em domínios fechados. Em domínios abertos, o maior problema reside na falta de exemplos anotados suficientes para o aprendizado.

Em contraponto às técnicas supervisionadas, encontram-se os algoritmos de agrupamento, que são completamente não supervisionados e que não dependem de dados anotados. Estes algoritmos detectam a organização de padrões similares em agrupamentos, que permitem a descoberta de similaridades e diferenças entre os padrões, e podem derivar conclusões úteis sobre os agrupamentos.

2.2.3. Segmentação de tópicos

A Segmentação de tópicos é a divisão de um dado textual em segmentos baseados no tópico ou assunto discutido, ou seja, é a delimitação do texto nos pontos em que o tópico que está sendo abordado muda. Um exemplo seria uma revista com cinco diferentes histórias (ou artigos) que, portanto, claramente divide o assunto em cinco diferentes tópicos. Outro exemplo seria um artigo longo que cobre um único tópico abrangente normalmente cobre uma série de subtópicos em seu conteúdo, examinando diferentes aspectos dentro do tópico principal, podendo ser segmentado em diferentes seções.

A atividade de segmentação de texto é muito aplicada na recuperação de informações, principalmente por (i) melhorar o critério de relevância: como o critério para estimar a similaridade entre uma expressão de busca e um documento leva em consideração a freqüência dos termos em relação aos demais termos do documento e da coleção, um documento extenso que aborde diversos assuntos (dentre eles o assunto pesquisado), tende a ser considerado menos relevante que um documento menor, que aborde menos assuntos; e (ii) possibilitar o acesso à informação diretamente no tópico de interesse, sem a necessidade de percorrer seqüencialmente o documento (REYNAR, 1999).

2.2.4. Agrupamento

Algoritmos de agrupamento, ou *clustering,* utilizam técnicas estatísticas que permitem uma geração de grupos de dados. Os vetores de características dos documentos da coleção são agrupados utilizando uma função que calcula a distância numérica ou similaridade entre cada par de documentos. O resultado deste agrupamento é uma partição da coleção em grupos de documentos relacionados (MANNING, RAGHAVAN e SCHÜTZE, 2008).

Os algoritmos de agrupamento podem ser classificados como planos quando os grupos produzidos não possuem nenhuma relação estrutural entre eles, e hierárquicos quando são criadas relações hierárquicas entre os grupos. Os algoritmos também podem ser classificados como rígidos, quando cada documento pertence a um grupo, ou suaves, quando um documento pode pertencer a mais de um grupo com certo grau de associação.

Os algoritmos de agrupamento utilizados em informações textuais possuem alguns importantes desafios em relação as demais aplicações de agrupamento (BEIL, ESTER, XU, 2002):

- Lidar com a grande dimensionalidade dos dados: como as informações textuais são representadas como vetores cujas dimensões são os termos, uma grande quantidade de termos implica em uma grande quantidade de dimensões, fazendo com que estes algoritmos possuam a habilidade de lidar com espaço de dados esparsos ou que utilizem métodos de redução desta dimensionalidade;
- Ligar com a grande quantidade de dados disponível, principalmente ao se lidar com informações da Internet, obrigando os algoritmos a serem eficientes e escaláveis para o uso em grandes coleções de dados;
- Os grupos de dados devem ser compreensíveis, isto é, devem possuir uma descrição que possibilite aos usuários uma navegação mesmo que não sejam especialistas.

2.3. Uso de expansão de consultas para tornar a recuperação de informação sensível ao contexto

Como existem diversos trabalhos relacionados a recuperação de informação sensível ao contexto, pretende-se nesta seção explicar as diferentes abordagens, citando alguns exemplos de trabalhos.

De forma geral, utiliza-se alguma representação de conhecimento para modelar algum tipo de contexto, que pode ser o domínio de alguma informação, o perfil do usuário, a atividade desenvolvida no momento, o histórico de pesquisa ou informações obtidas através de sensores. Após obtido o modelo do contexto, aplica-se algum método para expandir as consultas realizadas.

As diversas abordagens podem ser diferenciadas em relação à forma de modelar o contexto e na origem dos dados utilizados para gerar este modelo.

Com relação à forma, o modelo do contexto pode ser criado manualmente, com a interferência direta de especialistas ou de usuários, por exemplo, através de criação de ontologias, preenchimento de formulários de preferências ou marcação de documentos relevantes. Essa abordagem é utilizada em (CHANANA, GINIGE e MURUGESAN, 2004), que propõe uma metodologia de classificação de contexto para documentos, baseada no tipo de informação contida na coleção. Um exemplo do uso de ontologias pode ser encontrado em (JUN-FENG *et al*., 2005), que propõe um framework de recuperação de informações semânticas, que utiliza a ontologia e expansão de consultas para pesquisar documentos em uma base de dados RDF.

Embora apresentem bons resultados em tornar a busca mais eficiente, a criação manual do contexto apresenta como principais problemas o tempo e o esforço gastos, seja por especialistas ou pelos usuários finais, o que pode inviabilizar o uso das ferramentas de busca.

Outra possibilidade é a criação do contexto de forma automática, através de inferências baseadas no comportamento do usuário (análise de cliques, histórico de navegação e de consultas) ou informações obtidas do ambiente, de alguma base de conhecimento ou de algum *corpus*.

A técnica de mineração de textos é utilizada em apresentações selecionadas em (AMBRÓSIO, SILVA e NETO, 2009) para a criação automática do contexto, com o objetivo de recomendar documentos armazenados em diretórios e sugerir expansões de consultas a serem executadas em buscadores web. Outra forma de criação automática é apresentada em (SHEN, TAN e ZHAI, 2005). Um agente faz buscas na web com consultas expandidas com termos baseados em informações de consultas passadas, e reclassifica o resultado utilizando informações de cliques do mouse dados pelo usuário (*clickthrough*). O uso de cliques também é explorado em (WANG *et al.*, 2009), que

propõe um método que tenta prever quando a informação "clicada" pelo usuário contribuiu ou não para o atendimento da necessidade de informação. Este trabalho também utiliza algoritmo de agrupamento para reunir comportamentos (sequência de ações) similares.

A criação automática de contextos, embora não apresente os problemas da criação manual, pode acarretar resultados incorretos, caso o contexto inferido não reflita a necessidade de informação atual do usuário.

Uma possibilidade de reduzir estes problemas é tornar a criação de contexto uma atividade semi-automática, garantindo a qualidade do contexto com um esforço manual não tão grande. Em (WEN, ZHOU e AGGARWAL, 2007) é proposto um sistema específico para auxiliar tarefas investigativas. Neste trabalho, o contexto é capturado de forma automática com inferências a partir das atividades realizadas pelo usuário, mas é possível que o próprio usuário altere manualmente o contexto capturado.

A segunda possibilidade de classificação das abordagens de expansão de consulta é em relação à origem dos dados utilizados para criar o modelo do contexto. As abordagens são divididas naquelas cujo modelo é dependente de um *corpus*, ou seja, em que os termos da expansão são extraídos após a aplicação de algoritmos de análise e processamento nos documentos de uma coleção; e naquelas em que o modelo de contexto é independente de algum *corpus*, que utilize alguma estrutura pré-existente de conhecimento, como ontologias, dicionários e thesauros (BHOGAL, MACFARLANE e SMITH, 2007).

Dentre as abordagens cujo modelo de contexto é independente de *corpus*, a grande maioria utiliza ontologias como base de conhecimento. As principais vantagens encontradas no uso de ontologias são a grande disponibilidade de ontologias existentes para diversos domínios, o fato de incluírem nomes próprios (pessoas e lugares), a

existência de diversos softwares que suportam a manutenção de ontologias e o uso de uma linguagem padrão para a descrição das ontologias, facilitando a portabilidade (BHOGAL, MACFARLANE e SMITH, 2007). Porém o uso de ontologias também possui desvantagens, como o grande esforço necessário (de tempo, trabalho e aquisição de conhecimento) para a definição de uma ontologia caso ainda não exista uma para o domínio desejado. Um exemplo do uso de uma ontologia de domínio para apoiar a expansão de consultas em um ambiente educacional pode ser encontrado em (BAGHI, BILETSKIY e LI, 2008).

Alguns trabalhos cujo modelo de contexto é dependente de *corpus* e criado automaticamente, classificações similares a do presente trabalho, serão apresentados no Capítulo 7, sendo discutidas as principais diferenças entre as abordagens.

Capítulo 3 – Proposta

Neste capítulo serão apresentadas as arquiteturas conceitual, lógica e física da solução proposta. As arquiteturas são divididas em três módulos: configuração da base de conhecimento, extração de informações e busca.

3.1. Arquitetura conceitual

Para que a busca se torne sensível a um contexto de domínio, a ferramenta de expansão de consultas proposta neste trabalho se baseia no seguinte princípio: os termos encontrados com maior frequência em informações representativas de um domínio possuem maior probabilidade de também estarem presentes em outros documentos disponíveis na Internet que sejam relevantes e relacionados com este domínio. Logo, estes termos são em princípio bons candidatos a serem utilizados para expandir as consultas realizadas pelos usuários, com o objetivo de obter resultados de pesquisas mais úteis.

Seguindo este princípio, a arquitetura conceitual proposta prevê três camadas: uma camada de acesso às informações representativas do contexto, uma camada de obtenção de termos mais frequentes e finalmente uma camada de execução das consultas expandidas (Figura 4).

Porém, considerando que cada informação representativa de um contexto pode tratar em seu conteúdo de assuntos diferentes, embora todos relacionados ao contexto, a simples extração de termos baseada em ocorrências pode acabar combinando em uma

consulta termos de assuntos diferentes, possivelmente reduzindo as probabilidades de obtenção de resultados úteis na busca. Isto porque nos principais motores de busca da Internet a conjunção de palavras na expressão de busca é o principal critério de priorização de relevância, favorecendo os resultados de busca que contenham em seu conteúdo todos os termos digitados na expressão de busca (GOOGLE, 2011).

Para minimizar este risco, duas extrações de termos são realizadas: uma em que o cálculo dos pesos dos termos é baseado somente na frequência destes termos em toda a coleção, para obter os "termos gerais do contexto", e outra cujo cálculo leva em consideração a frequência dos termos nos documentos de cada assunto identificado no contexto, os "termos específicos por assunto". Cada extração de termos dá origem a diferentes expansões de consultas.

Figura 4 - Arquitetura conceitual.

A abordagem proposta considera que a expansão de consulta realizada com termos gerais do contexto retorna documentos mais superficiais, que abordam diversos assuntos dentro do contexto. Já com as expansões com termos de assuntos, a abordagem proposta considera que sejam retornados documentos mais específicos, que sejam relacionados somente ao assunto identificado dentro do contexto.

Para extrair os termos específicos por assunto, é necessário o processamento das informações do contexto com atividades de extração de informação. Para evitar que termos de assuntos diferentes sejam adicionados a uma consulta, o conteúdo de cada informação do contexto é segmentado em tópicos, garantindo que cada segmento trate de um assunto diferente, e sendo considerado como um documento independente dentro da coleção de informações do contexto (Figura 5).

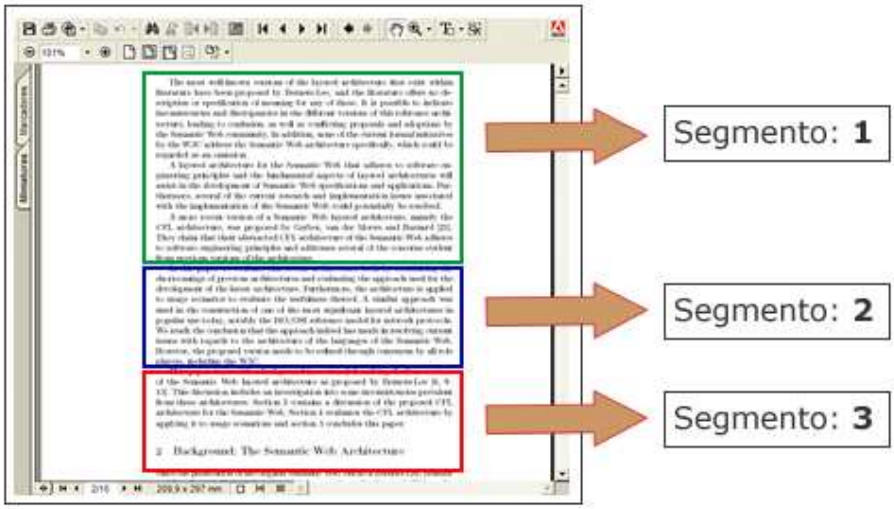


Figura 5 - Segmentação do texto em tópicos.

Após a segmentação de todas as informações, os segmentos são agrupados de acordo com a similaridade, formando subconjuntos de segmentos. Estes subconjuntos representam os diferentes assuntos identificados dentro do contexto e, para cada um deles, são extraídos os termos mais frequentes (Figura 6).

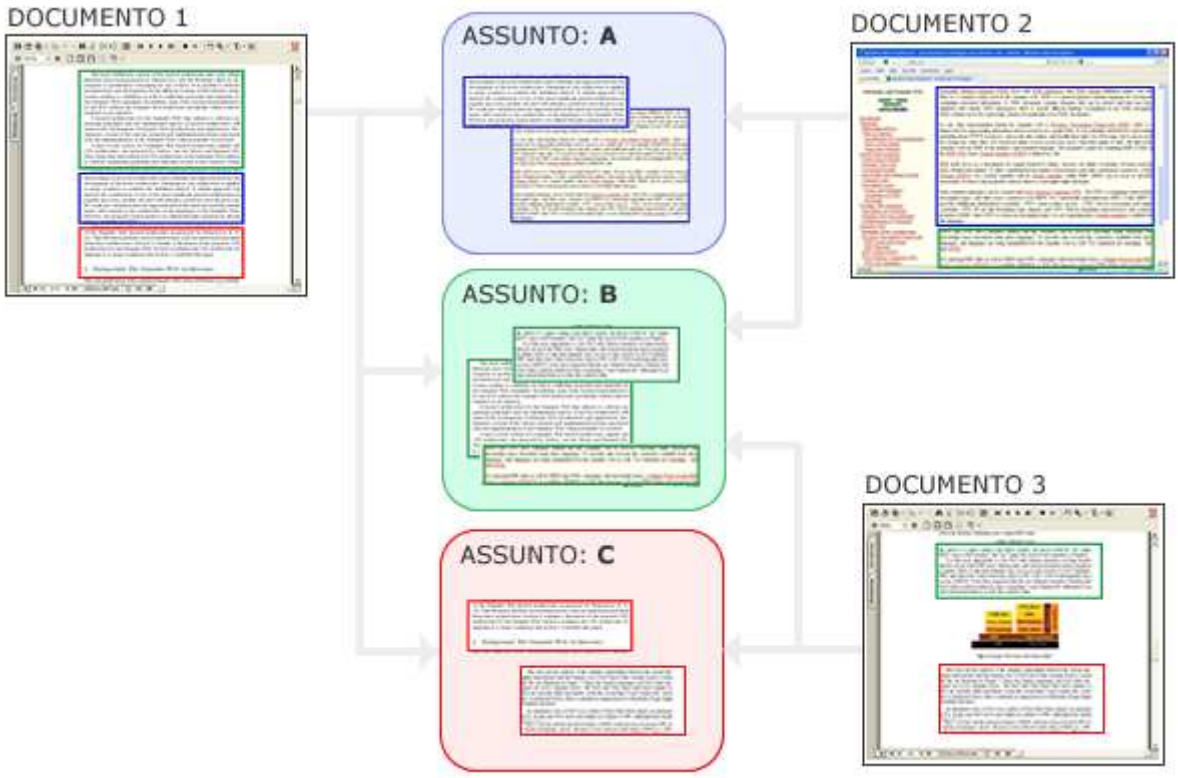


Figura 6 - Agrupamento de segmentos em assuntos.

3.2. Arquitetura lógica

A proposta de solução se baseia na arquitetura lógica apresentada na figura 7. A solução é dividida em três principais módulos: Configuração da base de conhecimento, Extração de informações e Busca, que serão detalhados a seguir.

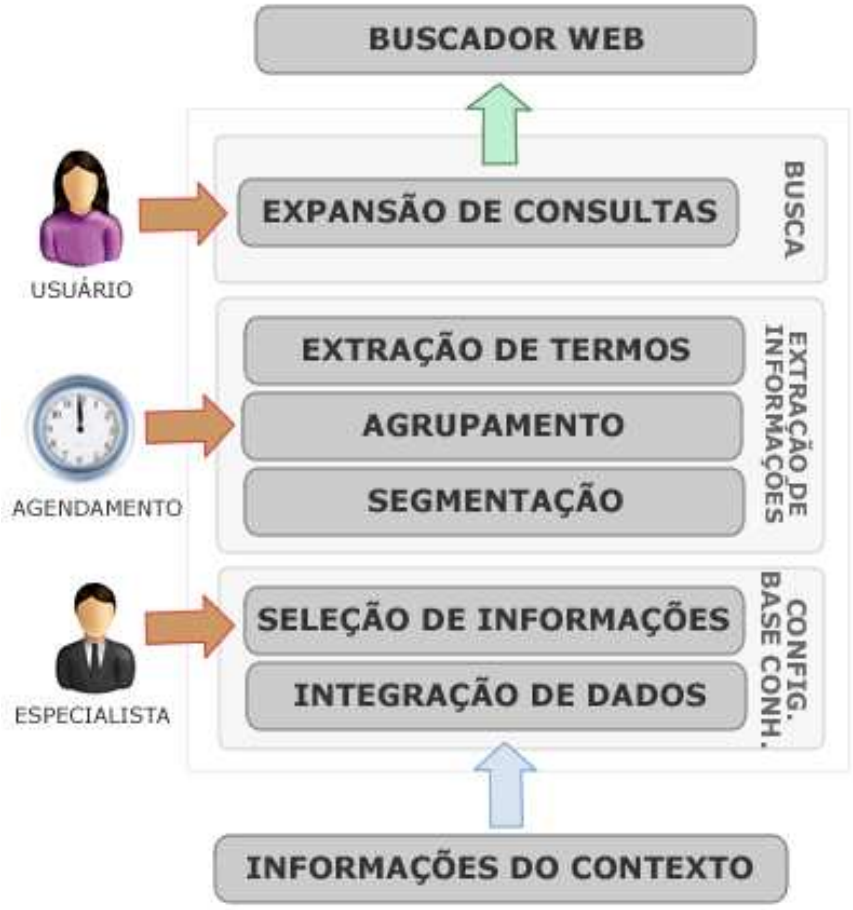


Figura 7 - Arquitetura lógica.

3.2.1. Módulo de configuração da base de conhecimento

A solução proposta considera que contexto de domínio deve ser modelado a partir de recursos pré-existentes, como por exemplo banco de dados, arquivos diversos (artigos, capítulos de livro, publicações em geral) ou sistemas de informação (incluindo sistemas como GEDs - gerenciadores eletrônicos de documentos – e integrações de dados realizadas através de *web services*, RMI - *Remote Method Invocation* – e outros tipos de conectores). Qualquer fonte de informação que contenha conteúdos textuais e cujas informações representem os assuntos abordados em um contexto de domínio pode ser utilizada.

Este módulo possui uma atividade puramente tecnológica, e outra intelectual. Os dados pré-existentes devem ser acessados através do sistema, logo é necessária a criação

de uma camada conectora, chamada de integração de dados, composta por *plugins* específicos que consigam ler cada tipo de informação disponível. Como exemplo de *plugins* é possível citar: *drivers* de conexão a banco de dados e leitores de arquivos PDF, XML e HTML.

A atividade intelectual é mostrada na camada seleção de informações. Um especialista no domínio deve, a partir de todas as fontes de informação existentes e acessíveis pelo sistema, selecionar aquelas que sejam boas representantes dos assuntos que compõem o contexto. Esta atividade é de extrema importância para que as consultas geradas consigam retornar resultados de pesquisa mais úteis, pois somente serão adicionados às consultas originais bons termos caso o ponto de partida seja a existência de boas fontes de informação para representar o contexto.

Além de selecionar as informações utilizadas para representar o contexto, o especialista também precisa identificar, para o grupo de documentos selecionados, a quantidade de diferentes assuntos abordados em seu conteúdo. Este número determinará, na etapa de extração de informações, quantos grupos de segmentos serão criados, e consequentemente a quantidade de expansões de consultas que serão geradas no módulo de busca.

Como o sistema trabalha com múltiplos contextos, é necessária a definição de informações para cada contexto configurado.

3.2.2. Módulo de extração de informações

O objetivo do módulo de extração de informações é identificar os principais termos dentre todas as informações do contexto obtidas do módulo de configuração da base de conhecimento, e disponibilizar estes termos para o módulo de busca. Conforme explicado no início da seção, são realizadas duas extrações de termos, uma para termos

mais frequentes em todo o contexto e outra para termos específicos de cada assunto identificado no contexto.

Em ambas as situações, são necessárias a aplicação de algumas atividades de preparação dos textos antes da extração dos termos: (i) a *Tokenization*, o processo de decomposição de um conteúdo textual em termos; (ii) a remoção das *stop words*, uma lista de termos comuns que ocorrem com frequência nos textos porém possuem pouco valor, como por exemplo preposições, pronomes e artigos; e (iii) *stemming*, o processo de reduzir um termo ao seu radical, removendo as desinências, afixos e vogais temáticas (MANNING, RAGHAVAN e SCHÜTZE, 2008).

A extração dos termos gerais do contexto é feita calculando-se os pesos dos termos e extraindo os *n* termos com maior peso, sendo *n* o número máximo de termos que possa ser utilizado para expansão de consultas.

A extração dos termos específicos de cada assunto identificado no contexto requer a aplicação de outras atividades de extração de informação. O primeiro passo é a aplicação de uma rotina de segmentação de tópicos, uma atividade de processamento de linguagem natural que objetiva identificar subtópicos dentro de um documento, delimitando os seus limites. O objetivo em aplicar esta segmentação é impedir que em um determinado conteúdo textual seja tratado mais de um assunto, ou seja, que existam neste conteúdo textual termos que possam tratar de assuntos diferentes.

Cada segmento identificado pode ser tratado como um documento independente. Mas como vários segmentos na coleção podem tratar do mesmo assunto ou de assuntos semelhantes, estes devem ser agrupados para que seja feita a extração em cada grupo de assunto.

Para isto considera-se o agrupamento, técnica estatística que permite uma geração automática de grupos de dados (documentos). Nesta técnica, os vetores de termos de

cada documento são agrupados de acordo com o cálculo de funções de similaridade e distância aplicadas a cada par de vetores.

Tendo todas as informações do contexto agrupadas por assunto, são calculados os pesos de todos os termos, com a diferença que o somatório de pesos para cada termo é feita considerando somente os documentos de cada grupo de segmentos. Depois de calculados os pesos, os *n* termos com maior peso para cada assunto são identificados e extraídos.

3.2.3. Módulo de busca

O módulo de busca recebe três informações dos usuários: a primeira delas é a indicação de qual contexto de domínio configurado na ferramenta é relacionado com a necessidade de informação atual; a segunda informação é a expressão de busca, composta por palavras-chave, que procura atender a necessidade de informação; e finalmente a seleção de qual ferramenta de busca será utilizada para executar as consultas.

De posse dessas três informações, o módulo de busca consulta os termos fornecidos pelo módulo de extração de informações, expande a consulta original com estes termos e executa a nova consulta no buscador web de preferência do usuário.

3.2.4. Atualização do contexto

Na arquitetura lógica proposta, a execução das atividades dos módulos de dados e de processamento deve ser realizada periodicamente, garantindo desta forma que os contextos capturados possam se manter atualizados. Para isso, esta periodicidade deve ser configurada na ferramenta.

Como exemplo, supondo que em um determinado contexto as informações são consultadas de uma base de conhecimento que é constantemente atualizada. A execução

periódica das atividades de captura e processamento garantirá que a lista de termos extraída estará sempre atualizada, evoluindo de acordo com as mudanças no contexto.

Um fator que pode influenciar a atualização dos contextos é a configuração correta dos filtros de informação utilizados no módulo de configuração da base de conhecimento. Caso o filtro configurado estabeleça critérios relacionados à data para a seleção das informações que serão utilizadas, é possível que seja retirado o peso de termos contidos em informações antigas, que não contribuem mais para o atendimento das necessidades de informação e que novos termos, contidos em informações mais atuais e que representem melhor as necessidades de informações atuais, possam passar a ser extraídos e adicionados nas consultas dos usuários.

Além da correta configuração dos filtros de informação, para que o agendamento das atividades possa atualizar os termos extraídos do contexto garantindo a qualidade das expansões de consulta, é necessário que seja definido um processo para a inclusão de novas informações na base de conhecimento, que assegure que as novas informações também sejam boas representantes do contexto configurado, contribuindo com bons termos para as futuras expansões.

3.3. Arquitetura física

A proposta de arquitetura lógica foi implementada na arquitetura física mostrada na Figura 8. O protótipo ContextSearch foi desenvolvido em Java e em sua implementação foi priorizada a utilização de soluções de código aberto (*open-source*) já disponíveis na comunidade. O desenvolvimento foi aplicado nos casos de funcionalidades em que não foi encontrada solução pré-existente ou na customização de funcionalidades em bibliotecas *open-source* utilizadas.

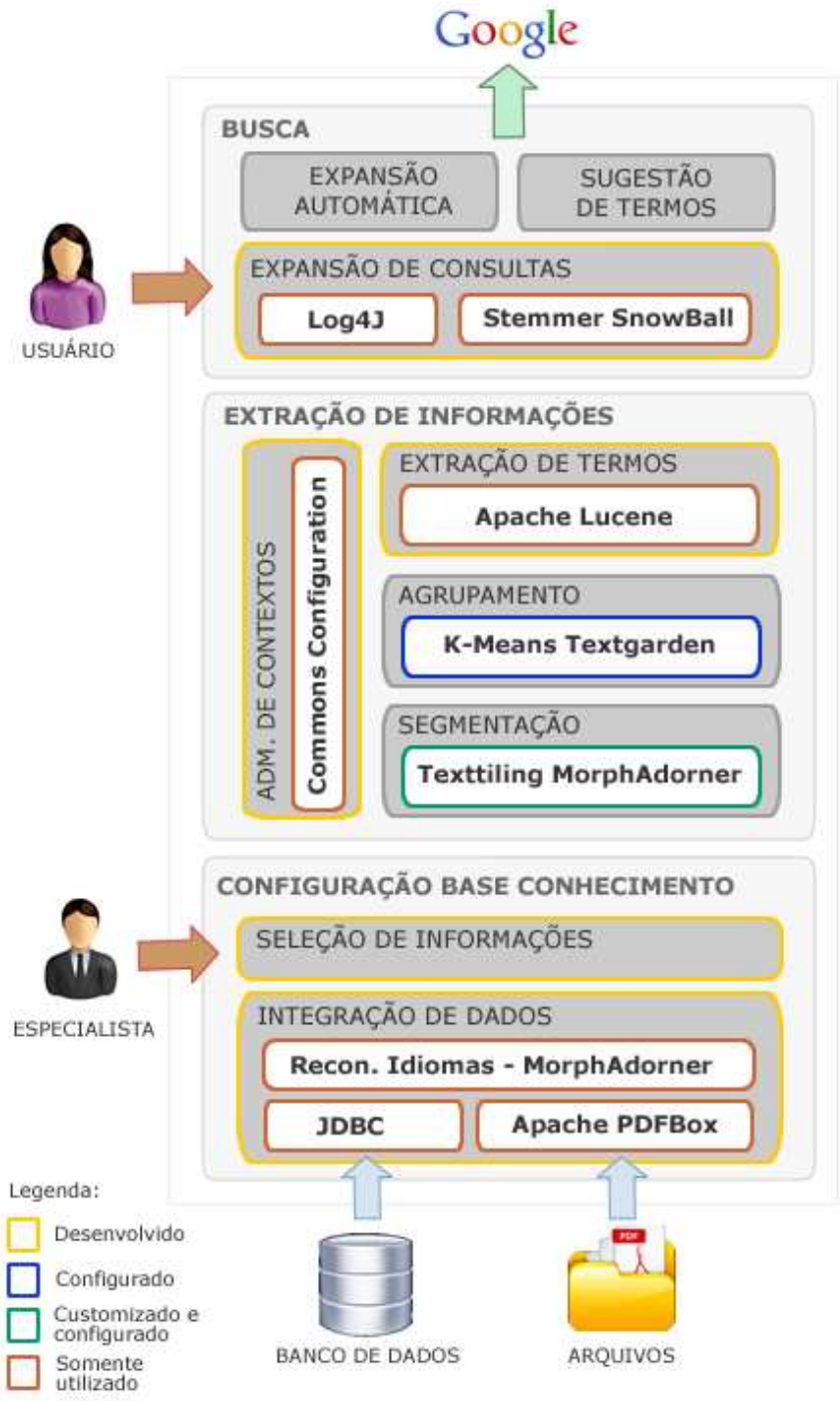


Figura 8 - Arquitetura física.

3.3.1. Módulo de configuração da base de conhecimento

O módulo de configuração da base de conhecimento necessita de uma camada conectora com *plugins* para os diferentes formatos de informação necessários para a captura do contexto. No protótipo, foram acessados dois formatos de informação: no

estudo de caso do ambiente corporativo, foram consultadas notícias em um *site* de informação, sendo implementada esta consulta através de acesso direto ao banco de dados do sistema de gestão de conteúdo utilizado para administrar este site. No estudo de caso do ambiente educacional, o acesso foi feito diretamente a arquivos PDF.

O provedor de banco de dados acessado foi o SQL Server e o acesso foi implementado através de JDBC (*Java Database Connectivity*). Já a leitura de arquivos PDF foi implementada com a utilização da biblioteca *open-source* Apache PDFBox[8]. Esta biblioteca permite a criação de documentos PDF, manipulação de documentos existentes e extração de conteúdo.

Para o estudo de caso no ambiente corporativo, em que foram definidos dois contextos (Energia e Meio Ambiente), a seleção das fontes de informações foi implementada com cláusulas SQL, cuja restrição foi feita a partir da classificação das notícias do site. Para capturar o contexto Energia, foram consultadas todas as notícias classificadas com os subtemas Biocombustíveis, Petróleo e Gás, e Energia Elétrica. Já para o contexto Meio Ambiente, foram consultadas as notícias classificadas como Desenvolvimento sustentável, Mudanças climáticas e Política ambiental.

3.3.1.1. Reconhecimento de idiomas

Uma atividade necessária antes do módulo de extração de termos é o reconhecimento de idiomas. Como nas fontes de informações utilizadas nos estudos de caso poderia haver conteúdos em outros idiomas, principalmente no site de informações utilizado no ambiente corporativo, foi necessário separar as informações por idioma, para que termos de idiomas diferentes não fossem extraídos e utilizados em uma mesma expansão de consulta.

[8] http://pdfbox.apache.org/

Foi utilizada a biblioteca de reconhecimento de idiomas da *suite* MorphAdorner, que embora não reconheça português, apresenta bons resultados para os outros idiomas, principalmente inglês e espanhol. Como nos estudos de caso havia a garantia de que todas as informações disponíveis estariam nestes três idiomas (português, inglês e espanhol), os textos em português eram determinados por exclusão.

As atividades executadas no módulo de configuração da base de conhecimento estão representadas no diagrama de atividades da figura 9. As fontes de dados configuradas são acessadas, o conteúdo a ser capturado é filtrado em duas atividades: de acordo com as regras definidas pelo especialista e pelo idioma português. O conteúdo textual da informação é extraído, e armazenado em arquivo texto, um para cada conteúdo.

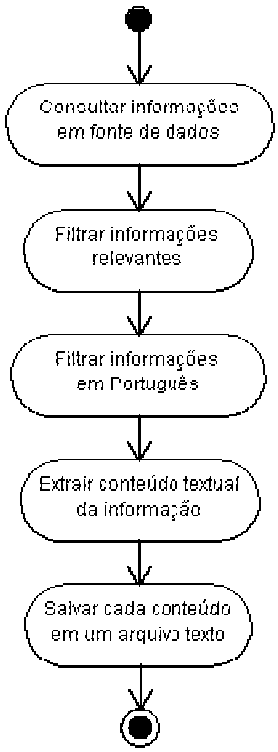


Figura 9 – Diagrama de atividades – Configuração da base de conhecimento.

3.3.2. Módulo de extração de informações

No módulo de extração de informações, os conteúdos identificados no módulo anterior foram segmentados e posteriormente agrupados em *clusters*. Após isso, o peso de cada termo foi determinado para que os termos com maior peso pudessem ser selecionados.

3.3.2.1. Segmentação de tópicos - Texttiling

A segmentação de tópicos foi implementada com o MorphAdorner[9], uma *suite* de bibliotecas de processamento de linguagem natural. O segmentador de tópicos do MorphAdorner implementa o método de segmentação linear Texttiling.

Texttiling é um algoritmo de segmentação de tópicos desenvolvido por Marti Hearst em 1994, baseado na delimitação do texto de acordo com padrões de co-ocorrência lexical e sua distribuição no texto, ou seja, baseado na hipótese de que sempre que há uma alteração no tópico abordado no texto, também há uma alteração significativa no vocabulário.

Baseando-se nisso, o sistema TextTiling pode utilizar três diferentes técnicas de verificação de co-ocorrência lexical para determinar as delimitações topicais de um texto: comparação de blocos adjacentes de textos, mudanças de vocabulário e cadeias lexicais.

A comparação de blocos adjacentes de texto explora as mudanças nas frequências das palavras obtidas na comparação entre cada par de blocos adjacentes. A mudança de vocabulário explora as novas palavras (sem ocorrência no bloco anterior) introduzidas em cada bloco. Para as cadeias lexicais, procuram-se seqüências de palavras similares distribuídas entre as pseudo-sentenças e assume-se que a extensão das cadeias de

[9] http://morphadorner.northwestern.edu/

palavras corresponde à extensão do segmento, caso as palavras similares estejam a uma distância máxima limite considerada no TextTiling.

O Texttiling é composto por três principais etapas: Delimitação das palavras no texto (*Tokenization)*, determinação do score léxico e identificação dos limites. Na etapa de delimitação das palavras, o texto é separado em termos, são removidas as palavras comumente encontradas em textos (conhecidas como *stop words*) e cada termo é reduzido a sua forma raiz com a aplicação de uma função de análise morfológica (*stemming*).

Na etapa de determinação do score léxico, divide-se o texto em pseudo-sentenças de tamanho fixo (blocos) e calcula-se uma nota para cada bloco de acordo com o método de verificação de co-ocorrência lexical escolhido. Com base nestas notas, a última etapa identifica as mudanças de tópicos e com isso a segmentação é finalizada (HEARST, 1997).

Embora o algoritmo Texttiling seja baseado em co-ocorrência e distribuição, a etapa de delimitação das palavras no texto (*Tokenization*) possui atividades sensíveis ao idioma, como a remoção de *stop words*. E como o segmentador de tópicos implementado no MorphAdorner não era preparado para português, foi necessária a sua customização, incluindo a remoção de *stop words* em português.

No diagrama de atividades apresentado na figura 10, são mostradas as diferentes etapas executadas durante a segmentação de tópicos: inicialmente os conteúdos a serem segmentados são extraídos dos arquivos textos gerados pelo módulo de configuração da base de conhecimento. Cada conteúdo é segmentado em tópicos, e cada tópico é armazenado em um arquivo texto.

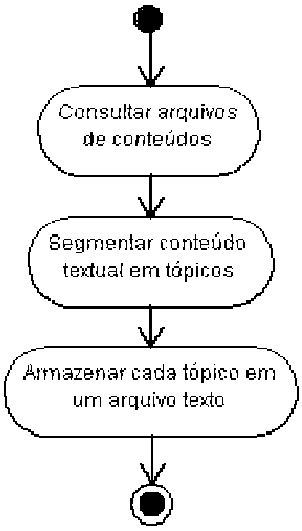


Figura 10 – Diagrama de atividades – Segmentação de tópicos.

3.3.2.2. Agrupamento - K-Means

Para o agrupamento, foi utilizada no protótipo uma implementação do algoritmo k-Means, disponível na *suite* de ferramentas de mineração de textos Textgarden[10].

K-Means é o mais importante algoritmo plano de agrupamento. Seu objetivo é minimizar a média da distância euclidiana ao quadrado dos vetores de documentos para o vetor que representa a média de seus grupos, também conhecido como *centroid* (MANNING, RAGHAVAN e SCHÜTZE, 2008).

A medida de quão bem os *centroids* representam os membros dos seus grupos é a soma residual dos quadrados, ou RSS, o quadrado da distância de cada vetor para seu *centroid*, somados para todos os vetores de documentos.

O primeiro passo do k-Means é selecionar um conjunto inicial de *K centroids* de grupos, escolhidos de forma aleatória. Estes centros iniciais selecionados são conhecidos como sementes. O algoritmo então repete 2 passos até que seja atingido o

[10] http://kt.ijs.si/dunja/textgarden/

critério de parada: cada documento é movido para um *cluster* cuja função de distância para seu centro seja a menor; e o centro de cada *cluster* é recalculado baseado nos seus membros atuais.

A condição de parada pode ser qualquer uma das seguintes: (i) um número fixo I de iterações foi completado; (ii) a atribuição de documentos aos *clusters* não sofreu alteração de uma iteração para outra; (iii) os centros de *clusters* não sofreram alterações de uma iteração para outra, ou (iv) quando o valor de RSS atinge um limite predefinido, que garante a qualidade dos grupos gerados.

Como o algoritmo k-Means exige como parâmetro de entrada o número K de grupos a serem gerados, este é um dos problemas em sua utilização: como garantir um valor K ideal para cada problema distinto.

No diagrama de atividades apresentado na figura 11, são mostradas as diferentes etapas executadas durante o agrupamento: o conteúdo de cada segmento gerado na atividade de segmentação de tópicos é consultado, e utilizados como entrada para o algoritmo de agrupamento, que reune os segmentos relacionados em grupos, e armazena cada conjunto de arquivos textos de segmentos em diretórios representando os diferentes grupos gerados.

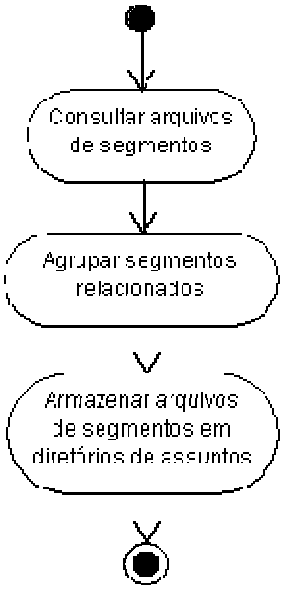


Figura 11 – Diagrama de atividades – Agrupamento.

3.3.2.3. Extração de termos

Após o agrupamento, a frequência dos termos deve ser determinada para cada grupo (representando os assuntos) de modo que os principais termos de cada grupo e de todo o contexto fossem extraídos. Para facilitar esta atividade, os conteúdos segmentados e agrupados foram indexados com a utilização da biblioteca *open-source* Apache Lucene[11], uma API de indexação e busca. Além de oferecer métodos simples para a determinação da frequência dos termos, a indexação do Lucene também já aplica ao conteúdo uma série de atividades necessárias para a extração dos termos, como por exemplo a decomposição do conteúdo textual em termos (*Tokenization*).

O cálculo do peso de cada termo em um documento (em nosso caso estamos consideramos cada segmento como um documento) foi realizado com a aplicação do logaritmo na fórmula de frequência de termos, com o objetivo de suavizar a influência desta frequência (MANNING, RAGHAVAN e SCHÜTZE, 2008):

[11] http://lucene.apache.org

$$Wft,d = \begin{cases} 1 + \log tft,d \text{ , se } \text{tft,d} > 0 \\ 0 \text{ , caso contrário} \end{cases}$$

Sendo considerados $wf_{t,d}$ o peso do termo t no documento d e $tf_{t,d}$ o número de ocorrências do termo no documento. Para calcular o peso do termo em toda a coleção é feito um somatório do peso deste termo em cada documento da coleção. No caso da extração de termos por assuntos, cada grupo de segmentos é considerado uma coleção independente.

As atividades executadas durante a extração de termos são apresentadas no diagrama de atividades abaixo (Figura 12). Para cada diretório de assunto gerado na atividade de agrupamento, são calculadas as frequências dos termos em cada arquivo de segmento. Após isso, são realizadas duas atividades de extração: uma em que são selecionados os termos mais frequentes de toda a coleção, e outra em que são selecionados os termos mais frequentes em cada diretório de assunto.

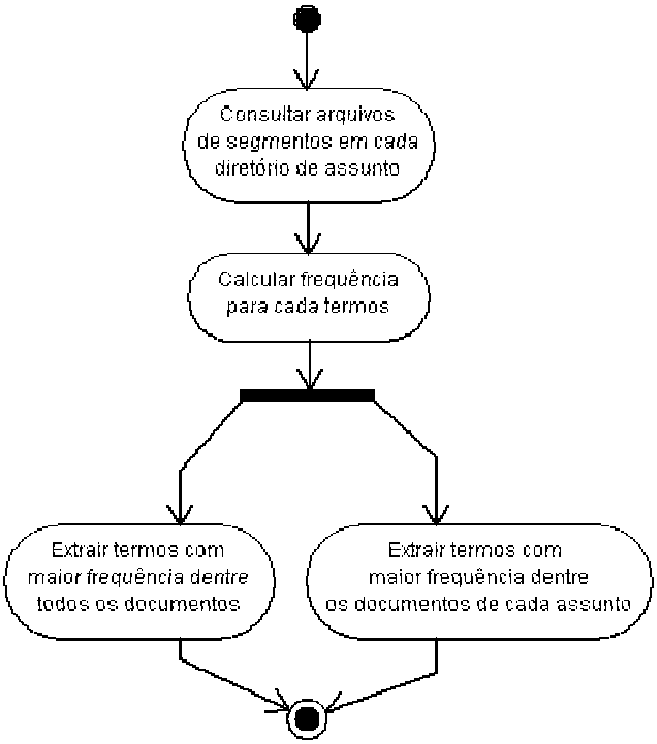


Figura 12 – Diagrama de atividades – Extração de termos.

3.3.2.4. Administração de contextos

Todos os algoritmos utilizados nas atividades do módulo de extração de informação permitem a configuração de parâmetros, além de ser necessário um sequenciamento entre estas atividades. Para esta finalidade foi desenvolvida a administração de contexto, uma interface que possibilita ao especialista controlar a execução das atividades e inserir parametrizações (que serão melhor explicados no Capítulo 4).

A aplicação permite a configuração e uso de múltiplos contextos, e armazena as configurações de parâmetros de cada contexto em um arquivo XML, conforme exemplo mostrado na figura 13. Este arquivo é atualizado pela interface de administração de contextos.

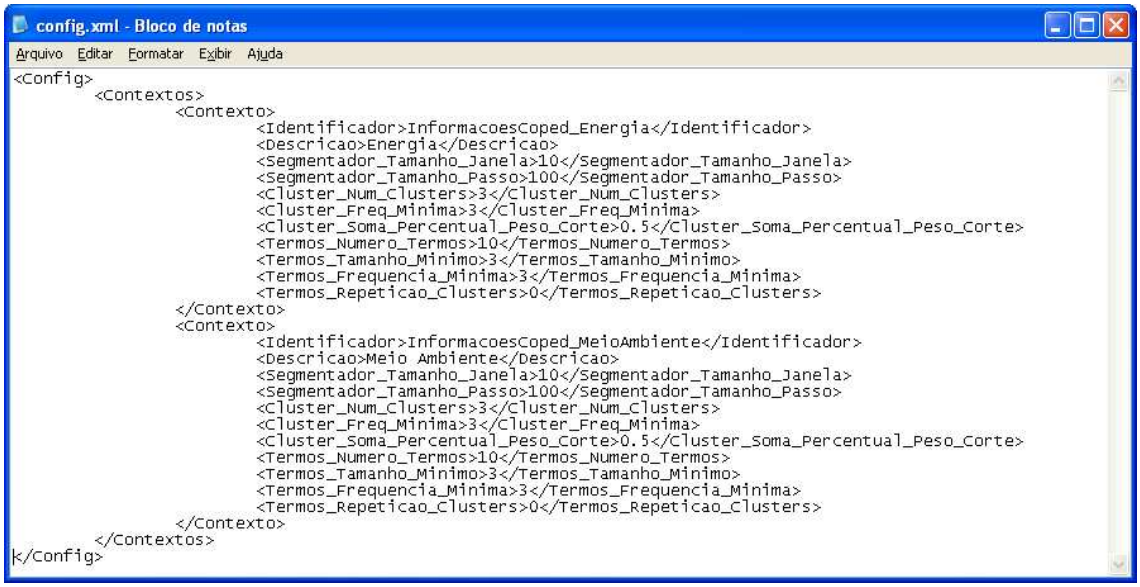


Figura 13 - Configuração dos contextos em XML.

Para a manipulação deste arquivo foi utilizada a biblioteca *open-source* Apache Commons Configuration[12], que provê uma interface de configuração genérica que permite a uma aplicação java acessar dados de configurações de uma variedade de tipos

[12] http://commons.apache.org/configuration/

(por exemplo, arquivos de propriedade, arquivos de inicialização do windows, XML, etc.).

3.3.3. Módulo de busca

Na arquitetura física proposta, a consulta original pode ser expandida de duas formas distintas. A primeira delas é a expansão automática, em que a consulta original é expandida pela ferramenta sem a interferência direta pelo usuário. São realizadas *n* expansões, sendo o valor de *n* igual a um mais o número de assuntos identificados no contexto selecionado. Uma expansão é feita com os termos gerais do contexto e as demais são expansões feitas com os termos específicos de cada assunto. Cada consulta expandida é executada no buscador web selecionado e os resultados são apresentados para o usuário em abas distintas.

A segunda modalidade de expansão considerada é a sugestão de termos, em que todos os termos extraídos (termos gerais do contexto e termos específicos dos assuntos) são apresentados para o usuário como uma sugestão em listas distintas. O usuário manualmente seleciona os termos de seu interesse para que sejam incorporados à consulta original, realizando uma única expansão. Para selecionar os termos, é possível combinar termos das diferentes listas exibidas, misturando termos gerais do contexto e termos dos diferentes assuntos identificados.

Duas funcionalidades foram implementadas no módulo de busca com a utilização de ferramentas *open-source*: a stemização e a expansão de termos com o uso do dicionário Wordnet [13].

A Stemização foi necessária para a expansão dos termos da consulta. Os termos utilizados na consulta original feita pelo usuário eram comparados com os termos

[13] http://wordnet.princeton.edu/

sugeridos para expansão, para que não fossem utilizados termos iguais. A stemização foi utilizada nesta comparação, evitando o uso de palavras derivadas, que por serem similares acabariam pouco influenciando o resultado. Para implementar a stemização foi utilizada a biblioteca Java SnowBall[14].

Além da expansão com termos do contexto, no estudo de caso piloto realizado com o protótipo (PRATES e SIQUEIRA, 2010) também foram realizadas expansões com sinônimos do termo, utilizando o dicionário Wordnet.

Por não haver uma implementação de um dicionário similar ao Wordnet em português e disponível na linguagem Java, nos outros estudos de caso a expansão com sinônimos não foi aplicada.

Outra biblioteca utilizada no módulo de busca é a Apache Log4j[15], utilizada para o registro das atividades de busca e avaliação realizadas na aplicação.

3.3.4. Interação entre usuário e especialista com a aplicação

Nas arquiteturas lógica e física são representadas, além das atividades já descritas, a interação da aplicação com dois atores: o Usuário e o Especialista. Nos diagramas de caso de uso apresentados a seguir essas interações serão detalhadas.

O especialista participa do caso de uso Configurar Contexto, responsável por fornecer todas as informações necessárias para que as atividades do módulo de Extração de Informações possam determinar todos os termos utilizados nas expansões de consulta (Figura 14).

[14] http://snowball.tartarus.org/
[15] http://logging.apache.org/log4j/

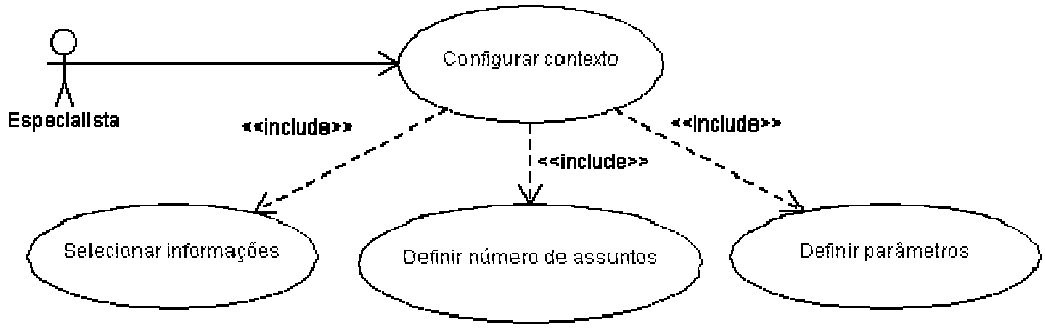


Figura 14 – Caso de uso Configurar Contexto.

A configuração do contexto inclui mais três casos de uso: "Selecionar informações", "Definir número de assuntos" e "Definir parâmetros". No "Selecionar informações", o especialista no domínio deve, a partir de todas as fontes de informação existentes e acessíveis pelo sistema, selecionar aquelas que sejam boas representantes dos assuntos que compõem o contexto.

No caso de uso "Definir número de assuntos", o especialista deve identificar a quantidade de diferentes assuntos abordados em todas as informações selecionadas para o contexto. Este número determinará, na etapa de extração de informações, quantos grupos de segmentos serão criados, e consequentemente a quantidade de expansões de consultas que serão geradas no módulo de busca.

Finalmente, no caso de uso "Definir parâmetros", o especialista deve configurar os parâmetros utilizados pelos algoritmos de extração de informação utilizados (atividades de segmentação de tópicos, agrupamento e extração de termos).

A interação do Usuário com a aplicação para a execução das consultas varia de acordo com a modalidade da busca utilizada. Na figura 15 é apresentado o caso de uso "Executar consulta" na modalidade expansão automática. O usuário elabora a consulta original, contendo palavras-chave relacionadas à sua necessidade de informação, e define quantos termos definidos pela aplicação serão incluídos na consulta, representado pelo caso de uso "Definir número de termos para a expansão".

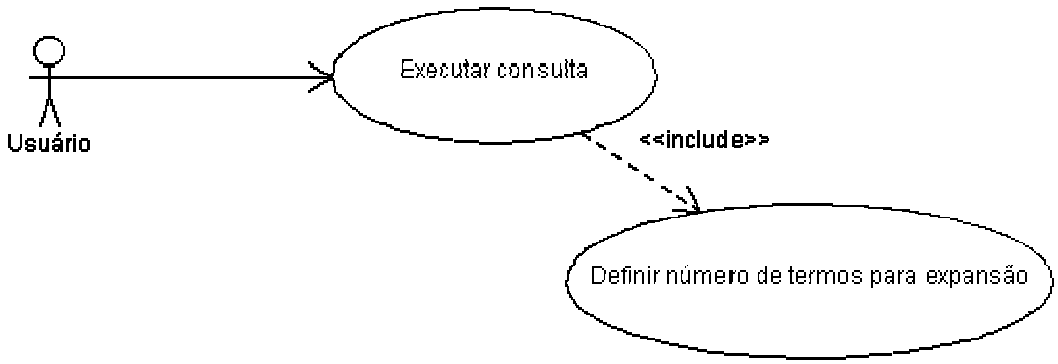


Figura 15 – Caso de uso Executar consulta – Expansão automática.

Na figura 16 é apresentado o caso de uso "Executar consulta" na modalidade sugestão de termos. Assim como na expansão automática, o usuário elabora a consulta original. Porém, no caso de uso "Selecionar termos para expansão", o usuário adiciona à sua consulta termos escolhidos a partir de listas de termos sugeridos pela aplicação.

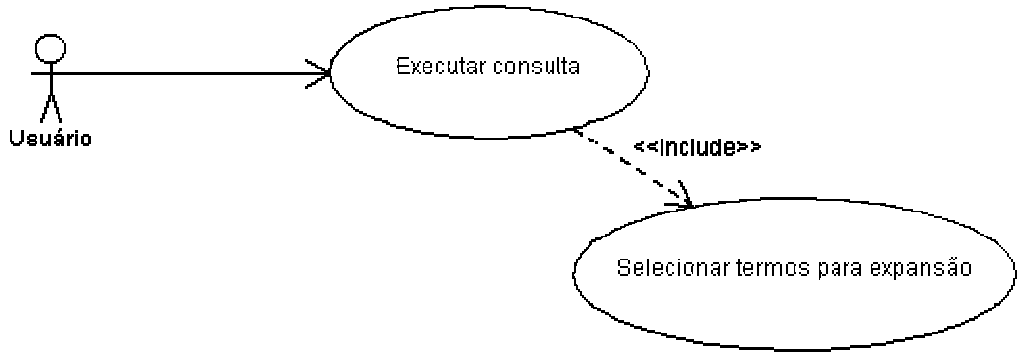


Figura 16 – Caso de uso Executar consulta – Sugestão de termos.

Capítulo 4 – Implementação

Para verificar a hipótese deste trabalho, foi desenvolvido em Java um protótipo funcional baseado na arquitetura física definida no Capítulo 3. Neste capítulo este protótipo será apresentado e detalhadas as funcionalidades de busca, avaliação e gerenciamento dos contextos.

4.1. O Protótipo

A aplicação desenvolvida, chamada *ContextSearch*, foi utilizada por usuários em dois estudos de caso, que serão detalhados no Capítulo 5. Para que as atividades realizadas pelos usuários pudessem ser registradas, na primeira tela do protótipo era necessário que o usuário se registrasse, conforme mostrado na Figura 17.

Figura 17 - Tela de identificação dos usuários.

Não foi criado na aplicação um mecanismo de autenticação, com o fornecimento de nome de usuário e senha. Entretanto era importante utilizar uma forma de identificação única para que as atividades dos diversos usuários pudessem ser diferenciadas. No caso do estudo de caso em ambiente corporativo, foi solicitado o

login (*userid*) utilizado na empresa. Já no ambiente educacional, foi solicitado o endereço de e-mail utilizado na universidade.

Após a identificação, o protótipo exibe a lista de contextos configurados, um campo texto para que o usuário informe a sua necessidade de informação atual e botões para iniciar uma das duas modalidades de busca disponíveis (Figura 18). A proposta do estudo de caso era que cada usuário utilizasse a ferramenta nas duas modalidades (expansão automática e sugestão de termos), para ajudar a encontrar resultados na Internet que pudessem satisfazer a uma necessidade de informação hipotética.

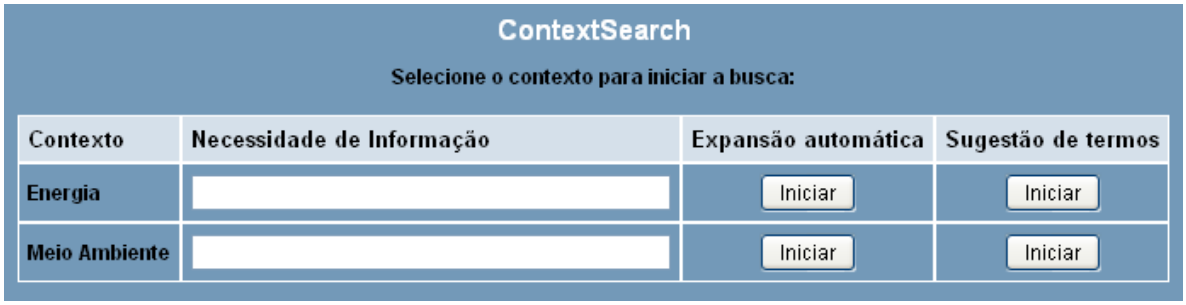


Figura 18 - Tela de listagem de contextos configurados.

4.1.1. Expansão automática

Na expansão automática, a pesquisa é muito parecida com as disponíveis nas ferramentas de busca on-line. O usuário digita a expressão de busca e seleciona em uma caixa de combinação quantos termos ele gostaria de obter automaticamente da ferramenta (Figura 19).

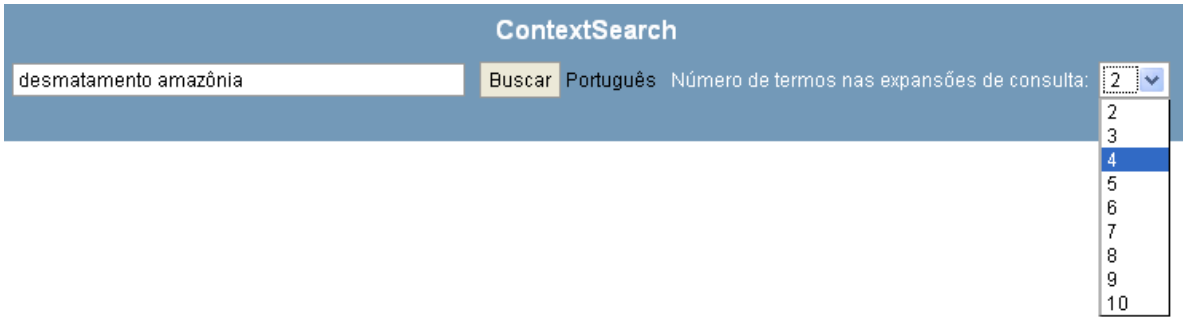


Figura 19 - Tela de busca, na modalidade expansão automática.

Após clicar em Buscar, a ferramenta realiza as expansões de consulta e as executa em um buscador web. No protótipo, as consultas eram redirecionadas para o Google, mas qualquer outro buscador que aceite a execução de buscas com a consulta sendo passada por parâmetro (seja pelo método *GET* ou *POST*) poderia ter sido utilizado.

Conforme ilustrado na Figura 20, a aplicação exibe os resultados de pesquisa em diferentes abas, bastando clicar nos botões exibidos logo abaixo da caixa de consulta para mudar os resultados exibidos. As abas exibidas são as seguintes:

- Consulta original: resultados da consulta original, sem expansão. Estes resultados são importantes para a comparação entre os resultados com e sem o uso das expansões de consulta;
- Consulta do contexto: resultados obtidos com a expansão da consulta original com termos gerais extraídos de todos os recursos que representam o contexto.
- Abas de assuntos: resultados obtidos com a expansão da consulta original com termos extraídos dos assuntos identificados dentro do contexto. Para cada assunto identificado, é feita uma expansão de consulta com os termos extraídos dos recursos que foram agrupados neste assunto. Nos dois estudos de caso realizados foram identificados três assuntos, sendo exibidas portanto três abas de assuntos.

Figura 20 - Exibição de resultados em diferentes abas.

4.1.2. Sugestão de termos

Na modalidade sugestão de termos, é necessário um esforço adicional do usuário para executar as buscas. A aplicação exibe uma lista contendo os termos gerais extraídos do contexto e os termos específicos de cada assunto, e cabe ao usuário decidir quais e quantos termos serão utilizados na expansão de consulta, bastando para isso marcar a caixa de seleção dos termos desejados. A ferramenta não impede que sejam marcados termos de assuntos diferentes (Figura 21).

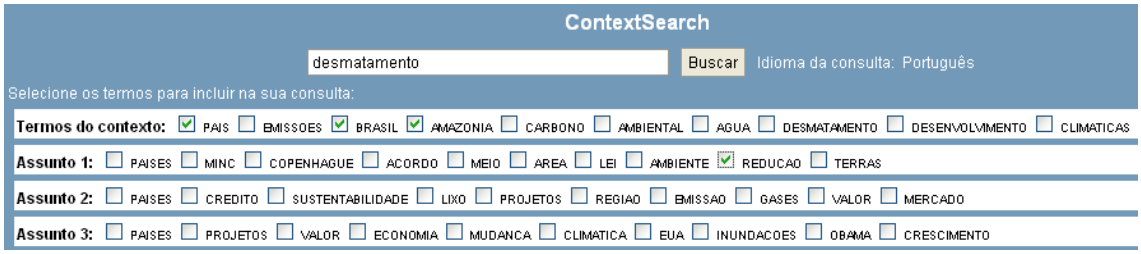


Figura 21 - Tela de busca, na modalidade sugestão de termos.

Como nesta modalidade só é realizada uma única expansão, a tela de resultados de busca é mais simples do que na modalidade expansão automática, pois somente são exibidas duas abas de resultado: uma com a consulta original e outra com a consulta expandida, composta dos termos da consulta original e os termos selecionados (Figura 22).

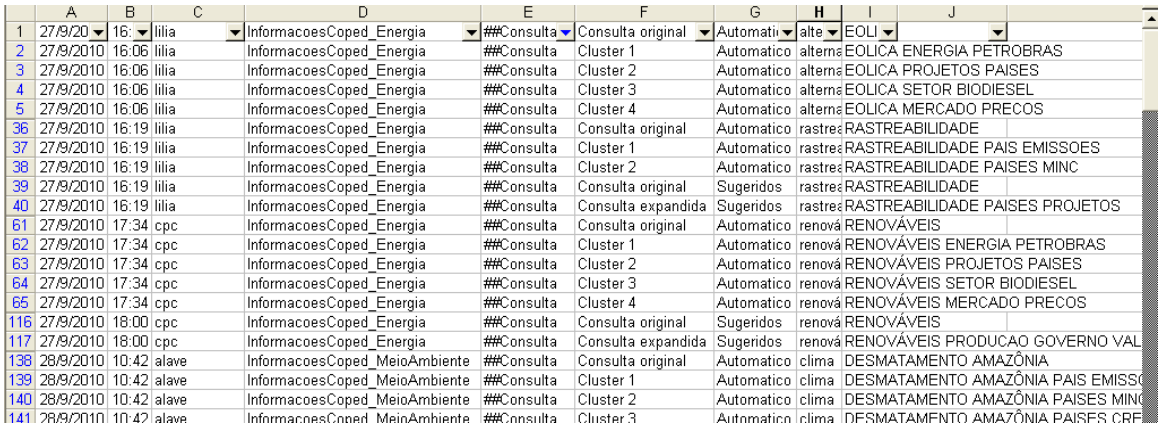
Figura 22 - Resultados de busca na modalidade sugestão de termos.

4.1.3. Avaliações

Uma etapa essencial na validação da abordagem é a avaliação de relevância nos resultados obtidos em cada aba, nas duas modalidades de pesquisa. Esta avaliação foi realizada na própria ferramenta, através de caixas de combinação exibidas ao lado de cada aba de resultado de busca (Figura 23). Os usuários deveriam avaliar a relevância dos 10 primeiros resultados de busca obtidos em cada aba, em uma escala de pontos que variava de 0 a 4. Mais explicações sobre os critérios de avaliação são apresentados no Capítulo 5, que fala sobre a metodologia da pesquisa.

Figura 23 - Avaliação de relevância dos resultados de busca.

Os resultados obtidos com a avaliação de relevância, juntamente com outras informações relevantes, tais como consultas realizadas e consultas expandidas, foram registrados e armazenados em arquivo texto no formato CSV (*comma-separated values*), permitindo a tabulação e análise posterior dos dados em planilha (Figura 24).



	A	B	C	D	E	F	G	H	I	J
1	27/9/20	16:	lilia	InformacoesCoped_Energia	##Consulta	Consulta original	Automati	alte	EOLI	
2	27/9/2010	16:06	lilia	InformacoesCoped_Energia	##Consulta	Cluster 1	Automatico	alterna	EOLICA ENERGIA PETROBRAS	
3	27/9/2010	16:06	lilia	InformacoesCoped_Energia	##Consulta	Cluster 2	Automatico	alterna	EOLICA PROJETOS PAISES	
4	27/9/2010	16:06	lilia	InformacoesCoped_Energia	##Consulta	Cluster 3	Automatico	alterna	EOLICA SETOR BIODIESEL	
5	27/9/2010	16:06	lilia	InformacoesCoped_Energia	##Consulta	Cluster 4	Automatico	alterna	EOLICA MERCADO PRECOS	
36	27/9/2010	16:19	lilia	InformacoesCoped_Energia	##Consulta	Consulta original	Automatico	rastrea	RASTREABILIDADE	
37	27/9/2010	16:19	lilia	InformacoesCoped_Energia	##Consulta	Cluster 1	Automatico	rastrea	RASTREABILIDADE PAIS EMISSOES	
38	27/9/2010	16:19	lilia	InformacoesCoped_Energia	##Consulta	Cluster 2	Automatico	rastrea	RASTREABILIDADE PAISES MINC	
39	27/9/2010	16:19	lilia	InformacoesCoped_Energia	##Consulta	Consulta original	Sugeridos	rastrea	RASTREABILIDADE	
40	27/9/2010	16:19	lilia	InformacoesCoped_Energia	##Consulta	Consulta expandida	Sugeridos	rastrea	RASTREABILIDADE PAISES PROJETOS	
61	27/9/2010	17:34	cpc	InformacoesCoped_Energia	##Consulta	Consulta original	Automatico	renová	RENOVÁVEIS	
62	27/9/2010	17:34	cpc	InformacoesCoped_Energia	##Consulta	Cluster 1	Automatico	renová	RENOVÁVEIS ENERGIA PETROBRAS	
63	27/9/2010	17:34	cpc	InformacoesCoped_Energia	##Consulta	Cluster 2	Automatico	renová	RENOVÁVEIS PROJETOS PAISES	
64	27/9/2010	17:34	cpc	InformacoesCoped_Energia	##Consulta	Cluster 3	Automatico	renová	RENOVÁVEIS SETOR BIODIESEL	
65	27/9/2010	17:34	cpc	InformacoesCoped_Energia	##Consulta	Cluster 4	Automatico	renová	RENOVÁVEIS MERCADO PRECOS	
116	27/9/2010	18:00	cpc	InformacoesCoped_Energia	##Consulta	Consulta original	Sugeridos	renová	RENOVÁVEIS	
117	27/9/2010	18:00	cpc	InformacoesCoped_Energia	##Consulta	Consulta expandida	Sugeridos	renová	RENOVÁVEIS PRODUCAO GOVERNO VAL	
138	28/9/2010	10:42	alave	InformacoesCoped_MeioAmbiente	##Consulta	Consulta original	Automatico	clima	DESMATAMENTO AMAZÔNIA	
139	28/9/2010	10:42	alave	InformacoesCoped_MeioAmbiente	##Consulta	Cluster 1	Automatico	clima	DESMATAMENTO AMAZÔNIA PAIS EMISS	
140	28/9/2010	10:42	alave	InformacoesCoped_MeioAmbiente	##Consulta	Cluster 2	Automatico	clima	DESMATAMENTO AMAZÔNIA PAISES MIN	
141	28/9/2010	10:42	alave	InformacoesCoped_MeioAmbiente	##Consulta	Cluster 3	Automatico	clima	DESMATAMENTO AMAZÔNIA PAISES CRE	

Figura 24 - Registros de consultas e avaliações feitas pelos usuários.

4.1.4. Administração dos contextos

O processo de extração dos termos a partir de recursos selecionados para representar o contexto envolve a execução de diversas técnicas de extração de informações, como segmentação de tópicos e agrupamento. Como estas técnicas exigem diferentes parametrizações em seus algoritmos, foi necessário o desenvolvimento de uma interface em que a administração simples destes parâmetros fosse permitida. Além disso, durante os experimentos que buscavam definir a influência destes parâmetros nas expansões de consultas realizadas, as configurações de contexto eram constantemente alteradas, sendo necessária também uma interface em que fosse possível disparar a execução de cada atividade necessária para a realização das extrações de termos.

Na tela inicial da interface administrativa, os contextos configurados na ferramenta são exibidos, sendo disponibilizada também uma opção rápida para a reconstrução de todos os contextos (Figura 25).

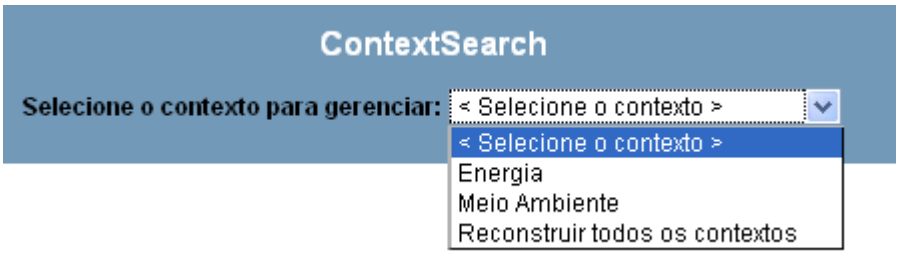


Figura 25 - Menu inicial da administração dos contextos.

Após selecionar o contexto desejado, dois formulários são exibidos, um para a administração dos principais parâmetros do contexto (Figura 26) e outro para a execução das atividades de extração de termos.

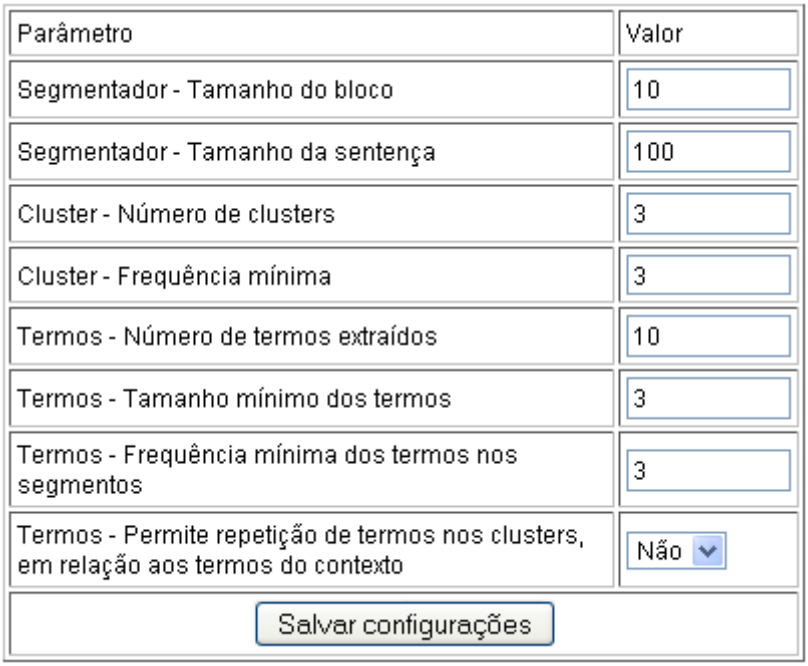


Figura 26 - Formulário de alteração de parâmetros.

Segue uma breve explicação de cada parâmetro utilizado e o seu valor configurado:

- **Segmentador – Tamanho do bloco:** Número de pseudo-sentenças contidas em cada bloco do algoritmo Texttiling (HEARST, 1997). Cada bloco será comparado aos seus blocos adjacentes para determinação das mudanças de tópicos.
- **Segmentador- Tamanho da sentença:** número de termos contidos em cada pseudo-sentença definida no algoritmo Texttiling (HEARST, 1997);
- **Cluster – Número de *clusters*:** define o número de *clusters* configurados para o contexto. Este parâmetro, definido pelo especialista no domínio do

contexto, representa o número de assuntos encontrados dentro dos recursos selecionados como representativos do contexto;

- **Cluster – Frequência mínima:** define a frequência de documento mínima dos termos que serão utilizados na representação do documento (na abordagem proposta, segmento) no algoritmo k-Means (MANNING, RAGHAVAN e SCHÜTZE, 2008), ou seja, só considera para a representação do documento os termos que ocorram em mais de *n* documentos da coleção, sendo *n* o valor do parâmetro;
- **Termos – Número de termos extraídos:** número máximo de termos extraídos do contexto e de cada assunto;
- **Termos – Tamanho mínimo dos termos:** tamanho mínimo em caracteres do termo a ser extraído;
- **Termos – Frequência mínima dos termos nos segmentos:** Frequência mínima do termo dentro do segmento para que possa ser extraído;
- **Termos – Permite repetição de termos nos *clusters*, em relação aos termos do contexto:** define se um termo já extraído como termo geral do contexto pode também ser extraído em algum assunto, podendo aparecer em mais de uma expansão de consulta simultaneamente, ou em mais de uma lista de sugestão de termos.

Os componentes que implementam os algoritmos de segmentação de tópicos e agrupamento permitem a definição de mais parâmetros dos que os especificados no formulário acima. Para estes parâmetros, foi utilizado sempre o valor padrão (valor adotado pelo componente em caso de não preenchimento).

No formulário de execução de atividades (Figura 27), é possível executar uma ou mais atividades dentro do processo de extração de termos. Por exemplo, caso haja uma

alteração somente no tamanho mínimo dos termos a serem extraídos, não é necessário reexecutar todas as atividades, mas somente a criação de expansões de consulta.

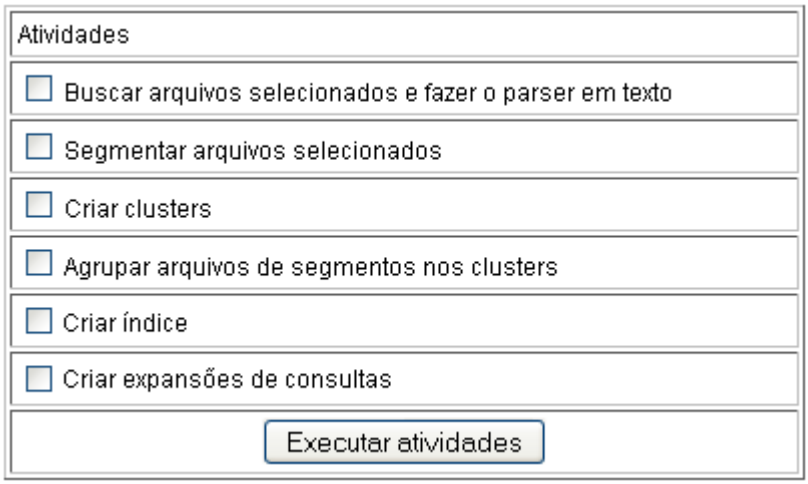


Figura 27 - Formulário de execução de atividades.

- **Buscar arquivos selecionados e fazer o *parser* em texto:** consulta a fonte da informação do contexto (banco de dados no caso do ambiente corporativo e arquivos PDF no caso do ambiente educacional) e extrai o conteúdo textual das informações, armazenando-o em arquivos de texto (formato txt);
- **Segmentar arquivos selecionados:** segmenta todos os arquivos texto do contexto, armazenando cada segmento em um novo arquivo;
- **Criar *clusters*:** percorre todos os segmentos do contexto, identificando qual segmento deve ser agrupado em cada assunto;
- **Agrupar arquivos de segmentos nos clusters:** copia os arquivos de segmento para o diretório de assunto correspondente;
- **Criar índice:** cria um índice para cada assunto identificado com o auxílio da biblioteca Apache Lucene[16];

[16] http://lucene.apache.org

- **Criar expansões de consulta:** utiliza cada índice criado para calcular os pesos de termos, ordenando-os de forma decrescente e extraindo os *n* termos com maior valor, sendo *n* o parâmetro "número de termos extraídos".

4.1.4.1. Definição dos parâmetros

Os parâmetros utilizados nos algoritmos de segmentação, agrupamento e extração de termos podem variar de acordo com a natureza das informações de entrada, principalmente o seu tamanho (quantidade de termos). Como exemplo, em documentos pequenos, contendo poucos termos, caso seja selecionado um valor alto para o parâmetro que define a frequência mínima de um termo para a extração, é grande a probabilidade de nenhum ou somente poucos termos serem extraídos.

Para a definição dos valores dos parâmetros, foram realizados diversos experimentos em laboratório, variando para cada atividade (segmentação, agrupamento e extração de termos) os seus respectivos parâmetros e verificando os resultados obtidos a partir de coleções de documentos de natureza e tamanhos diversos (artigos, notícias, capítulos de livro, etc.).

Os parâmetros utilizados nos três contextos dos dois estudos de caso (ambientes corporativo e educacional) foram configurados praticamente com os mesmos valores, com poucas variações. Os parâmetros do segmentador, após variações em testes que não ocasionaram melhoria nos resultados obtidos, foram mantidos com os valores iniciais utilizados na implementação do Texttiling, ou seja, tamanho do bloco igual a dez (10) e tamanho da sentença igual a cem (100).

Por coincidência, foi considerada a existência de três grandes assuntos nos três contextos trabalhados, fazendo com que o parâmetro número de *clusters* fosse configurado com o valor igual a três (3).

O valor da frequência mínima de documentos para os termos foi definido arbitrando um valor inicial alto e reduzindo este valor até que fosse retornada a quantidade de termos desejada. No contexto de Meio Ambiente este valor foi igual a cinco (5), enquanto no contexto Energia, que possuía menos informações de contexto, este valor foi três (3) e no contexto de gestão de conhecimento, com menos informação ainda, foi dois (2).

O número de termos extraídos foi definido com valor dez (10) para todos os contextos.

O tamanho mínimo dos termos foi definido como sendo de três caracteres para todos os contextos, por não existir nenhum termo relevante e frequente nestes três contextos com este tamanho. Já com quatro caracteres foram encontradas palavras relevantes que não poderiam ser desprezadas.

A frequência mínima dos termos em cada segmento foi definida de forma similar à frequência de documentos, sendo repetidos os valores deste parâmetro. A única exceção foi o contexto Meio ambiente, cujo valor foi definido em três (3).

Não foi permitida a extração nos assuntos de termos previamente extraídos de todo o contexto.

Capítulo 5 – Estudos de Caso

Neste capítulo serão descritos os métodos de pesquisa utilizados, os objetivos esperados com cada método e as justificativas para sua adoção. Serão apresentados os estudos de caso, realizados em diferentes contextos, com usuários de diferentes perfis e nas duas modalidades disponíveis de expansão de consulta.

5.1. Metodologia

Após o término da construção do protótipo *ContextSearch* e da definição dos parâmetros a serem utilizados, um estudo de caso piloto foi realizado em uma disciplina de mestrado da UNIRIO (PRATES e SIQUEIRA, 2010) contando com a participação de sete (7) estudantes. Segundo YIN (2005), a execução de um estudo de caso piloto auxilia o pesquisador a refinar os seus procedimentos de coleta e registro de dados, lhe conferindo a oportunidade de testar os procedimentos estabelecidos para esta finalidade.

Ainda segundo YIN (2005), em geral a conveniência, os acessos aos dados e a proximidade geográfica podem ser os principais critérios na hora de selecionar o caso piloto.

Após alterações na ferramenta e no processo de avaliação, decorrente das observações realizadas no estudo de caso piloto, foram realizados dois estudos de caso em ambientes diferentes, um educacional (PRATES e SIQUEIRA, 2011) e outro corporativo. Este método de pesquisa foi escolhido por possibilitar a observação da aplicação da ferramenta *ContextSearch* em atividades de busca reais, por usuários com

diferentes formações e necessidades de informação, permitindo desta forma comparar, sob diversos aspectos, as pesquisas realizadas em um buscador web com a consulta originalmente informada e as consultas expandidas através da ferramenta.

O método estudo de caso é utilizado para investigar um fenômeno contemporâneo dentro do seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não são claramente definidos (YIN, 2005). Segundo (STAKE, 1995), um estudo de caso é utilizado quando se busca os detalhes das interações com o seu contexto.

Outra justificativa para a escolha do estudo de caso como método de pesquisa foi a dificuldade de se encontrar um paradigma de avaliação automática para ferramentas de busca sensíveis ao contexto. A área de recuperação de informação possui mecanismos para a avaliação automática de ferramentas de busca, através da aplicação destas ferramentas em coleções de testes, ou *corpus* anotados.

Contudo, estas coleções de testes não podem ser aplicadas em ferramentas de busca adaptativas, que utilizam outras informações além da necessidade de informação para definir quais documentos devem ser retornados em uma pesquisa. Este paradigma de avaliação não leva em consideração os diferentes contextos existentes no julgamento de relevância dos documentos da coleção, não sendo aplicado desta forma em ferramentas de busca sensíveis ao contexto. Segundo VOORHEES (2008), a avaliação de ferramentas de busca adaptativas é um problema de pesquisa em aberto.

Em cada estudo de caso, foi proposto aos participantes que elaborassem necessidades de informações, expressões de busca que traduzissem estas necessidades, e utilizassem o ContextSearch para fazer pesquisas nas suas duas modalidades disponíveis: expansão automática e sugestão de termos. O número de pesquisas a ser realizado foi deixado em aberto, sendo definida a obrigação de avaliar a relevância dos

resultados de pelo menos uma pesquisa. Este julgamento de relevância foi realizado nos 10 primeiros resultados de cada consulta executada no buscador web (consulta original e consultas expandidas). Além destas avaliações de caráter quantitativo, os participantes também responderam a um questionário (Anexo I) cujo objetivo foi identificar os critérios qualitativos da avaliação da ferramenta.

A aplicação dos estudos de caso em ambientes diferentes teve como objetivo verificar se a motivação das pesquisas e o perfil dos participantes envolvidos trariam diferenças significativas nas avaliações da ferramenta. Embora resultados positivos em ambos os ambientes não possam levar a uma conclusão de generalização, um resultado positivo somente em um contexto poderia indicar a utilização em um ambiente específico, com condições específicas. A generalização esperada com os estudos de caso em diferentes ambientes é o que YIN (2005) considera como "generalização analítica", em que o pesquisador não procura casos representativos de uma população para a qual pretende generalizar os resultados, mas a partir de um conjunto particular de resultados pode gerar proposições teóricas que sejam aplicadas a outros contextos.

5.2. Estudo de caso 1: Ambiente educacional

O primeiro estudo de caso foi realizado na UNIRIO (PRATES e SIQUEIRA, 2011), com alunos do primeiro período do curso de Bacharelado em Sistemas de Informação, matriculados na disciplina "Fundamentos de Sistemas de Informação". Nesta disciplina, o professor indicou alguns livros como referência e antes de cada aula os alunos deveriam ler os capítulos dos livros indicados que fossem relacionados com os assuntos da aula.

Para a realização do estudo de caso, foi definido como contexto a aula sobre Gestão do Conhecimento e Sistemas Especializados. Nesta aula, os alunos debateram

sobre os diversos assuntos envolvidos com a mediação do professor e, posteriormente, utilizaram a ferramenta ContextSearch para pesquisar sobre algum dos temas discutidos durante a aula.

A fonte de informação utilizada para capturar o contexto foi o capítulo entitulado Gestão do conhecimento e sistemas especializados de informação, extraído do livro (STAIR e REYNOLDS, 2005) indicado como livro-texto pelo professor da disciplina.

Os alunos contaram com um tempo de 1 hora para a realização de atividades definidas em um roteiro (Anexo II). Os alunos utilizaram a ferramenta nas suas duas modalidades disponíveis (expansão automática e sugestão de termos) em pelo menos uma necessidade de informação, que fosse referente a um dos três assuntos de pesquisa discutidos na sala de aula: Sistemas de gestão de conhecimento, Visão geral da inteligência artificial e Visão geral dos sistemas especialistas. A ferramenta foi configurada para trabalhar com 3 assuntos.

Na modalidade expansão automática, os alunos deveriam formular uma expressão de busca que atendesse a necessidade de informação e selecionar o número de termos que a ferramenta iria incluir na consulta original. Foi permitido que os alunos reexecutassem a consulta quantas vezes fosse necessário, testando diferentes quantidades de termos, até que fosse desejado iniciar a avaliação de relevância dos resultados.

Ao iniciar a avaliação, os alunos deveriam julgar a relevância dos 10 primeiros resultados exibidos em todos os painéis de resultado (consulta original, consulta do contexto e três painéis de assuntos), utilizando uma escala de cinco (5) pontos. O valor "0" indicando que o resultado não possui relevância e o valor "4" indicando alta relevância. O critério utilizado para julgar a relevância foi se o resultado trazia informação considerada útil para o participante no atendimento da necessidade de

informação definida. Também deveria ser levada em consideração no julgamento se a informação desejada pode ser encontrada diretamente no resultado obtido ou se foram necessárias navegações no resultado para obter a informação desejada.

Na modalidade sugestão de termos, as atividades foram muito semelhantes, com a diferença de que não deveria ser selecionado o número de termos, mas sim quais termos deveriam ser acrescentados à consulta original, a partir de uma lista composta por termos gerais do contexto e termos específicos de cada assunto configurado na ferramenta.

Após o término das atividades, os alunos preencheram um questionário com questões qualitativas sobre a utilização da ferramenta ContextSearch.

5.3. Estudo de caso 2: Ambiente corporativo

O Banco Nacional de Desenvolvimento Econômico e Social (BNDES)[17], empresa pública federal, atua no financiamento de longo prazo para a realização de investimentos em todos os segmentos da economia. Para analisar os diversos projetos de financiamento, é importante que o corpo técnico do banco tenha acesso a informações sobre diferentes setores que possam auxiliá-lo na tomada de decisões.

Neste processo de gestão de conhecimento, dois atores exercem um papel muito importante e complementar. As gerências setoriais, existentes em todas as áreas operacionais do banco, especializadas em setores econômicos e que atuam como consultoras para as demais equipes que trabalham com projetos do setor; e a gerência Coped, da área de Planejamento, responsável por capturar e fornecer a estas gerências principalmente, e aos demais técnicos do banco, informações sobre os diversos setores econômicos em que o banco atua.

[17] http://www.bndes.gov.br

Um dos importantes trabalhos realizados pelo Coped é a gestão de um site de informações (Figura 28), disponibilizado na intranet, em que são publicadas informações diversas sobre os setores econômicos apoiados pelo banco. A maior parte dessas informações são notícias veiculadas na Internet, que são pesquisadas e cadastradas diariamente no site pelas bibliotecárias da gerência. Além de consulta no próprio site, os técnicos do banco tem acesso a este conteúdo através de *newsletters* enviadas diariamente.

Figura 28 – Site de Informações do Coped

Neste estudo de caso, por se tratar de um cenário mais complexo que o anterior, as avaliações da ferramenta foram realizadas com diferentes perfis de usuários e dois contextos diferentes: Energia e Meio Ambiente. Em relação ao perfil, o estudo de caso pretendia avaliar como a ferramenta proposta poderia ou não apoiar a realização das atividades de busca dependendo do conhecimento sobre o assunto a ser pesquisado e a formação de cada participante. Foram selecionados dois grupos de usuários: de um lado as bibliotecárias do Coped, especialistas em informação, que realizam pesquisas na Internet diariamente e que portanto possuem maior afinidade com mecanismos de

busca; e de outro lado funcionários das gerências setoriais, especialistas no domínio da informação, principais consumidores das informações sobre setores econômicos.

A fonte de informação utilizada para capturar os contextos foi o site de informações mantido pelo Coped. Foram utilizadas 310 de notícias classificadas como Meio Ambiente e 1.788 classificadas como Energia. Todas estas notícias foram segmentadas e posteriormente agrupadas, seguindo a abordagem proposta nesta Dissertação, em três assuntos para cada contexto.

A duração total do estudo de caso foi de duas semanas, sendo que cada usuário pôde escolher o momento mais apropriado para a sua realização, nos dias e horários em que pudesse ser melhor conciliado com as suas atividades profissionais. No caso das bibliotecárias, o prazo foi necessário para que as mesmas pudessem se dedicar a avaliação da ferramenta sem comprometer o trabalho da gerência.

Assim como no primeiro estudo de caso, os usuários utilizaram a ferramenta nas suas duas modalidades disponíveis (expansão automática e sugestão de termos) em pelo menos uma necessidade de informação. A principal diferença entre os dois estudos foi em relação aos contextos. Os funcionários das gerências setoriais de Meio Ambiente, deveriam realizar pesquisas referentes a um dos seguintes assuntos: Desenvolvimento sustentável, Mudanças climáticas e Política ambiental. Já os funcionários das gerências setoriais de Energia pesquisariam nos assuntos Biocombustíveis, Petróleo e Gás e Energia Elétrica. As bibliotecárias do Coped realizaram avaliações nos dois contextos, podendo pesquisar em quaisquer assuntos dentre os citados.

As atividades a serem realizadas foram as mesmas do primeiro experimento: uso da ferramenta nas modalidades expansão automática e sugestão de termos, avaliações de relevância nos resultados de busca e preenchimento de questionário, com as mesmas

perguntas. O questionário enviado para as bibliotecárias possuia um pergunta adicional, que procurava comparar a realização do estudo nos dois diferentes contextos.

5.4. Os questionários

O objetivo do questionário utilizado nos dois estudos de caso foi avaliar qualitativamente a utilização da ferramenta ContextSearch. As perguntas foram apresentadas aos usuários agrupadas em 3 seções, referentes a perguntas sobre cada uma das modalidades de pesquisa avaliadas (expansão automática e sugestão de termos), e perguntas gerais sobre o experimento, que procuravam comparar as duas modalidades.

A maioria das perguntas do questionário foi de múltipla escolha, sendo que em grande parte destas foram apresentadas afirmações sobre a utilização da ferramenta, e que os usuários deveriam se posicionar em relação a estas afirmações, marcando a alternativa mais adequada dentro da escala de 5 posições fornecida, mostrada na Tabela 1.

Tabela 1: opções de resposta do questionário

Opção
Discordo totalmente
Discordo parcialmente
Neutro
Concordo parcialmente
Concordo totalmente

As perguntas discursivas e não obrigatórias, apresentadas na seção de perguntas gerais, tiveram o objetivo de capturar possíveis informações sobre o estudo de caso não previstas nas perguntas de múltipla escolha. Além de buscar a opinião dos usuários sobre a ferramenta e as suas modalidades, também foi esperado obter sugestões de melhoria para a interface de pesquisa. Todas as questões do formulário utilizado nos estudos de caso podem ser consultadas no Anexo I, bem como o roteiro entregue aos

participantes (Anexo II), contendo uma breve explicação sobre a pesquisa, os objetivos do estudo e a lista de atividades a serem realizadas.

No próximo capítulo serão apresentadas as avaliações dos resultados obtidos nos dois estudos de caso realizados.

Capítulo 6 – Avaliação

Neste capítulo serão apresentadas as métricas utilizadas para a avaliação dos estudos de caso, os resultados obtidos em avaliações quantitativas e qualitativas nas duas modalidades da ferramenta: expansão automática e sugestão de termos; nos diferentes ambientes: educacional e corporativo; e com os diferentes perfis de usuários: estudantes, funcionários de gerências setoriais e bibliotecárias.

6.1. Métricas utilizadas na avaliação

Os resultados obtidos foram avaliados de acordo com três métricas propostas por diferentes pesquisadores e consolidadas por (TANG e SUN, 2003) para a avaliação de buscadores web: Precisão total dos 10 primeiros resultados, profundidade da busca e correlação de *ranking*.

Precisão é uma métrica largamente utilizada em recuperação de informações, representando a fração dos documentos relevantes dentre todos os recuperados (MANNING, RAGHAVAN e SCHÜTZE, 2008). Porém, o julgamento binário de relevância adotado nas avaliações tradicionais não leva em consideração a diferente quantidade de informações relevantes encontradas em cada documento. A medida de precisão total (CHIGNELL, GWIZDKA e BODNER, 1999) tenta somar a quantidade de informações relevantes encontradas nos primeiros resultados.

A métrica de precisão total aplicada na pesquisa original de (TANG e SUN, 2003) considerava os 20 primeiros resultados. Nesta dissertação foi feita uma adaptação na

métrica de precisão total, que foi aplicada nos 10 primeiros resultados. A alteração do número de resultados avaliados se justifica por estudos que mostram o comportamento ao realizar buscas na Internet, em que a maioria dos usuários acessa somente a primeira página de resultados e em geral se limitando aos 5 primeiros resultados (SPINK e JANSEN, 2004) e (JANSEN *et al*., 1998). Como os principais buscadores web (Google, Yahoo e Bing) utilizam como configuração padrão a exibição de 10 resultados por página, foram avaliados somente os resultados correspondentes à primeira página de resultado.

Os usuários julgaram cada resultado obtido, dentre os 10 primeiros, de acordo com sua relevância em uma escala de 5 pontos, usando "0" para indicar sem relevância e "4" indicando alta relevância. Estes julgamentos foram utilizados para gerar a medida de precisão total dos 10 primeiros resultados, calculada com a aplicação da seguinte fórmula:

$$\text{Precisão Total dos } 10 \text{ primeiros resultados} = \frac{\sum_{i=1}^{10} \text{julgamentoi}}{10 \times 4}$$

Onde julgamento$_i$ corresponde ao julgamento de relevância atribuído ao i-ésimo resultado, 10 é o número de resultados considerados e 4 é o valor máximo de relevância atribuído para cada resultado.

A segunda métrica utilizada foi a profundidade da busca (COOPER, 1968), que reflete o número de documentos não relevantes que o usuário deve avaliar até que encontre *x* documentos consecutivos considerados relevantes. Logo, quanto menor o valor, menor é o esforço do usuário para encontrar resultados relevantes. Assim como em (TANG e SUN, 2003), nesta pesquisa a profundidade da busca foi definida como o número de documentos avaliados até que fossem encontrados 2 resultados consecutivos com avaliação de relevância maior ou igual a 3. Como neste trabalho estão sendo

avaliados os 10 primeiros documentos, ao invés dos 20 conforme a pesquisa original, esta métrica tende a ser um pouco mais rigorosa, desprezando os casos em que a profundidade de busca seja maior do que 10.

A última métrica utilizada, correlação de ranking (SU, CHEN e DONG, 1998), objetiva comparar a correlação entre a priorização obtida na busca com uma priorização ideal, em que os resultados obtidos sejam ordenados em ordem decrescente de relevância. Quanto maior é a correlação entre a relevância dos resultados de busca e a priorização ideal, mais eficiente é a ferramenta de busca. Dado que foram avaliados os 10 primeiros resultados, e a escala de avaliação é de 5 posições, a priorização ideal a ser considerada possui nas primeiras posições dois documentos com avaliação 4, dois com avaliação 3, e assim sucessivamente até os dois últimos resultados com avaliação 0. Para o cálculo da correlação foi utilizado o coeficiente de correlação de Pearson (KENDALL e STUART, 1973).

De acordo com (KOWALSKI, 1999), a relevância de um resultado de pesquisa (i) possui natureza subjetiva, dependendo do julgamento de um usuário específico; (ii) é situacional, relacionado aos requisitos do usuário; (iii) é cognitivo, dependendo da percepção e comportamento humano; (iv) é temporal, variando com o tempo; e (v) é mensurável, observável em pontos no tempo.

Dadas estas características dos julgamentos de relevância, e como cada julgamento foi realizado por um usuário diferente sobre necessidades de informações diferentes, não foi aplicado nenhum cálculo estatístico para agrupar os dados obtidos aplicando as métricas de avaliação nos julgamentos de relevâncias.

Para consolidar os dados obtidos nos estudos de caso, os valores obtidos com as métricas foram utilizados para determinar, para cada avaliação, se os resultados foram melhores na consulta original ou nas consultas expandidas, variando o número de

expansões e consequentemente comparações de acordo com a modalidade avaliada (expansão automática e sugestão de termos).

Os resultados são apresentados em percentuais, representando a quantidade de vezes em que o resultado obtido com a consulta original ou com as consultas expandidas foi melhor que as demais. No caso das consultas expandidas com termos dos assuntos, foi considerado que a expansão foi melhor se em pelo menos uma das consultas o resultado obtido foi melhor do que os resultados da consulta original. Os percentuais não totalizam 100%, pois existem casos em que os resultados obtidos na consulta original foram equivalentes aos resultados obtidos nas consultas expandidas.

6.2. Número de avaliações consideradas

A avaliação quantitativa consistia em julgar a relevância dos 10 primeiros resultados de busca obtidos na consulta original e nas consultas expandidas. Para a modalidade expansão automática, eram julgados os resultados da consulta expandida com termos do contexto e os resultados das três consultas expandidas com termos de assuntos, totalizando 50 documentos em 5 abas de resultados. Já na modalidade sugestão de termos, só é realizada uma expansão de consulta, totalizando desta forma 20 resultados em 2 abas.

Tabela 2: Resultados avaliados

Modalidade	Consulta original	Expansão contexto	Expansão assuntos	Total
Expansão automática	10	10	30	50
Sugestão de termos	10	10	-	20

Durante o estudo de caso, somente foram consideradas as avaliações em que todos os resultados de busca exibidos em todas as abas e nas duas modalidades (expansão automática e sugestão de termos) tiveram a sua relevância atribuída. Alguns usuários

avaliaram somente uma modalidade de forma completa (avaliando todas as abas de resultado desta modalidade). Neste caso, as avaliações foram parcialmente consideradas.

No estudo de caso no ambiente educacional, vinte e seis estudantes participaram das avaliações, sendo que deste total, oito avaliações foram totalmente consideradas e sete avaliações foram parcialmente consideradas (em alguma das duas modalidades), totalizando doze avaliações para a modalidade expansão automática e doze avaliações para a modalidade sugestão de termos. Duas avaliações foram descartadas, embora tenham sido consideradas completas: na primeira a pesquisa foi completamente fora do contexto (a expressão de consulta utilizada foi "Lady Gaga"), enquanto na segunda todos os resultados de todas as abas foram avaliados com relevância máxima (4).

No estudo de caso no ambiente corporativo, todos os resultados foram aproveitados. Participaram do estudo cinco bibliotecárias e seis funcionários de gerências setoriais, três de cada contexto (Meio Ambiente e Energia). As bibliotecárias deveriam fazer duas avaliações, uma para cada contexto, sendo que uma bibliotecária fez somente em um (Energia). Dois funcionários de gerências setoriais, um de cada contexto, realizaram duas avaliações. No total, foram obtidas oito avaliações no contexto Meio Ambiente e nove avaliações no contexto Energia.

6.3. Análise quantitativa

Os resultados obtidos nos dois estudos de caso serão apresentados na forma de gráficos de barras, representando percentuais obtidos em cada consulta executada nas seguintes opções:

Modalidade expansão automática

- Consulta original: Percentual de avaliações em que a aba de resultados da consulta original apresentou melhores valores na métrica em questão;

- Expansão contexto: Percentual de resultados em que a aba de resultados da consulta expandida com termos do contexto apresentou melhores valores do que a consulta original na métrica em questão;
- Expansão assuntos: Percentual de resultados em que pelo menos uma das três abas de resultados da consulta expandida com termos dos assuntos apresentou melhores valores do que a consulta original na métrica em questão;
- Qualquer expansão: Percentual de resultados em que pelo menos uma das abas de resultados de consulta expandida apresentou melhores valores do que a consulta original na métrica em questão.

Modalidade sugestão de termos

- Consulta original: Percentual de avaliações em que a aba de resultados da consulta original apresentou melhores valores na métrica em questão;
- Consulta expandida: Percentual de resultados em que a aba de resultados da consulta expandida com termos selecionados apresentou melhores valores do que a consulta original na métrica em questão.

Os resultados gerais das duas modalidades foram praticamente os mesmos nos dois estudos de caso. Na modalidade expansão automática, em todas as métricas a expansão com termos gerais do contexto e com termos dos assuntos apresentou melhores resultados do que os obtidos com a consulta original.

Na precisão total, nos dois ambientes foi observada uma melhora em aproximadamente 75% dos casos em pelo menos uma das expansões, enquanto o resultado da consulta original só apresentou melhores resultados em 25% e 11,76% dos casos, nos ambientes educacional e corporativo respectivamente. Nos dois ambientes, foi observada a melhora percentual tanto na comparação dos resultados da consulta

original com os resultados da consulta expandida com termos do contexto, quanto na comparação dos resultados da consulta original com os resultados das consultas expandidas com termos dos assuntos. A diferença observada entre os dois ambientes foi em relação a qual tipo de expansão obteve os melhores resultados. Enquanto no ambiente educacional a expansão com termos do contexto obteve melhores resultados que a expansão com termos do assunto, no ambiente corporativo o resultado se inverteu (Figura 29).

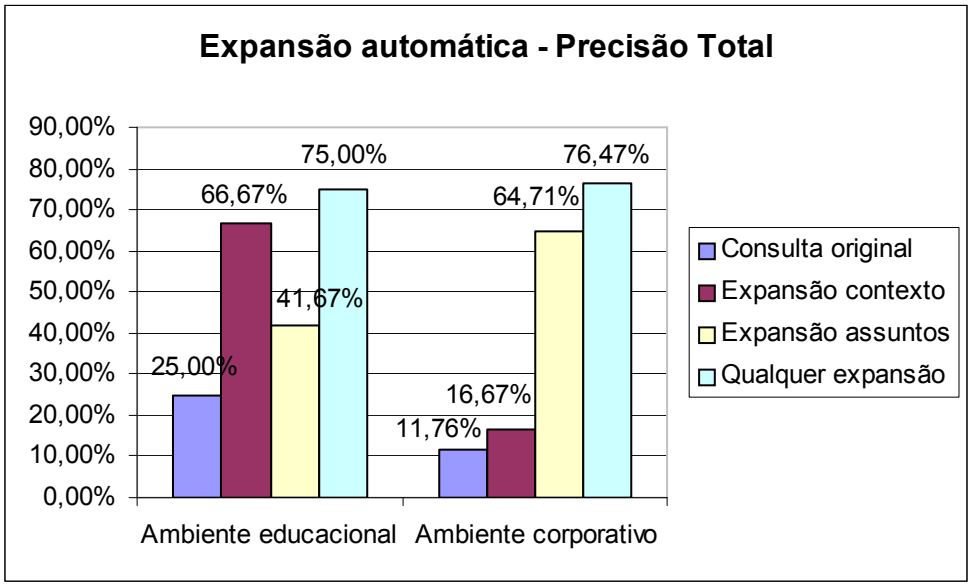


Figura 29 – Resultados da métrica Precisão total na modalidade Expansão automática

Na métrica profundidade de busca foi observado o mesmo comportamento apresentado na precisão total, porém com resultados menos discrepantes. Assim como na métrica precisão total, foi observada a mesma diferença nos dois ambientes em relação a qual expansão obteve melhores resultados. No ambiente educacional, a expansão com termos do contexto apresentou resultados melhores do que a expansão com termos dos assuntos, enquanto que no ambiente corporativo o resultado foi inverso (Figura 30).

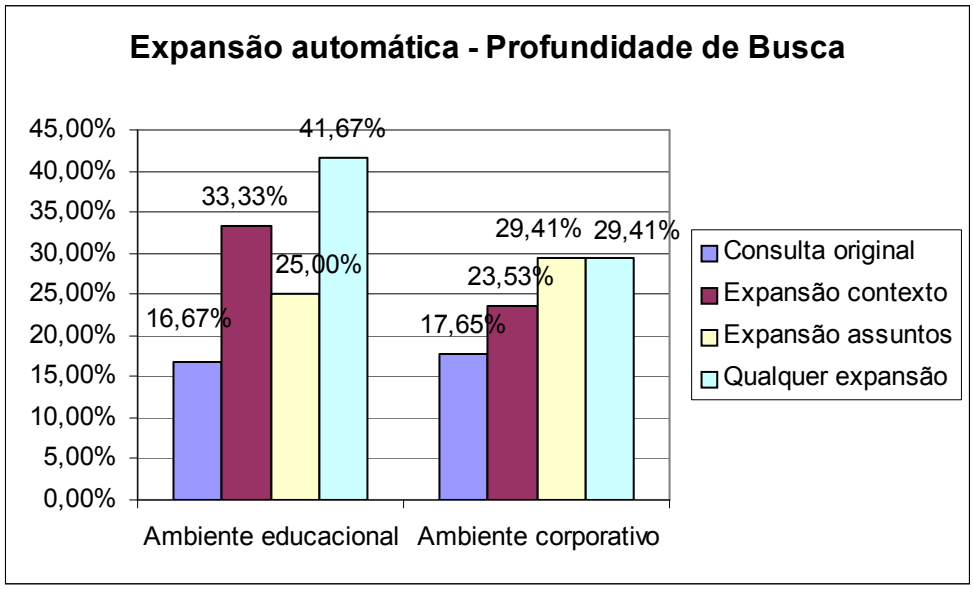


Figura 30 – Resultados da métrica Profundidade de busca na modalidade Expansão automática

Já na terceira métrica, correlação de *ranking*, a expansão de consulta com termos dos assuntos apresentou, em pelo menos em uma de suas abas de resultados de pesquisa, melhores índices de correlação com o *ranking* ideal. O valor da métrica na expansão com termos do contexto apresentou o único resultado desfavorável em relação a consulta original, no estudo de caso realizado no ambiente educacional (Figura 31).

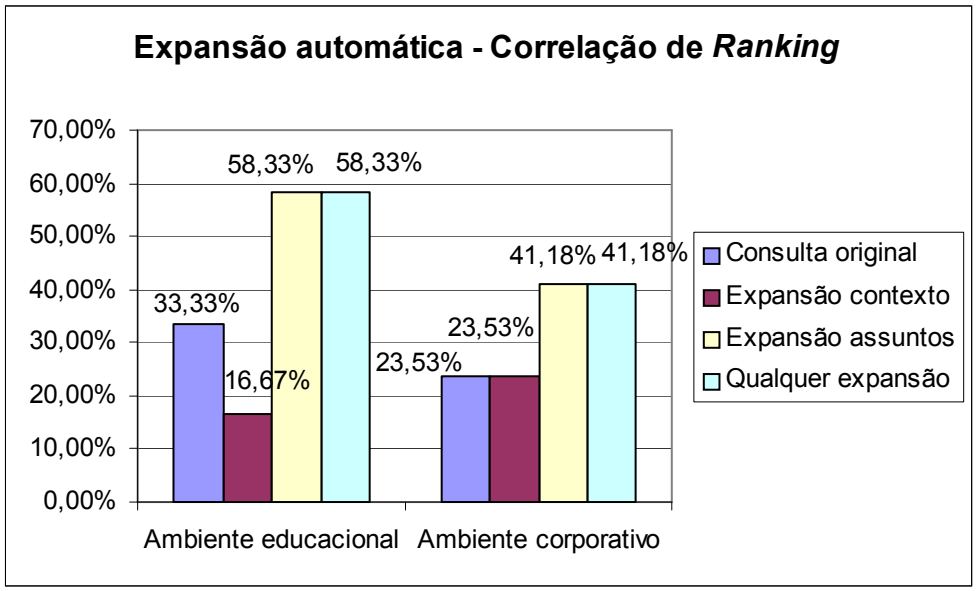


Figura 31 – Resultados da métrica Correlação de ranking na modalidade Expansão automática

Na modalidade sugestão de termos, diferentemente do ocorrido com a expansão automática, em todas as métricas a expansão com termos do contexto apresentou piores resultados do que os obtidos com a consulta original. E a diferença observada entre os percertuais foi muito grande, chegando em alguns casos a superar 7 vezes o percentual da consulta expandida.

A precisão total dos 10 primeiros resultados apresentou, nos dois ambientes, uma ampla vantagem das consultas originais em relação à consulta expandidas. Enquanto no ambiente educacional, o percentual de resultados em que a consulta original foi melhor do que a consulta expandida foi de 75%, no ambiente corporativo este valor foi 58,82% (Figura 32).

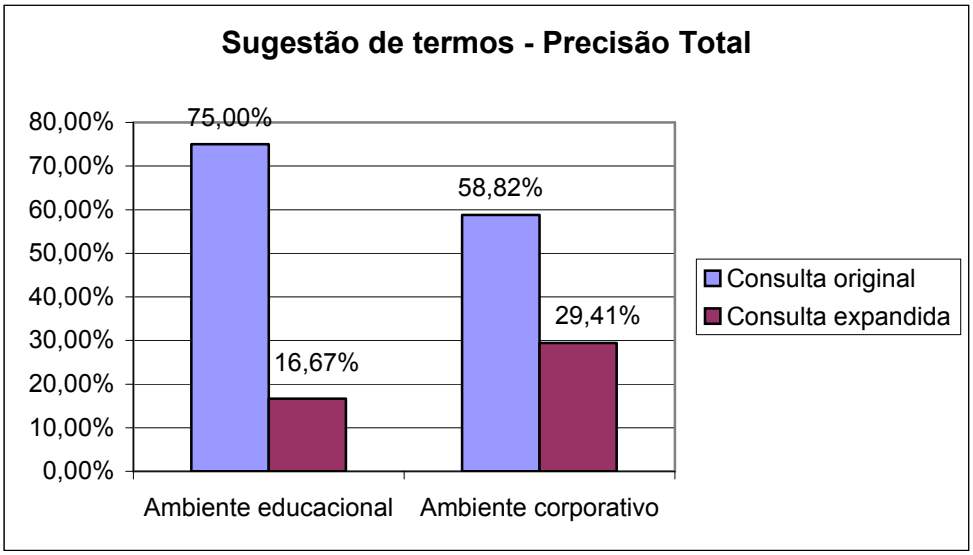


Figura 32 – Resultados da métrica Precisão total na modalidade Sugestão de termos

Também na profundidade da busca foi constatada a maior eficiência das consultas originais em relação às consultas expandidas. No ambiente educacional, somente 25% dos resultados obtidos com as consultas originais foram melhores do que os obtidos

com as consultas expandidas, enquanto que no ambiente corporativo este valor foi de 47,06% (Figura 33).

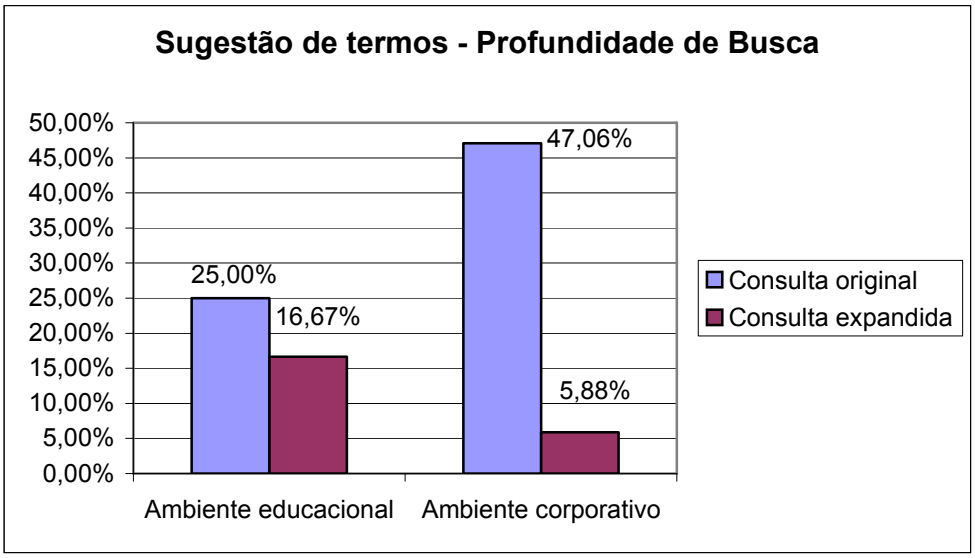


Figura 33 – Resultados da métrica Profundidade de busca na modalidade Sugestão de termos

Na terceira métrica avaliada (correlação de *ranking*), no ambiente educacional, foi observada a única situação em que a consulta expandida obteve melhores resultados (50%). No ambiente corporativo, o resultado foi compatível com o obtido nas demais métricas, em que 70,59% dos resultados obtidos com a consulta original foram superiores aos obtidos com as consultas expandidas (Figura 34).

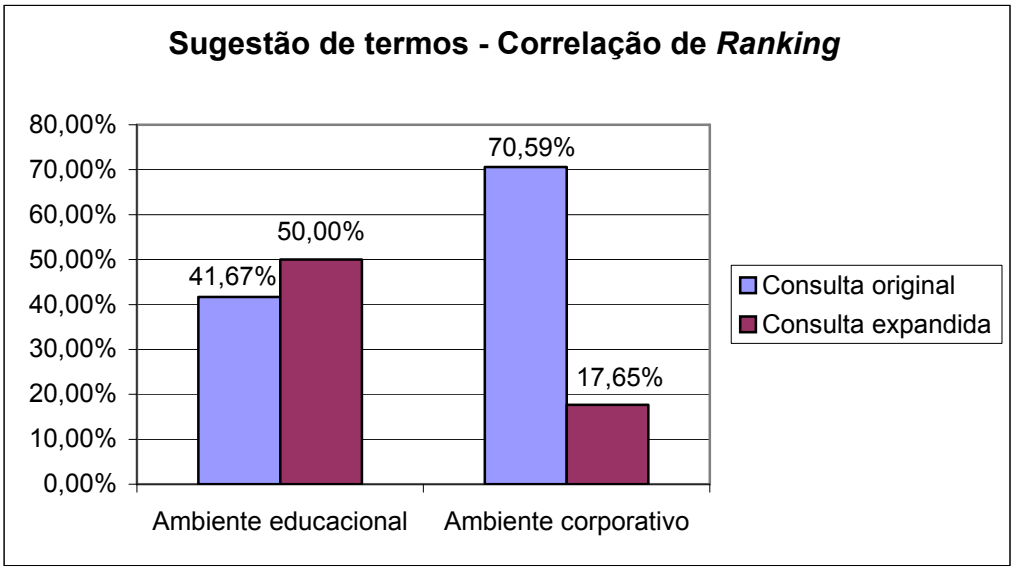


Figura 34 – Resultados da métrica Correlação de ranking na modalidade Sugestão de termos

6.3.1. Justificativas

A hipótese desta pesquisa é que, a utilização dos principais termos do contexto, sendo estes termos gerais ou específicos por subtemas do contexto, na expansão de consultas elaboradas por usuários, torna os resultados de pesquisas mais úteis. Os resultados gerais obtidos nas pesquisas mostram que a princípio a hipótese vale para a modalidade expansão automática, porém não vale para a sugestão de termos.

Os resultados das avaliações de relevância e os questionários foram analisados, e algumas justificativas foram encontradas para a diferença tão grande entre os valores obtidos nas três métricas nas duas modalidades.

6.3.1.1. Número de termos selecionados

Nas duas modalidades o usuário indicava quantos termos seriam incluídos na consulta original, porém de formas distintas: enquanto na expansão automática o usuário influenciava mais claramente esta quantidade, pois era obrigado a selecionar em uma combo este valor; na sugestão de termos essa influência era indireta, pois os termos eram sugeridos (40 termos no total) e o usuário os incluia de acordo com o seu julgamento se o termo seria ou não relevante para atender a sua necessidade de informação, sem se preocupar em limitar o número de termos que seriam incluídos na consulta.

Esta diferença acabou impactando significativamente o número de termos utilizados nas consultas. Enquanto na modalidade expansão automática os usuários em média expandiram a consulta com aproximadamente 3,5 termos, na modalidade sugestão de termos os usuários selecionaram em média aproximadamente 6,9 termos, praticamente o dobro. Já a quantidade de termos utilizada nas consultas originais foi

praticamente a mesma nas duas modalidades. Na expansão automática o tamanho médio foi de aproxidamente 2,65 termos e na sugestão de termos, aproximadamente 2,37.

A diferença de termos utilizados na expansão de forma isolada não justifica a diferença nos resultados, mas apresenta grande contribuição. A modalidade expansão automática apresentou diferenças relevantes de acordo com o número de termos utilizados na expansão (Seção 6.3.2).

6.3.1.2. Escolha de palavras-chave

Um dos problemas da busca tradicional baseada em palavras-chave, relatado na introdução deste trabalho, é que nesta abordagem o usuário precisa conhecer as palavras-chaves que devem estar presentes no conteúdo dos documentos a serem recuperados, para elaborar boas expressões de busca e consequentemente obter bons resultados. Porém, identificar as palavras-chave que estarão presentes em bons resultados de pesquisa não é uma atividade trivial. Trabalhos na área de expansão de consultas interativa (IQE - *Interactive Query Expansion*) mostram que usuários normalmente consideram uma atividade difícil selecionar termos para expansão a partir de uma lista de sugestões (JOHO, SANDERSON e BEAULIEU, 2004).

Os resultados mostraram que a seleção dos termos a partir de regras automáticas baseadas na frequência destes termos em relação aos documentos que compõem o contexto apresenta melhores resultados do que a seleção de termos manual, feita pelos usuários, embora os termos apresentados na lista de sugestão também tenham sido selecionados com base na frequência.

Uma indicação de como a dificuldade de selecionar termos está relacionada com os resultados ruins obtidos com a modalidade sugestão de termos pode ser comprovada analisando os diferentes resultados obtidos na métrica precisão total com perfis diferentes de usuários.

No ambiente educacional, os estudantes selecionaram termos de um assunto que conheciam a menos de uma semana, quando receberam a tarefa de ler o capítulo do livro referente a gestão de conhecimento. Em 75% dos resultados, as consultas originais obtiveram melhores resultados do que as consultas expandidas, enquanto em somente 16,67% dos casos as consultas expandidas foram melhores.

No ambiente corporativo, foram analisados os resultados dos dois diferentes perfis de usuários. Os funcionários das gerências setoriais, que conhecem o domínio da informação, porém não são especialistas em informação e busca, obtiveram resultados um pouco melhores (do que os dos alunos). 75% de resultados melhores na consulta original, porém 25% dos resultados foram melhores com as consultas expandidas.

Já no grupo das bibliotecárias, especialistas em informação, cujo trabalho diário envolve a captura e cadastramento de informações dos contextos avaliados, os resultados foram ainda melhores. Enquanto as consultas originais obtiveram melhores resultados em 55,56% dos casos, as consultas expandidas foram melhores em 44,44% dos casos.

6.3.1.3. Seleção de termos do contexto e diferentes assuntos

A terceira e última justificativa identificada para os maus resultados na modalidade sugestão de termos foi a possibilidade de combinar termos de diferentes assuntos em uma mesma expansão de consulta. Na expansão automática eram realizadas expansões separadas, uma para cada assunto e para os termos gerais do contexto.

Analisando os resultados do questionário, foi observado que a possibilidade de misturar os termos foi aplicada na prática. 51,72% dos participantes revelaram que na maioria das consultas, utilizaram termos do contexto e dos assuntos ao mesmo tempo para elaborar a consulta. 34,48% dos participantes não concordaram com esta afirmação.

Em outra questão do questionário, 72,41% dos participantes concordaram que os termos sugeridos para cada assunto realmente identificavam diferentes assuntos dentro do contexto. Como somente são retornados pela busca documentos que contenham todos os termos da expressão de busca, a utilização de termos de subtemas diferentes, embora todos referentes ao mesmo contexto, pode ter ocasionado piores resultados.

6.3.2. Impacto das variáveis nos julgamentos de relevância

Além das avaliações dos resultados gerais, também foi avaliada a influência das variáveis independentes e intervenientes nos julgamentos de relevâncias realizados pelos participantes. Como variável independente foi considerado o número de termos utilizados na expansão, selecionado pelos usuários nas duas modalidades. Já como variáveis intervenientes foram considerados: o perfil dos usuários selecionados para cada estudo de caso (estudantes, bibliotecárias e especialistas no domínio) e os diferentes contextos utilizados no ambiente corporativo (Energia e Meio Ambiente).

O maior impacto percebido foi o número de termos utilizados para a expansão das consultas. Na modalidade expansão automática, caso sejam considerados somente os resultados em que o número de termos utilizados para expansão é menor ou igual ao número de termos utilizados na consulta original feita pelo usuário, observamos uma grande melhora na métrica precisão total, conforme mostrado na figura 35.

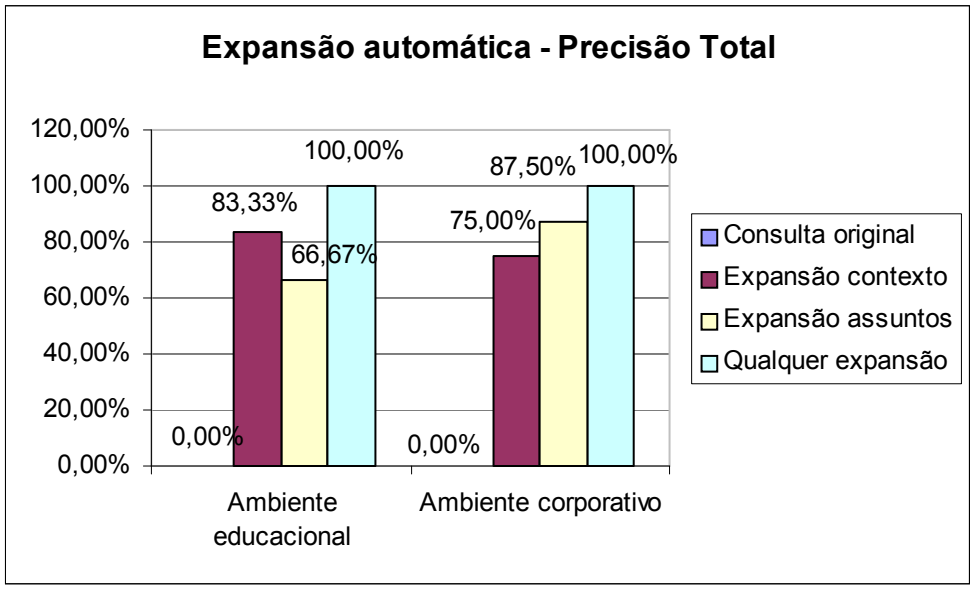


Figura 35 – Resultados da métrica Precisão total na modalidade Expansão automática, limitando o número de termos utilizados na expansão

Em nenhum dos casos analisados a consulta original obteve resultados melhores ou iguais aos obtidos com as consultas expandidas. A diferença observada entre os ambientes foi qual dos dois tipos de expansão obteve melhores resultados. Enquanto no ambiente educacional a expansão com termos de contexto obteve melhores resultados (83,33% contra 66,67% da expansão com termos dos assuntos), no ambiente corporativo pelo menos uma das expansões com termos dos assuntos mostraram melhores resultados (87,50% contra 75% da expansão com termos do contexto). Em todos os casos e nos dois ambientes alguma das expansões foi melhor do que a consulta original.

Pode-se observar com estes resultados que obtemos maior precisão quando os termos digitados pelo usuário possuem maior peso (ou igual) aos termos expandidos automaticamente. Os termos gerais do contexto e dos assuntos fazem com que a consulta obtenha resultados mais úteis, porém estes termos não devem ter na consulta final uma importância maior do que os termos elaborados pelos usuários, presentes na consulta original. Isto pode ser confirmado ao analisar somente os casos em que o

tamanho da expansão, isto é, o número de termos utilizados nas expansões, é maior do que o tamanho da consulta original feita pelos usuários. No ambiente corporativo foi observada uma piora dos resultados, aumentando o número de situações em que a consulta original obteve resultados mais precisos. Já no ambiente educacional, a piora foi ainda mais acentuada, tornando a consulta original tão precisa quanto a consulta com termos do contexto, e melhor do que as consultas expandidas com termos dos assuntos (Figura 36).

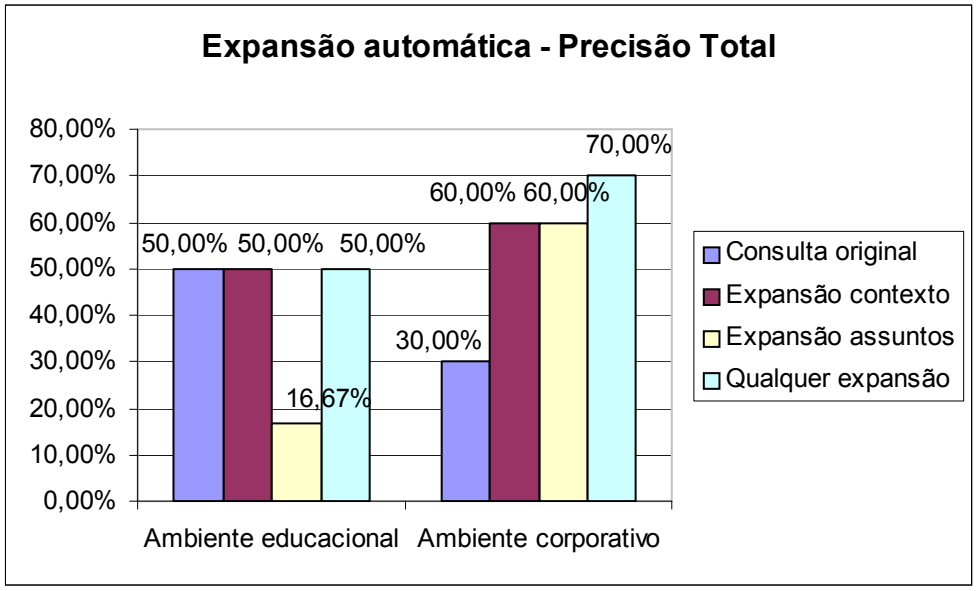


Figura 36 – Resultados da métrica Precisão total na modalidade Expansão automática, com tamanho da expansão maior que o tamanho da consulta original

A diferença entre os perfis de participantes também apresentou influência na avaliação dos resultados, porém se considerarmos apenas qual modalidade de expansão foi superior ou inferior aos resultados das consultas originais, em todos os perfis foram obtidas as mesmas respostas. Na expansão automática, os três perfis obtiveram melhores resultados com as consultas expandidas do que com as consultas originais. Já na modalidade sugestão de termos, todos obtiveram resultados piores.

Os diferentes contextos utilizados no estudo de caso no ambiente corporativo não influenciaram as avaliações de relevância. Os resultados obtidos nas três métricas não

apresentaram diferenças significativas, embora parte dos grupos de avaliadores tenham sido diferentes (as bibliotecárias avaliaram os dois contextos, porém os funcionários de gerências setoriais só avaliaram o contexto relacionado com sua área de atuação).

6.4. Análise qualitativa

As questões apresentadas no questionário tinham como objetivo a avaliação qualitativa das duas modalidades de expansão de consulta, em outros critérios além da relevância dos resultados. A seguir são apresentados os resultados mais significativos.

Uma das questões do questionário procurava comprovar uma informação obtida nos experimentos laboratoriais, em que se observou que os resultados obtidos com as consultas expandidas faziam mais referências a publicações, tais como artigos, relatórios, pesquisas, etc.. Ao serem questionados se este fato foi observado durante o uso da ferramenta, 65,52% dos participantes concordaram em relação à modalidade expansão automática, enquanto 58,62% concordaram em relação à modalidade sugestão de termos (Tabela 3).

Tabela 3: Referências a publicações

Os resultados das consultas expandidas trazem mais referências a publicações do que os resultados da consulta original					
Modalidades	**Discordo totalmente**	**Discordo parcialmente**	**Neutro**	**Concordo parcialmente**	**Concordo totalmente**
Expansão automática	6,9%	27,59%	0%	34,48%	31,03%
Sugestão de termos	10,34%	17,24%	13,79%	37,93%	20,69%

Outra questão em que houve concordância dos particiantes foi em relação à identificação de diferentes assuntos nas abas de resultados ou nos termos sugeridos, respectivamente na modalidade expansão automática e sugestão de termos. Uma das

questões afirmava que na expansão automática os resultados das abas de assuntos apresentavam diferentes visões sobre o tema pesquisado; outra afirmava que os termos dos diferentes assuntos realmente representavam diferentes assuntos dentro do contexto. Enquanto na primeira a concordância obtida foi de 65,52%, na segunda este valor foi de 72,4% (Tabela 4).

Tabela 4: Identificação de assuntos

Os resultados apresentados nas abas de assuntos representam diferentes visões sobre o tema pesquisado. Os termos sugeridos para cada assunto realmente identificam diferentes assuntos dentro do contexto.					
Modalidades	**Discordo totalmente**	**Discordo parcialmente**	**Neutro**	**Concordo parcialmente**	**Concordo totalmente**
Expansão automática	6,9%	13,79%	13,79%	44,83%	20,69%
Sugestão de termos	6,9%	17,24%	3,45%	51,72%	20,69%

A avaliação dos participantes em relação à ferramenta em geral, e em particular para cada modalidade também foi muito boa. Ao serem questionados se gostariam de ter uma ferramenta como a avaliada para apoiar na realização de suas atividades de pesquisa, 79,31% concordaram em relação à modalidade expansão automática, enquanto 82,76% concordaram em relação a modalidade sugestão de termos. Avaliando a ferramenta como um todo, 79,31% concordaram que a ferramenta é útil para a realização de pesquisas.

6.4.1. Comparação entre as modalidades

Embora na análise quantitativa a avaliação da modalidade expansão automática tenha sido largamente mais favorável do que a avaliação da modalidade sugestão de termos, na avaliação qualitativa este cenário não se repetiu, não sendo possível afirmar que uma modalidade apresentou melhores resultados que a outra.

Ao serem questionados sobre qual abordagem apresentou resultados mais úteis, 44,83% dos participantes apontaram a expansão automática, e 55,17% escolheram a sugestão de termos. Já ao opinar sobre qual das duas apresenta o melhor custo-benefício, o resultado se inverteu: 58,62% preferiram a modalidade expansão automática, contra 41,38% que optaram pela sugestão de termos. Ao analisar o custo-benefício, foi considerado que com a expansão automática a pesquisa pode ser executada de forma mais rápida do que na sugestão de termos, em que é necessário um esforço adicional de ler todos os termos e selecionar os mais adequados.

A maioria dos participantes (51,72%) também concordou com a afirmativa de que existem situações em que é mais indicado utilizar a expansão automática e outras em que é melhor utilizar a sugestão de termos.

Embora esta diferença entre os resultados obtidos com as avaliações de relevância e os questionários, relativa à comparação entre as duas modalidades de pesquisa, não possa ser plenamente justificada, ao analisar algumas respostas abertas do questionário é possível levantar algumas hipóteses:

(i) na modalidade sugestão de termos o usuário possui o controle da expressão de busca a ser gerada, enquanto na expansão automática o usuário não sabe quais termos serão utilizados;

(ii) na modalidade sugestão de termos é possível adicionar ou removar termos com facilidade após as execuções de consultas, facilitando o refinamento dos resultados. Ao receber o primeiro resultado de pesquisa, é possível melhorá-lo após tentativas de inclusão / exclusão de termos;

(iii) a visualização do resultado na modalidade sugestão de termos é única, enquanto na expansão automática foram exibidas quatro abas de

resultados (uma aba para termos do contexto e três abas para termos dos assuntos);

(iv) quando os termos escolhidos automaticamente não possuem relação com a necessidade de informação, os resultados acabam sendo muito piores do que a consulta original. Já na sugestão de termos, isto não ocorre pois é o próprio usuário quem faz a seleção. Caso nenhum termo sugerido seja relacionado, basta não selecionar nenhum.

Capítulo 7 – Conclusão

Nesta dissertação foi apresentado o desenvolvimento de uma abordagem para melhorar a relevância dos resultados obtidos em motores de busca na Internet. Esta abordagem consiste em um método de expansão de consultas cujos termos utilizados são extraídos de recursos predefinidos com a aplicação de técnicas de segmentação de tópicos e agrupamento, fazendo com que os resultados de busca sejam mais contextualizados e consequentemente mais relevantes dos que os obtidos com as consultas originais.

Além de tornar os resultados de pesquisa mais relevantes, a abordagem requer baixo custo para a definição do contexto, exigindo pouco tempo e esforço de especialistas. O trabalho envolvido por parte do especialista se resume a identificar uma coleção de documentos que seja representativa do contexto a ser utilizado.

Na ferramenta desenvolvida foram implementadas duas modalidades distintas para a expansão das consultas. A primeira delas é a expansão automática, em que a consulta original é expandida pela ferramenta sem a interferência direta pelo usuário. São realizadas *n* expansões, sendo o valor de *n* igual a um (1) mais o número de assuntos identificados no contexto selecionado. Uma expansão é feita com os termos gerais do contexto e as demais são expansões feitas com os termos específicos de cada assunto. Cada consulta expandida é executada no buscador web selecionado e os resultados são apresentados para o usuário em abas distintas.

A segunda modalidade de expansão implementada é a sugestão de termos, em que todos os termos extraídos (termos gerais do contexto e termos específicos dos assuntos) são apresentados para o usuário como uma sugestão em listas distintas. O usuário manualmente seleciona os termos de seu interesse para que sejam incorporados à consulta original, realizando uma única expansão. Para selecionar os termos, é possível combinar termos das diferentes listas exibidas, misturando termos gerais do contexto e termos dos diferentes assuntos identificados.

Para avaliar a abordagem proposta, foram conduzidos dois estudos de caso em ambientes distintos: um educacional e um corporativo; com participantes de diferentes perfis: estudantes, especialistas em informação (bibliotecárias) e especialistas de domínio. No ambiente corporativo também foram analisados dois diferentes contextos: Energia e Meio ambiente.

Todos os participantes utilizaram a ferramenta desenvolvida para efetuar buscas e julgar a relevância dos resultados obtidos, respondendo a um questionário ao final do estudo de caso. Os julgamentos de relevância foram avaliados a partir de métricas consolidadas por (TANG e SUN, 2003) para a avaliação de buscadores web.

Os resultados obtidos nos estudos de caso mostraram que a hipótese de trabalho desta pesquisa é válida para a modalidade de busca expansão automática, porém os resultados das métricas não foram bons para a modalidade de busca sugestão de termos. Como justificativas para a diferença entre os resultados obtidos nas diferentes modalidades estão: (i) a diferença entre o número de termos utilizados para expansão nas duas modalidades, (ii) a dificuldade encontrada pelos usuários para selecionar palavras-chave para expansão e (iii) a seleção de termos de diferentes assuntos em uma mesma expansão.

O uso das diferentes expansões de consulta, com termos gerais do contexto e termos específicos dos assuntos, pode ser justificada pela diferença obtida nos ambientes educacional e corporativo. Em cada ambiente um dos tipos de expansão obteve melhores resultados do que no outro, nas métricas precisão total e profundidade de busca.

Além das avaliações dos resultados gerais, também foi avaliada a influência das variáveis de pesquisa nos julgamentos de relevâncias realizados pelos participantes. O maior impacto percebido foi o número de termos utilizados para a expansão das consultas na modalidade expansão automática. Considerando somente os resultados em que o número de termos utilizados para expansão foi menor ou igual ao tamanho da consulta feita pelo usuário, foi observada uma grande melhora na métrica precisão total. Pode-se observar com estes resultados que se obtem maior precisão quando o peso do conjunto de termos digitados pelo usuário é maior ou igual ao peso do conjunto de termos expandidos automaticamente. Os termos gerais do contexto e dos assuntos fazem com que a consulta obtenha resultados mais relevantes, porém estes termos não devem ter na consulta final uma importância maior do que os termos elaborados pelos usuários, presentes na consulta original.

As demais alterações em variáveis não ocasionaram impacto relevante. A diferença entre os perfis de participantes apresentou influência na avaliação dos resultados, porém ao considerar apenas qual modalidade de expansão foi superior ou inferior aos resultados das consultas originais, em todos os perfis foram obtidas as mesmas respostas. Os diferentes contextos utilizados no estudo de caso no ambiente corporativo não influenciaram as avaliações de relevância.

7.1. Trabalhos relacionados

Nesta dissertação, a criação do modelo de contexto é feita de forma automática e o modelo é dependente de um *corpus*. Dentre os trabalhos com a mesma classificação, os mais semelhantes são os que utilizam atividades de extração de informação, tais como stemização, agrupamento e co-ocorrência para a criação do modelo de contexto. A co-ocorrência é utilizada na criação do modelo de contexto em (XU e CROFT, 2000), em que é proposta uma técnica chamada análise de contexto local que seleciona os termos para expansão com base em co-ocorrência entre os termos da consulta e os termos dos *n* primeiros documentos retornados na consulta original.

Alguns trabalhos que, assim como o presente trabalho, utilizam a expansão da consulta com agrupamento são apresentados em (BHOGAL, MACFARLANE e SMITH, 2007) e (JONES, 1971), e baseiam na suposição de que, sendo criados grupos de documentos semelhantes, os documentos de um mesmo grupo são relevantes para os mesmos pedidos. Logo, ao classificar os termos da consulta em um ou mais grupos, os termos destes grupos seriam usados para expandir a consulta. Como principais críticas em relação a estes trabalhos estão os resultados ruins obtidos em casos em que a coleção de documentos utilizada para a geração de grupos é pequena, ou em que a diferença de vocabulários entre documentos relevantes e não relevantes é insuficiente. Outro problema apontado é em relação a premissa de que um documento só pode pertencer a um único grupo. Outro trabalho proposto por (KANG *et al*., 2010), aplica a técnica de agrupamento nos *n* primeiros documentos retornados pela consulta original, a partir dos quais são gerados termos para cada grupo e exibidos para que o usuário selecione o grupo de termos mais apropriado para a expansão.

Como principal diferença encontrada neste trabalho em comparação com os trabalhos anteriores que utilizam agrupamento está a aplicação da segmentação de

tópicos em toda a coleção de documentos antes da etapa de agrupamento. Desta forma, é considerado que um documento pode tratar de assuntos diferentes, contribuindo com segmentos para diferentes grupos, e consequentemente fornecendo termos para as expansões de consulta a partir de diferentes grupos.

Outra diferença importante é a realização de diversas expansões de consultas, uma para cada grupo. Neste caso, foi levado em consideração que é difícil identificar apenas um *cluster* candidato para a expansão somente considerando os termos da consulta informada pelo usuário, principalmente quando ela é composta por poucos termos.

Finalmente, neste trabalho é dada uma grande importância à etapa de seleção do *corpus* (e o papel do especialista nesta atividade), garantindo uma qualidade nos recursos selecionados e consequentemente na geração dos grupos e na extração dos termos para a expansão.

7.2. Contribuições

Como principais contribuições deste trabalho, podem ser citadas:

- Definição de método de expansão de consultas que pode ser utilizado em sistemas de informação para apoiar as ferramentas de busca web, tornando-as sensíveis ao contexto de domínio da necessidade de informação dos usuários, requerendo para isto um baixo custo de tempo e esforço de um especialista para a definição do contexto;
- Definição de uma sistemática para a extração de informações em conteúdo não estruturado com a aplicação de atividades de segmentação de tópicos e agrupamento que gera termos para a expansão de consultas;
- Desenvolvimento de uma ferramenta em Java para implementar o método proposto;

- Aplicação desta ferramenta em estudos de caso realizados em dois ambientes distintos: um corporativo e um educacional.

7.3. Limitações da abordagem e trabalhos futuros

A abordagem proposta neste trabalho depende fortemente das informações selecionadas para a definição do contexto, pois somente serão adicionados às consultas originais bons termos caso o ponto de partida seja a existência de boas fontes de informação que sejam representativas do domínio do contexto a ser pesquisado. Outra atividade realizada manualmente pelo especialista é a determinação do número de assuntos abordado no conteúdo destas informações selecionadas como representante do contexto. Este valor é utilizado para determinar a quantidade de grupos de segmentos gerados e, consequentemente, a quantidade de expansões de consultas geradas na busca. Enquanto um número menor de assuntos pode levar à geração de grupos de segmentos que contenham assuntos diferentes, um número elevado pode levar à geração de expansões de consultas desnecessárias.

Fontes de informação escritas em diferentes idiomas não são tratadas de forma conjunta para a expansão de consultas, isto é, embora as informações do contexto possam ser complementares, não podem ser combinadas para a geração de uma mesma expansão de consulta. Os segmentos extraídos das informações do contexto são separados por idioma identificado, dando origem a diferentes conjuntos de termos. No momento da execução da busca é identificado o idioma da consulta do usuário e as expansões são executadas com termos do mesmo idioma.

Como trabalhos futuros, podem ser destacados os seguintes:

- A análise dos termos da expressão de busca original como critério para a escolha dos termos a serem utilizados na expansão de consultas. O critério atual só leva

em consideração a frequência dos termos nos documentos representantes do contexto. A fórmula de cálculo de pesos pode levar em consideração, por exemplo, a correferência entre os termos da consulta e os termos dos documentos da coleção;

- Avaliação pelos usuários das expansões de consulta executadas, de acordo com a atribuição de relevância aos resultados obtidos e aplicação de pesos diferenciados aos termos de acordo com as avaliações recebidas;
- A identificação dos assuntos mais relacionados com a expressão de busca original, para que as abas e consequentemente os resultados de busca, possam ser priorizadas de acordo com este relacionamento;
- Aplicação do método em diferentes tipos de contexto, utilizando para a definição do contexto, ao invés de informações de domínio, informações relacionadas ao perfil do usuário ou às atividades profissionais a serem desempenhadas;
- Aplicação do método em contextos com informações de diferentes idiomas;
- Implementação de novas técnicas para a melhoria de resultados na modalidade sugestão de termos. Uma possibilidade seria uma melhor forma de organização dos termos sugeridos, facilitando o entendimento e consequentemente a seleção dos termos pelos usuários. Como exemplo, os termos poderiam ser agrupados de forma hierárquica, conforme sugerido em (JOHO, SANDERSON e BEAULIEU, 2004);
- Identificação automática da quantidade de diferentes assuntos abordados nas informações de contexto;
- O uso de outras formas de representação de conhecimentos, tais como ontologias ou thesauros, como fonte de informação para a captura do contexto.

Neste caso seriam alteradas as atividades de extração de informação necessárias, possivelmente só sendo utilizado o agrupamento de conceitos;

- Utilização de outros algoritmos de segmentação de tópicos e agrupamento, e avaliações em relação a diferença nos resultados de busca e na performance em capturar os textos do contexto.

Referências

AMBRÓSIO, A. P., SILVA, L. O., NETO, V. G. “Automatic Retrieval of Complementary Learning Material for Slide Presentations,” In: *International Conference on Interactive Computer Aided Blended Learning*, Nov 2009.

BAGHI, H., BILETSKIY, Y., LI, H. "A blended approach for search of learning objects". In: *Electrical and Computer Engineering*, p.p. 1805-1810, May. 2008.BEIL, F., ESTER, M., XU X., “Frequent term-based text clustering”. In: *Conference on Knowledge Discovery and Data Mining*, 2002.

BHOGAL, J., MACFARLANE, A., SMITH, P. "A review of ontology based query expansion", *Information Processing and Management: an International Journal*, v. 43 issue 4, Jul. 2007.

CHANANA, V., GINIGE, A., MURUGESAN, S. “Improving information retrieval effectiveness by assigning context to documents”. In: International Symposium on Information and Communication Technologies, 2004.

CHIGNELL, M. H., GWIZDKA, J., BODNER, R. C. “Discriminating meta-search: A framework for evaluation”, *Information Processing and Management: an International Journalt*, v. 35, issue 3, pp-337-362, 1999.

COOPER, W. S. “Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems”, *Journal of American Society of Information Science*, v. 19, issue 1, pp-30-41, 1968

DEY, A. K., ABOWD, G. D. “Towards a better understanding of context and contextawareness”. In: *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, pp. 304-307. Springer-Verlag, 1999.

EXPERIAN HITWISE, Searches statistics. fev, 2010. Available at http://www.hitwise.com/us/press-center/press-releases/google-searches-feb-10/

FALOUTSOS, C., OARD, D.W. “A survey of information retrieval and filtering methods”, *Technical Report CS-TR-3514*, University of Maryland, College Park, MD, Aug. 1995.

FUHR N., “Probabilistic models in information retrieval”, *The Computer Journal - Special issue on information retrieval*. v. 35, issue 3, Jun. 1992.

GOOGLE, *Ajuda do Pesquisa na Web*, abr. 2011. Disponível em: http://www.google.com.br/support/websearch/?hl=br

GREENGRASS, E. *Information Retrieval: A Survey,* 2000

HEARST, M.A. “TextTiling: Segmenting text into multi-paragraph subtopic passages”, *Computational Linguistics*, pp 33-64, 1997.

JANSEN, B. J., SPINK, A., BATEMAN, J., SARACEVIC, T. “Real life information retrieval: A study of user queries on the Web”, *ACM SIGIR Forum*, 32(1), pp 5-17, 1998.

JOHO, H., SANDERSON, M., BEAULIEU, M. "A study of user interaction with a concept-based interactive query expansion support tool". In: *Proceedings of the European Conference on IR Research*, Springer , pp. 42-56, Apr. 2004

JONES, K. S., *Automatic keyword classification for information retrieval*. Archon Books, 1971

JUN-FENG, S., WEI-MING, Z., WEI-DONG, X., GUO-HUI, L., ZHEN-NING, X. "Ontology based semantic information retrieval". In: *IEEE International Conference on e-Technology, e-Commerce and e-Service*, 2005.

KANG, J. W., KANG, H., KO, M., JEON, H. S., NAM, J. "A Term Cluster Query Expansion Model Based on Classification Information in Natural Language Information Retrieval". In: *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, pp. 172 – 176, Oct. 2010.

KENDALL, M.G., STUART, A., *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Griffin, 1973

KOWALSKI, G. J., *Information retrieval systems: theory and implementation*. Kluwer Academic Publishers, 1999

LEE, J.H. "Properties of extended boolean models in information retrieval". In: *Proceedings Annual International ACM SIGIR Conference on Research and Development in Information*. pp. 182-190, 1994

MANNING, C.D., RAGHAVAN, P., SCHÜTZE, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008

MOENS, M. F., *Information Extraction: Algorithms and Prospects in a Retrieval Context.* The Information Retrieval Series. Vol. 21. 1 ed, Softcover, 2006

NEDJA, N., MOURELLE, L. M., KACPRZYK J., FRANÇA, F. M. G., SOUZA, A. F., *Intelligent Text Categorization and Clustering*. Springer ,2008.

NETCRAFT, *Web Server Survey*. jan, 2011. Available at http://news.netcraft.com/archives/2011/01/12/january-2011-web-server-survey-4.html

PEW RESEARCH CENTER, Search Engine Use. 2008. Available at http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx

PRATES, J. C., SIQUEIRA, S. W. M. “Uso de recursos educacionais para melhorar a eficiência das buscas na Internet”, *Simpósio Brasileiro de Informática na Educação*, João Pessoa, Brasil, nov. 2010.

PRATES, J. C., SIQUEIRA, S. W. M. “Using educational resources to improve the efficiency of Web searches for additional learning material”. In: *11th IEEE International Conference on Advanced Learning Technologies*, Athens, Georgia, USA, jul. 2011 – Aceito para publicação.

REYNAR, J. C. “Statistical models for topic segmentation”. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 1999.

RIJSBERGEN, C. J. V., CRESTANI F., LALMAS M., *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*. Kluwer Academic Publishers, 1998.

SALTON, G., *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley, 1989.

SARAWAGI, S., *Information Extraction.* Now Publishers Inc, 2008.

SHEN X., TAN, B., ZHAI, C. “Implicit User Modeling for Personalized Search”. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, New York, NY, USA, 2005

SPINK, A., JANSEN, B.J. “A study of Web search trends”, *Webology*, 1(2), article 4, 2004. Available at: http://www.webology.ir/2004/v1n2/a4.html

STAIR, R. M., REYNOLDS, G. W., *Princípios de sistemas de informação*. Pioneira Thomson, 2005

STAKE, R. E., *The Art Of Case Study Research*. SAGE Publications, 1995

SU. L. T., CHEN, H. L., DONG, X. Y., “Evaluation of Web-based search engines from an end-user's perspective: A pilot study”. In: *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, Pittsburgh, PA., pp-348-361, 1998.

TANG, M.C., SUN, Y. “Evaluation of Web-Based Search Engines Using User-Effort Measures”, *LIBRES Research Electronic Journal*, 13(2), 2003.

VOORHEES, E.M. “On test collections for adaptative information retrieval”, *Information Processing and Management: an International Journal*, 44(6), 2008.

XU, J., CROFT, W. B. “Improving the Effectiveness of Information Retrieval with Local Context Analysis”, *Journal ACM Transactions on Information Systems*, TOIS Homepage archive, v 18, issue 1, Jan. 2000.

YATES, R. B., NETO, B. R., *Modern Information Retrieval*. 1 ed, Addison Wesley, 1999

WANG, C., CHANG, G., WANG X., MA, Y., MA, H. “A User Motivation Model for Web Search Engine”. In: *Internation conference on Hybrid Intelligent Systems*, 2009

WEN, Z., ZHOU, M., AGGARWAL, V. Context-Aware, adaptive information retrieval for investigative tasks . In: *International Conference on Intelligent User Interfaces*, 2007

YAHOO, *O novo Yahoo Search*, abr. 2001. Disponível em: http://br.docs.yahoo.com/searchtour.html

YIN, R. K., *Estudo de Caso - Planejamento e Métodos*. 3ed, Bookman, 2005

ZHANG, G. Q., ZHANG, G. Q., CHENG, S. Q., ZHOU, T. “Evolution of the Internet and its cores”, *New Journal of Physics*, v. 10, issue 12, pp. 123027, 2008.

Anexo I – Questionário para os estudos de caso

Neste anexo serão listadas todas as perguntas do questionário respondido pelos participantes dos dois estudos de caso. As perguntas marcadas com * são de preenchimento obrigatório.

I.1. Identificação

O objetivo deste questionário é avaliar qualitativamente a utilização da ferramenta ContextSearch. As perguntas serão agrupadas em 3 seções, referentes a perguntas gerais sobre o experimento e perguntas específicas para a duas modalidades de pesquisa: expansão automática e sugestão de termos. Os participantes desta pesquisa terão suas identidades mantidas em sigilo. As informações obtidas nesta pesquisa serão divulgados exclusivamente pelo pesquisador e seu orientador na literatura especializada, incluindo possivelmente revistas e eventos científicos da área.

Nome:

I.2. Perguntas sobre a modalidade expansão automática

Para as perguntas e/ou afirmações abaixo, marque a alternativa mais adequada.

1 - Obtive informações mais úteis nos resultados de pesquisa obtidos com a expansão dos termos do contexto (segunda aba) do que nos resultados obtidos com a expansão dos termos extraídos dos assuntos (abas de assuntos) *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

2 - Os resultados das consultas expandidas (consulta do contexto e de assuntos) trazem mais referências a publicações (como notícias, artigos, relatórios, pesquisas, etc.) do que os resultados da consulta original *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

3 - Os resultados exibidos nos painéis de assuntos representam diferentes visões em relação ao tema pesquisado *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

4 - Os resultados de pesquisa exibidos na aba “Consulta do contexto” apontam para documentos mais superficiais do que os exibidos nas abas dos assuntos. Ou seja, falam sobre o assunto pesquisado, porém não se aprofundam no assunto *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

5 - As expansões por assuntos poderiam ser retiradas da ferramenta (3ª, 4ª e 5ª abas), permanecendo somente a expansão com termos do contexto (2ª aba) *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

6 - Ao realizar várias pesquisas com a ferramenta, passei a informar expressões de busca menores *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

7 - Os resultados de uma das abas de assunto eram relacionados com a minha necessidade de informação. Os outros não tinham nenhuma relação *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

8 - Gostaria de ter disponível este tipo de ferramenta para a realização das atividades de pesquisa *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

9 - Com qual número de termos você considera ter obtido melhores resultados? *

( ) De 2 a 4 termos
( ) De 5 a 7 termos
( ) De 8 a 10 termos
( ) O número de termos variou de acordo com a necessidade de informação

I.3. Perguntas sobre a modalidade sugestão de termos

Para as perguntas e/ou afirmações abaixo, marque a alternativa mais adequada.

10 - Os resultados da consulta expandida (2ª aba) trazem mais referências a publicações (como notícias, artigos, relatórios, pesquisas, etc.) do que os resultados da consulta original *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

11- Os termos sugeridos para cada assunto realmente identificam diferentes assuntos dentro do contexto pesquisado *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente

( ) Concordo totalmente

12 - Na maioria das vezes, somente selecionei termos do contexto *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

13 - Na maioria das vezes, misturei termos do contexto e dos diferentes assuntos para montar minha consulta *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

14 - Exibir resultados sem abas de assuntos simplificou muito a análise dos resultados da busca *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

15 - As sugestões de termos por assuntos poderiam ser retiradas da ferramenta, permanecendo somente as sugestões de termos do contexto *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente

( ) Concordo totalmente

16 - Gostaria de ter disponível este tipo de ferramenta para a realização das atividades de pesquisa *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

I.4. Perguntas gerais sobre o experimento

Para as perguntas e/ou afirmações abaixo, marque a alternativa mais adequada.

17 - Os resultados de pesquisa exibidos na consulta original (1ª aba) apontam para documentos mais superficiais do que os exibidos nas demais abas de consultas expandidas. Ou seja, falam sobre o assunto pesquisado, porém não se aprofundam no assunto *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro
( ) Concordo parcialmente
( ) Concordo totalmente

18 - Os resultados da consulta original na sua maioria apontam para páginas "índice", isto é, páginas que não contém informações sobre o assunto pesquisado, mas sim links para outras páginas que tratam do assunto *

( ) Discordo totalmente
( ) Discordo parcialmente
( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

19 - Os resultados das consultas expandidas raramente apontam para páginas "índice" *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

20 - A exibição de resultados em diferentes abas é confusa *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

21 - Identifico algumas situações em que seria melhor utilizar a expansão automática, e outras em que seria melhor utilizar a sugestão de termos *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

22 - Considero a ferramenta útil para as atividades de pesquisa realizadas no trabalho *

( ) Discordo totalmente

( ) Discordo parcialmente

( ) Neutro

( ) Concordo parcialmente

( ) Concordo totalmente

23 - Quais os pontos positivos e negativos na utilização do ContextSearch?

24 - Possui alguma sugestão para melhorar a interface da ferramenta nas duas modalidades disponíveis?

25 - Qual das modalidades apresentou resultados mais úteis? *

( ) Expansão automática

( ) Sugestão de termos

26 - Considerando que com a expansão automática a pesquisa pode ser executada de forma mais rápida do que na sugestão de termos, em que é necessário um esforço adicional de ler todos os termos e selecionar os mais adequados, qual das modalidades apresenta o melhor custo-benefício? *

( ) Expansão automática

( ) Sugestão de termos

27 - Faça uma comparação entre as duas modalidades, apontando vantagens e desvantagens em cada uma.

28 – (Pergunta aplicada exclusivamente as bibliotecárias participantes do estudo de caso do ambiente corporativo) Você reparou alguma diferença relevante na qualidade dos resultados exibidos nos dois contextos (Energia e Meio ambiente)? Explique.

Anexo II – Roteiros dos estudos de caso

Neste anexo serão exibidos os dois roteiros utilizados nos estudos de caso do ambiente corporativo e educacional. Os participantes utilizaram este roteiro como informativo sobre o estudo de caso e para seguir todas as atividades propostas.

Como os roteiro são muito parecidos, variando somente a explicação sobre qual é o contexto a ser utilizado e os assuntos identificados no contexto, será apresentado somente o roteiro utilizado no ambiente corporativo pelas bibliotecárias, que utilizaram no estudo de caso dois contextos (Energia e Meio ambiente).

II.1. Roteiro

Este documento está estruturado em 4 seções: Na primeira seção o problema de pesquisa é explicado de forma resumida; na segunda é explicado o objetivo do estudo de caso; a terceira seção é um guia rápido sobre a aplicação; e a última seção contém todas as atividades do estudo de caso, passo a passo.

Introdução

A busca por informações sobre assuntos relacionados aos setores econômicos apoiados pelo BNDES é uma atividade realizada de forma periódica pelos funcionários do banco, com o apoio de ferramentas de busca disponíveis em sites da Internet, tais como: Google, Bing, Yahoo, etc..

Nestas ferramentas de busca, o usuário informa a necessidade de informação digitando palavras-chave ou expressões. Embora esta abordagem apresente como grande vantagem a facilidade em expressar a necessidade de informação, ela possui alguns problemas, tais como:

- o usuário precisa conhecer as palavras-chave que devem estar presentes no conteúdo dos recursos a serem recuperados;
- palavras-chave podem ter diferentes significados e a desambiguação destes conceitos deve ser feita pelo próprio usuário, através de realização de novas consultas com a inclusão ou alteração de palavras-chave;
- não é considerado na pesquisa aspectos pessoais do usuário e do ambiente em que esta consulta está sendo executada.

Outra questão importante a ser observada é o comportamento do usuário ao realizar buscas na Internet. Se por um lado os usuários acessam poucas páginas de resultado, por outro as estatísticas mostram que as expressões de busca normalmente são compostas de poucos termos. 65% das buscas realizadas na Internet contém de 1 a 3 termos.

Para tornar os resultados de busca mais relevantes considerando este comportamento dos usuários ao elaborar uma expressão, é utilizada uma técnica de recuperação de informações conhecida como expansão de consultas, que consiste em adicionar termos na consulta original feita pelo usuário em uma tentativa de prover uma maior contextualização, recuperando documentos mais úteis.

O Estudo de caso

O estudo de caso proposto consiste em utilizar uma ferramenta que se propõe a "melhorar" as consultas realizadas pelos usuários, adicionando a esta consulta termos extraídos de uma base de conhecimentos definida para o contexto em que se deseja

realizar a pesquisa. Neste caso, a base de conhecimentos utilizada é o portal de informações e conhecimento mantido pelo COPED (http://informacoescoped).

A hipótese é que, considerando que a base de conhecimento utilizada para capturar o contexto possui documentos considerados relevantes, os termos que ocorrem com maior frequência nesta coleção provavelmente ocorrerão também em outros documentos relevantes existentes na Internet. Logo, adicionando estes termos à consulta original, aumentam as chances de obter documentos úteis como resultados de busca.

O estudo de caso é uma etapa essencial do projeto de pesquisa de mestrado, em desenvolvimento na Universidade Federal do Estado do Rio de Janeiro (UNIRIO), realizado por João Carlos Prates dos Santos, e orientado pelo professor Sean Wolfgand Matsui Siqueira. Os participantes desta pesquisa terão suas identidades mantidas em sigilo. As informações obtidas nesta pesquisa, incluindo os textos e os dados obtidos no estudo serão divulgados exclusivamente pelo pesquisador e seu orientador na literatura especializada, incluindo possivelmente revistas e eventos científicos da área.

A Ferramenta

Conforme já citado na seção anterior, a aplicação ContextSearch utiliza informações de contextos para "aumentar" as consultas realizadas pelos usuários, objetivando com isso trazer resultados mais úteis. E como as informações utilizadas para modelar um contexto podem tratar de assuntos diferentes, a ferramenta leva em consideração esta divisão por assuntos. Como exemplo, informações sobre energia elétrica e biocombustíveis, ambos assuntos diferentes dentro do contexto Energia.

As pesquisas no ContextSearch podem ser realizadas em duas modalidades:

- Expansão automática: Em que o usuário digita a expressão de busca (de forma similar ao realizado em buscadores web) e seleciona quantos termos

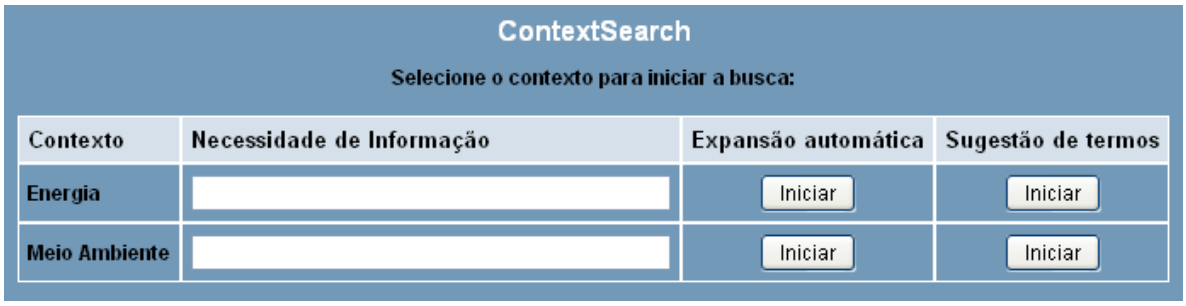
ele quer que o ContextSearch adicione à sua consulta. Os termos são adicionados automaticamente.

- Sugestão de termos: O usuário também digita a expressão de busca, porém escolhe quais termos serão adicionados à sua consulta, dentre uma lista de termos que é sugerida pela ferramenta.

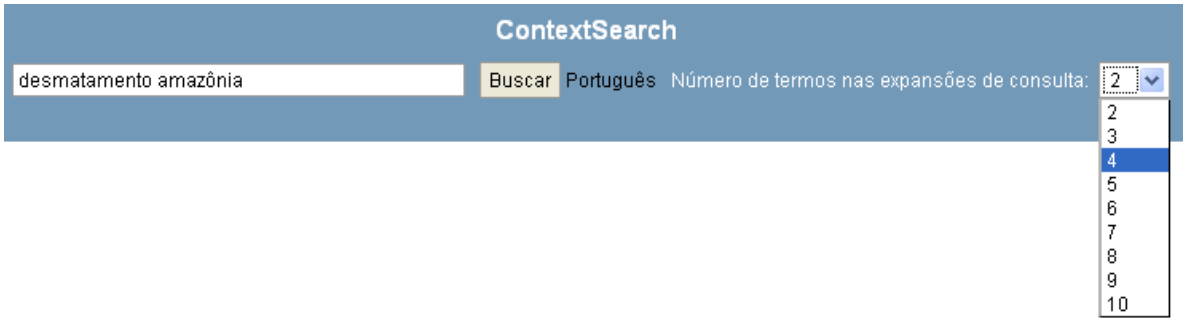
Para entrar na aplicação, acesse em seu navegador o endereço http://<IP>/ContextSearch/. A primeira tela a ser exibida é a de identificação do usuário. Digite a sua *Userid* e clique no botão Ok. A identificação é necessária somente para que os resultados das avaliações possam ser agrupados.

Após a identificação, a aplicação exibirá a lista dos dois contextos disponíveis para a realização do estudo de caso: Energia e Meio Ambiente. Para iniciar a pesquisa é necessário informar a necessidade de informação, que nada mais é do que um resumo em poucas palavras do objetivo da busca que será iniciada. Por exemplo: encontrar pesquisas sobre mudanças climáticas.

Depois de informada a necessidade de informação, basta clicar no botão iniciar correspondente a uma das modalidades disponíveis: a expansão automática ou a sugestão de termos. A tela de pesquisa de cada uma será aberta em uma nova janela.



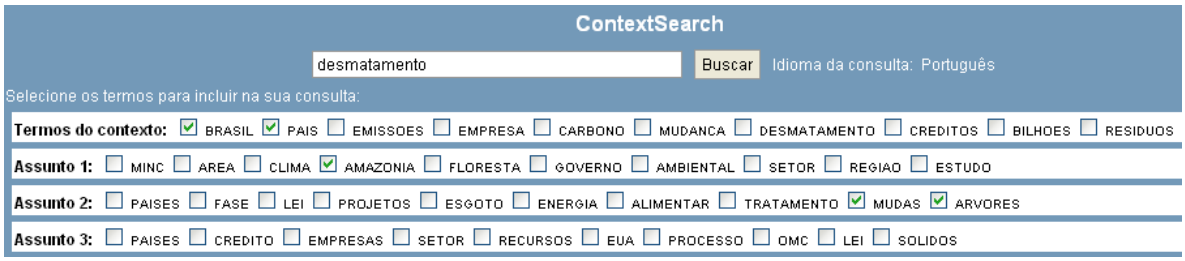
Na expansão automática, a pesquisa é muito parecida com as disponíveis nas ferramentas de busca *online*. O usuário digita a expressão de busca, e seleciona em uma caixa de combinação quantos termos ele gostaria de obter automaticamente da ferramenta.



Após clicar em Buscar, a ferramenta utiliza a expressão digitada para realizar diversas consultas no buscador Google, e exibe todos os resultados em abas, conforme ilustrado na figura abaixo. São exibidos resultados para a consulta original, a consulta expandida com termos comuns a todos os documentos do contexto, e um resultado para cada assunto identificado dentro do contexto. No exemplo abaixo, foram identificados três assuntos, e em cada um destes resultados a consulta original foi expandida com termos comuns de cada assunto.

Já na modalidade sugestão de termos, os termos do contexto e de cada assunto do contexto são exibidos, e cabe ao usuário selecionar os termos que podem ser úteis para compor a expressão de busca. É possível selecionar termos de assuntos diferentes.

Os resultados são exibidos somente em duas abas: uma com a consulta original, e outro com a consulta composta dos termos da consulta original e os termos selecionados.



Uma etapa muito importante no estudo de caso é avaliar a relevância nos resultados obtidos em cada aba, nas duas modalidades. Esta avaliação será realizada na própria ferramenta, logo ao lado do painel que exibe os resultados. Basta selecionar a pontuação para cada um dos 10 primeiros resultados e clicar no botão Enviar. Os critérios a serem utilizados nas avaliações serão melhor detalhados na seção a seguir.

Atividades do estudo de caso

O estudo de caso pode ser dividido em duas principais etapas: utilização da ferramenta e preenchimento de questionário.

A ferramenta deve ser utilizada nos dois contextos configurados (Energia e Meio Ambiente), e nas duas modalidades disponíveis (expansão automática e sugestão de termos). A aplicação deve ser utilizada para pesquisar pelo menos duas necessidades de informações distintas.

Contexto Energia

Para o contexto Energia estão sendo considerados três assuntos de pesquisa: Biocombustíveis, Petróleo e Gás e Energia Elétrica. As necessidades de informação devem ser relacionadas a um destes três assuntos (ou a mais de um deles). Seguem as atividades do estudo de caso:

1) Conforme explicado na seção anterior, acesse a ferramenta e se identifique. Na tela de listagem de contextos, preencha a necessidade de informação relacionada com Energia no campo correspondente ao contexto, e clique no botão Iniciar correspondente a expansão automática.

2) Digite a expressão de busca e selecione o número de termos que deverão ser adicionados automaticamente à sua expressão. Clique no botão Buscar.

3) Consulte todas as abas disponíveis de resultados, analisando superficialmente os resultados obtidos. É possível reexecutar a busca selecionando mais ou menos termos quantas vezes forem necessárias, para testar a ferramenta, até que se deseje avaliar os resultados.

4) Avalie a relevância dos 10 primeiros resultados exibidos em todos os painéis de resultado, utilizando uma escala de 5 pontos. O valor "0" indica que o resultado não possui relevância e o valor "4" indica alta relevância. O critério utilizado para julgar a relevância é se o resultado traz informação considerada útil para o participante no atendimento da necessidade de informação definida. Também deve ser levada em consideração no julgamento se a informação desejada pode ser encontrada diretamente no resultado obtido ou se foram necessárias navegações no resultado para obter a informação desejada.

5) Verifique se em todas as abas de resultados a mensagem "Resultado armazenado!" está sendo exibida.

6) Após o término das avaliações, feche a janela da pesquisa – modalidade expansão automática. O navegador estará exibindo novamente a tela de listagem de contextos, com sua necessidade de informação já preenchida. Caso esta janela tenha sido acidentalmente fechada, basta repetir as instruções iniciais do passo 1, digitando novamente a mesma necessidade de informação.

7) Clique no botão iniciar referente a modalidade sugestão de termos.

8) Digite a expressão de busca e verifique na lista de termos listados abaixo da caixa de listagem, se algum deles poderia ser adicionado à consulta original para que

sejam obtidos melhores resultados. Marque quantos termos julgar necessários, podendo ser combinados termos de diferentes listas (termos do contexto e dos assuntos). Clique no botão Buscar. Obs.: é necessário selecionar pelo menos um termo.

9) Consulte as duas abas de resultados, da consulta original e da consulta expandida.

10) Avalie a relevância dos 10 primeiros resultados exibidos nos dois painéis de resultado, utilizando uma escala de 5 pontos. O valor "0" indica que o resultado não possui relevância e o valor "4" indica alta relevância. O critério utilizado para julgar a relevância é se o resultado traz informação considerada útil para o participante no atendimento da necessidade de informação definida. Também deve ser levada em consideração no julgamento se a informação desejada pode ser encontrada diretamente no resultado obtido ou se foram necessárias navegações no resultado para obter a informação desejada.

11) Verifique se nas duas abas de resultados a mensagem "Resultado armazenado!" está sendo exibida.

12) Após o término das avaliações, feche a janela da pesquisa – modalidade sugestão de termos.

Refaça os passos de 1 a 12 para uma segunda necessidade de informação. Para incluir uma nova necessidade de informação, basta alterar o campo correspondente na página de listagem de contextos.

Contexto Meio Ambiente

Para o contexto Meio Ambiente estão sendo considerados três assuntos de pesquisa: Desenvolvimento sustentável, Mudanças climáticas e Política ambiental. As

necessidades de informação devem ser relacionadas a um destes três assuntos (ou a mais de um deles). Seguem as atividades do estudo de caso:

1) Conforme explicado na seção anterior, acesse a ferramenta e se identifique. Na tela de listagem de contextos, preencha a necessidade de informação relacionada com Meio Ambiente no campo correspondente ao contexto, e clique no botão Iniciar correspondente a expansão automática.

2) Digite a expressão de busca e selecione o número de termos que deverão ser adicionados automaticamente à sua expressão. Clique no botão Buscar.

3) Consulte todas as abas disponíveis de resultados, analisando superficialmente os resultados obtidos. É possível reexecutar a busca selecionando mais ou menos termos quantas vezes forem necessárias, para testar a ferramenta, até que se deseje avaliar os resultados.

4) Avalie a relevância dos 10 primeiros resultados exibidos em todos os painéis de resultado, utilizando uma escala de 5 pontos. O valor "0" indica que o resultado não possui relevância e o valor "4" indica alta relevância. O critério utilizado para julgar a relevância é se o resultado traz informação considerada útil para o participante no atendimento da necessidade de informação definida. Também deve ser levada em consideração no julgamento se a informação desejada pode ser encontrada diretamente no resultado obtido ou se foram necessárias navegações no resultado para obter a informação desejada.

5) Verifique se em todas as abas de resultados a mensagem "Resultado armazenado!" está sendo exibida.

6) Após o término das avaliações, feche a janela da pesquisa – modalidade expansão automática. O navegador estará exibindo novamente a tela de listagem de contextos, com sua necessidade de informação já preenchida. Caso esta janela tenha

sido acidentalmente fechada, basta repetir as instruções iniciais do passo 1, digitando novamente a mesma necessidade de informação.

7) Clique no botão iniciar referente a modalidade sugestão de termos.

8) Digite a expressão de busca e verifique na lista de termos listados abaixo da caixa de listagem, se algum deles poderia ser adicionado a consulta original para que sejam obtidos melhores resultados. Marque quantos termos julgar necessários, podendo ser combinados termos de diferentes listas (termos do contexto e dos assuntos). Clique no botão Buscar. Obs.: é necessário selecionar pelo menos um termo.

9) Consulte as duas abas de resultados, da consulta original e da consulta expandida.

10) Avalie a relevância dos 10 primeiros resultados exibidos nos dois painéis de resultado, utilizando uma escala de 5 pontos. O valor "0" indica que o resultado não possui relevância e o valor "4" indica alta relevância. O critério utilizado para julgar a relevância é se o resultado traz informação considerada útil para o participante no atendimento da necessidade de informação definida. Também deve ser levada em consideração no julgamento se a informação desejada pode ser encontrada diretamente no resultado obtido ou se foram necessárias navegações no resultado para obter a informação desejada.

11) Verifique se nas duas abas de resultados a mensagem "Resultado armazenado!" está sendo exibida.

12) Após o término das avaliações, feche a janela da pesquisa – modalidade sugestão de termos.

Refaça os passos de 1 a 12 para uma segunda necessidade de informação. Para incluir uma nova necessidade de informação, basta alterar o campo correspondente na página de listagem de contextos.