UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

LUCIANA OLIVEIRA E SILVA

BOOKISH

Uma ferramenta para contextualização de documentos utilizando mineração de textos e expansão de consulta

Goiânia
2009

LUCIANA OLIVEIRA E SILVA

BOOKISH

Uma ferramenta para contextualização de documentos utilizando mineração de textos e expansão de consulta

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Computação.

Área de concentração: Ciência da Computação.

Orientadora: Profa. Dra. Ana Paula Laboissìere Ambrósio

Goiânia
2009

LUCIANA OLIVEIRA E SILVA

BOOKISH

Uma ferramenta para contextualização de documentos utilizando mineração de textos e expansão de consulta

Dissertação defendida no Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do título de Mestre em Computação, aprovada em 14 de Agosto de 2009, pela Banca Examinadora constituída pelos professores:

Profa. Dra. Ana Paula Laboissière Ambrósio
Instituto de Informática – UFG
Presidente da Banca

Prof. Dr. Fábio Moreira Costa
Instituto de Informática – UFG

Prof. Dr. Sean Wolfgand Matsui Siqueira
Departamento de Informática Aplicada – UNIRIO

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador(a).

Luciana Oliveira e Silva

Graduada em Ciência da Computação pelo Instituto de Informática (INF) da Universidade Federal de Goiás (UFG). Especialista em Tecnologia da Informação pela Universidade Salgado de Oliveira (UNIVERSO). Atualmente atua como docente na Universidade Estadual de Goiás (UEG) e desenvolve soluções de integração e sistemas WEB para a Celg Distribuição S.A.

Dedico este trabalho à DEUS, que me manteve firme no propósito e me ajudou a enfrentar toda sorte de problemas durante os últimos meses de execução deste trabalho.

Agradecimentos

Meu mais sincero e sublime agracecimento à professora Dra. Ana Paula L. Ambrósio, por sua sabedoria, serenidade, competência, seriedade e doçura. Por ter acreditado, por ter investido seu tempo e sobretudo sua paciência durante o desenvolvimento deste trabalho.

Aos professores Dr.Fábio Moreira Costa e Dr. Fábio Nogueira de Lucena pelas orientações e constante estímulo.

Ao Instituto de Informática da UFG, seus professores e funcionários, sempre dispostos a cooperar e cujo contato foi indispensável para a minha formação acadêmica e para a vida.

A Celg Distribuição S.A., pela liberação de horário para estudos. Aos meus chefes, João Guilherme Nalon, Fernando Quinta e Roberto Craveiro Curado, que entenderam a minha difícil caminhada, fico eternamente grata pelo apoio.

Aos meus colegas de mestrado que, assim como eu, exerceram um esforço enorme para conseguir conciliar trabalho, estudo e família. Fico grata pela amizade, companheirismo e solidariedade nos momentos difíceis e nos momentos de alegria e descontração. Em especial aos amigos Ms.Rogério Arantes Gaioso, Ms.Walid Jradi, Ms.José Mauro da Silva, Ms.Márcio Balian e Luciana Nishi, vocês me ajudaram muito nessa caminhada.

Meus agradecimentos especiais ao meu grande amigo Eng. de Computação Chayner Cordeiro Barros, pelo apoio na materialização e enriquecimento de minhas idéias, ao qual, tenho especial estima e gratidão.

A DEUS, por tudo.

"Sentir-se saturado pela existência de tantas coisas na rede é como se sentir saturado pela massa de magníficas paisagens no campo. Você não é obrigado a visitá-las, mas é agradável saber que elas estão lá. Especialmente pela liberdade e variedade."

Tim Berners-Lee,
O Relógio do Longo Agora, de Stewart Brand.

Resumo

e Silva, Luciana Oliveira. **BOOKISH**. Goiânia, 2009. 92p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

O contínuo desenvolvimento da tecnologia e sua disseminação em todas as áreas têm provocado mudanças significativas na sociedade e na educação. É preciso buscar a formação necessária às novas competências do mundo globalizado e considerar que o momento proporciona uma oportunidade de aproximar novas tecnologias ao processo educativo como possibilidade de melhorar os sistemas de ensino tradicionais. O foco deve ser a busca da informação significativa e da pesquisa, o desenvolvimento de projetos e não predominantemente a simples transmissão de conteúdo. Ao ministrar conteúdo de determinada disciplina, o professor muitas vezes disponibiliza fontes complementares que ajudam na compreensão do tema e auxiliam os alunos na execução de atividades. Já o aluno, dentro de uma abordagem pró-ativa, deve ser capaz de interpretar e identificar outras fontes que melhor complementem e expandam assunto. No entanto, um dos desafios atuais é a sobrecarga de informação - são muitos documentos à disposição e poucas formas eficientes de tratá-los. O sistema BOOKISH, proposto neste trabalho, busca auxiliar os alunos na atividade de identificar e filtrar informações relevantes e dentro do contexto que está sendo estudado em sala de aula. A partir de apresentações em forma de slides disponibilizados pelos professores, a ferramenta identifica documentos eletrônicos contextualmente semelhantes e os disponibiliza para os alunos. É objetivo minimizar o tempo gasto nas atividades de busca por material complementar relevante e direcionar o aluno para o conteúdo do qual necessita. A ferramenta apresentada neste trabalho utiliza técnicas de mineração de textos e expansão automática de consultas com esta finalidade.

Palavras–chave

Mineração de Textos, Expansão de Consulta, Tecnologias de Comunicação e Informação.

Abstract

e Silva, Luciana Oliveira. **BOOKISH**. Goiânia, 2009. 92p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

The continuous development of technology and its dissemination in all domains have caused significant changes in society and in education. The new global society demands new skills and provides an opportunity to introduce new technologies into the educational process, improving traditional education systems. The focus should be on the search for information, significant research, and on the development of projects, rather than on the pure transmission of content. When delivering a lecture about a given content, teachers often provide additional sources that will help students deepen their understanding of the subject and carry out activities. Furthermore, it is desirable to have proactive students, capable of interpreting and identifying other sources of information that complement and expand the subject being studied. However, one of the challenges today is information overload - there are many documents available and few effective ways to treat them. Every day, large numbers of documents are stored and made available. These documents contain a lot of relevant information. However finding that knowledge is a difficult task. The BOOKISH system, proposed in this work, assists students in their search activity. Analyzing PowerPoint slide presentations, the tool identifies contextually similar electronic documents, minimizing the time spent in searching for additional relevant material and directing the student to the content he needs. The tool presented in this document uses text mining techniques and automatic query expansion.

Keywords

Text Mining, Query Expansion, Information and Communication Technologies.

Sumário

1 Introdução **12**
- 1.1 Contextualização Automática de Documentos 15
- 1.2 Mineração de Textos 15
- 1.3 Expansão Automática de Consulta 16
- 1.4 O Problema de Pesquisa 16
- 1.5 Objetivos 17
  - 1.5.1 Objetivos Gerais 17
  - 1.5.2 Objetivo Específico 17
- 1.6 Estrutura da Dissertação 17

2 Mineração de Textos **18**
- 2.1 Pré-processamento 19
  - 2.1.1 Análise dos Dados 20
  - 2.1.2 *Case Folding* 20
  - 2.1.3 Eliminação de *Stopwords* 20
  - 2.1.4 *Stemming* 20
  - 2.1.5 Tesauro 20
  - 2.1.6 Representação dos Textos 21
    - 2.1.6.1 Atribuição de Pesos 22
    - 2.1.6.2 *Term-Frequency – Inverse Document Frequency* (TF-IDF) 23
    - 2.1.6.3 Dimensão da Representação 24
    - 2.1.6.4 Lei de Zipf 25
    - 2.1.6.5 Significância das Palavras de Luhn 26
    - 2.1.6.6 Decomposição de Valores Singulares 27
- 2.2 Indexação 30
  - 2.2.1 Indexação *Full Text* 31
  - 2.2.2 Indexação Temática 31
  - 2.2.3 Indexação em Listas Invertidas 31
  - 2.2.4 Indexação Semântica Latente 31
- 2.3 Mineração 35
  - 2.3.1 Sumarização 35
  - 2.3.2 Categorização 35
  - 2.3.3 Clusterização de Documentos 35
- 2.4 Análise 35
  - 2.4.1 Métricas de Avaliação de Resultados 36
- 2.5 Considerações Finais 36

3 Expansão de Consulta **38**
3.1 Abordagens para Expansão de Consulta 39
3.2 Expansão Manual de Consulta 40
3.2.1 Blocos de Construção - *Building Block* 40
3.2.2 Crescimento de Pérolas - *Pearl Growing* 41
3.2.3 Frações Sucessivas - *Successive Fractions* 41
3.3 Expansão Automática de Consulta 42
3.3.1 Expansão Automática de Consulta Baseada em Documentos 42
3.3.2 Expansão Automática de Consulta Baseada em Ontologia 43
3.3.3 Expansão Automática de Consulta baseada em Tesauro 43
3.4 Considerações Finais 44

4 Trabalhos Relacionados **45**
4.1 Sistemas de Gestão da Aprendizagem 45
4.1.1 Moodle 46
4.1.2 Amadeus 47
4.1.3 TelEduc 48
4.2 Ferramentas de Apresentação de Slides 49
4.2.1 Classroom Presenter 50
4.2.2 Ubiquitous Presenter 51
4.2.3 iPH 51
4.2.4 Virtual Multiboard 52
4.2.5 Tablet Mylar Slides 53
4.3 Considerações Finais 53

5 A Ferramenta BOOKISH **54**
5.1 Tecnologias Utilizadas na Implementação 55
5.2 Funcionamento da Ferramenta 56
5.2.1 Processamento da Apresentação 58
5.2.2 Processamento de Documentos 61
5.2.3 Processamento da Consulta 62
5.2.3.1 Consulta à Internet 62
5.2.3.2 Consulta ao Repositório 65
5.3 Considerações Finais 66

6 Resultados **68**
6.1 Cenário dos Testes 68
6.2 Estratégia de Teste 69
6.3 Ambiente de Teste 70
6.4 Estatísticas 70
6.4.1 Ganhos Obtidos com a Ferramenta 75

7 Conclusão **78**
7.1 Trabalhos Publicados 80
7.2 Contribuições 80
7.3 Trabalhos Futuros 81

Referências Bibliográficas **82**

CAPÍTULO 1

Introdução

Nas últimas décadas o mundo passou por um processo de grandes e constantes mudanças: a abertura dos mercados mundiais, o crescente desenvolvimento e popularização de tecnologias, a revolução nos sistemas de comunicações, entre outros. Esse acelerado avanço tecnológico e a facilidade de acesso às novas tecnologias têm provocado mudanças significativas no modo de vida da população mundial. Essas transformações refletem-se nos mais variados setores, inclusive na educação.

Para [15], a escola contemporânea sofre com o desenvolvimento acelerado que ocorre à sua volta, onde as informações são atualizadas em frações de segundo, ocasionando, de certa forma, o comprometimento das ações voltadas para o aprimoramento do ensino, fazendo com que a sala de aula se torne um ambiente de pouca relevância para a consolidação do conhecimento, tornando a vivência social o requisito primordial para a busca de aprendizado.

As instituições de ensino e também os professores são forçados a se adaptar a essa nova realidade. A visão educacional tradicional, fundamentada no conceito-chave de que o professor deve atuar como fonte transmissora de informação aos alunos, é considerada ultrapassada. Alguns autores [6, 68, 102] defendem que é preciso substituir essa visão por um enfoque alicerçado em processos de construção, gerenciamento e disseminação do conhecimento, com ênfase no "aprender a aprender" e na educação ao longo da vida.

Segundo [6], educar é fazer emergir vivências do processo de conhecimento. Para o autor, o produto da educação deveria ser chamado de experiências de aprendizagem e não simplesmente aquisição de conhecimentos prontos e disponíveis para simples transmissão. Ainda segundo [6], a educação só consegue bons resultados quando se preocupa com a geração de experiências de aprendizagem, criatividade para construir novos conhecimentos e habilidades para saber acessar fontes de informação sobre os mais variados assuntos.

É preciso inserir o aluno nesta nova "sociedade do conhecimento", onde o conhecimento é caracterizado pela capacidade de agir. Cabe a este novo aluno a busca por conhecimento e atualização. É preciso que ele dê sentido à informação que possui,

transformando-a em conhecimento, novo motor da economia. O papel dos educadores nessa nova sociedade passa a ser o de formar pessoas com habilidade para identificar e interpretar informações relevantes, e convertê-las em conhecimento para a vida.

No mundo contemporâneo é importante para o aluno saber interpretar uma informação; ele deve ser capaz de fazer uma análise crítica sobre o material que está lendo, classificar sua relevância e, se for o caso, extrair do documento as informações úteis para auxiliar na construção do seu conhecimento.

Para [16], o material didático colabora para a transformação social na medida em que favorece a elaboração constante do conhecimento como resultado de experiências interativas, propiciando o crescimento de um cidadão criativo, crítico e produtivo, pronto para enfrentar a vida com mais segurança.

Com a ampla utilização de tecnologias pelos alunos, sobretudo da Internet, a obtenção da informação depende cada vez menos do professor. Os alunos conseguem obter muitos dados, imagens e documentos de forma rápida e atraente. Infelizmente a quantidade não está diretamente relacionada à qualidade. Encontrar um documento relevante pode se tornar uma atividade trabalhosa e demorada, desestimulante para a maioria dos alunos, que acabam não recorrendo a material complementar para enriquecer seu aprendizado.

Assim, freqüentemente, a recuperação e a seleção deste material complementar fica a cargo do professor, que disponibiliza textos, apostilas, artigos e tutoriais, através de mídia, rede local, repositório ou ambiente de aprendizagem, com o objetivo de sedimentar ou complementar a aprendizagem.

A disponibilização de material complementar exige que o professor esteja constantemente selecionando material e atualizando aqueles já disponíveis. Isto implica em tempo e trabalho, gerando o inconveniente de que a quantidade de documentos disponibilizados para os alunos acaba sendo pequena.

A recuperação de material relevante, direcionado aos objetivos de aprendizagem, de forma simples e rápida é desejável. Neste sentido, a utilização de ferramentas que identificam automaticamente material relevante às necessidades dos alunos é extremamente importante para minimizar o tempo gasto na localização de material didático adequado e, sobretudo, maximizar as chances de sucesso no processo de assimilação de conteúdo (interpretação da informação).

Já existem várias ferramentas que auxiliam a busca de informações na Internet e nos sistemas de arquivos. A maior parte destas ferramentas permitem a localização eficiente de documentos textuais a partir de palavras chave ou diretórios de assuntos. Os problemas mais graves associados à utilização deste tipo de ferramenta estão na dificuldade que estas possuem em “entender” a semântica contida nos textos e em contextualizar as buscas realizadas pelos usuários, limitando a recuperação à documentos

que contenham especificamente as palavras informadas.

Algumas propostas apresentam soluções interessantes para este problema, como por exemplo a utilização de tesauros e ontologias usados para contextualizar o domínio no qual a consulta ocorre. Com isso, a pesquisa pode ser ampliada, recuperando documentos que contenham os termos originais da consulta e também termos relacionados, identificados através de estruturas auxiliares.

No caso específico da recuperação de material didático complementar, para a identificação das necessidades dos alunos é preciso, inicialmente, conhecer os seus objetivos de aprendizagem ou seja, o domínio de conhecimento ou contexto no qual a aprendizagem vai ocorrer. Este contexto é freqüentemente formalizado pelos professores através de apresentações que organizam e resumem o conteúdo a ser ensinado, sendo muitas vezes disponibilizados para os alunos. Estas apresentações acabam servindo como referência para os alunos na hora de estudar.

Nesta dissertação é apresentada a ferramenta BOOKISH, que tem como objetivo analisar e identificar automaticamente o contexto de apresentações e associar o seu conteúdo a documentos eletrônicos e à páginas da Internet, permitindo uma recuperação de material didático contextualmente semelhante. Para isso, foram utilizadas técnicas de mineração de textos e expansão automática de consulta.

Apesar das técnicas utilizadas serem frequentemente usadas na recuperação de documentos, durante o desenvolvimento deste estudo, não foram encontradas ferramentas voltadas especificamente para a recuperação de material didático complementar. O uso dos slides de apresentação de conteúdo usado por professores apresenta uma abordagem invodora, na mesma linha de outras ferramentas como o Classroom Presenter[1], que visam enriquecer a comunicação e a interação aluno-professor, permitindo a construção do conhecimento de forma dinâmica.

Para validar o sistema proposto, um protótipo foi desenvolvido. Foram realizadas pesquisas de avaliação com 2 turmas de disciplinas diferentes. Como fonte de material didático, o ambiente de testes foi montado com uma coleção com 60 documentos de diversos domínios de conhecimento e acesso à Internet. Foram disponibilizadas apresentações com conteúdo específico para cada uma das turmas, sendo estas as fontes primárias para identificação das necessidades dos alunos. Os resultados obtidos foram considerados satisfatórios, indicando que 83% dos entrevistados utilizariam a ferramenta em sala de aula ou em casa para auxiliar na recuperação de material didático complementar.

[1]CLASSROOM PRESENTER: http://classroompresenter.cs.washington.edu/

1.1 Contextualização Automática de Documentos

Uma vez que identificar o contexto da busca é essencial para melhorar o processo de recuperação, mecanismos automáticos de contextualização podem ajudar a resolver diversos problemas relacionados à localização eficiente de documentos.

Para os alunos, o principal benefício da contextualização está relacionada à transformação de grandes volumes de dados em conhecimento útil. Para os professores, a contextualização facilita a condução das atividades em sala de aula, já que os alunos serão direcionados a documentos relevantes quando realizarem suas buscas.

Um dos principais problemas da contextualização automática de documentos não-estruturados está relacionado à identificação correta dos contextos, ou seja, como classificar os termos presentes em um documento de forma que eles permitam uma classificação automática.

Neste trabalho, a definição dos contextos é feita através da identificação dos conceitos-chave existentes nas apresentações. Isto é feito utilizando técnicas da mineração de textos. Os conceitos-chave identificados são então usados para recuperar documentos relevantes e contextualmente relacionados existentes em um repositório específico ou na Internet.

Como a ferramenta executa a análise no momento em que a apresentação é aberta, o contexto corresponde à atualidade e o material apresentado é relevante para esse contexto, podendo retornar documentos recentemente adicionados ao repositório sem a necessidade do professor fazer esta disponibilização manualmente.

1.2 Mineração de Textos

Mineração de textos é o nome dado às técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam os textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, como o casamento de palavras.

A principal motivação para aplicação da mineração de textos neste trabalho é que ela permite extrair conhecimento a partir de dados textuais não-estruturados. Uma vez que os documentos submetidos ao sistema proposto estão neste formato, este tipo de técnica pode ser utilizada.

1.3 Expansão Automática de Consulta

Expansão automática de consulta é o nome dado ao processo de complementar uma consulta feita pelo usuário com termos adicionais. Os termos que são adicionados à consulta original podem ser selecionados de forma manual, interativa ou automática, com o apoio de tesauros e ontologias, por exemplo.

A aplicação dessa técnica permite a contextualização da consulta do usuário dentro do contexto do conteúdo de aprendizagem sendo abordado. Isto pode garantir melhores resultados na recuperação.

1.4 O Problema de Pesquisa

Sabemos que a Internet se tornou uma vasta biblioteca de consulta para toda a população e também uma fonte primária de pesquisa dos alunos. O grande volume de informações nem sempre garante o acesso a um conjunto de informações corretas ou específicas para a pesquisa que está sendo realizada. Porém, essas informações não podem simplesmente ser descartadas. Elas precisam ser selecionadas, avaliadas, compiladas e processadas para que se transformem em conhecimento válido, relevante e necessário para o crescimento do homem como ser humano em um mundo sustentável [19].

Por outro lado, informação não é sinônimo de conhecimento. O conhecimento implica em uma gestão criativa dessa informação e subentende a percepção das formas de acesso, seleção, articulação e organização das informações, a apreensão e concepção de contextos globais na compreensão do seu caráter multidimensional e das relações entre o todo e cada uma das partes [11].

A construção do conhecimento pode ser desenvolvida pelo fato do aluno ter que buscar novos conteúdos e estratégias para incrementar o nível de conhecimento que já dispõe sobre o assunto que está sendo tratado, e essa tarefa pode ser auxiliada via computador.

Neste sentido, o trabalho apresentado tem como proposta **analisar e aplicar técnicas de recuperação de informações que levam em consideração semântica e contexto. O estudo será apoiado com a implementação de uma ferramenta com o objetivo de recuperar, de forma automática, material didático complementar de alta relevância na formação dos alunos**.

1.5 Objetivos

1.5.1 Objetivos Gerais

Este trabalho visa a realização de um um estudo de técnicas, apoiado pelo desenvolvimento de uma ferramenta facilitadora no processo de pesquisa e de auto-aprendizagem, através de um aplicativo que forneça ao aluno o acesso a um conjunto de materiais didáticos complementares automaticamente contextualizados e dentro do assunto que está sendo apresentado em sala de aula através de apresentações em slides. Espera-se que a ferramenta seja capaz de auxiliar o aluno em suas atividades de estudo e/ou complementação de conteúdo.

Por outro lado, a ferramenta visa ser um instrumento também para o professor, no sentido de facilitar a disponibilização de conteúdo. Cabe ao professor selecionar material de qualidade que é armazenado no repositório, sem a necessidade de identificar explicitamente a qual conteúdo didático ele deve ser associado. Esta identificação é realizada automaticamente pelo sistema.

1.5.2 Objetivo Específico

O objetivo específico deste trabalho é verificar a exeqüibilidade de algoritmos de mineração de textos e de expansão automática de consulta como ferramenta de análise e identificação de contextos de documentos.

Para análise da viabilidade de aplicação das técnicas mencionadas foi implementada uma ferramenta que atua como interface para o aluno na recuperação de material didático.

1.6 Estrutura da Dissertação

Este documento encontra-se estruturado em sete capítulos. O Capítulo 2 fornece uma breve revisão bibliográfica dos conceitos básicos de mineração de textos, seguida por uma breve revisão bibliográfica do conceito de expansão de consulta no Capítulo 3. O Capítulo 4 apresenta ferramentas relacionadas, direta ou indiretamente com este trabalho e que serviram como fonte de inspiração para a ferramenta BOOKISH, apresentada no Capítulo 5 junto com sua implementação. O Capítulo 6 discute os resultados das avaliações da ferramenta implementada. Por fim, o Capítulo 7 apresenta as conclusões do trabalho e projetos futuros que podem ser realizados a partir do estado atual da pesquisa.

CAPÍTULO 2

Mineração de Textos

Com o crescimento do número de usuários de computadores e, principalmente, do volume de usuários da Internet, um volume cada vez maior de documentos é criado e colocado à disposição das pessoas. A grande quantidade de documentos, por um lado, é extremamente interessante, pois torna possível inferir muito conhecimento. Por outro lado, encontrar informação que leve ao conhecimento não é uma tarefa fácil [23]. Um número elevado de documentos disponíveis e nenhuma forma eficiente de tratá-los gera o que alguns autores chamam de "sobrecarga de informação" [23, 78].

A análise de grandes quantidades de textos pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas [78]. Por este motivo é imprescindível o desenvolvimento de ferramentas que auxiliem o ser humano nesta tarefa.

A mineração de textos, também conhecida como mineração de dados textuais ou descoberta de conhecimento em textos (*knowledge discovery from text* – KDT), surgiu com a finalidade de tratar dados e informações textuais, ou não-estruturadas, considerando o alto nível de complexidade envolvido nesse tipo de representação de informação [38].

O *Text Mining Research Group* [1] apresenta a seguinte definição: "Mineração de textos é a procura por padrões em um texto em linguagem natural e pode ser definida como o processo de análise do texto para extrair informação dele para um propósito particular".

Desta forma, podemos definir o termo mineração de textos como sendo o processo utilizado para descobrir conhecimento útil em uma coleção de documentos textuais através da identificação e exploração de padrões interessantes nesses documentos.

É interessante ressaltar que a principal diferença entre mineração de textos e mineração de dados é que a primeira utiliza técnicas avançadas para explorar uma grande coleção de dados textuais, provavelmente desestruturados, enquanto a segunda enfatiza a descoberta dentro de coleções estruturadas em bancos de dados [50].

Neste capítulo são apresentados os conceitos básicos relacionados ao processo de mineração de textos e às etapas que compõem esse processo, conforme demonstrado na Figura 2.1.

[1] Grupo de pesquisa em KDT da University of Waikato (http://textmining.zcu.cz/)

Figura 2.1: *Etapas da mineração de textos*

2.1 Pré-processamento

Em geral, documentos textuais possuem características como:

- Serem constituídos na forma de arquivos desestruturados, ou seja, sem qualquer construção formal, como ocorre com arquivos de banco de dados. Desta forma, é necessário modificá-los para uma forma estruturada para que seja possível a aplicação de técnicas de mineração.
- Possuírem centenas ou milhares de palavras em seu conteúdo, muitas repetidas, expressando o mesmo significado ou com significados irrelevantes. Todas essas palavras devem ser tratadas de forma a minimizar as repetições e eliminação de "ruídos" dos documentos.

A preparação dos textos é a primeira etapa do processo de descoberta de conhecimento e envolve a seleção do conjunto daqueles que constituem os documentos de interesse, bem como a tentativa de seleção das informações que melhor expressam o conteúdo dos textos ou termos relevantes.

Nessa etapa, é possível determinar a similaridade em função da morfologia[2] ou do significado dos termos, além de promover a redução da dimensionalidade da representação dos documentos (Seção 2.1.6.3), feita através de seus termos relevantes.

[2]Morfologia é o estudo da estrutura, da formação e da classificação das palavras. A morfologia estuda as palavras de forma isolada e não dentro da sua participação em uma frase ou período. A morfologia está agrupada em dez classes, denominadas classes de palavras ou classes gramaticais: substantivo, artigo, adjetivo, numeral, pronome, verbo, advérbio, preposição, conjunção e interjeição [99].

2.1.1 Análise dos Dados

O objetivo da etapa de análise dos dados é facilitar a identificação de similaridades de significado entre as palavras. Nessa etapa podem ser realizadas diferentes atividades, de acordo com o objetivo de cada sistema e não necessariamente englobam todas os itens descritos a seguir.

2.1.2 *Case Folding*

É o processo de converter todos os caracteres de um documento em um mesmo padrão de letras, deixando todas maiúsculas ou minúsculas. Esse procedimento é útil para acelerar comparações no processo de indexação.

2.1.3 Eliminação de *Stopwords*

É o processo de eliminação do que não deve ser considerado nos próximos passos do processamento de textos. Nesta atividade, os sistemas tentam remover todos os termos que não constituem conhecimento. Para isso, uma lista com palavras que devem ser descartadas é formada.

Formalmente, define-se *stopword* como palavras sem conteúdo semanticamente significante no contexto em que ela está inserida. Portanto, são palavras consideradas irrelevantes na análise de textos [111] e [42].

As *stopwords* são normalmente palavras auxiliares ou conectivas (artigos, proposições, pronomes, palavras curtas e comuns) e que não fornecem nenhuma informação discriminativa na expressão do conteúdo dos textos.

2.1.4 *Stemming*

O processo de *stemming* é realizado sobre cada palavra separadamente, na tentativa de reduzi-la a sua palavra raiz (*stem*). Esse processo tem a vantagem de eliminar sufixos, indicando formas verbais ou plurais. Como desvantagem, esse processo utiliza lingüística e é dependente de idioma.

Infelizmente, talvez pela complexidade do idioma, poucos são os algoritmos de radicalização disponíveis para a língua portuguesa: adaptações para os algoritmos de Porter, Orengo e Pegastemming são alguns exemplos [104].

2.1.5 Tesauro

O termo em inglês *thesaurus* significava originalmente “estoque de tesouros” e começou a ser utilizado como sinônimo para dicionário. Um tesauro não é apenas uma

lista de palavras-chave (ou termos) e seus sinônimos: é também a realização de uma hierarquia global de termos relacionados.

Tesauros têm como função "evitar ou diminuir a flexibilidade da linguagem e descrever um conceito de maneira unívoca em um sistema de informação", orientando assim sobre o termo mais adequado para representar um conceito [29].

Segundo [95], a construção de tesauros pode ser realizada de três maneiras:

- manual - realizada por especialistas que possuem um alto nível de conhecimento sobre um determinado assunto, sendo responsáveis por definir um domínio e criar o tesauro referente a ele;
- automática - baseada em aplicativos de software que utilizam coleções de documentos eletrônicos, construindo os tesauros por mecanismos de inferência.
- Semi-automática - onde ferramentas de software são utilizadas por especialistas como auxílio na construção de tesauros.

O uso de tesauros na implementação de estratégias semânticas de recuperação de informação é atualmente bastante amplo e documentado na literatura científica. Entretanto, grande parte dos esforços desenvolvidos pelas comunidades acadêmicas de diversas áreas do conhecimento encontra uma limitação de ordem prática, que é a necessidade de construção de tesauros específicos para as áreas em questão.

Os tesauros são utilizados na indexação das informações e na recuperação da informação. A sua metodologia de elaboração parte da idéia de sistema. Neste tipo de sistema, os conceitos se relacionam entre si e são representados por termos. Cada termo obrigatoriamente tem uma vinculação com outro termo, e esta vinculação é que forma a estrutura do tesauro. Assim, tesauros são principalmente utilizados para promover o controle de vocabulário em sistemas de recuperação da informação. São utilizados pelos indexadores no momento da indexação e devem ser disponibilizados para uso no momento da recuperação da informação.

2.1.6 Representação dos Textos

Uma das formas mais utilizadas para representar os termos dos documentos é através da conversão em tabelas. Este formato permite a aplicação de diversas técnicas desenvolvidas para dados estruturados.

Essa conversão deve ser realizada após as etapas de eliminação das *stopwords* (2.1.3) e *stemming* (2.1.4), quando temos um conjunto de dados reduzido em relação ao original.

No modelo espaço vetorial, cada documento de uma coleção deve ser representado como um vetor de termos contidos no mesmo. Com o objetivo de identificar todos os

termos presentes em um documento, a tokenização é realizada. Normalmente são utilizados os espaços em branco, tabulações e sinais de pontuação como separadores dos termos de um documento.

O conjunto dos diferentes termos que ocorrem em uma coleção de documentos é chamado de dicionário da coleção de documentos.

Seja uma coleção de documentos D = $d_1, d_2, \ldots, d_n$ e o conjunto T = $t_1, t_2, .., t_m$ de todos os diferentes termos presentes em D. Essa coleção pode ser representada como uma tabela, onde cada documento d_i corresponde a uma coluna na tabela e cada termo t_j corresponde a uma linha. A tabela atributo-valor que corresponde a representação de documentos é apresentada na tabela 2.1.

Tabela 2.1: *Representação estruturada de documentos*

	d_1	...	d_k	...	d_m
t_1	v_{11}	...	v_{1k}	...	v_{1m}
...	...	...	...	...	...
t_j	v_{j1}	...	v_{jk}	...	v_{jm}
...	...	...	...	...	...
t_n	v_{n1}	...	v_{nk}	...	v_{nm}

O valor v_{jk} denota a importância relativa do termo t_k em relação ao documento d_j. Para quantificar esta importância, normalmente é utilizada uma representação vetorial, onde para a atribuição de valores aos termos, emprega-se medidas estatísticas baseadas na freqüência dos termos nos documentos [108].

2.1.6.1 Atribuição de Pesos

No modelo de espaço vetorial, cada documento é representado como um vetor cujas dimensões são os termos presentes na coleção inicial de documentos. Cada coordenada do vetor é um termo e tem um valor numérico que representa sua relevância para o documento. Geralmente, valores maiores implicam em maior relevância.

Este processo de associar valores numéricos a coordenadas de vetores é referenciado como atribuição de pesos ou *weighting* [63]. Portanto, a atribuição de pesos é o processo de destacar os termos mais importantes.

Existem diversas medidas de atribuição de pesos. Entre as mais populares estão [12]:

- Binária: O esquema binário utiliza os valores 1 e 0 para identificar a existência ou não de um termo no documento, respectivamente.
- *Term-Frequency* (TF): Conta as ocorrências de um termo em um documento e usa esse contador como medida numérica. As medidas são normalizadas para valores no intervalo [0,1]. Isto é feito independentemente para cada documento,

dividindo-se cada medida de coordenada pela medida de coordenada mais alta do documento considerado. Este procedimento ajuda a resolver problemas associados com o tamanho do documento. Sem a normalização, um termo pode ter a medida maior em um vetor-documento simplesmente porque o documento correspondente é muito grande.

- *Term-Frequency – Inverse Document Frequency* (TF-IDF): O TF-IDF trabalha com a comparação do cálculo da freqüência relativa de palavras em um documento na proporção inversa às palavras contidas no conjunto de documentos. O procedimento para implementação do TF-IDF pode variar de acordo com a finalidade da aplicação, mas a forma geral pelo qual o método trabalha é descrita na Seção 2.1.6.2.

Sobre o uso de pesos no modelo vetorial, [7] apresenta algumas considerações:

- Pesos não binários podem considerar mais adequadamente *matchings* (casamentos) parciais;
- Esses pesos são utilizados para calcular um grau de similaridade entre a consulta e o documento;
- A fórmula com que são calculados os pesos varia entre as implementações.

2.1.6.2 *Term-Frequency – Inverse Document Frequency* (TF-IDF)

Dada uma coleção de documentos D, uma palavra w e um documento específico d, onde $d \in D$.

$$w_d = f_{w,d} * log(|D|/f_{w,D})$$

onde:

- $f_{w,d}$ representa do número de vezes que a palavra w aparece no documento específico d;
- $|D|$ representa o tamanho da coleção de documentos;
- $f_{w,D}$ representa o número de documentos em que w aparece em D.

O código para implementar TF-IDF é considerado simples. Dada uma pesquisa **q** composta por um conjunto de palavras w_i, calcula-se $w_{i,d}$ (peso da palavra no documento) para cada palavra w_i em todos documentos pertencentes à D [88].

De uma forma simples, isto pode ser feito percorrendo toda a coleção de documentos e realizando, durante a varredura, os cálculos de $\Sigma f_{w,d}$ e $\Sigma f_{w,D}$. Uma vez concluídos esses somatórios, pode-se facilmente calcular $w_{i,d}$.

Uma vez que todas as $w_{i,d}$ foram encontradas, o sistema retorna o conjunto D^* contendo os documentos específicos d tais que maximizem a equação $\sum_i w_{i,d}$. O usuário

ou o próprio sistema podem determinar o tamanho de D^* antes de iniciar a pesquisa. Desta forma, os documentos são recuperados em ordem decrescente de relevância.

2.1.6.3 Dimensão da Representação

Um conceito importante na recuperação de informação é a dimensionalidade dos dados. Cada documento da coleção possui características ou termos mais ou menos apropriados para caracterizá-lo. A escolha das características mais relevantes é determinante para a representação individualizada e sem perda de informação de cada documento da coleção [110].

Tendo em vista que cada termo é um atributo do documento e é utilizado para caracterizá-lo, o problema da alta dimensionalidade é típico do processo de KDT e a busca por melhor representatividade dos documentos, sem perda de informação, e pela eficiência de processos computacionais é uma necessidade.

Independentemente da medida escolhida, a matriz atributo-valor terá um número para indicar a presença do termo e o número zero para indicar a ausência dele, conforme demonstrado na tabela 2.2.

Tabela 2.2: *Exemplo de alta dimensionalidade – termo por documento da coleção*

	d_1	...	d_k	...	d_m
t_1	0	...	1	...	0
...	...	...	...	...	...
t_j	0	...	0	...	1
...	...	...	...	...	...
t_n	1	...	0	...	1

[94] chama a atenção para a ocorrência de dois problemas quando se adota essa abordagem:

- a ocorrência de um grande número de termos: ao transformar um documento em termos, que são dispostos nas colunas da matriz atributo-valor, o número de termos tende a ser elevado.
- a ocorrência de uma grande quantidade de zeros: estima-se que os termos identificados em um documento não sejam os mesmos identificados nos demais, especialmente se a análise for realizada em documentos extensos. Desta forma, estima-se que a matriz atributo-valor terá uma grande quantidade de colunas com valores zero, indicando a ausência do termo no documento.

2.1.6.4 Lei de Zipf

A Lei de Zipf, é também conhecida como a lei do mínimo esforço e consiste em medir a freqüência do aparecimento das palavras em vários textos, gerando uma lista ordenada de termos de uma determinada disciplina ou assunto [103].

Segundo esta lei, os termos encontram-se distribuídos nos textos obedecendo a um padrão, de maneira que existe uma relação na frequência de alguns termos e suas posições em uma lista ordenada.

A aplicação dessa abordagem sobre uma coleção de textos consiste na contagem dos termos f e na sua ordenação r. O produto da frequência de cada termo f e sua ordem r na lista de termos é aproximadamente uma constante k, tal que:

$$f * r = k$$

Um exemplo de comparação da aplicação dessa abordagem é demonstrada na Tabela 2.3.

Tabela 2.3: *Exemplo da Lei de Kipf*

Palavra	Frequência (f)	Ordenação (f)	Constante $K = f * r$
A	1000	1	1000
Informação	500	2	1000
Busca	333	3	1000
Ciência	250	4	1000
...	...	...	...
Texto	1	1000	1000

Analisando o exemplo da Tabela 2.3, pode-se concluir que, se o termo mais frequente se repete mil vezes, então o segundo termo mais frequênte se repetirá k/2 vezes que é 500 e, por conseguinte, o terceiro termo seria contado k/3 vezes, totalizando 333 e assim por diante.

Assim, a Lei de Zipf é uma constatação empírica e apresenta uma descrição da distribuição de frequência de palavras na linguagem: existem poucos termos muito comuns, uma quantidade média de termos de frequência intermediária e muitos termos que ocorrem poucas vezes, conforme ilustrado na Figura 2.2.

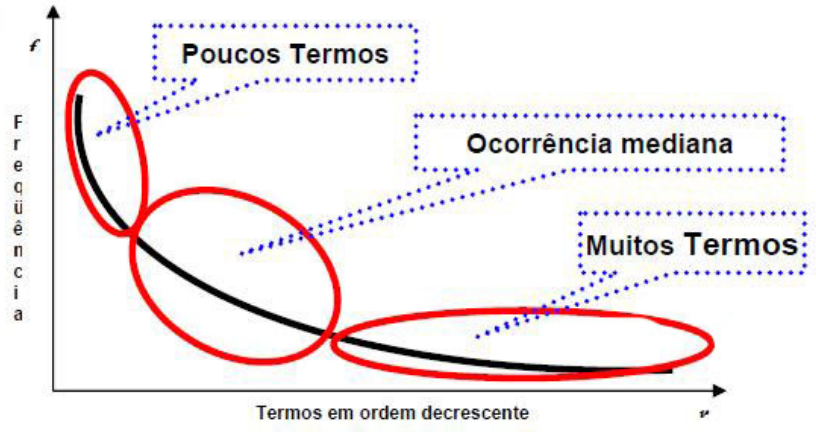

Figura 2.2: *Representação Gráfica da Lei de Zipf*

A análise do gráfico da Figura 2.2, pode auxiliar na seleção de pontos de corte, com o objetivo de excluir palavras com baixo poder de discriminação e, assim, reduzir a dimensionalidade. A escolha do ponto de corte é feita de maneira arbitrária, principalmente através da experiência do analista para que se mantenha o menor número de termos, sem perda de informação.

2.1.6.5 Significância das Palavras de Luhn

Para [64], a divisão de textos em capítulos, parágrafos, orações, frases, etc. é uma manifestação física da associação de idéias do escritor. Assim, na linguagem escrita, as idéias mais associadas intelectualmente são implementadas por palavras mais associadas fisicamente.

Segundo [64], um autor normalmente repete palavras à medida que avança ou varia sua argumentação e, assim, elabora os aspectos de seu assunto. A frequência de um termo em um documento fornece uma medida útil para determinar a significância de uma palavra. Essa abordagem não leva em consideração as relações lógicas ou semânticas.

Para [70], foi Luhn quem descobriu que padrões de distribuição de termos poderiam fornecer informação significativa sobre o conteúdo de um documento. Altas frequências de termos tendem a ser comuns e não são relevantes para destacar o conteúdo. Por outro lado, uma ou duas ocorrências de um termo em textos relativamente extensos também podem não fornecer informação relevante na descoberta do assunto apresentado no documento.

Dessa forma, a abordagem de Luhn é utilizada para criar uma lista de termos em ordem decrescente de frequência e, então, identifica a sua relevância em função do assunto do documento. A idéia de Luhn é que existem pontos de corte que podem ser calculados através de métodos estatísticos ou atribuídos pela experiência de analistas de

domínio. Esses pontos delimitam os termos que são significativos para a identificação do tema.

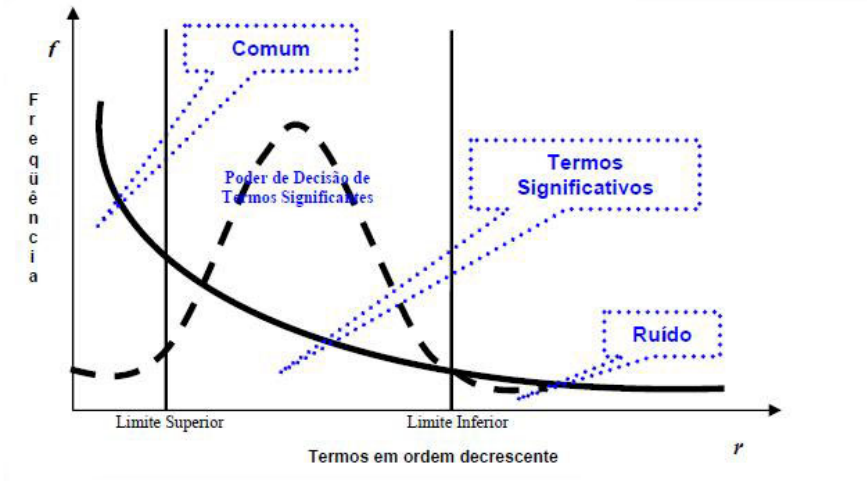

Figura 2.3: *Significância de Termos - [64]*

Conforme ilustrado na Figura 2.3, a abordagem de Luhn propõe que os termos à esquerda do limite superior são comuns e aqueles à direita do limite inferior são raros e não contribuem significativamente para o conteúdo do texto. Assim, os termos que contribuem significativamente estão entre os limites superior e inferior.

Existe também uma curva, que Luhn chamou de poder de decisão de termos significantes, que expressam a capacidade de discriminar o conteúdo, ilustrando que os termos, em uma ordem de significância imaginária que se inicia próxima de zero, vão crescendo em habilidade de discriminação até atingirem o pico na metade entre os limites superior e inferior e então começam a diminuir simetricamente até o último termo.

Nessa abordagem, os limites de corte também são estabelecidos de forma arbitrária.

É possível identificar a relação entre a curva de Zipf e o conceito de Luhn, na identificação do ponto onde os termos significantes estão; ambas apontam os termos de baixa representatividade nas extremidades da distribuição dos termos [70].

2.1.6.6 Decomposição de Valores Singulares

Segundo [67], as técnicas de redução de dimensionalidade extraem um grupo de objetos que existem no espaço com muitas dimensões e os representa no espaço com poucas dimensões com a finalidade de visualização.

O modelo de espaço vetorial é uma representação frequente na recuperação de informação, principalmente por sua simplicidade conceitual e utilização de proximidade espacial para denotar similaridade semântica entre documentos.

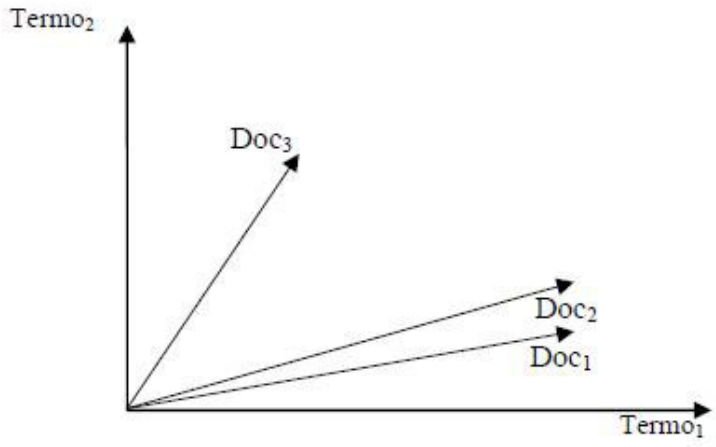

Figura 2.4: *Espaço Vetorial em 2 Dimensões*

Esse modelo é uma maneira de representar documentos espacialmente por meio das palavras que o compõem. Cada documento é representado como um vetor em relação ao vocabulário que contém.

A Figura 2.4, mostra a representação, em duas dimensões, correspondente aos termos 1 e 2 e a três documentos no espaço. A proximidade dos vetores é calculada pelo ângulo, ou seja, quanto menor o ângulo entre dois vetores, mais próximos semanticamente eles são. Essa proximidade é obtida através da análise dos termos presentes nos documentos.

Na Figura 2.4, os documentos 1 e 2 possuem uma proximidade espacial muito maior entre si do que em relação ao documento 3. Isto indicaria que os documentos 1 e 2 são semanticamente mais similares.

A decomposição em valores singulares (DVS) é uma técnica matemática de redução de dimensionalidade que visa formar novas variáveis que são combinações lineares das variáveis originais. A finalidade é utilizar um número consideravelmente menor de novas variáveis que contêm a informação das variáveis originais, isto é, utilizam-se poucas variáveis, sem perda de informação.

Essa técnica transforma a matriz atributo-valor que representa os documentos indexados em outras 3 (três) matrizes distintas (U, S, V), confome demonstrado na Figura 2.5. A multiplicação dessas matrizes reconstitui a matriz original.

- **Matriz U**: contém todos os termos da coleção original;
- **Matriz S**: contém os valores dos pesos de cada termo na coleção (os valores singulares da matriz);
- **Matriz V**: contém os documentos da coleção original.

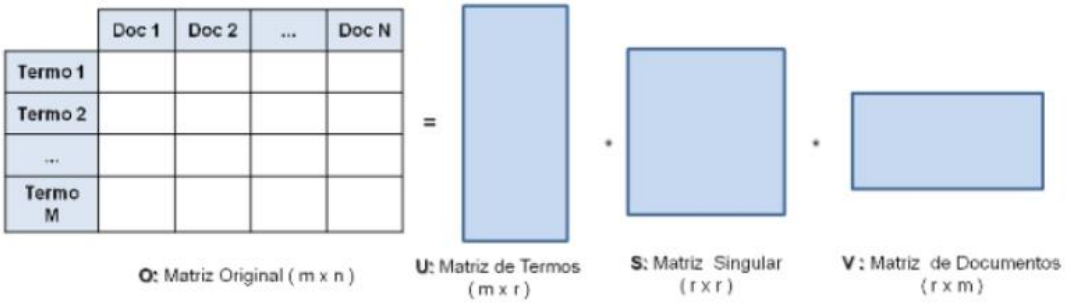

Figura 2.5: *Matriz atributo-valor original (O) e sua representação após a decomposição (U, S, V). Onde m representa o número de termos, n representa o número de documentos e r representa a dimensão necessária para representação dos valores ponderados dos termos da coleção*

As dimensões dessas matrizes são geralmente muito grandes, já que o número de documentos e de termos indexados é extenso.

Pela visão do usuário, quanto mais informação ele tiver acesso, melhor. Isto significa que, para o usuário, utilizar um processo onde há perda de informação, acarretará perda de valor. Já para a recuperação de informação, essa visão não é exatamente adequada uma vez que ela busca formas de generalização e, para generalizar, é preciso concentrar a atenção em fatores estatisticamente relevantes.

Por esta razão, o processo realiza um outro procedimento: uma redução no tamanho da matriz de valores singulares. Ao invés de mantê-la com o tamanho (r x r) é realizada uma redução para um tamanho (k x k), onde *k* é um número muito menor que *r*.

Ou seja, uma vez criadas as 3 (três) matrizes é escolhido um tamanho *k* adequado (chamado de nível *k*) e, através deste valor, são criadas 3 (três) novas matrizes U´, S´ e V´, todas de tamanho *k*. Qualquer valor acima de *k* será zerado. Isto permitirá que a matriz de valores singulares tenha uma versão mais “enxuta”, contendo somente os valores mais significativos.

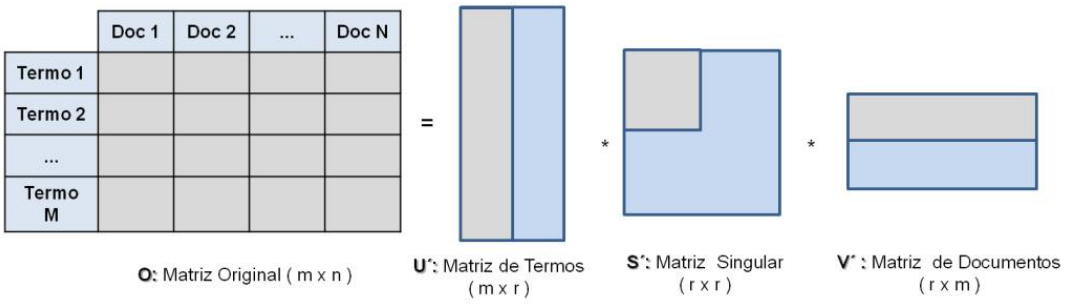

Figura 2.6: *Matrizes resultantes após aplicação da redução para dimensão k*

A Figura 2.6 mostra, nas áreas acizentadas, os valores significativos presentes nas matrizes U, S e V. Esta redução de área irá montar uma nova matriz atributo-valor que possuirá elementos bastante diferentes da matriz original. Essa operação de redução possibilita a redução da influência individual dos termos, preservando os padrões principais de uso desses termos. Cada coluna da nova matriz será uma descrição (um vetor) que representará diferentes padrões de uso das palavras mais relevantes desse documento. Isto permite que seja possível comparar documentos de acordo com um critério estatisticamente interessante.

A aplicação da DVS na área de recuperação da informação é chamada de indexação semântica latente (LSI), apresentada na Seção 2.2.4. Essas novas dimensões são uma melhor representação de documentos e de consultas. O nome “latente” é uma metáfora devido ao fato de que essas novas dimensões são a representação verdadeira, pois a indexação semântica latente (LSI) recupera a estrutura semântica original do espaço e suas dimensões originais.

2.2 Indexação

Grande parte dos problemas na manipulação de textos está relacionada com a representação adequada dos dados, de forma que o desempenho do modelo selecionado atenda aos critérios de qualidade e eficiência. Existem várias tarefas que podem ser realizadas, onde cada tipo extrai um tipo diferente de informação dos textos.

A indexação permite uma busca eficiente por documentos relevantes, sem que seja preciso examinar documentos inteiros. A indexação de documentos é semelhante à indexação de banco de dados, onde é permitido que toda a tabela seja percorrida para a recuperação de linhas de dados.

Os tipos mais comuns de indexação são descritos a seguir.

2.2.1 Indexação *Full Text*

Segundo [63], a indexação de textos completos deve ocorrer automaticamente quando os documentos são carregados. Os índices guardam informações sobre a localização dos termos dentro do texto, de forma que operadores de proximidade possam ser utilizados em consultas no texto completo. Os operadores mais comuns são:

- operadores booleanos: AND, OR e NOT;
- operadores de proximidade: NEAR e WITHIN.

2.2.2 Indexação Temática

Esse tipo de indexação é dependente de tesauro [63], por isso, permite encontrar documentos relevantes rapidamente, através de uma estrutura hierárquica e de generalizações e especializações de termos [85].

2.2.3 Indexação em Listas Invertidas

Índice invertido, segundo [7], é a estrutura mais comum para indexar informação de maneira a permitir um melhor desempenho durante uma tarefa de busca.

O índice invertido contém a lista de todas as palavras indexadas. Cada palavra fica associada à lista dos identificadores dos documentos em que ela ocorre. Esta estrutura permite que um único termo aponte para vários documentos.

As listas devem ser armazenadas de forma ordenada para garantir a eficiência dos algoritmos de busca.

2.2.4 Indexação Semântica Latente

Para [40], a indexação semântica latente (LSI) tenta superar as deficiências da recuperação por combinação de termos, tratando a falta de confiabilidade dos dados associados a uma relação termo-documento ou documento-documento como um problema estatístico.

Este método assume que há uma estrutura semântica oculta (latente), subjacente aos dados. Para o autor, esta semântica é esquecida parcialmente pela aleatoriedade da escolha da palavra no que se refere à recuperação, pelo fato de que se escolhem palavras individuais para serem recuperadas, indexadas, etc. A descrição dos termos e dos documentos baseados na estrutura semântica latente é utilizada tanto para a indexação como para a recuperação.

Entende-se por “estrutura semântica” a estrutura de correlação entre as palavras individuais que aparecem nos documentos. O termo “semântico” implica o fato de que

os termos, em um documento, possam ser tomados como referentes ao documento ou ao assunto desse documento. Esta técnica de análise da semântica de palavras em documentos distintos é automática: essa é a diferença principal que existe entre o LSI e os outros modelos existentes [40].

Na recuperação de informação, a LSI é utilizada para calcular a similaridade entre os documentos de uma coleção e, também, calcular a similaridade entre os termos de consulta do usuário e os documentos de uma coleção, conforme demonstrado na Figura 2.7.

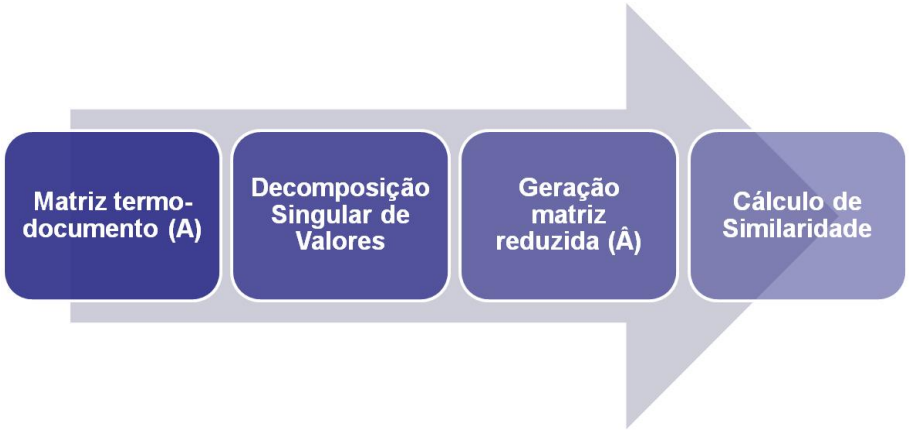

Figura 2.7: *Etapas da LSI na recuperação de informação*

O modelo matemático utilizado para criar a estrutura semântica corresponde à DVS, descrita na Seção 2.1.6.6. O resultado da aplicação desse modelo, após realizadas operações matriciais, é uma matriz reduzida, aproximada à matriz original.

Na Figura 2.8 é possível observar a representação da decomposição DVS para uma matriz termo-documento, conforme apresentado na Seção 2.1.6.6. Onde *m* representa as linhas (termos da coleção) e *n* representa as colunas (documentos da coleção).

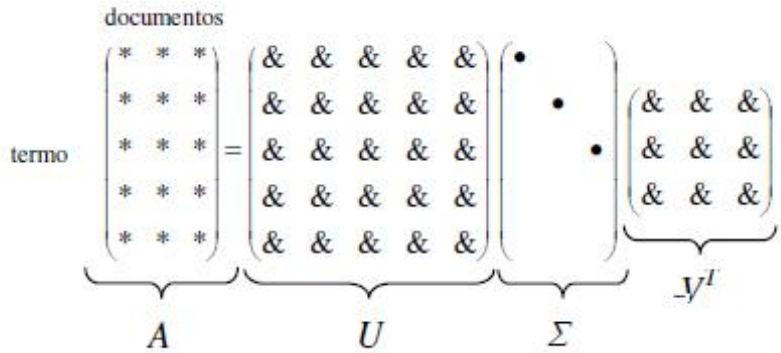

Figura 2.8: *Representação da decomposição DVS sobre a matriz atributo-valor A*

Matematicamente este resultado pode ser interpretado como uma configuração na qual o produto cosseno entre vetores representa a similaridade estimada entre dois documentos.

Na recuperação de informação, a DVS é interpretada como uma técnica para gerar um conjunto de indexações não-correlacionadas de variáveis ou fatores; cada relação é representada por seu vetor de valores [30].

Formalmente, a LSI trabalha com a matriz na qual relacionam-se termos e documentos (matriz atributo-valor), exemplificada pela Figura 2.9.

	documento 1	documento 2	documento 3	documento 4	documento 5	documento 6
java	1	0	1	0	0	0
herança	0	1	0	0	0	0
polimorfismo	1	1	0	0	0	0
paradigma	1	0	0	1	1	0
programação	0	0	0	1	0	1

Figura 2.9: *Representação da matriz atributo-valor A*

Essa matriz é analisada, processada e decomposta em 3 novas matrizes. Ainda como parte do processo de DSV, a dimensão dessas matrizes sofre uma redução. Após a multiplicação das 3 novas matrizes reduzidas é gerada uma matriz de menor dimensão. Assume-se que essa nova matriz, de tamanho k, é a melhor aproximação em relação à matriz original, ou seja, possui os termos e documentos mais significativos da coleção.

O processo realizado pela LSI é ilustrado pelas figuras 2.9, 2.10, 2.11, 2.12, 2.13 e 2.14.

java	-0,44	-0,30	0,57	0,58	0,25
herança	-0,13	-0,33	-0,59	0,00	0,73
polimorfismo	-0,48	-0,51	-0,37	0,00	-0,61
paradigma	-0,70	0,35	0,15	-0,58	0,16
programação	-0,26	0,65	-0,41	0,58	-0,09

Figura 2.10: *Matriz de termos U, resultante da aplicação da DVS sobre a matriz original A*

2,16	0,00	0,00	0,00	0,00
0,00	1,59	0,00	0,00	0,00
0,00	0,00	1,28	0,00	0,00
0,00	0,00	0,00	1,00	0,00
0,00	0,00	0,00	0,00	0,39

Figura 2.11: *Matriz S de valores singulares de A*

	documento 1	documento 2	documento 3	documento 4	documento 5	documento 6
java	-0,75	-0,28	-0,20	-0,45	-0,33	-0,12
herança	-0,29	-0,53	-0,19	0,63	0,22	0,41
polimorfismo	0,28	-0,75	0,45	-0,20	0,12	-0,33
paradigma	0,00	0,00	0,58	0,00	-0,58	0,58
programação	-0,53	0,29	0,63	0,19	0,41	-0,22

Figura 2.12: *Matriz de documentos V, resultante da decomposição DVS sobre a matriz original A*

Como continuação do processo da DVS, são calculados os maiores valores singulares da matriz S, de onde é obtida a matriz S´, ilustrada pela Figura 2.13.

2,16	0,00	0,00	0,00	0,00	0,00
0,00	1,59	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00

Figura 2.13: *Matriz S´, com os maiores valores singulares da matriz S*

Através da multiplicação de U, S´ e V, obtemos a matriz reduzida A´, que expressa a melhor representação da estrutura semântica de certo domínio, podendo ser este domínio uma coleção de documentos ou um banco de dados.

	documento 1	documento 2	documento 3	documento 4	documento 5	documento 6
java	-1,62	-0,60	-0,44	-0,97	-0,70	-0,26
herança	-0,46	-0,84	-0,30	1,00	0,35	0,65
polimorfismo	0,00	0,00	0,00	0,00	0,00	0,00
paradigma	0,00	0,00	0,00	0,00	0,00	0,00
programação	0,00	0,00	0,00	0,00	0,00	0,00

Figura 2.14: *Matriz parcial A´, resultante da multiplicação das matrizes U, S´ e V*

Para calcular a similaridade entre a consulta e os documentos da coleção, uma das formas mais utilizadas é a conversão da consulta do usuário em um vetor de termos, semelhante à representação da matriz atributo-valor. Esse novo vetor é comparado a cada documento da coleção, como forma de obtenção dos elementos semanticamente mais semelhantes. Esta similaridade entre a consulta do usuário e os documentos da coleção é obtida através da fórmula:

$$\cos v_j = \frac{(A_k\, e_j)^T\, q}{\|(A_k\, e_j)\|_2\, \|q_2\|}$$

Onde:

- A_k - matriz aproximada (reduzida) de A
- e_j - vetor que representa a coluna j da matriz matriz singular de valores
- q - vetor que representa a consulta do usuário
- $||\ ||_2$ - norma euclidiana do vetor

2.3 Mineração

A extração de informação (EI) é uma área de pesquisa que utiliza as abordagens de mineração de textos, processamento de linguagem natural e aprendizado de máquina.

Trata-se do processo de extrair informações específicas sobre objetos e relacionamentos (nomes próprios ou siglas, por exemplo) entre eles, a partir de documentos textuais. Os objetivos principais da EI são: dividir um documento em partes relevantes e irrelevantes e preencher modelos pré-definidos com a informação extraída [109].

2.3.1 Sumarização

Consiste nas atividades realizadas para a criação de uma descrição compacta de um documento ou uma coleção de documentos [83].

2.3.2 Categorização

Dado um conjunto pré-definido de categorias ou classes, o objetivo da categorização é induzir um classificador que possa predizer se um documento pertence ou não a uma categoria [111]. Como um documento pode pertencer a várias categorias, a tarefa geralmente consiste em predizer se o documento pertence a cada uma das categorias em separado.

2.3.3 Clusterização de Documentos

Consiste em identificar um conjunto de agrupamentos (*clusters*) a partir da coleção de documentos. Após esse processo, os documentos são distribuídos em um número de *clusters*, onde os documentos presentes em um mesmo cluster são similares e distintos de documentos presentes em outros *clusters* [112]

2.4 Análise

Esta etapa consiste na validação das descobertas realizadas pela etapa de processamento de dados e na visualização dos resultados encontrados. Para auxiliar na consoli-

dação dos resultados, podem ser utilizadas métricas de avaliação de resultados, ferramentas de visualização e também conhecimento de especialistas.

2.4.1 Métricas de Avaliação de Resultados

As métricas de avaliação de desempenho de um sistema são baseadas na noção de relevância, ou seja, se um documento atende à necessidade de informação do usuário.

As medidas de avaliação mais comuns são:

- Precisão
  Mede a habilidade do sistema manter os documentos irrelevantes fora do resultado de uma consulta. É obtida pela fórmula:

  $$\text{Precisão} = \frac{R}{|R|}$$

  Onde R representa o número de itens relevantes recuperados e |R| representa o número total de itens recuperados.

- *Abrangência*
  Mede a habilidade do sistema para recuperar os documentos mais relevantes para o usuário, com base no termo ou expressão utilizado na formulação de sua busca. É calculado através da execução da fórmula:

  $$\text{Recall} = \frac{R}{|R_r|}$$

  Onde R representa o número de itens relevantes recuperados e $|R_r|$ representa o número de itens relevantes na coleção.

2.5 Considerações Finais

A mineração de textos faz parte da área de descoberta de conhecimento, provendo técnicas efetivas de descoberta de conhecimento em bases de dados não estruturados. Uma vez que a maioria dos dados disponibilizados, não somente na Internet, mas nas empresas em geral, é armazenado neste formato, este tipo de técnica possui um vasto campo de aplicação. Apesar disso, essa área de pesquisa enfrenta grandes problemas, como a falta de técnicas efetivas de análise semântica de textos. Isto ocorre principalmente pelo fato de que análise semântica de textos é muito difícil de ser realizada de forma automática, justamente pelas características destes textos.

Dentro do processo de mineração, encontra-se a Indexação Semântica Latente (LSI - *Latent Semantic Analysis*), uma das técnicas que se propõem a solucionar o problema de análise de semântica de textos. A LSI objetiva realizar a análise de documentos,

identificar a semântica dos documentos e, ainda, os relacionamentos ocultos entre estes documentos, estabelecidos através da variação de termos utilizados nos documentos da coleção.

Este trabalho utiliza algumas técnicas da mineração de textos como forma de preparação dos documentos a serem processados e a LSI para validar a capacidade da técnica para determinar a similaridade entre documentos de uma coleção e apresentações em forma de slides, contendo conteúdo teórico abordado em sala de aula.

O Capítulo a seguir apresenta a técnica de expansão de consulta, utilizada neste trabalho para melhor alimentar os mecanismos de busca e recuperação de material didático.

CAPÍTULO 3

Expansão de Consulta

Segundo [54], um dos grandes problemas dos sistemas de recuperação da informação está no fato de que os termos utilizados nas consultas são curtos ou imprecisos, o que significa dizer que eles trazem pouca informação a respeito da real necessidade do usuário. Outro fator que dificulta a eficácia desse tipo de sistema é a falta de conhecimento a respeito do funcionamento dos mecanismos de busca por parte dos usuários, o que acarreta na má formulação de consultas, gerando resultados insatisfatórios [69] [96] e pior, a incerteza por parte do usuário quanto ao que ele próprio deseja.

Portanto, os sistema de recuperação da informação tem a difícil tarefa de encontrar informações relevantes. Para isso, a utilização de elementos adicionais nesse ambiente deve ser considerada na tentativa de melhorar os resultados de recuperação, já que o simples casamento das palavras-chaves pode não ser suficiente [69].

Nesta metodologia, um conjunto de documentos é recuperado utilizando a consulta original do usuário. Esses documentos são automaticamente considerados relevantes, mesmo sem a intervenção do usuário. Esses documentos podem ser utilizados para construir uma nova consulta, desta vez expandida, que é executada com a finalidade de tentar recuperar um conjunto de documentos com maior grau de relevância para o usuário [69].

Um dos problemas dessa abordagem está no fato de que se os documentos automaticamente assumidos como relevantes são, na verdade, irrelevantes, então as palavras adicionadas à consulta expandida têm grande probabilidade de não estar relacionadas à real necessidade do usuário e o resultado será uma recuperação de documentos insatisfatória. Por outro lado, se os documentos inicialmente recuperados são considerados efetivamente relevantes, então a probabilidade dos termos adicionados após a expansão serem adequados às necessidades do usuário será maior e os documentos recuperados tendem a ser altamente relevantes. Logo, a realimentação *adhoc* parece ser capaz de aumentar ou reduzir a qualidade para diferentes pesquisas [69].

Neste capítulo serão apresentadas as principais abordagens utilizadas para o processo de expansão de consulta e as atividades que compõem o processo.

3.1 Abordagens para Expansão de Consulta

Expansão de consulta (*query expansion*) é o processo de complementar a consulta original com termos adicionais. Ele pode ser considerado um método para melhorar o desempenho da recuperação e é aplicável a qualquer situação, independentemente da técnica de recuperação utilizada [34]. A consulta inicial fornecida pelo usuário pode ser uma representação insuficiente ou incompleta da necessidade de informação do usuário por si só ou em relação às idéias presentes nos documentos. Essa abordagem pode ser aplicada na formulação inicial da consulta, no estágio de reformulação da mesma ou em ambos os casos.

A expansão de consulta, representada na Figura 3.1, pode ser executada manualmente, automaticamente ou interativamente (também conhecida como semi-automática ou mediada pelo usuário).

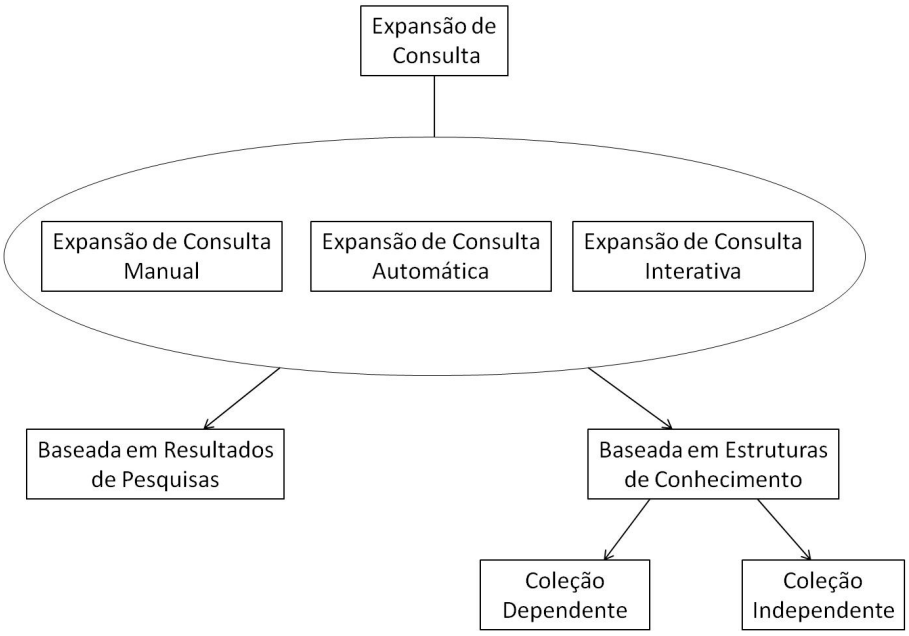

Figura 3.1: *Métodos e fontes da Expansão de Consulta [34].*

Para [34], dois elementos chave devem ser considerados quando aplicada qualquer forma de expansão de consulta:

- a fonte, que irá fornecer os termos para a execução da expansão.
- o método, que será utilizado para selecionar os termos para serem usados na expansão.

Um tipo de fonte é baseado no resultado da pesquisa. Nesse caso, os documentos recuperados em interações anteriores devem ser identificados como relevantes, tornando-se fontes para a expansão de consulta. Outro tipo de fonte é baseado em alguma estrutura

de conhecimento (como um dicionário, por exemplo), independente do processo de pesquisa.

3.2 Expansão Manual de Consulta

Elaborar uma estratégia de busca consiste em formular uma tática para recuperar informações armazenadas em um banco de dados. Uma estratégia de busca será eficiente se as informações recuperadas atenderem às necessidades do usuário. A eficiência do usuário, por sua vez, depende de sua capacidade em oferecer ao sistema elementos suficientes para que sejam selecionados, a partir da totalidade das informações armazenadas, um conjunto de itens que constituam a resposta que procura.

Segundo [77], para planejar a estratégia de busca, o usuário deve decidir qual é a melhor base de dados para seu tema, selecionar os termos de busca adequados e formular a estratégia.

Diversas estratégias foram identificadas na literatura. Algumas das mais citadas são: blocos de construção (*building block*), crescimento de pérolas (*pearl growing*) e frações sucessivas (*successive fractions*).

3.2.1 Blocos de Construção - *Building Block*

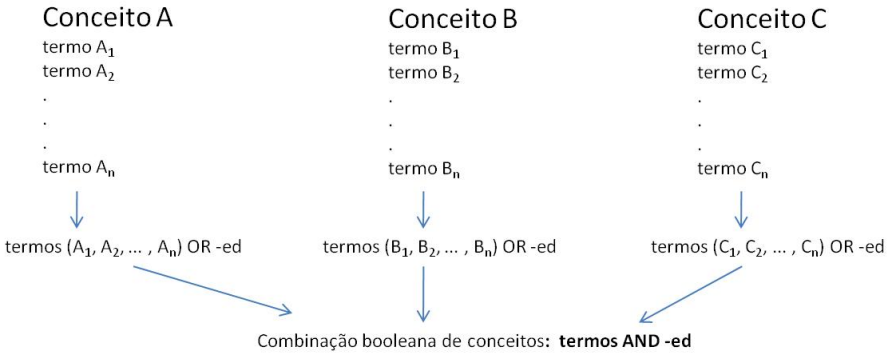

Figura 3.2: *Exemplo da metodologia building block*

O *building block* é uma metodologia eficaz de pesquisa ainda muito utilizada. A estratégia consiste em um refinamento da pesquisa original [81] e [82] seguindo, basicamente, os seguintes passos:

1. A pesquisa é quebrada em conceitos;
2. Cada conceito é novamente decomposto, visando formar agrupamentos de termos;

3. São estabelecidos relacionamentos entre os grupos usando os operadores booleanos OR (toda e qualquer citação usando todo e qualquer termo), NOT (somente citações do primeiro bloco que não mencionam termos do próximo bloco) e AND (somente citações onde termos do primeiro bloco aparecem como termos do outro bloco).

4. Os blocos de conceitos são preenchidos com termos que são sinônimos ou equivalentes, ou seja, termos ligados por um OR.

3.2.2 Crescimento de Pérolas - *Pearl Growing*

Para [81] e [82], se o usuário encontra um elemento relevante, este pode ser útil para a realização de novas pesquisas, incluindo termos localizados neste elemento. A estratégia consiste, basicamente, nos seguintes passos:

1. Início da pesquisa a partir de um termo ou documento altamente relevante (chamada pérola). Se esse elemento de alta relevância não é conhecido, será preciso realizar uma busca limitando-se aos termos presentes nos títulos;

2. Pesquisa na base de dados, tentando encontrar registros para os referidos termos;

3. Revisão dos descritores atribuídos ao registro;

4. Realização de uma nova busca, utilizando os descritores da pérola;

5. Exame dos novos registros recuperados;

6. Revisão dos descritores atribuídos aos novos registros recuperados;

7. Realização de uma nova busca, utilizando qualquer descritor adicional presente nos novos registros recuperados.

8. Repetição do processo (passos de 5 a 7), com mais registros relevantes, até que o material desejado seja encontrado.

3.2.3 Frações Sucessivas - *Successive Fractions*

A abordagem de frações sucessivas (*successive fractions*) [81] e [82] consiste em modificar o conjunto inicial de pesquisa, iniciando pela seleção de um grande conjunto, que englobe a maior quantidade possível de referências para o tópico da pesquisa. Então, basta continuar adicionando conceitos, dos mais amplos até os mais específicos. Cada pesquisa sucessiva utiliza um operador lógico AND para cortar o elemento menos relevante do conjunto. Ao final do processo, o usuário terá o melhor conjunto que poderia encontrar.

De acordo com [81] e [82], podem ser utilizadas variações destas três estratégias básicas, dependendo da situação específica.

A formulação de estratégias de pesquisa é um problema altamente desestruturado e requer um amplo leque de conhecimento. Embora esteja sendo sistematicamente estudada, ainda não é um processo bem entendido. Conseqüentemente, o processo de formulação de pesquisa é bastante difícil de automatizar.

3.3 Expansão Automática de Consulta

Estudos encontrados na literatura mostram que a expansão de consulta manual, que utiliza técnicas de realimentação de palavras relevantes, normalmente produzem resultados melhores que técnicas de expansão automática [7]. Para [69], no entanto, as técnicas de realimentação manual de palavras relevantes costumam não ser bem sucedidas em sistemas de busca, pois os usuários raramente fornecem informação para realimentar o sistema. Nestes casos, a melhor solução para expandir as consultas é através de técnicas de expansão automática.

Portanto, a expansão automática de consulta é uma alternativa eficaz para contornar os problemas ocasionados por uma consulta com termos "pobres". O objetivo dos algoritmos de expansão de consultas é formular uma consulta mais elaborada a partir da consulta feita inicialmente pelo usuário.

Um sistema que utiliza métodos de expansão pode ser capaz de recuperar documentos que estejam relacionados com a consulta inicial do usuário mesmo que não haja palavras da consulta no documento.

O casamento entre as palavras da consulta e as palavras dos documentos é muitas vezes utilizado como filtro para selecionar o conjunto de documentos que pode ser mostrado como parte do conjunto de resposta de um sistema de recuperação de informação. Características extra dos documentos deste conjunto inicial podem então ser estudadas para que se produza a ordenação final das respostas a serem apresentadas aos usuários.

O processo de expansão de consultas deve sempre levar em conta:

- como selecionar os termos a serem utilizados na expansão;
- quais e quantos termos devem ser utilizados e onde obtê-los.

3.3.1 Expansão Automática de Consulta Baseada em Documentos

Com relação à expansão de consultas a partir da coleção de documentos, existem duas técnicas distintas: a análise global e a análise local. Na global, os

relacionamentos existentes entre os termos são obtidos utilizando-se todos os documentos da coleção. Na análise local, apenas os documentos de topo, aqueles que aparecem melhor colocados na ordenação das respostas, são utilizados para a obtenção de novos termos. Ou seja, assume-se que os documentos do topo são mais importantes. Essa abordagem é denominada pseudo-realimentação de palavras relevantes.

Uma das técnicas de análise global é o agrupamento de termos, através da qual os grupos de termos são encontrados com base na co-ocorrência de termos em todos os documentos da coleção. Tal procedimento é computacionalmente complexo e não produz resultados efetivos [46]. Além disso, para [42], estruturas globais não se adaptam bem ao contexto local de uma consulta, pois correlações existentes na coleção inteira podem não valer no contexto específico de uma consulta, especialmente para coleções genéricas.

3.3.2 Expansão Automática de Consulta Baseada em Ontologia

Em geral, a expansão semântica de consultas baseada em ontologias analisa conceitos e relacionamentos de um domínio específico para adicionar termos à consulta original ou alterar suas condições. Assim, a consulta é modificada com base nos conceitos da ontologia que estejam semanticamente associados aos termos especificados pelo usuário. Para obter tais conceitos, são aplicados mecanismos de raciocínio automático sobre ontologias, permitindo que informações inferidas sejam consideradas durante o processo de expansão.

Segundo [48], o uso de ontologias e mecanismos de inferência no processo de recuperação implica que os resultados recuperados sejam relevantes para a consulta do usuário. Os autores declaram, ainda, que a revocação também é melhorada, uma vez que o raciocínio sobre ontologias possibilita a derivação de novas informações relevantes com base no conhecimento de domínio. Desta forma, os resultados obtidos com o uso de ontologias tendem a ser melhores que aqueles recuperados pela abordagem baseada na co-ocorrência estatística de palavras-chave em documentos do domínio. Diante destas razões, muitos sistemas têm utilizado ontologias para realizar a expansão de consulta, com o intuito de obter resultados mais eficazes.

3.3.3 Expansão Automática de Consulta baseada em Tesauro

Os tesauros, apresentados na Seção 2.1.5, não devem ser confundidos com ontologias. Um tesauro é definido como uma "linguagem documentária caracterizada pela especificidade e complexidade existente no relacionamento entre os termos que comunicam o conhecimento especializado", ao passo que a ontologia é um "modelo de representação do conhecimento que, a exemplo do tesauro, é utilizada para representar e recuperar informação por meio de estruturas conceituais" [69].

A expansão é feita através do acréscimo de termos nos argumentos da consulta por intermédio de um tesauro (sinônimos e palavras derivadas). O processo de adicionar termos e adequar uma pesquisa a um determinado contexto é facilitado com a utilização do tesauro, pois as relações semânticas nele existentes ajudam a encontrar as palavras que melhor representam um determinado conceito. Um sistema de consulta pode, por exemplo, incluir em uma pesquisa os sinônimos e os termos mais específicos das palavras pesquisadas.

Juntamente com a expansão de consulta, ou mesmo com outros métodos de RI, um tesauro pode ser útil para medir a relevância dos documentos encontrados. Assim, os documentos mais relevantes serão exibidos como melhores opções para o usuário.

3.4 Considerações Finais

A expansão de consulta consiste em uma poderosa ferramenta para auxiliar o usuário na recuperação de informação relevante. Quando realizada de forma automática pode garantir resultados ainda mais satisfatórios, pois fica independente do conhecimento do usuário sobre sistemas de recuperação.

Entretanto, para ser realizada de forma automática, a expansão de consulta necessita de alguma estrutura de apoio de onde ela pode inferir conceitos relacionados. Estas estruturas podem ser tesauros, ontologias e até mesmo documentos de uma coleção.

Neste trabalho utilizamos a expansão automática de consulta com o objetivo de complementar a consulta do usuário com termos relevantes e, assim, obter resultados mais satisfatórios na recuperação de páginas da Internet (páginas estas com conteúdo semanticamente semelhante ao que está sendo estudado pelo aluno).

Para tanto, através das técnicas de mineração de textos apresentadas no Capítulo 2, são identificados os termos relevantes de apresentações utilizadas pelo professor para expor conteúdo em sala de aula. Estes termos são utilizados como estrutura de apoio, de onde são obtidos os termos para realização da expansão automática de consulta. Com isso, a probabilidade de recuperação de páginas relevantes é aumentada, mesmo que o usuário tenha informado termos inadequados para consulta.

O capítulo a seguir apresenta algumas ferramentas que, direta ou indiretamente relacionadas, contribuiram como fonte de pesquisa e inspiração para o desenvolvimento deste trabalho.

CAPÍTULO 4

Trabalhos Relacionados

O objetivo deste trabalho é desenvolver uma ferramenta que seja capaz de colaborar com a ampliação do conhecimento dos alunos sobre determinado assunto.

Propomos a utilização de apresentações em forma de slides, com a matéria abordada durante a realização de uma aula, para realizar uma ligação automática entre o conteúdo destes slides e um conjunto de material didático complementar. Esse material servirá de apoio para realização de estudos e/ou desenvolvimento de atividades em sala de aula ou mesmo individualmente.

No decorrer deste trabalho foram encontrados alguns aplicativos que utilizam as apresentações em forma de slides, sendo que uns enfatizam a interação entre professor e aluno em ambientes presenciais, uns favorecem a interação em ambientes de ensino a distância e outros tentam se adaptar ao ensino misto (*blended learning*)[1].

É importante ressaltar que os trabalhos avaliados não possuem o mesmo objetivo do sistema proposto neste estudo. Assim sendo, os sistemas apresentados neste capítulo serviram como fonte de inspiração para o desenvolvimento das idéias relacionadas a este trabalho.

4.1 Sistemas de Gestão da Aprendizagem

Um sistema de gestão da aprendizagem (*Learning Management Systems* - LMS) é um sistema de gestão que possui um conjunto de funcionalidades para promover o aprendizado. Essas funcionalidades relacionam-se com distribuição, acompanhamento, monitoramento e administração de conteúdo de aprendizagem e com o progresso e interações dos alunos, entre outros [44]. O termo LMS pode ser aplicado a sistemas simples de gerenciamento de cursos ou a ambientes distribuídos complexos.

Um LMS tem, como um dos objetivos, simplificar a administração dos programas de treinamento e ensino em uma organização. O sistema auxilia no planejamento

[1] *Blended Learning*: sistema de formação onde a maior parte dos conteúdos é transmitido em curso a distância, normalmente pela Internet, embora inclua necessariamente situações presenciais.

dos processos de aprendizagem e ainda permite que os participantes colaborem entre si através da troca de informações e conhecimentos.

Esses sistemas auxiliam na análise, na disponibilização das informações, no rastreamento de dados, e na geração de relatórios sobre o progresso dos aprendizes. A maioria dos sistemas do tipo LMS não possuem recursos que permitam a rápida e simples criação de conteúdos instrucionais. Este é um dos principais motivos pelos quais a maioria das empresas fornecedoras tem procurado oferecer ferramentas complementares, ou trabalhar com parceiros de conteúdos.

Segundo [89], as principais funcionalidades de sistemas do tipo LMS, são:

- Criar e administrar cursos;
- Oferecer ferramentas de comunicação, tais como lista de discussão, chats e mensagens instantâneas;
- Administrar grades curriculares e listas de espera;
- Fornecer tarefas, avaliações e exercícios;
- Monitorar os acessos do usuário;
- Gerar relatórios e informações sobre o desempenho dos alunos, etc

Os sistemas de gestão de aprendizagem (LMS) permitem que o professor disponibilize material didático para seus alunos. No entanto, é preciso que esse material seja manualmente categorizado ou identificado de forma que o aluno consiga identificar seu conteúdo. Os sistemas LMS possuem mecanismos de busca, porém, grande parte são baseados no modelo booleano, ou seja, busca por palavras chave e não realizam a avaliação da semântica em que está sendo empregado.

Dentre os diversos sistemas de LMS estudados, destacamos: Moodle, Amadeus e TelEduc.

4.1.1 Moodle

O *Modular Object-Oriented Dynamic Learning Environment*[2] (Moodle) é um software livre de apoio à aprendizagem executado num ambiente virtual. Trata-se de um sistema de administração de atividades educacionais destinado à criação de comunidades on-line, em ambientes virtuais de aprendizagem colaborativa.

Os recursos disponíveis no moodle são: disponibilização de material, avaliação do curso, chat, diálogo, diário, fórum, glossário, pesquisa de opinião, questionário, tarefa, Wiki, entre outros.

[2]MOODLE: http://moodle.org

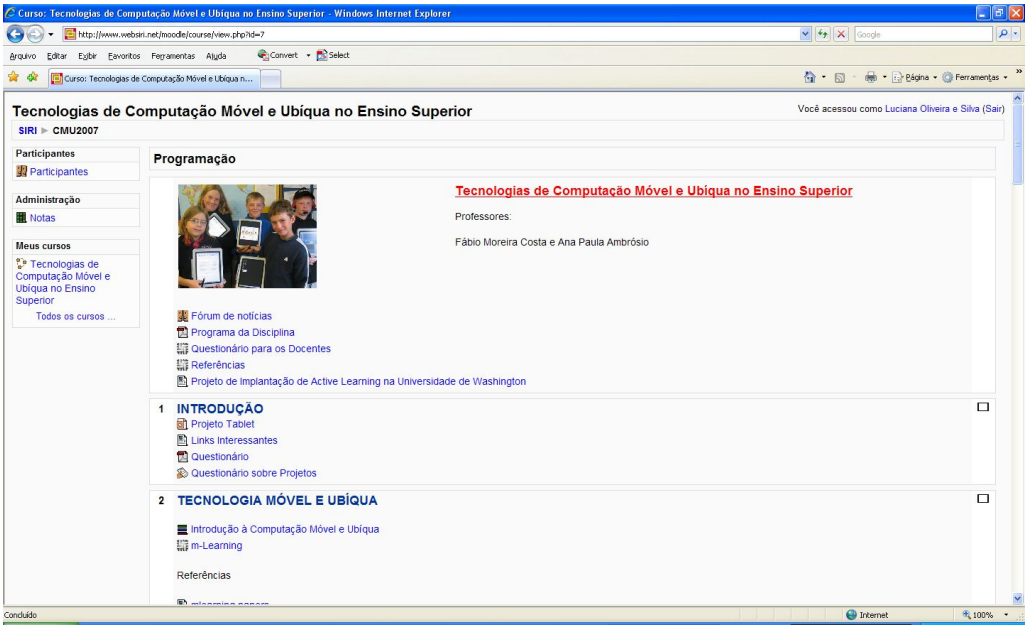

Figura 4.1: *Ambiente de disciplina do Mestrado INF/UFG no Moodle*

Muitas instituições de ensino superior, como o INF/UFG, e centros de formação utilizam o Moodle para cursos totalmente virtuais, para apoio a cursos presenciais, para formação de grupos de estudo, para capacitação de professores e para o desenvolvimento de projetos.

Embora seja um sistema de gerenciamento de curso bastante abrangente, possui certa limitação quando analisamos a recuperação de material didático. O moodle permite que o professor disponibilize inúmeros documentos para consulta por parte dos alunos, porém não possui um mecanismo de recuperação eficiente. A recuperação é realizada através de palavras-chave e cabe ao aluno identificar os documentos ou páginas mais adequados à sua necessidade.

4.1.2 Amadeus

A plataforma livre Amadeus[3] foi desenvolvida pelo Centro de Informática (CIn), da Universidade Federal de Pernambuco (UFPE). Trata-se de uma plataforma de ensino criada para facilitar e difundir o uso da tecnologia por parte de professores e alunos brasileiros. Ela foi idealizada após pesquisas sobre plataformas da mesma categoria (LMS) e análise do uso dessas plataformas no contexto de ensino a distância.

[3] AMADEUS: http://amadeus.cin.ufpe.br

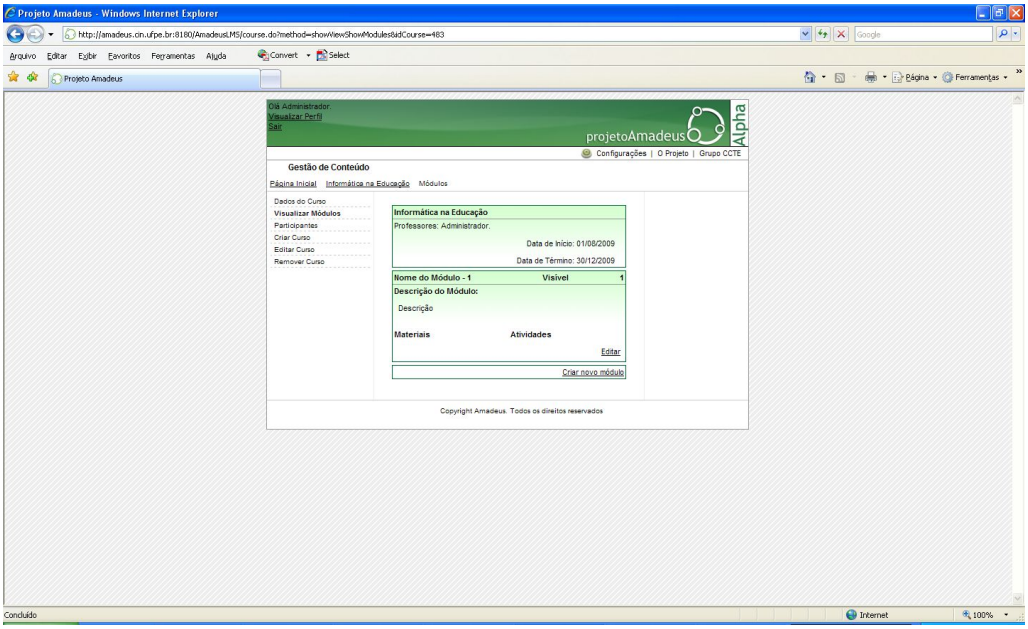

Figura 4.2: *Interface da plataforma Amadeus*

Dentre os recursos que o Amadeus disponibiliza temos: disponibilização de material, avaliação do Curso, variadas formas de avaliação da aprendizagem, chat, fórum, pesquisa de opinião, questionário, SCORM, tarefa e trabalho com revisão, gestão de conteúdos, geração e gestão de questões em Base de Dados, sondagens, glossários e suporte multi-idioma.

Apesar de permitir a disponibilização de material didático por parte do professor, a ferramenta não disponibiliza mecanismos de busca específicos a estes documentos. A busca é realizada em todo o conteúdo do curso.

4.1.3 TelEduc

O TelEduc[4] é um ambiente com licença de software livre para a criação, participação e administração de cursos na Internet.

O TelEduc foi desenvolvido de forma participativa, ou seja, todas as suas ferramentas foram idealizadas, projetadas e depuradas segundo necessidades relatadas por seus usuários. Com isso, ele apresenta características que o diferenciam dos demais ambientes para educação a distância disponíveis no mercado, como a facilidade de uso por pessoas não especialistas em computação, a flexibilidade quanto à forma de usá-lo, e um conjunto enxuto de funcionalidades.

[4]TELEDUC: http://www.teleduc.org.br/

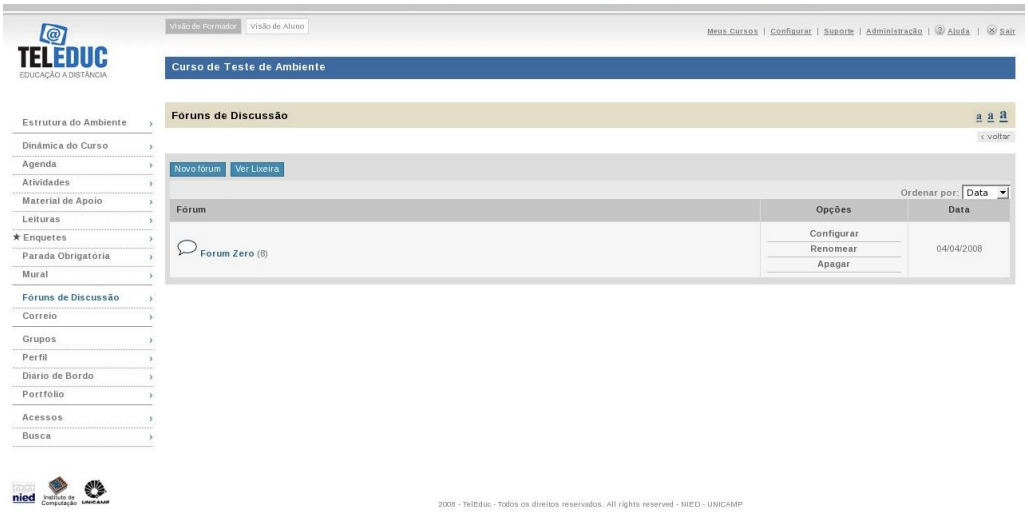

Figura 4.3: *Ambiente de disponibilização de material no TelEduc*

Os recursos disponíveis para o desenvolvimento das atividades no TelEduc são: dinâmica do curso, agenda, perfil dos participantes, mural de avisos, chat, Webfórum, correio eletrônico, material de apoio, leituras, gestão de grupos, diário de bordo, parada obrigatória, portfólio de participantes e grupos, aviso automático por email de mudanças no site, estatísticas de acesso e funções administrativas e de configuração.

Semelhante aos sistemas LMS apresentados anteriormente, possui mecanismo de busca baseado em palavras-chave, não havendo análise do contexto da busca.

4.2 Ferramentas de Apresentação de Slides

A seguir são apresentadas algumas ferramentas avaliadas no decorrer do desenvolvimento deste trabalho. Estas ferramentas objetivam, principalmente, promover a integração de alunos e professores em sala de aula, através do uso de apresentações em forma de slides e de dispositivos de tecnologias móveis (notebooks, palms, Tablet PC, entre outros).

As ferramentas avaliadas focam a troca de informação, a interação. Material didático para consulta deve ser disponibilizado e gerenciado de forma independente do sistema.

4.2.1 Classroom Presenter

O Classroom Presenter (CP)[5] foi desenvolvido pela Universidade de Washington e é o aplicativo que inspirou o desenvolvimento da aplicação proposta nesse trabalho. O CP foi concebido para uso em Tablet PCs e tem como objetivo facilitar o ensino em sala de aula, disponibilizando uma interface em que todos os participantes visualizam o mesmo conteúdo que o participante controlador (mestre) da apresentação [4].

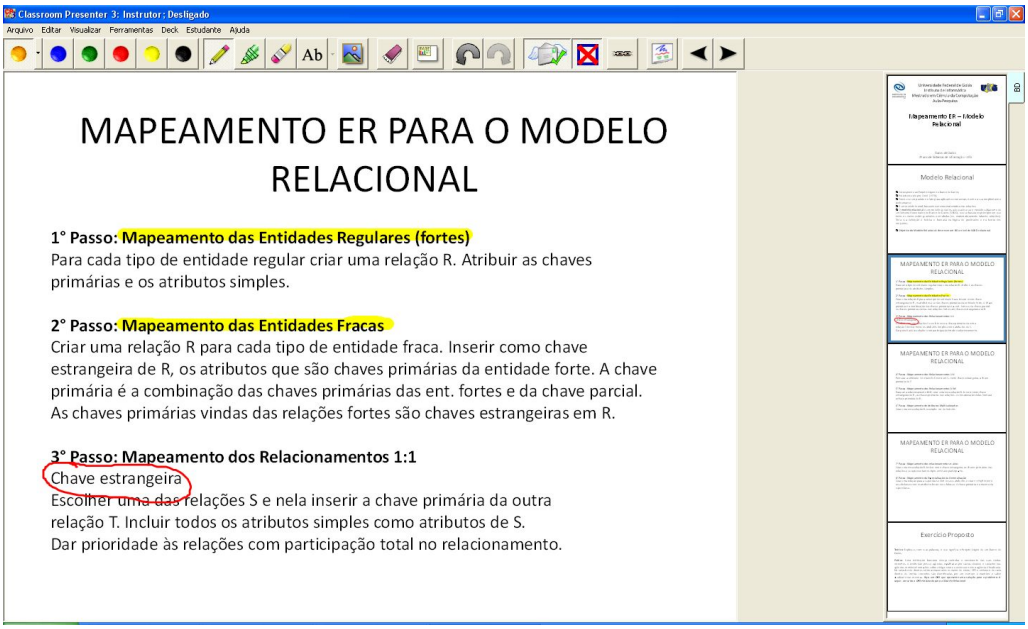

Figura 4.4: *Demonstração da tela do Classroom Presenter*

Todas as anotações realizadas pelo mestre são instantaneamente visualizadas pelos outros participantes. Os outros participantes também podem editar os slides da apresentação em seus tablet PCs e enviar suas colaborações para o mestre, que poderá decidir se irá disponibilizá-la para todos os demais participantes.

Uma vantagem do CP é seu suporte para apresentações do tipo PowerPoint, ferramenta de apresentação comumente utilizada em salas de aula.

Outra vantagem é a série de funcionalidades que o CP possui que melhoram a interatividade com o usuário, como:

- A proposição de perguntas de múltipla-escolha para os participantes. Estas perguntas são visualizadas pelos alunos, que escolhem uma resposta e enviam de volta ao professor;
- A possibilidade de gravar uma sessão, e posteriormente visualizar todas as anotações (contribuições) feitas.

[5]CLASSROOM PRESENTER: http://classroompresenter.cs.washington.edu/

A maior limitação do CP é que esta ferramenta está restrita a notebooks, tablet PCs e computadores desktop, mais especificamente, apenas os dispositivos capazes de executar o .Net Framework. Não existe uma versão para o uso em handhelds ou via Web.

O foco do CP é o trabalho desenvolvido em sala de aula. Não possui mecanismos de busca ou de associação com material para estudo.

4.2.2 Ubiquitous Presenter

As pesquisas sobre o CP continuaram, sendo um dos frutos dessas pesquisas o Ubiquitous Presenter (UP) [6], o qual fornece uma interface Web do CP para os participantes.

As grandes vantagens do UP são:

- Os alunos necessitam somente de um navegador Web (*browser*) e a URL do curso.
- O CP utiliza multicast para a colaboração entre os participantes, enquanto que o UP utiliza um Web service para a comunicação. Esta arquitetura cliente-servidor facilita a implantação e utilização do UP, se comparado com o CP, visto que este necessita de instalação e depende de multicast.

Apesar de ser um sistema Web, não tem foco na recuperação de informação, o foco é a interação entre os participantes de um curso de ensino a distância.

4.2.3 iPH

O *Interactive Presenter for Handhelds* (iPH)[7] foi desenvolvido dentro do projeto *Active Classroom Project*, mantido pelo *Laboratory for Advanced Collaboration* (LAC) da Pontifícia Universidade Católica do Rio de Janeiro (PUC/Rio) [66].

O iPH é uma aplicação que suporta o compartilhamento e a co-edição de apresentações entre professor e alunos em uma sala de aula. O iPH pode ser executado em diferentes tipos de dispositivos como: Tablet PCs, notebooks e handhelds. Segundo [66], seu grande diferencial é que se trata de uma ferramenta sensível ao contexto, ou seja, ele também pode acessar informações de contexto computacional do dispositivo (por exemplo, nível de energia do dispositivo, memória disponível, qualidade do enlace da conexão sem fio, entre outras) e realizar auto-adaptações de modo à melhorar a interação com o usuário.

[6] UBIQUITOUS PRESENTER: http://up.soe.ucsc.edu/
[7] IPH: http://www.lac.inf.puc-rio.br/iph/

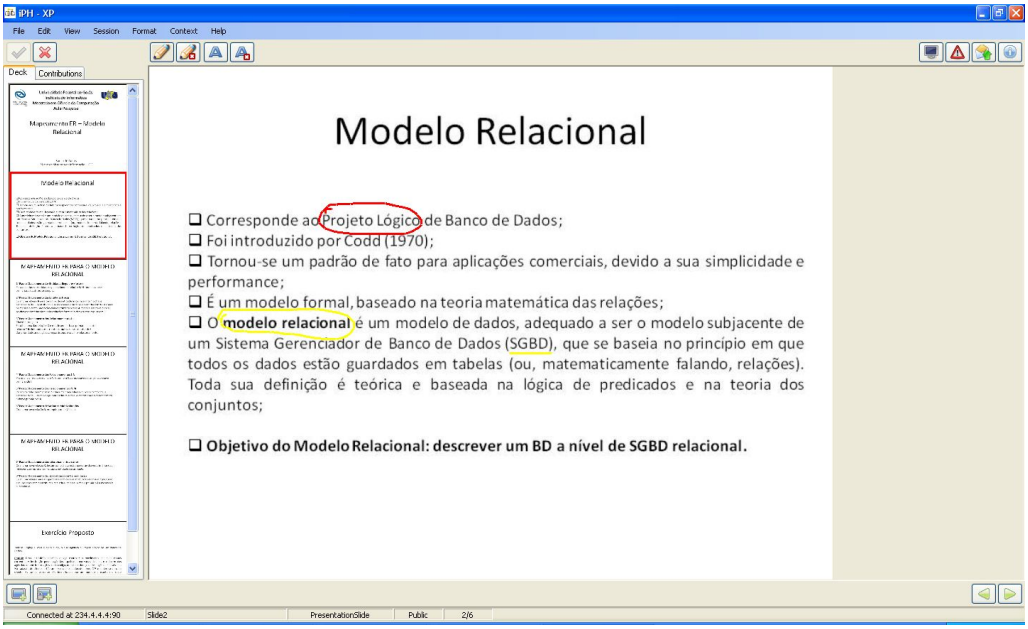

Figura 4.5: *Demonstração de funcionalidades do iPH como ambiente colaborativo*

Não existe sistema de busca, sendo o objetivo principal desta ferramenta a interação entre alunos e professor através da troca de informação através da edição de slides.

4.2.4 Virtual Multiboard

O Virtual Multiboard (VMB) é um sistema que permite o gerenciamento das aplicações utilizadas em uma apresentação, e funcionalidades de anotação e gravação de uma aula. Ao contrário das outras aplicações que fazem uso diretamente das apresentações PowerPoint, exportando os quadros destas a um formato próprio para utilização, o VMB executa paralelamente uma apresentação PowerPoint, ou qualquer outro tipo de apresentação. Devido a este desacoplamento das apresentações utilizadas, as anotações feitas durante a apresentação são inseridas em um painel transparente (glass pane) do próprio VMB.

A partir de seus notebooks ou palmtops, os participantes podem enviar perguntas ao controlador e participar de questionários eletrônicos. O VMB oferece suporte à visualização de múltiplos quadros de uma apresentação simultaneamente, pois considera que os participantes de uma apresentação podem perder a linha de aprendizado caso visualizem sempre um único quadro. Foram desenvolvidas aulas e palestras onde os últimos dois quadros da apresentação eram exibidos juntamente com o quadro atual, ajudando os participantes a entenderem melhor o conteúdo da apresentação.

4.2.5 Tablet Mylar Slides

O uso do Tablet Mylar Slides (TMS)[8] é semelhante à maneira que o PowerPoint é utilizado nas salas de aula: o professor utiliza um computador conectado a um projetor, que exibe a apresentação para os participantes da aula.

O diferencial do TMS para o PowerPoint são: o tamanho dos slides que, no TMS, pode ser infinita, facilitando a contribuição do professor, visto que este sempre terá espaço para fazer suas anotações; a exportação da apresentação editada para formato HTML, de modo que possa ser distribuída aos participantes que estiveram ausentes; e o conceito de caneta secreta (*hidden ink*). Com esta caneta o professor realiza suas anotações e estas não são exibidas pelo projetor.

4.3 Considerações Finais

Os estudos realizados levam a conclusão de que os sistemas de apoio a aprendizagem possuem focos distintos: alguns sistemas auxiliam na organização e no gerenciamento de cursos, enquanto outros facilitam a interação entre alunos e professores, seja em ambiente presencial ou de ensino à distância.

No entanto, nenhum dos sistemas avaliados leva em consideração a necessidade dos alunos de buscar novos conhecimentos ou mesmo de ampliar seu conhecimento sobre determinado assunto. Os sistemas de gestão de aprendizagem permitem aos professores disponibilizar material didático, mas estes não possuem a capacidade de identificar o contexto em que o aluno está inserido como forma de contextualizar a busca e obter resultados mais eficientes.

No capítulo a seguir apresentamos a ferramenta proposta para solucionar essa defasagem dos sistemas de interação: a capacidade de identificar a semântica dos slides e realizar a ligação destes slides a um conjunto de material didático, de forma que estes fiquem à disposição para estudo por parte dos alunos.

[8]TABLET MYLAR SLIDES: http://www.cs.umd.edu/ egolub/TabletMylarSlides/

CAPÍTULO 5

A Ferramenta BOOKISH

Como forma de validação desta dissertação, este capítulo apresenta uma ferramenta, denominada BOOKISH [1]. Conforme descrito na Seção 1.5, o objetivo proposto neste trabalho é utilizar técnicas de mineração de textos e expansão automática de consulta para contextualizar documentos.

O principal objetivo do BOOKISH é, através da análise de apresentações em forma de slides, identificar o assunto que está sendo abordado em sala de aula e realizar a ligação direta destas apresentações com um conjunto de materiais didáticos relevantes para que o aluno possa aprofundar seu conhecimento sobre o assunto e desenvolver tarefas propostas pelo professor, entre outras atividades.

A Figura 5.1 ilustra, em alto nível, o funcionamento do BOOKISH. A ferramenta inicia sua atuação sobre apresentações do Microsoft PowerPoint (em formato .ppt e .pptx), utilizadas pelo professor para exposição de conteúdo. Essas apresentações devem ser disponibilizadas pelo professor através da rede ou outro recurso disponível.

Como ilustrado pela Figura 5.1, as atividades que o BOOKISH realiza podem ser divididas em: *processamento da apresentação*, *processamento de documentos* e *processamento de consultas*, descritas em detalhes nas próximas seções deste capítulo.

As técnicas de mineração de textos e de indexação semântica latente utilizadas nesta ferramenta foram totalmente implementadas para o BOOKISH, não sendo utilizada nenhuma biblioteca específica para estes objetivos.

Para fins de comprovação de eficiência da abordagem proposta, foram realizados experimentos com dois grupos de alunos. Os resultados desses testes são apresentados no Capítulo 6.

[1] O termo BOOKISH refere-se ao termo no idioma inglês cujo significado é “estudioso”. Trata-se de uma alusão aos objetivos do sistema, no sentido de facilitar o acesso do aluno a um conjunto de materiais didáticos, ampliando o domínio sobre o assunto que está sendo estudado.

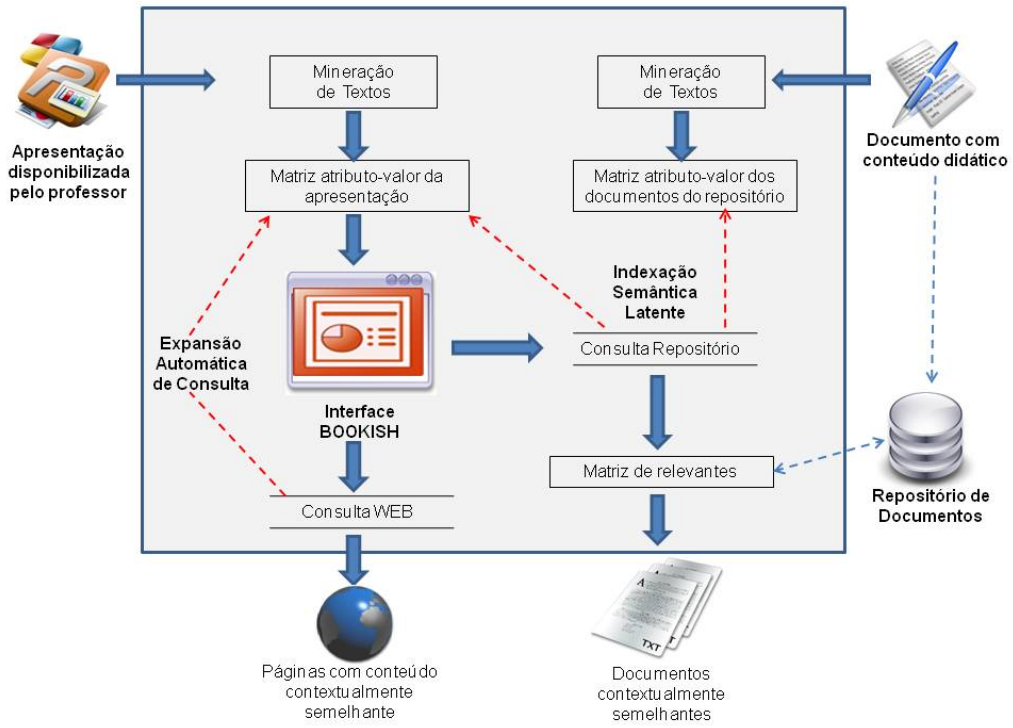

Figura 5.1: *Visão dos processos executados pelo BOOKISH*

5.1 Tecnologias Utilizadas na Implementação

A decisão por utilização de apresentações no formato Microsoft PowerPoint foi considerada adequada por esta se tratar de uma das ferramentas mais utilizadas pelos professores. Este formato também foi considerado por ser o padrão adotado pelas ferramentas Classroom Presenter e Ubiquitous Presenter com as quais esperamos, no futuro, fazer uma integração.

Por esse motivo, as soluções tecnológicas selecionadas para implementar a proposta foram:

- **Microsoft .Net Framework** [2]: Utilizada para desenvolver a ferramenta apresentada neste trabalho, foi selecionada por permitir, de forma nativa, a programação de add-in[3] para o PowerPoint. O .Net é uma iniciativa da Microsoft que visa ser uma plataforma única para desenvolvimento e execução de sistemas e aplicações. Isto quer dizer que todo e qualquer código gerado pode ser executado em qualquer dispositivo que possua o framework da plataforma. A idéia é semelhante à da

[2] Mais informações em: http://msdn.microsoft.com

[3] Um add-in permite aos desenvolvedores estenderem as funcionalidades de alguns de seus aplicativos, neste caso aqueles pertencentes ao pacote Microsoft Office. Um add-in permite a alteração dos objetos da aplicação: personalização de eventos, componentes e código, por exemplo.

plataforma Java, onde o código deixa de ser escrito para um dispositivo específico, podendo ser executado em qualquer ambiente que possua o framework. Por esse motivo, não é necessária nenhuma licença especial para execução de aplicativos desenvolvidos com a tecnologia.

- **C#**: linguagem de programação orientada a objetos desenvolvida pela Microsoft como parte da plataforma .Net. Sua sintaxe foi baseada na linguagem C++, mas inclui diversas influências de outras linguagens, como Java. A linguagem C# foi criada junto com a arquitetura .Net. Embora existam várias outras linguagens que suportam essa tecnologia, C# é considerada a linguagem símbolo do .Net. Como algumas razões para isto, podemos citar: a linguagem foi criada praticamente do zero para funcionar na nova plataforma, sem preocupações de compatibilidade com código legado e também porque a maior parte das classes do .Net Framework foram desenvolvidas em C#. Esta linguagem foi selecionada por ser nativa ao .Net Framework.
- **SQLite**: O SQLite[4] é um software gratuito, multiplataforma, desenvolvido em C padrão (ANSI) e pode ser definido como uma ferramenta - mais precisamente, uma biblioteca – que pode ser integrada a programas escritos em diferentes linguagens com o intuito de possibilitar a manipulação de dados através de instruções SQL. O SQLite funciona como um "mini-SGBD", capaz de criar um arquivo em disco e ler e escrever diretamente sobre este arquivo. Foi utilizada neste projeto como SBGD, armazenando todas as estruturas relevantes dos documentos, slides, stopwords, entre outros.

5.2 Funcionamento da Ferramenta

Ao iniciar a execução do BOOKISH, o aluno deve selecionar a apresentação que será utilizada. A ferramenta aceita como entrada arquivos em formato .ppt, .pps e .pptx, isto é, apresentações feitas através do Microsoft PowerPoint versão 2003 ou 2007.

Ao abrir a apresentação, o BOOKISH realiza operações de mineração de textos sobre os slides, com o objetivo de identificar os termos mais relevantes, isto é, os termos que representam com mais propriedade o(s) assunto(s) abordado(s) pela apresentação. Esses termos mais representativos constituem a matriz representativa da apresentação e servirão como base para a realização das atividades posteriores da ferramenta, conforme ilustrado pela Tabela 5.1.

[4]SQLite: http://www.sqlite.org

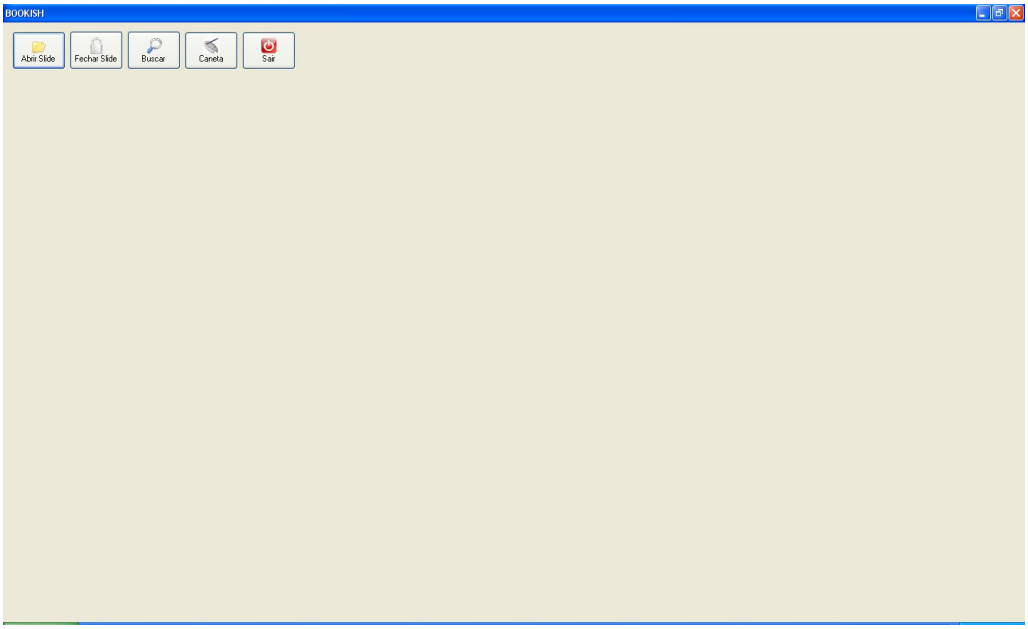

Figura 5.2: *Tela inicial do BOOKISH*

Tabela 5.1: *Exemplo de matriz representativa de uma apresentação*

termo	peso
java	2
orientação	1
...	...
polimorfismo	2

A partir da identificação do "assunto" da apresentação, o BOOKISH disponibiliza ao usuário a opção de buscar material didático na Internet ou em repositório de documentos. A interface do BOOKISH apresenta todos os termos relevantes da apresentação em forma de *hyperlinks*, sendo que estes, quando selecionados, redirecionam para um site de busca, conforme descrito na Seção 5.2.3.1.

Caso seja opção do usuário realizar a busca de material didático no repositório, o BOOKISH utiliza a indexação semântica latente para realizar a comparação entre a matriz representativa da apresentação e a matriz representativa do conjunto de documentos do repositório para calcular e recuperar os documentos com maior similaridade.

O repositório acadêmico é, na verdade, um diretório compartilhado na rede, com acesso de escrita restrito ao BOOKISH e acesso de leitura aos demais usuários da rede. Neste diretório serão armazenados todos os arquivos considerados interessantes pelos professores, independente do assunto abordado, formando, ao longo do tempo uma base multidisciplinar de arquivos (tutoriais, apostilas, artigos, ebooks, entre outros).

Todo documento deve ser inserido a partir da ferramenta. Assim, uma vez que o BOOKISH analisa e indexa cada um dos documentos do repositório, ela será capaz de

realizar uma comparação mais eficiente e contextualizada de documentos. Além disso, o fato de cada documento ser individualmente analisado pela ferramenta, garante que os alunos tenham acesso à base sempre atualizada, sem que haja a necessidade de atualizações no arquivo ou nova indexação por parte do professor.

5.2.1 Processamento da Apresentação

Quando uma apresentação é selecionada, o BOOKISH faz a análise de todo o conteúdo da apresentação com o objetivo de identificar, dentre os elementos textuais presentes, aqueles de maior relevância.

Para processamento das apresentações, foi desenvolvido um *add-in* para o Microsoft PowerPoint, seguindo as especificações da MSDN Library para o PowerPoint. Foram utilizadas as bibliotecas específicas do PowerPoint, incluídas no pacote Microsoft.Office.Interop.PowerPoint[5].

As funcionalidades do add-in são implementadas dentro dos tratadores de eventos **ThisAddIn_Startup** e **ThisAddIn_Shutdown**, localizados na classe **ThisAddIn.cs**. No evento ThisAddIn_Startup, devem ser inseridas as atividades a serem realizadas durante a inicialização do PowerPoint. O evento ThisAddIn_Shutdown trata todas as atividades que devem ser executadas quando o PowerPoint for encerrado.

No evento ThisAddIn_Startup, foi configurado o evento Application.PresentationOpen, que inclui as atividades a serem realizadas quanto uma apresentação é aberta no PowerPoint. O algoritmo que representa as atividades realizadas pelo evento Application.PresentationOpen é apresentado no Código 5.1.

[5]Microsoft PowerPoint para desenvolvedores disponível em http://msdn.microsoft.com/en-us/library/bb265982.aspx

Algoritmo 5.1: Algoritmo de extração de relevantes das apresentações

Entrada: Arquivo de slides

Saída: Matriz atributo-valor da apresentação

```
1  foreach slide da apresentação do
2      foreach componente da apresentação do
3          if É componente do tipo texto then
4              foreach palavra do componente do
5                  if Converte para minúsculo (case folding) then
6                      if Não é stopword then
7                          Adiciona termo à matriz termo-valor
8                      end
9                  end
10             end
11         end
12     end
13 end
14 Calcula peso de cada termo da matriz termo-valor
15 Ordena matriz atributo-valor através do peso calculado
16 Aplica ponto de corte de Luhn
17 foreach Elemento da nova matriz termo-valor do
18     Converte termo para hyperlink;
19 end
```

Para processamento das apresentações, selecionamos algumas atividades da mineração de textos: a conversão de maiúsculas em minúsculas (*case folding*) e a eliminação de *stopwords*. Após o processo de mineração, o BOOKISH gera a matriz termo-valor, com a seleção de termos que melhor representam o arquivo analisado.

O cálculo do peso de cada termo é obtido através da fórmula:

$$v = P_B + P_{TF}$$

Onde:

- P_B - representa o peso do termo, atribuído pela significância estimada do termo na apresentação, conforme Tabela 5.2;
- P_{TF} - representa o peso do termo, calculado através do método TF.

Como a métrica TF estabelece valores dentro do intervalo [0,1] e os termos representativos da apresentação recebem uma ponderação dentro do intervalo [0,2], os

Tabela 5.2: *Tabela de ponderação de termos pela significância na apresentação*

Característica do Termo	Peso
Título	2
Negrito, Itálico ou Sublinhado	1
Texto simples (sem formatação)	0

termos relevantes da apresentação podem ser ponderados dentro do intervalo [0,3], onde valores maiores representam termos mais significativos.

Conforme o conceito de Luhn, apresentado na Seção 2.1.6.5, podem ser utilizados métodos para identificar termos relevantes cuja presença não é determinante para indexação de um documento. Dessa forma, podem ser aplicados cortes de termos dentro dos limites inferior e superior da matriz representativa.

Para garantir que todas as apresentações tenham uma análise semelhante, independente do número de termos relevantes presentes, a ferramenta adota uma abordagem onde a quantidade inicial de elementos da matriz determina o ponto de corte. Assim, seja i o número total de termos na matriz, serão eliminados os $i/10$ primeiros elementos e os $i/10$ últimos elementos, o que, na prática, representa a eliminação de 10% dos elementos iniciais e 10% dos elementos finais do vetor. Esses valores foram determinados após vários testes, sendo os que apresentaram melhor resultado de eliminação de termos não-discriminantes.

Assim, após a execução do corte, a ferramenta atualiza a matriz termo-valor com a representação final dos termos mais significativos da apresentação. Essa tabela está ilustrada em 5.3, onde as linhas representam os termos relevantes identificados após aplicação do ponto de corte e as colunas guardam o valor v_i, representando o peso do termo i na apresentação.

Tabela 5.3: *Matriz termo-valor da apresentação analisada*

t_1	v_1
t_2	v_1
...	...
t_n	v_n

A matriz resultante dessa operação será utilizada como referência para a realização da expansão automática de consulta, onde os termos adicionados à consulta original do usuário são os termos presentes na matriz, ordenados pela relevância.

A interface do BOOKISH exibe para o usuário a apresentação onde os termos considerados como relevantes, estão no formato de *hyperlink*. Essa interface está ilustrada na Figura 5.3.

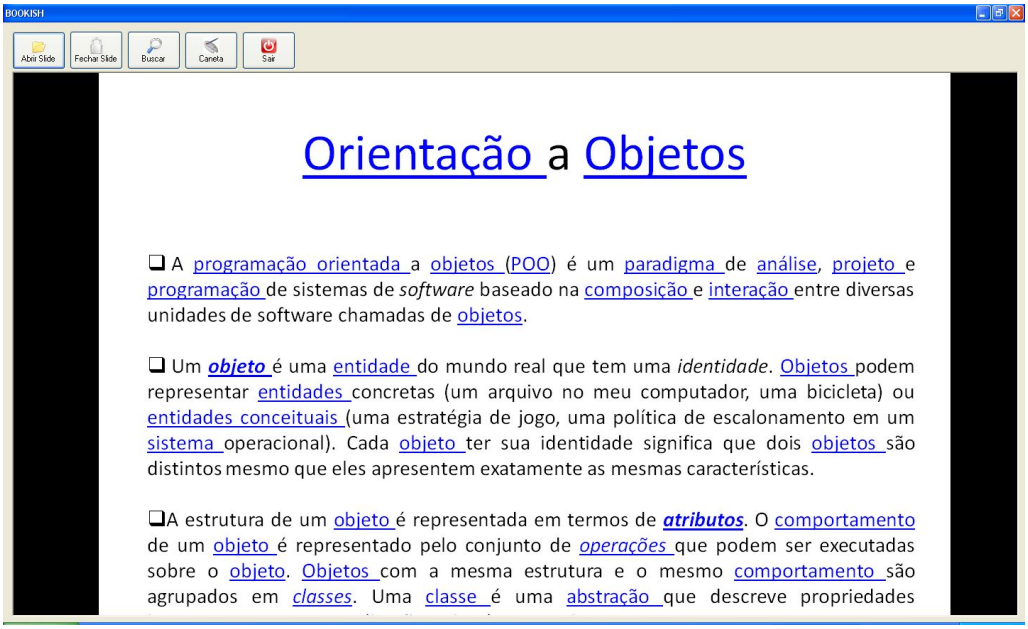

Figura 5.3: *Exemplo de identificação de relevantes na apresentação*

A conversão automática dos termos relevantes para *hyperlinks* é realizada com o objetivo de facilitar a navegação do usuário entre a ferramenta e a Internet. A forma como a ferramenta atua na busca de informação na Internet está descrita na Seção 5.2.3.

É importante ressaltar que a matriz termo-valor que representa a apresentação é armazenada em banco de dados, esta matriz permanece salva enquanto a apresentação está ativa. Assim que o usuário fecha a apresentação, o conteúdo desta tabela é apagado.

5.2.2 Processamento de Documentos

Uma das atividades essenciais para o bom funcionamento da ferramenta é a alimentação do repositório de documentos. Um repositório formado por uma boa base de documentos garantirá uma boa fonte de consulta à disposição dos alunos em suas interações com o sistema.

Um documento, para ser disponibilizado para os alunos, deve ser inserido no repositório obrigatóriamente através do BOOKISH. Cada documento, ao ser inserido no repositório, passa por processos básicos da mineração de textos, como case folding e eliminação de stopwords. Os termos resultantes desse processo formam a matriz termo-documento que representa a coleção de documentos, ilustrada na Tabela 5.4.

Nesta tabela, estão representadas a coleção de documentos D = $d_1, d_2, \ldots, d_n$ e o conjunto dos termos relevantes distintos presentes em D, T = $t_1, t_2, .., t_m$. O valor v_{jk} representa o peso do termo t_k em relação ao documento d_j. Para quantificar essa importância, foi utilizado a métrica TF-IDF, que trabalha com a comparação do cálculo

da freqüência relativa de palavras em um documento na proporção inversa às palavras contidas no conjunto de documentos. Essa técnica de ponderação está descrita na Seção 2.1.6.2.

Tabela 5.4: *Matriz termo-documento representado os documentos da coleção*

	d_1	...	d_k	...	d_m
t_1	v_{11}	...	v_{1k}	...	v_{1m}
...	...	...	...	...	...
t_j	v_{j1}	...	v_{jk}	...	v_{jm}
...	...	...	...	...	...
t_n	v_{n1}	...	v_{nk}	...	v_{nm}

Cada vez que um documento é inserido no repositório, a matriz termo-documento A é atualizada. A cada atualização da matriz, a ferramenta chama o método responsável pela atividade de indexação semântica latente, que irá gerar a matriz reduzida de A, conforme detalhado na Seção 2.2.4. E esta matriz reduzida A´ é utilizada especificamente para cálculo de similaridade entre a apresentação ativa e os documentos do repositório.

Tanto a estrutura das matrizes A e A´ são armazenadas em banco de dados para processamento fututo, sendo que a matriz A reflete a estrutura completa dos termos e documentos do repositório e a matriz reduzida A´ será utilizada para cálculo de similaridade.

A execução do algoritmo de LSI é da ordem de O(n^3). Este custo pode ser considerado alto, levando-se em consideração que documentos são compostos de um número elevado de termos.

5.2.3 Processamento da Consulta

A partir do processamento dos slides de uma apresentação, o BOOKISH identifica seus termos mais significativos. Com isso, o usuário pode utilizar a ferramenta para realizar consulta de material de estudo complementar. Para isto, o BOOKISH permite que as consultas sejam realizadas na Internet ou no repositório de documentos.

5.2.3.1 Consulta à Internet

A consulta à internet pode ser iniciada com o clique do usuário em um dos termos relevantes, destacados como *hyperlinks* na apresentação. Para obtenção de melhores resultados, a ferramenta utiliza a técnica de expansão automática de consulta para complementar a consulta original do usuário. Assim, ao termo selecionado pelo usuário são

acrescidos os termos presentes na matriz termo-valor (obtida pela execução das atividades descritas em 5.2.1).

Após o clique, o equipamento do usuário executa automaticamente o navegador padrão, redirecionando-o para o site de busca Google[6]. O Google foi uma escolha natural, em virtude de sua popularidade e reconhecida eficiência na localização de informações, pois utiliza técnicas sofisticadas de indexação e está em constante evolução.

As Figuras 5.4 e 5.6 ilustram a realização da consulta de um termo na Internet. A Figura 5.4 exibe parte de uma apresentação escrita para uma aula de programação, onde o tópico abordado foi "introdução à orientação a objetos com ênfase em Java". A partir do slide exibido, o usuário clica, por exemplo, sobre o termo "polimorfismo" com o objetivo de realizar uma busca na Internet. A Figura 5.6 demonstra o resultado dessa busca, onde são exibidos os termos selecionados para a expansão automática de consulta, assim como os resultados obtidos através dessa busca expandida.

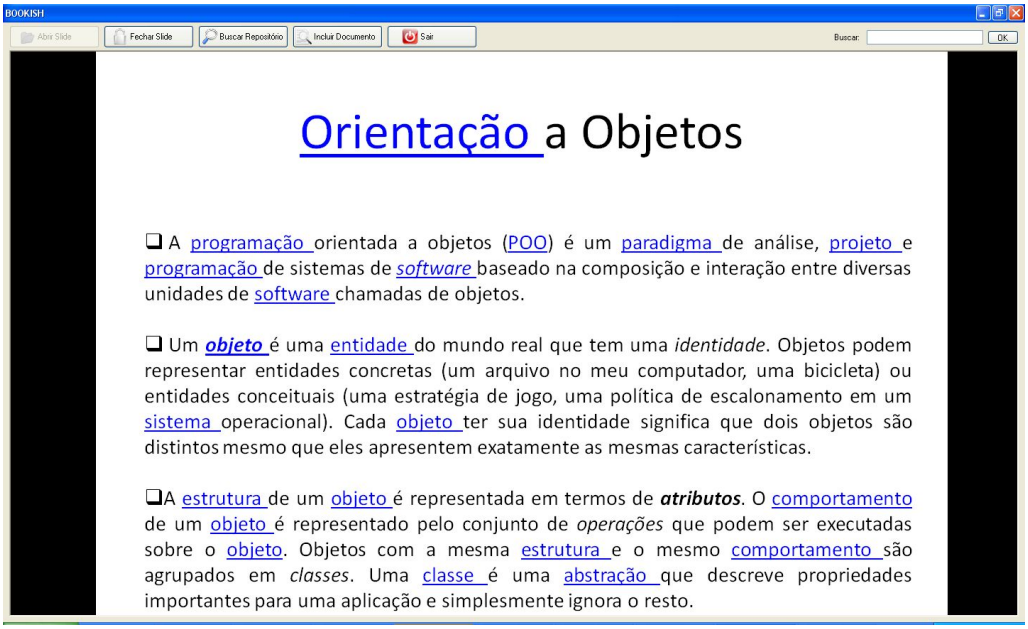

Figura 5.4: *Exemplo de termos identificados como relevantes pelo BOOKISH*

[6]http://www.google.com.br

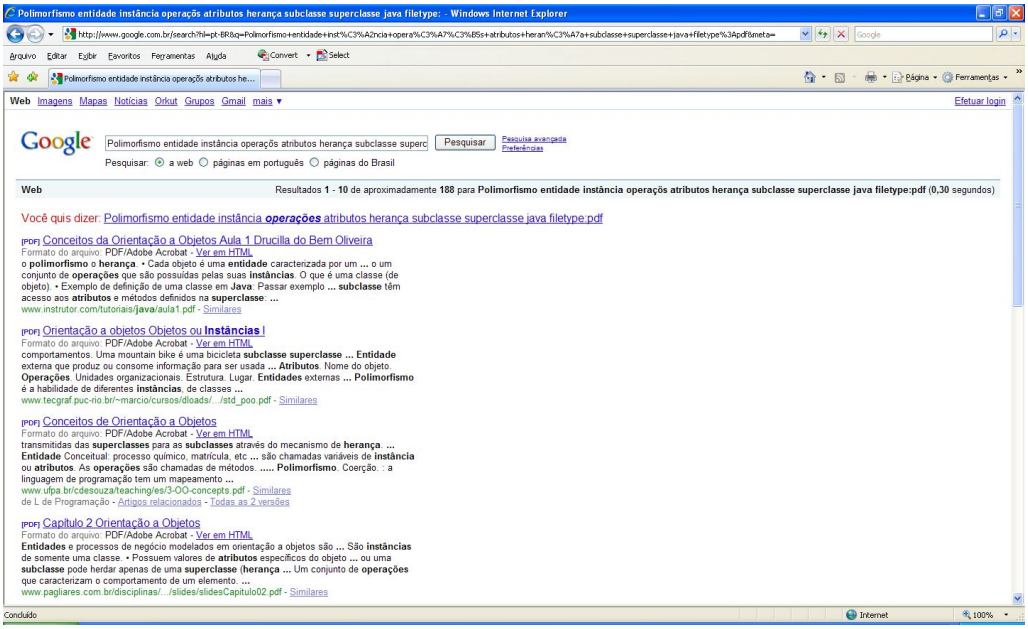

Figura 5.5: *Exemplo da recuperação com expansão automática de consulta na internet, a partir da ferramenta*

Além da pesquisa através dos *hyperlinks*, a ferramenta permite que o usuário digite um termo à sua escolha, conforme demonstrado na figura XXX. Esse termo será acrescido dos termos relevantes da apresentação e a busca redirecionada para o site da Google. Essa opção dá mais flexibilidade à consulta do usuário, já que os *hyperlinks* estão limitados ao vocabulário dos slides.

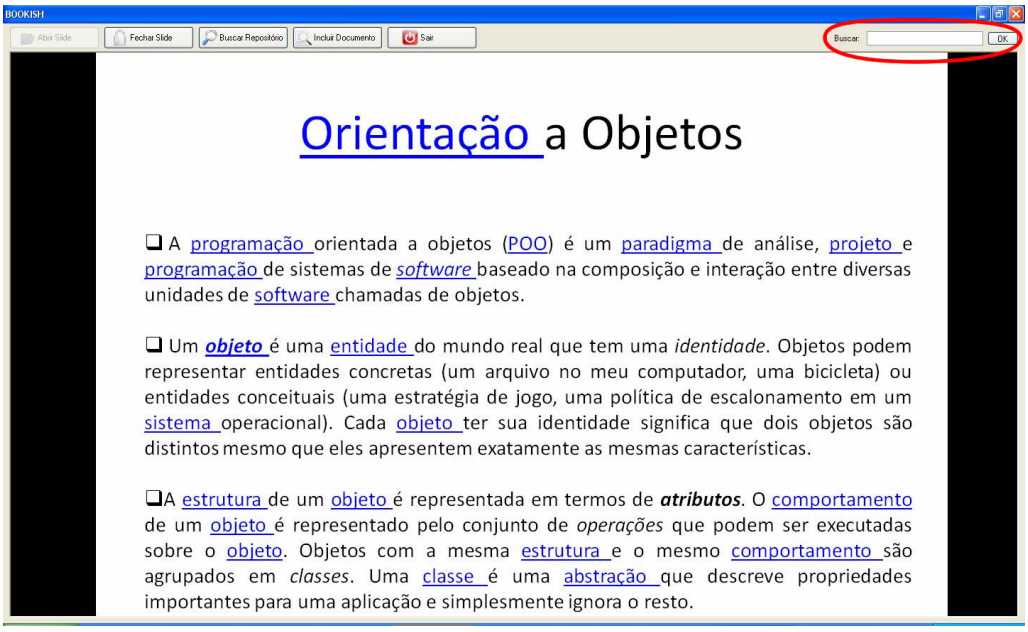

Figura 5.6: *Demonstração da caixa de digitação de termos, para consulta personalizada à Internet*

A utilização da expansão automática de consulta para recuperação de páginas na internet foi considerada bastante positiva. Os termos expandidos conseguiram fornecer maior riqueza de detalhes ao site de busca, recuperando conteúdos contextualmente semelhantes aos conteúdos das apresentações.

5.2.3.2 Consulta ao Repositório

Conforme apresentado na Seção 5.2.2, os documentos inseridos no repositório são automaticamente indexados, de forma que o material disponibilizado ao aluno estará sempre atualizado em relação à base.

Para testar a viabilidade do método para a finalidade proposta pela ferramenta, o BOOKISH utiliza a indexação semântica latente para estabelecer a co-relação entre os documentos da coleção e a apresentação analisada.

Como a ferramenta está programada para gerar a matriz reduzida de A toda vez que um novo documento for inserido no repositório, a matriz A´ está sempre atualizada em relação a base e pode ser utilizada para o cálculo de similaridade.

Para realizar a busca ao repositório, ao invés solicitar ao usuário a digitação de termos, a ferramenta utiliza os termos presentes na matriz representativa da apresentação como elementos de consulta.

Conforme apresentado na Seção 2.2.4, a similaridade entre um documento d e uma apresentação q é determinada através do modelo vetorial, expressa pela correlação entre os vetores d e q. De fato, é realizado o cálculo do cosseno do ângulo formado por estes dois vetores e é expressa pela equação:

$$\cos v_j = \frac{(A_k\, e_j)^T\, q}{\|(A_k\, e_j)\|_2\, \|q_2\|}$$

Após o cálculo é retornada uma lista com os 10 documentos com maior similaridade, em ordem descrescente.

As Figuras 5.7 e 5.8 apresentam um exemplo de recuperação de documentos de um repositório com a utilização da ferramenta. A Figura 5.4 exibe parte de uma apresentação escrita para uma aula de programação, onde o tópico abordado foi “introdução à orientação a objetos com ênfase em Java”.

A partir de qualquer página dessa apresentação, quando o usuário clica sobre o botão “Buscar Repositório”, o processo de comparação é iniciado. Os documentos mais relevantes encontrados pelo BOOKISH são exibidos em uma lista para o usuário, em ordem decrescente de grau de similaridade. Caso o usuário queira consultar o conteúdo do material indicado, ele deve selecionar o nome do arquivo na lista e clicar sobre o botão “Abrir Documento”.

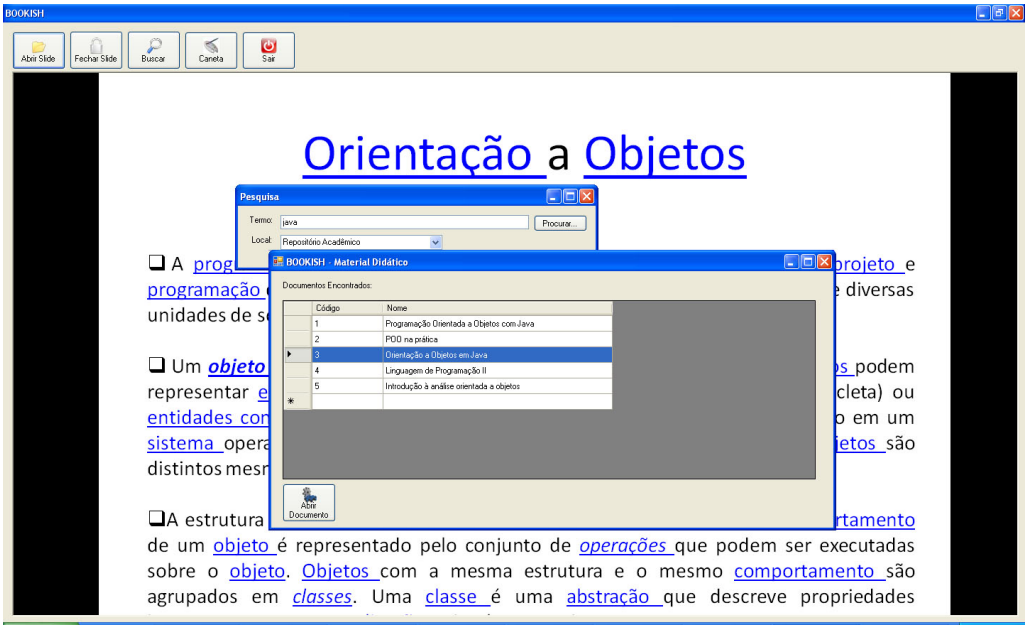

Figura 5.7: *Exemplo de lista de arquivos recuperados pelo BOOKISH*

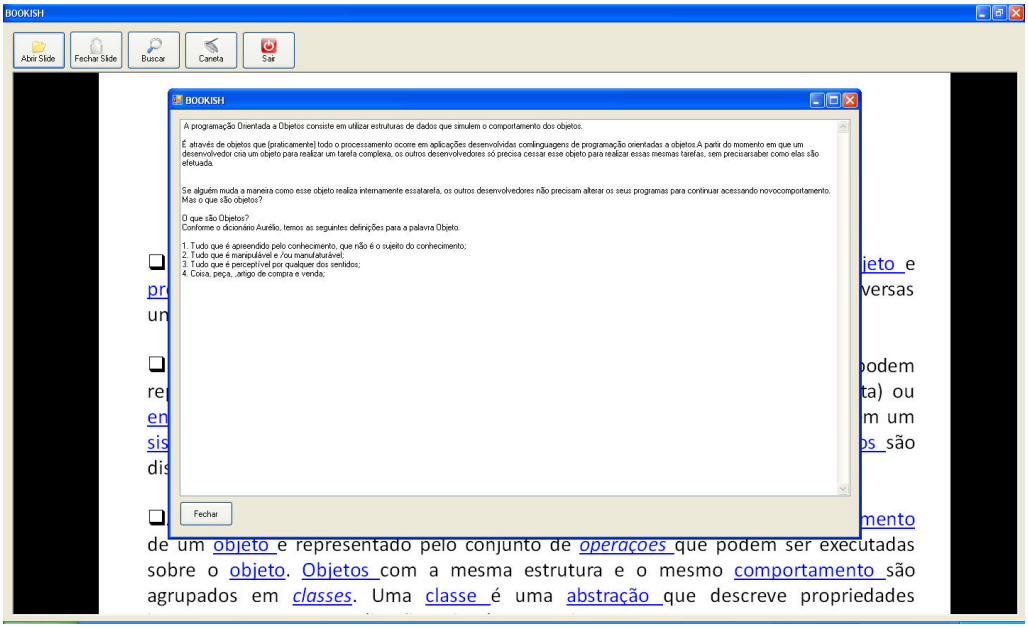

Figura 5.8: *Conteúdo de um dos arquivos recuperados pelo BOOKISH*

5.3 Considerações Finais

A expansão automática de consulta é uma técnica considerada eficiente por alguns autores . Com o apoio de uma estrutura auxiliar, no caso deste trabalho a utilização

de uma matriz representativa de termos relevantes da apresentação que está sendo exibida, demonstrou ser uma técnica bastante eficiente. O seu processamento levou à resultados satisfatórios a baixo custo de processamento.

Já a indexação semântica latente para recuperação de documentos demonstrou em nossos testes ser uma técnica eficiente. Apesar de seus resultados satisfatórios cabe ressaltar o inconveniente de seu alto custo de processamento.

De um modo geral, a forma como a ferramenta foi implementada, garantiu a recuperação de documentos semanticamente semelhantes ao conteúdo abordado pelos slides.

O capítulo a seguir apresenta como a ferramenta foi testada, assim como os resultados obtidos na fase de testes.

CAPÍTULO 6

Resultados

Neste capítulo, são apresentados os resultados obtidos na avaliação experimental do protótipo BOOKISH. Foi realizada uma avaliação qualitativa da ferramenta, com o objetivo de verificar se a abordagem adotada é realmente eficaz no auxílio à recuperação de material didático por parte dos alunos.

Deseja-se, sobretudo, verificar se a abordagem, através da ferramenta implementada, oferece:

- capacidade de identificação do contexto dos slides;
- recuperação rápida e precisa de documentos relevantes e dentro do contexto dos slides;
- recuperação rápida e precisa de páginas da Internet relevantes e dentro do contexto dos slides;
- manipulação intuitiva e de fácil compreensão.

Além dos testes em laboratório, realizados durante a implementação da ferramenta, foram realizados experimentos práticos em sala de aula, com a utilização do BOOKISH por grupos distintos de alunos e o preenchimento de questionário de avaliação. Os procedimentos para os experimentos estão descritos com mais detalhes a seguir.

6.1 Cenário dos Testes

A avaliação realizou-se por meio de experiência em laboratório. Visando atingir diferentes níveis, experiências, expectativas e necessidades, os testes foram realizados com alunos de 2 turmas do curso de Sistemas de Informação da Universidade Estadual de Goiás - Unidade Itaberaí, sendo uma turma da disciplina de Banco de Dados (2ª série) e outra turma da disciplina de Programação II (3ª série).

O ambiente de simulação foi montado de acordo com os pré-requisitos de funcionamento do BOOKISH: computadores com acesso a Internet e a um repositório de documentos.

O repositório de testes foi previamente montado de forma a conter documentos referentes a cada área específica de estudo das turmas, entre outros documentos de diversas áreas de conhecimento. Foram realizados experimentos sobre um conjunto de 60 arquivos no formato textual (.txt) e também sobre o universo de páginas recuperadas da Internet.

Com relação aos textos, buscou-se a montagem de um repositório com exemplos de diferentes áreas de atuação, de forma a possibilitar uma simulação de execução do protótipo em diversas áreas de conhecimento. Os domínios foram:

- Programação Orientada a Objetos
- Introdução a Linguagem Java
- Introdução a Banco de Dados
- Engenharia de Software
- Introdução a Redes de Computadores
- Introdução à Programação

6.2 Estratégia de Teste

Para execução dos testes, foram elaboradas apresentações de acordo com a grade curricular de cada turma, sendo:

- Banco de Dados: Mapeamento ER para Modelo Relacional;
- Programação II: Introdução à Orientação Objetos com Java.

No primeiro momento do processo de teste, foi realizada a exposição da motivação e objetivos do trabalho, uma breve descrição da ferramenta e uma explicação de sua interface e seus recursos. A seguir, cada turma assistiu a exposição de conteúdo específico, de acordo com a disciplina matriculada, utilizando somente o Microsoft PowerPoint para acompanhamento dos slides.

Em um terceiro momento, os alunos receberam como atividade uma lista de exercícios teóricos e práticos para serem resolvidos em duplas. Neste ponto, os alunos foram orientados a buscar material didático de forma livre, isto é, na Internet ou em um diretório na rede, onde os documentos do repositório estavam armazenados (sem indicação de conteúdo). A seguir, a utilização do BOOKISH foi estimulada, como fonte de recuperação de material de apoio para resolução dos exercícios.

Para avaliar a opinião dos alunos, foi elaborado um questionário (apresentado no Apêndice A, preenchido pelos alunos após a resolução dos exercícios propostos. Além de perguntas específicas, o questionário possui um campo aberto para realização de críticas e sugestões sobre o trabalho.

6.3 Ambiente de Teste

Os testes foram realizados em um laboratório com 15 (quinze) computadores padrão com as seguintes características: processador Intel Pentium 3, 512 MB de memória RAM e processador de 500 MHz. Todos os equipamentos estavam ligados em rede e com acesso à internet através de proxy.

O sistema operacional utilizado nos testes foi o Microsoft Windows XP Professional, Service Pack 3. Como pré-requisito para execução do BOOKISH, os equipamentos contavam com o Microsoft Office 2003 instalado, Internet Explorer versão 6 e .Net Framework 3.5.

6.4 Estatísticas

Os gráficos a seguir demonstram os resultados dos testes realizados.

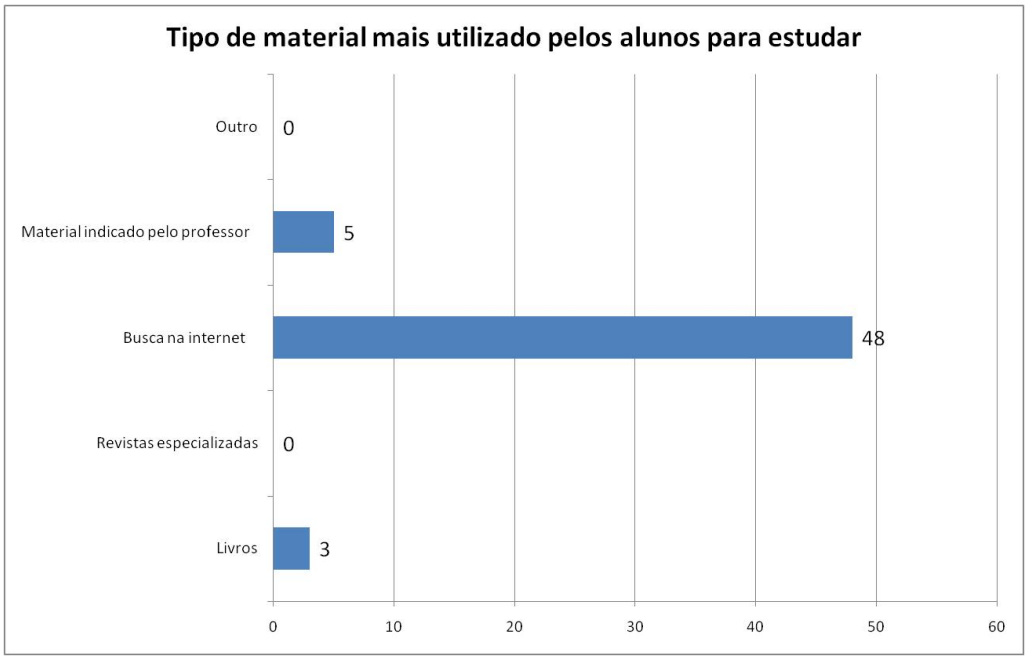

Figura 6.1: *Avaliação das principais fontes utilizadas pelos alunos para estudar*

A Figura 6.1 demonstra que o perfil do aluno é voltado para o uso de tecnologia, inclusive no momento de procurar por material didático para estudar, apesar do resultado demonstrado pela Figura 6.2, onde a maioria avalia que os livros são a melhor fonte de pesquisa.

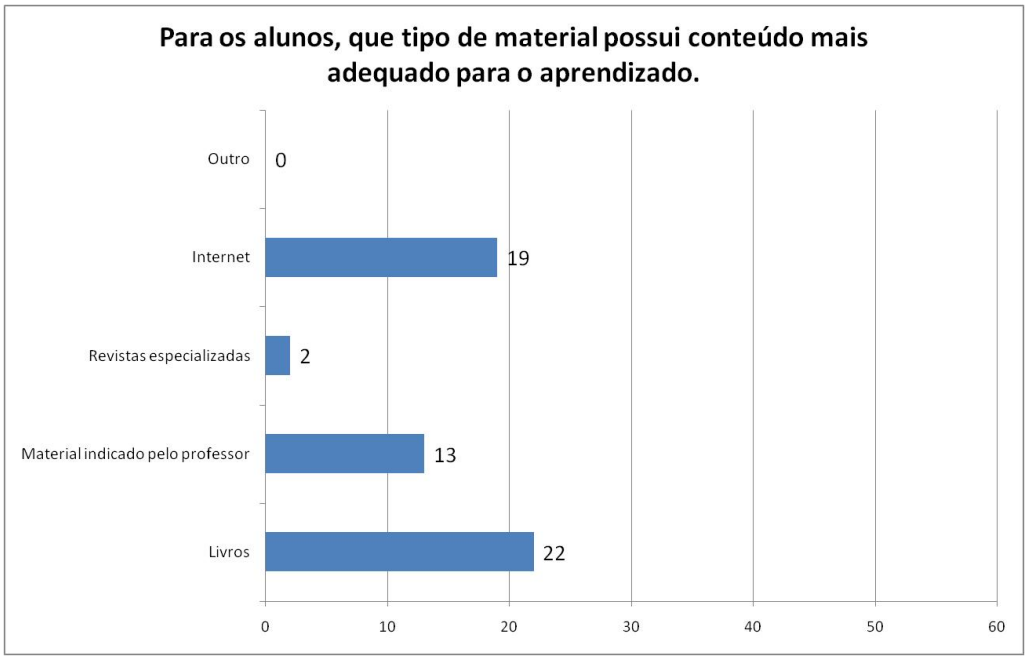

Figura 6.2: *Avaliação de material didático mais adequado para estudo*

Os gráficos das Figuras 6.3 e 6.4 demonstram que os alunos encontram mais dificuldade para analisar e encontrar informações específicas em um documento impresso, dificuldade que diminui quando a fonte de consulta é a Internet.

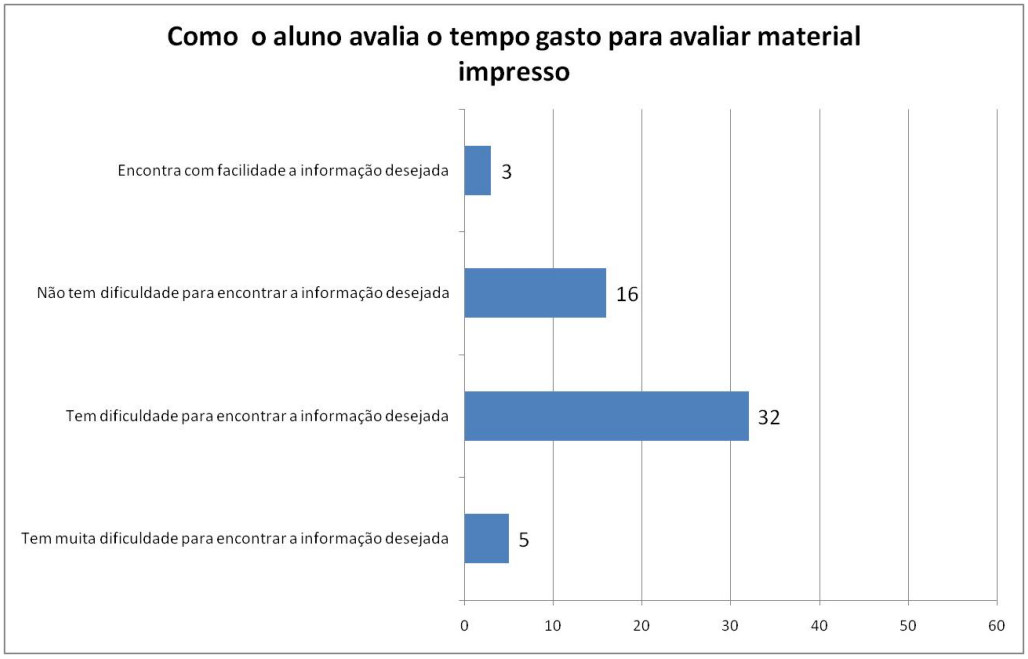

Figura 6.3: *Avaliação do tempo utilizado para localizar uma informação em material impresso*

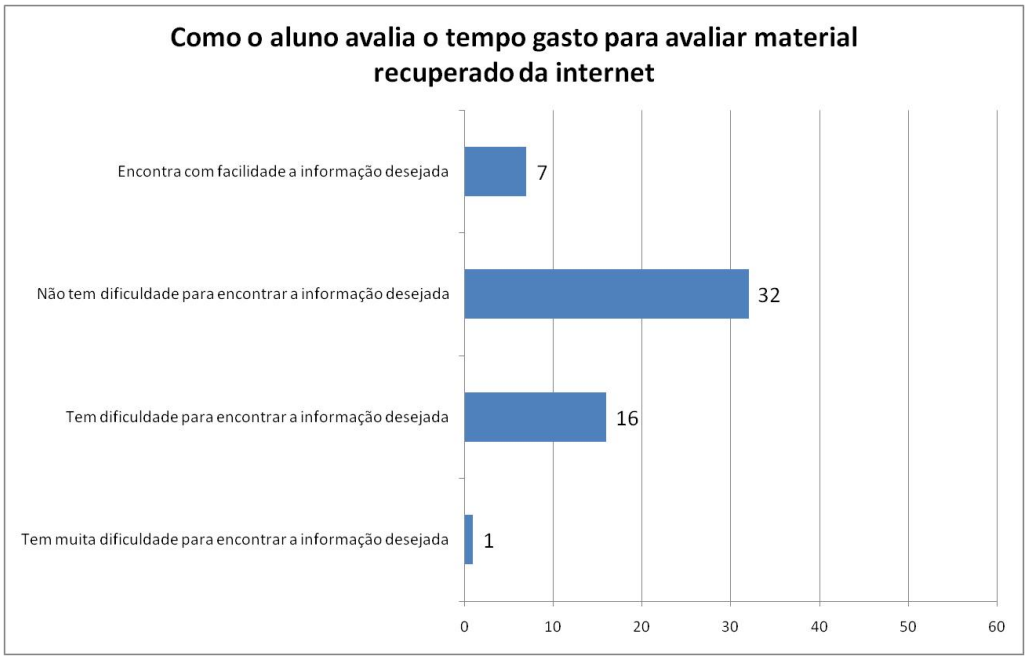

Figura 6.4: *Avaliação do tempo utilizado para localizar uma informação na Internet*

Os gráficos das Figuras 6.5 e 6.6 trazem os resultados da utilização do BOOKISH no ambiente simulado. As estatísticas demonstram que a ferramenta atinge seus objetivos específicos, no sentido de recuperar documentos de um repositório e páginas da Internet dentro do contexto que estava sendo apresentado pelo slide.

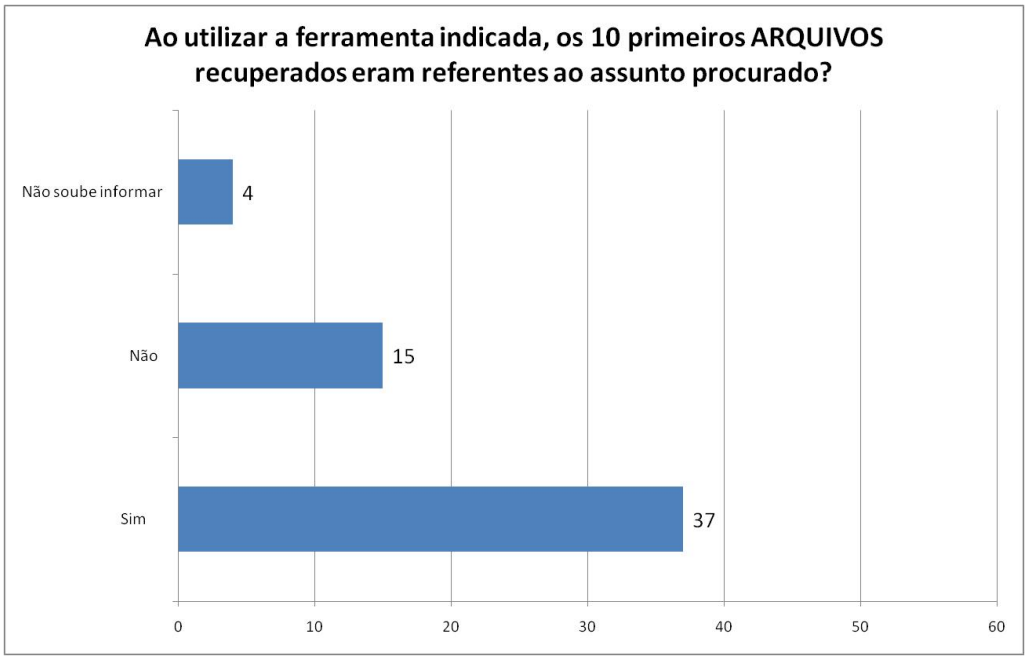

Figura 6.5: *Avaliação do contexto dos arquivos recuperados pelo BOOKISH*

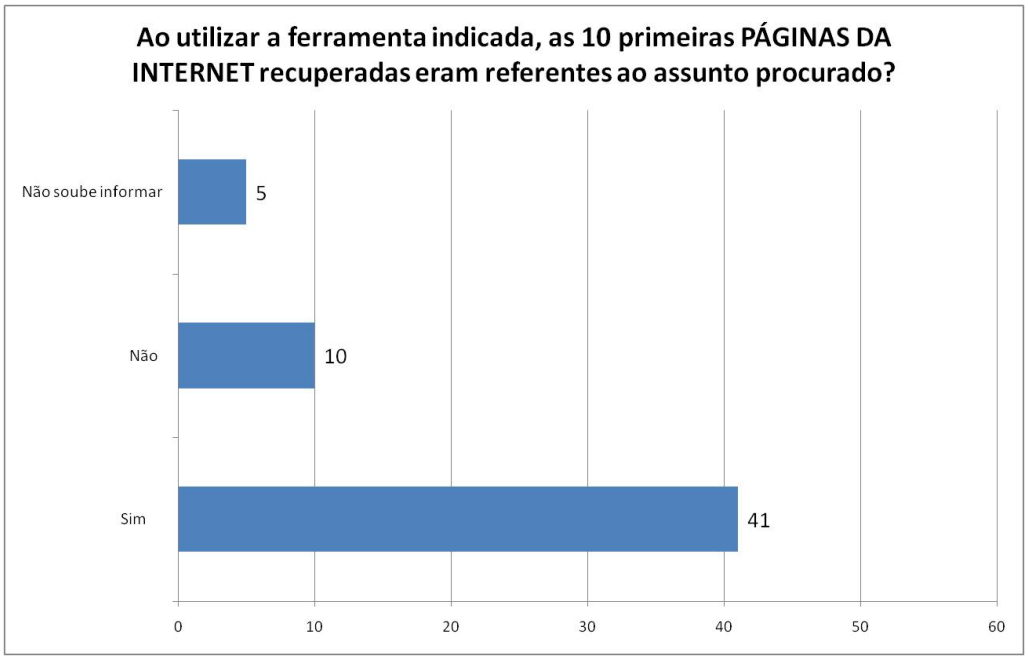

Figura 6.6: *Avaliação do contexto das páginas da internet recuperadas pelo BOOKISH*

Os gráficos das Figuras 6.7 e 6.8 demonstram que a ferramenta atingiu os objetivos estabelecidos quando recupera para os alunos um conjunto satisfatório de documentos.

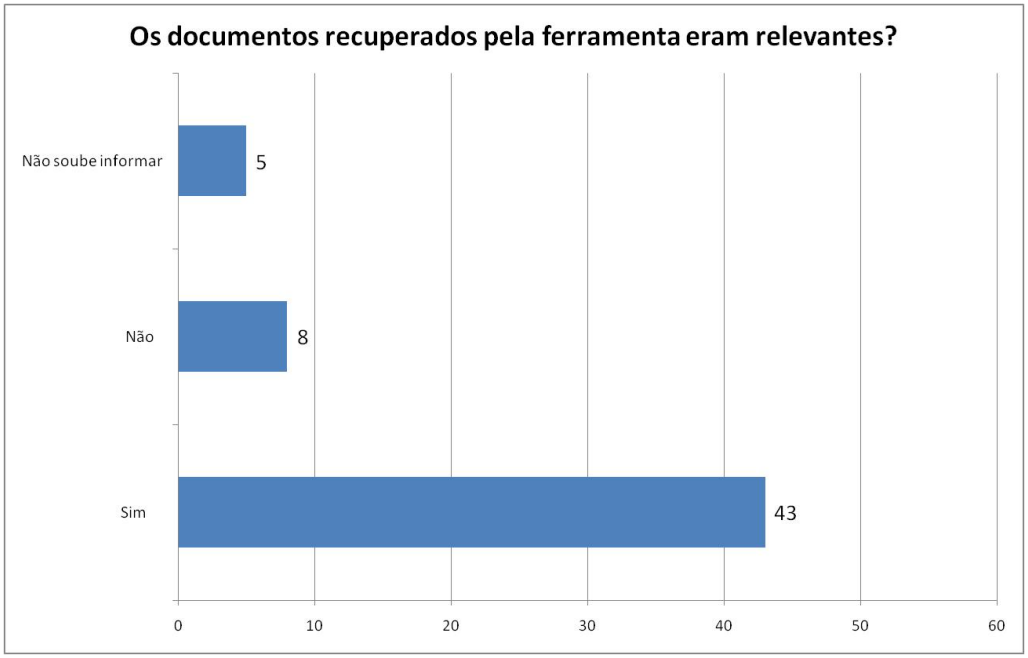

Figura 6.7: *Avaliação do conteúdo dos documentos do repositório recuperados pelo BOOKISH*

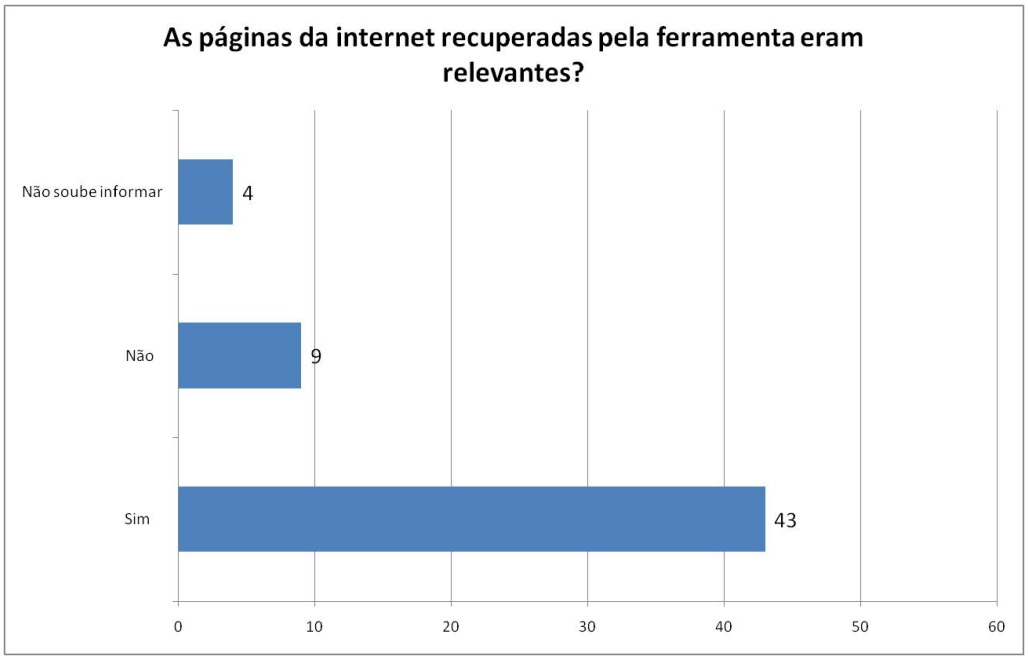

Figura 6.8: *Avaliação do conteúdo das páginas da internet recuperadas pelo BOOKISH*

O gráfico da Figura 6.9 demonstra que a maioria dos alunos considera adequada a utilização desse tipo de ferramenta para auxílio do desenvolvimento das atividades de pesquisa e complementação de conteúdo em sala de aula.

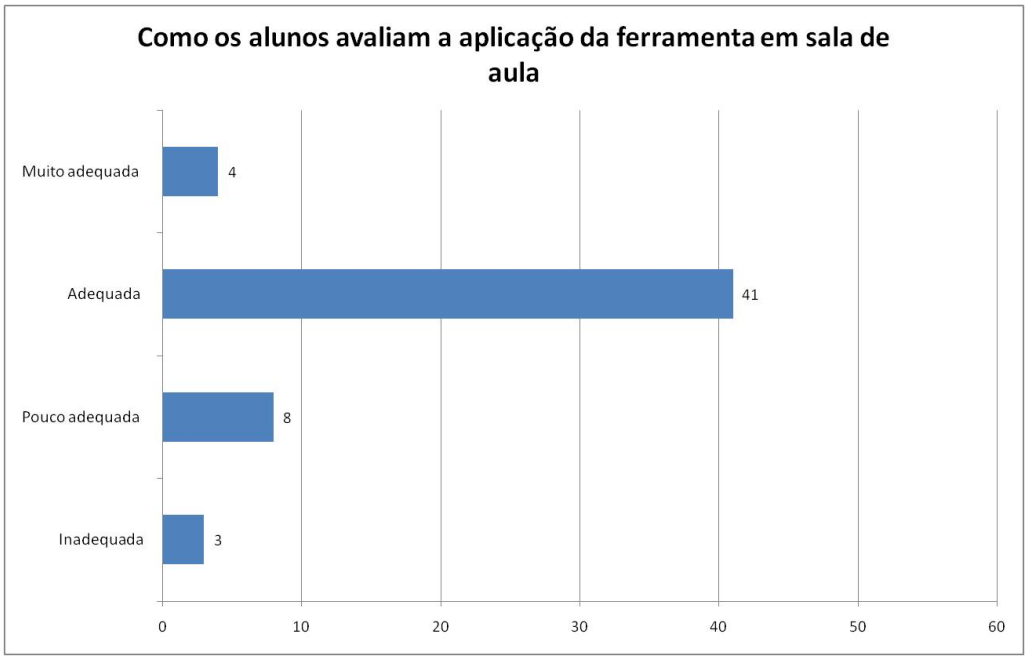

Figura 6.9: *Avaliação da aplicação do BOOKISH em sala de aula*

Conforme demonstrado na Figura 6.10, grande parte dos alunos entrevistados utilizaria o BOOKISH como ferramenta de apoio à pesquisa em sala de aula ou em casa.

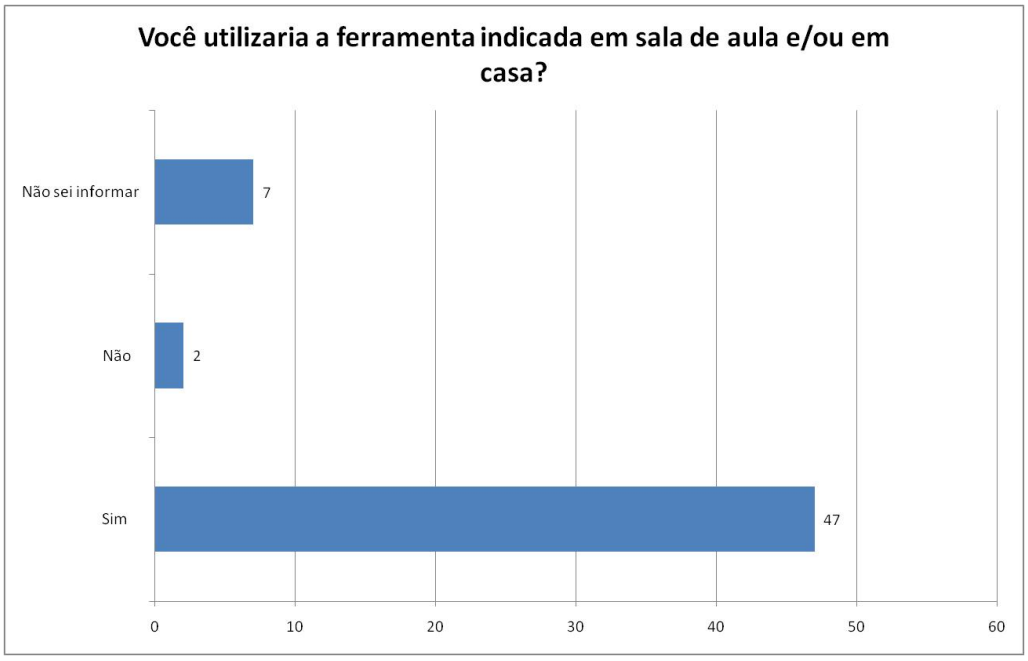

Figura 6.10: *Avaliação da utilização do BOOKISH pelos alunos em sala de aula ou em casa*

6.4.1 Ganhos Obtidos com a Ferramenta

Com o intuito de verificar se os objetivos da ferramenta foram atingidos (recuperar conteúdos relevantes e dentro do contexto desejado), os alunos foram incentivados a fazer dois tipos de testes: buscas independentes, ou seja, cada um usando seus próprios termos e critérios pessoais e também buscas com a utilização do BOOKISH. Nos dois testes, os alunos foram instruídos a realizar consultas tanto na Internet quanto em um diretório com documentos, buscando material que os auxiliasse na complementação de conhecimento sobre o conteúdo da aula.

Sem a utilização da ferramenta, os alunos indicavam seus próprios termos no site de busca e utilizavam seus próprios métodos para consultar o conteúdo dos documentos do repositório - alguns alunos utilizavam a busca de conteúdo, disponibilizado pelo sistema operacional instalado no computador, outros abriam cada documento da coleção em busca do material desejado.

Foi solicitado a cada usuário que analisasse os dez primeiros resultados encontrados, com e sem a utilização do BOOKISH, com relação à relevância do documento - se o documento possuía informações importantes sobre o assunto estudado, e sobre o contexto do documento - se o conteúdo do documento estava dentro do assunto abordado.

Os alunos receberam um arquivo com uma apresentação cujo tema era “introdução à programação orientada a objetos com Java”. O grupo foi orientado a ler a apresen-

tação, analisar seu conteúdo, realizar suas buscas (manuais e com a ferramenta) e, dentre outras questões, avaliar os tópicos:

- relevância: os alunos deveriam analisar se os resultados recuperados tratavam exatamente do **o(s) tópico(s)** presente(s) na apresentação (o que é objeto, paradigma da orientação a objetos, como programar objetos com Java, etc.);
- contexto: aqui, os alunos deveriam verificar se os resultados recuperados abordavam **o assunto** "programação orientada a objetos" como um todo, independentemente do tópico que estava sendo tratado.

A Tabela 6.1 apresenta o resultado obtido com a aplicação dos métodos propostos pelo BOOKISH quanto ao critério **relevância**.

Tabela 6.1: *Resultados da pesquisa de relevância*

Avaliação da Relevância	Busca Manual	Busca BOOKISH	Ganho
Relevância dos 10 primeiros arquivos recuperados	45%	77%	32%
Relevância das 10 primeiras páginas recuperadas	64%	77%	13%

O ganho obtido com o BOOKISH, no que tange à recuperação de documentos, foi de 32%. Isso se deve ao fato de que cada documento teve seu nome descaracterizado, sendo renomeado para *arquivo1.txt, arquivo2.txt*, e assim, sucessivamente. Esta operação foi realizada com o objetivo de não fornecer ao aluno, através do nome do arquivo, a indicação de seu conteúdo. Logo, para conhecer o conteúdo de cada documento o mesmo precisava ser aberto e lido ou pesquisado através de uma ferramenta disponível no sistema operacional.

O ganho não foi tão expressivo no que tange à recuperação na Internet, porque os alunos entrevistados possuem, pela experiência adquirida (são alunos do 2º e 3º do curso de sistemas de informação), uma capacidade intuitiva de selecionar bons termos para pesquisa e também de realizar a expansão de termos de forma a melhorar os resultados obtidos. Além disto, o tema da apresentação escolhida já está bem contextualizado, os termos geralmente usados para as consultas não geram ambiguidade. O ganho na contextualização será mais evidente quando a apresentação usar conceitos mais gerais, como "linguagem" e "alfabeto", por exemplo.

A Tabela 6.2 apresenta o resultado obtido com a aplicação da abordagem proposta quanto ao critério "**contexto**".

O ganho obtido com o BOOKISH foi de 25%, quanto à contextualização dos documentos. Este resultado se deve ao fato de que os alunos tinham à disposição um conjunto maior de documentos contextualmente semelhantes ao documento procurado, já que

Tabela 6.2: *Resultados da pesquisa de contextualização*

Avaliação do Contexto	Busca Manual	Busca BOOKISH	Ganho
Contexto adequado dos 10 primeiros arquivos recuperados	41%	66%	25%
Contexto adequado das 10 primeiras páginas recuperadas	59%	73%	14%

a coleção montada era composta por textos que abordavam, dentre outros assuntos, programação orientada a objetos, introdução à linguagem Java e introdução à programação.

Já nas buscas na Internet, houve um ganho de 14% na recuperação de páginas contextualizadas, já que, como foi dito, os alunos que participaram da pesquisa possuem conhecimento suficiente para realizar suas buscas de forma autônoma na Internet, sem a necessidade de auxílio por parte de uma ferramenta.

Esses resultados demonstram que a ferramenta pode ser bastante útil para alunos que possuem pouca experiência ou dificuldade na manipulação de mecanismos de busca.

Com base nestes resultados, concluímos que a utilização das técnicas abordadas por estes trabalho trouxeram uma boa colaboração para a recuperação de informação relevante e contextualizada. Para garantir que a ferramenta proposta tenha resultados ainda mais eficazes, é importante que a mesma possua mecanismos de *feedback* por parte dos usuários. A implementação deste tipo de mecanismo já está prevista nos trabalhos futuros a serem desenvolvidos sobre a ferramenta.

CAPÍTULO 7

Conclusão

Com a explosão do número de usuários de computadores e usuários da Internet, modelos para a recuperação precisa de informações passaram a ter um papel ainda mais importante. Na sociedade do conhecimento e da informação, as pessoas precisam ser capazes de obter, rapidamente, a maior quantidade possível de informações relevantes.

Neste trabalho, apresentamos uma abordagem cujo objetivo principal é facilitar a atividade de consulta de material didático relevante e contextualmente semelhante a partir da análise de apresentações em forma de slides. Essas apresentações são utilizadas pelos professores para ministrar conteúdos de suas disciplinas em sala de aula.

Os slides são processados através da mineração de textos e uma matriz representativa dos termos relevantes da apresentação é gerada. Esta matriz servirá como base para as atividades de recuperação de material didático. Podem ser recuperados documentos através da Internet ou em um repositório específico, sendo que para cada opção de consulta, uma técnica é utilizada.

Para isto, foi desenvolvida uma ferramenta, fundamentada nessa abordagem, denominada BOOKISH. A ferramenta desenvolvida utiliza expansão automática de consulta, com o objetivo de melhorar a recuperação de documentos através da Internet e indexação semântica latente, como método para identificar os documentos com maior grau de similaridade com o conteúdo da apresentação.

Para o desenvolvimento da aplicação proposta, foi realizado um levantamento de requisitos com o objetivo de identificar as funcionalidades necessárias para um software de apoio ao aluno e ao professor, com foco no tratamento e consulta de material didático.

A aplicação foi desenvolvida utilizando a plataforma Microsoft .Net Framework, que permite o desenvolvimento de novas funcionalidades para as ferramentas que compõem o pacote Microsoft Office, em especial o Microsoft PowerPoint. Além disso, a plataforma amplia a capacidade de integração com as aplicações Classroom Presenter e Ubiquitous Presenter, já que ambas utilizam a mesma tecnologia. Essa característica é importante, pois essas ferramentas foram selecionadas para serem integradas ao BOOKISH, em virtude de suas funcionalidades e ampla aceitação por parte dos usuários.

Um dos objetivos do BOOKISH é o de melhorar a recuperação de documentos na Internet. Para tanto, a escolha da técnica de expansão automática de consulta visa garantir uma seleção automática de termos mais adequados a uma recuperação eficiente.

O ponto de partida para seleção dos termos que serão utilizados na expansão de consulta é a apresentação que está sendo manipulada em sala de aula. Desta forma, evita-se a necessidade de utilização de dicionários, ontologias ou outras estruturas de apoio, embora estas possam contribuir quando anexadas à ferramenta.

Outro objetivo do BOOKISH é realizar a recuperação de documentos relevantes em uma coleção genérica. A utilização da indexação semântica latente possibilita a recuperação de informação textual a partir de textos semanticamente associados aos termos usados na consulta. Isso se dá pela significativa redução de informação e manutenção dos termos essenciais.

Após implementado, o BOOKISH conseguiu encontrar respostas relevantes para os usuários. O ganho médio obtido na recuperação de documentos relevantes foi de 22% e de 20% com relação à contextualização.

Embora estes resultados possam ser considerados modestos para um sistema de recuperação de informação, eles são considerados satisfatórios por várias questões, entre as quais podemos citar:

- Conforme o Capítulo 6, os alunos que participaram dos testes são estudantes do 2º e 3º ano de Sistemas de Informação e possuem experiência e habilidade da manipulação de computadores e sistemas. Isto significa também, que grande parte desses alunos já conhece, mesmo que intuitivamente, técnicas de mineração de textos, o que facilita e torna suas buscas bastante objetivas;
- Como demonstrado pela literatura e também pelo perfil dos entrevistados, os alunos "contemporâneos" possuem grande predileção pela Internet, principalmente por este ser um ambiente extremamente dinâmico e livre. Apesar de considerar os livros e o material indicado pelo professor como mais importante para sua formação (Seção 6.4), para os alunos, a internet oferece muito mais fontes sobre a mesma informação dos livros, só que livres para manipulação.
- Cada aluno é diferente um do outro e todos possuem necessidades de aprendizagem diferentes. Essas necessidades não são tratadas pelos sistemas de recuperação de informação, seus resultados são os mesmos para os diferentes alunos (considerando a mesma entrada de dados). Os alunos gostam de possuir autonomia para localizar conteúdos que mais se adaptam às suas necessidades, ou seja, eles gostam de procurar por textos que trazem a informação numa linguagem que eles consigam assimilar melhor, e essa variedade é encontrada na Internet. Evidenciando ainda mais essa afirmação, em nossa pesquisa, quando questionados sobre a localização

de informação em material didático impresso, 57% dos alunos afirmaram que têm dificuldade para encontrar o que precisam, mesmo quando o material é indicado pelo professor.

Para verificar essa teoria, novos testes devem ser aplicados a grupos de alunos de outras áreas, como Ciências humanas e biológicas.

7.1 Trabalhos Publicados

A partir das pesquisas realizadas para elaboração desta dissertação foram publicados, até o momento, dois artigos:

- **PowerInf - Collaborative Learning Environment**
  Este artigo, apresentado no XVII Workshop sobre Educação em Computação do XXIX Congresso Brasileiro da Sociedade Brasileira de Computação de 2009, em Bento Gonçalves/RS, apresenta um ambiente eletrônico de sala de aula que oferece suporte aos alunos nas atividades de investigação, permitindo que ele tenha fácil e rápido acesso a uma variedade de materiais didáticos complementares, dentro do contexto dos slides apresentados em sala de aula [33].
- **Digital Ink as a collaborative Learning Support**
  Este artigo, apresentado na International Conference on Computer Supported Education 2009, em Lisboa - Portugal, apresenta um ambiente de aprendizagem colaborativa como técnica utilizada nas tarefas de resolução de problemas, onde a aprendizagem ocorre por meio de intercâmbio de conhecimentos entre alunos e professores. O ambiente apresentado introduz várias ferramentas computadorizadas para apoio a esses ambientes de aprendizagem, usando tinta digital para facilitar a expressividade e criatividade [3].

7.2 Contribuições

As principais contribuições do trabalho são, portanto, as seguintes:

- Um Sistema de Recuperação de Informação, que atua como ferramenta de suporte aos alunos quanto à recuperação de material didático contextualizado ao conteúdo apresentado em sala de aula, conteúdo este identificado através da análise das apresentações fornecidas pelos professores;
- Uma técnica de busca automática, capaz de avaliar o conteúdo de um documento e extrair o seu contexto.

7.3 Trabalhos Futuros

Como trabalhos futuros, propomos:

- Possibilitar a análise de documentos com outros formatos dentro do repositório;
- Desenvolver a funcionalidade de *feedback*, de forma a aprimorar continuamente os resultados encontrados pela ferramenta;
- Integrar a ferramenta a um "ambiente de sala de aula eletrônica", como o Classroom Presenter e Ubiquitous Presenter, de modo que o aluno possa acompanhar, de forma interativa, a apresentação dos slides e, ainda, realizar anotações, entre outras atividades disponíveis numa ferramenta dessa categoria;
- Desenvolver uma interface WEB capaz de avaliar o conteúdo de uma apresentação e orientar os alunos nos momentos de estudo individual sem a necessidade de instalação da aplicação;
- Incluir recursos de tinta digital com funcionalidades específicas, como permitir que o usuário selecione um termo do slide para utilizar em uma consulta.

Referências Bibliográficas

[1] **Review of "natural language understanding"by james allen. benjamin/cummings 1987.** Comput. Linguist., 14(4):96–97, 1988. Reviewer-Allen, James.

[2] ALVES, R. **Será que a leitura dos jornais nos torna estúpidos?** Folha de S. Paulo, 2001.

[3] AMBROSIO, A. P. L; COSTA, F. M; ALMEIDA, C. G; GONDIM, H. W. A. S; PROVENSI, L; SILVA, L. O. **Digital ink as a collaborative learning support**. International Conference on Computer Supported Education, Proceedings of CSEDU 2009, 2009.

[4] ANDERSON, R; ANDERSON, R; DAVIS, P; LINNELL, N; PRINCE, C; RAZMOV, V; VIDEON, F. **Classroom presenter: Enhancing interactive education with digital ink**. Computer, 40(9):56–61, 2007.

[5] ANTONELLO, C. S. **Aprendizagem na ação revisitada e seu papel no desenvolvimento de competências**. Aletheia, (26):146–167, 2007.

[6] ASSMANN, H. **Reencantar a educação: rumo à sociedade aprendente**. Editora Vozes, 2001.

[7] BAEZA-YATES, R; RIBEIRO-NETO, B. **Modern Information Retrieval**. Addison-Wesley Longman Publishing Co., 1999.

[8] BARTHES, R. **S/Z**. Editora Nova Fronteira, 1992.

[9] BASSO, C. **Algumas reflexões sobre o ensino mediado por computadores**. http://www.ufsm.br, 2000.

[10] BELKIN, N. J; CROFT, W. B. **Information filtering and information retrieval: two sides of the same coin?** Commun. ACM, 35(12):29–38, 1992.

[11] BELLUZZO, R. C. B. **A aprendizagem ao longo da vida: um desafio para a educação na sociedade do conhecimento**. In: Rivero, C. M. L; Gallo, S, editors, A

FORMAçãO DE PROFESSORES NA SOCIEDADE DO CONHECIMENTO, p. 145–159. EDUSC, Bauru, 2004.

[12] BEPPLER, F. D. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. PhD thesis, Universidade Federal de Santa Catarina, Florianópolis/SC, 2008.

[13] BERRY, M. W; BROWNE, M. **Understanding Search Engines: Mathematical Modeling and Text Retrieval**. Soc for Industrial & Applied Math, 1999.

[14] BERRY, M. W; DUMAIS, S. T; O'BRIEN, G. W. **Using linear algebra for intelligent information retrieval**. SIAM Rev., 37(4):573–595, 1995.

[15] BRANQUINHO, L. A. **A prática pedagógica da educação atual**. http://www.meuartigo.brasilescola.com/pedagogia/a-pratica-pedagogica-educacao-atual.htm, 2008.

[16] BROMBERG, M. C. **O material didático e sua importancia**. http://www.hiperatividade.com.br/article.php?sid=90, 2007.

[17] CHAKRABARTI, S. **Mining the Web: Discovering Knowledge from Hypertext Data**. Morgan-Kauffman, 2002.

[18] COLLINGS, P; WALKER, D. **Applications to support student group work**. In: CSCL '95: THE FIRST INTERNATIONAL CONFERENCE ON COMPUTER SUPPORT FOR COLLABORATIVE LEARNING, p. 75–79, Hillsdale, NJ, USA, 1995. L. Erlbaum Associates Inc.

[19] CORTELAZZO, I. B. C. **Pedagogia e as novas tecnologias**. http://www.utp.br/mestradoemeducacao/pubonline/cortelazzoart.html, 2002.

[20] DA PONTE, J. P; OLIVEIRA, H; VARANDAS, J. M. **O contributo das tecnologias de informação e comunicação para o desenvolvimento do conhecimento e da identidade profissional**. Mercado de Letras, 2003.

[21] DA SILVA, J. A. **Modernização no planejamento e na estrutura de tic da secretaria de estado de saúde de minas gerais para garantir a qualidade na prestação dos serviços**, 2008.

[22] DA SILVA BORGES, M. R; CAVALCANTI, M. C. R; CAMPOS, M. L. M. **Suporte por computador ao trabalho cooperativo**. XV Congresso Brasileiro da Sociedade Brasileira de Computação. Jornada de Atualização em Informática, 1995.

[23] DAVIES, R. **The creation of new knowledge by information retrieval and classification**. Journal of Documentation, v.45, n.4, 1989.

[24] DE FREITAS BULCÃO NETO, R. **Um processo de software e um modelo ontológico para apoio ao desenvolvimento de aplicações sensíveis a contexto**. 2006.

[25] DE MEDEIROS GUIMARÃES, J. M. **Educação, globalização e educação a distância: uma reflexão sobre as políticas educacionais adotadas no brasil**. In: REVISTA LUSóFONA DE EDUCAçãO, volume 9. 2007.

[26] DE MOURA, A. M. M; DE AZEVEDO, A. M. P; MEHLECKE, Q. **As teorias de aprendizagem e os recursos da internet auxiliando o professor na construção do conhecimento**.

[27] DE OLIVEIRA, M. A. A. **Reflexões Sobre Conhecimento e Educação**. EDUFAL - Editora da Universidade Federal de Alagoas, 2000.

[28] DE OLIVEIRA VEDOOTTO, D. **A importância da formação de leitores críticos**. 2009.

[29] DE SALES, R; CAFÉ, L. **Semelhanças e diferenças entre tesauros e ontologias**. Datagramazero, 2008.

[30] DEERWESTER, S; DUMAIS, S. T; FURNAS, G. W; LANDAUER, T. K; HARSHMAN, R. **Indexing by latent semantic analysis**. Journal of the American Society for Information Science, 41:391–407, 1990.

[31] DELORS, J. **Educação: um tesouro a descobrir**. Editora Cortez, 1998.

[32] DORRE, J; GERSTL, P; SEIFFERT, R. **Text mining: finding nuggets in mountains of textual data**. In: KDD 99: PROCEEDINGS OF THE FIFTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, p. 398–401, New York, NY, USA, 1999. ACM.

[33] E SILVA, L. O; AMBRÓSIO, A. P. **Powerinf collaborative learning**. XVII Workshop de Educação em Informática - XXIX Congresso Brasileiro da Sociedade Brasileira de Computação, 2009.

[34] EFTHIMIADIS, E. N. **Query Expansion**. Annual Review of Information Systems and Technology, Seattle, WA, USA, 1996.

[35] ENS, R. T. **Relação professor, aluno, tecnologia: um espaço para o saber, o saber fazer, o saber conviver e o saber ser**. p. 37–44, 2002.

[36] FAYYAD, U. M; PIATETSKY-SHAPIRO, G; SMYTH, P. **From data mining to knowledge discovery: an overview**. p. 1–34, 1996.

[37] FAYYAD, U. M; PIATETSKY-SHAPIRO, G; SMYTH, P. **The kdd process for extracting useful knowledge from volumes of data**. Commun. ACM, 39(11):27–34, 1996.

[38] FELDMAN, R; DAGAN, I. **Knowledge discovery in textual databases**. In: IN PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-95, p. 112–117. AAAI Press, 1995.

[39] FELDMAN, R; HIRSH, H. **Exploiting background information in knowledge discovery from text**. J. Intell. Inf. Syst., 9(1):83–97, 1997.

[40] FORONDA, D. A. H. **Estudo exploratório da indexação semântica latente e das funções peso**. Master's thesis, Porto Alegre, 2005.

[41] FOUCAMBERT, J. **A leitura em questão**. Editora Artmed, 1994.

[42] FRAKES, W. B; BAEZA-YATES, R. **Information Retrieval: data structures and algorithms**. Prentice Hall, 1992.

[43] GADOTTI, M. **Boniteza de um sonho. Ensinar-e-aprender com sentido**. Editora Cortez, 2002.

[44] GOÑI, J. L; RIVERA, L. A. **Um lms-paradigmático para a customização de sistemas de gerenciamento de aprendizagem usando objetos de aprendizado**, 2008.

[45] GOUVÊA, M. T; MOTTA, C. L. R; SANTORO, F. M. **Estimulando a participação em um ambiente virtual de aprendizagem colaborativa**. XVIII Simpósio Brasileiro de Informática na Educação, 2006.

[46] GREENBERG, J. **Automatic query expansion via lexical-semantic relationships**. J. Am. Soc. Inf. Sci. Technol., 52(5):402–415, 2001.

[47] GREENGRASS, E. **Information retrieval: A survey**. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1855, 2001.

[48] GUARINO, N; MASOLO, C; VETERE, G. **Ontoseek: Content-based access to the web**. IEEE Intelligent Systems, 14(3):70–80, 1999.

[49] HAMZE, A. **O professor e o mundo contemporâneo**. http://www.educador.brasilescola.com/gestao-educacional/professor-mundo.htm, 2004.

[50] HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2000.

[51] HARASIM, L. **Network learning: What have we learned and what does it mean?** In Collaboration, Communication, and Computers: What do we think we know about networks and learning?, 1997.

[52] HOUAISS, A; VILLAR, M. D. S. **Minidicionário Houaiss da Língua Portuguesa**. Objetiva, 2004.

[53] JACÓ, C. **O uso do material didático impresso em ead**. http://aguarras.com.br/2008/04/02/o-uso-do-material-didatico-impresso-em-ead, 2008.

[54] JANSEN, B. J; SPINK, A; BATEMAN, J; SARACEVIC, T. **Real life information retrieval: a study of user queries on the web**. SIGIR Forum, 32(1):5–17, 1998.

[55] KENSKI, V. M. **Tecnologias e ensino presencial e a distância**. Papirus Editora, 2004.

[56] KHAN, L; MCLEOD, D; HOVY, E. **Retrieval effectiveness of an ontology-based model for information selection**. The VLDB Journal, 13(1):71–85, 2004.

[57] KHATTREE, R; NAIK, D. **Multivariate Data Reduction and Discrimination with SAS Software**. SAS Press and John Wiley Sons Inc., 2000.

[58] KLEMM, W. **Benefits of collaboration software for on-site classes**. 1997.

[59] LANGVILLE, A. N; MEYER, C. D. **Google's PageRank and beyond: the science of search engine rankings**. Princeton University Press, 2006.

[60] LAROCQUE, D; FAUCON, N. **Me, myself and... you? collaborative learning: why bother?** Teaching in the Community Colleges Online Conference: Trends and Issues in Online Instruction, 1997.

[61] LARSEN, D. **Samplers: Nine vicious little hypertexts**. http://www.eastgate.com/catalog/Samplers.html, 2000.

[62] LEE, J. H. **Properties of extended boolean models in information retrieval**. In: SIGIR '94: PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 182–190, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[63] LOPES, M. C. S. **Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português**. PhD thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro/RJ, 2004.

[64] LUHN, H. P. **The automatic creation of literature abstracts**. IBM Journal of Research and Development, 1958.

[65] ÁLVAREZ, A. C. **Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem**. Master's thesis, ICMC-USP, 2007.

[66] MALCHER, M. G; ENDLER, M. **iph: Uma aplicação para compartilhamento e co-edição de apresentações em sala de aula**. Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2009).

[67] MANNING, C. D; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. MIT Press, 1999.

[68] MASETTO, M. T; MORAN, J. M; BEHRENS, M. A. **Novas tecnologias e mediação pedagógica**. Editora Papirus, 2006.

[69] MITRA, M; SINGHAL, A; BUCKLEY, C. **Improving automatic query expansion**. In: SIGIR '98: PROCEEDINGS OF THE 21ST ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 206–214, New York, NY, USA, 1998. ACM.

[70] MOENS, M.-F. **Automatic Indexing and Abstracting of Document Texts**. Kluwer Academic Pub, 2000.

[71] MORAN, J. M. **Educação e tecnologias: Mudar para valer!** http://www.senado.gov.br.

[72] MORAN, J. M. **Novas tecnologias e o re-encantamento do mundo**. In: Revista Tecnologia Educacional, Rio de Janeiro, 1995.

[73] MORAN, J. M. **Especialista em projetos inovadores na educação presencial e a distância**. In: Revista Informática na Educação: Teoria & Prática, Porto Alegre, 1997.

[74] MORAN, J. M. **Tendências da educação online no Brasil**. Qualitymark, 2005.

[75] NEDER, M. L. C. **Educação e Comunicação em Educação a Distância**, volume 1. UFPR, Curitiba, 2001.

[76] NITZKE, J. A; CARNEIRO, M. L. F; GELLER, M; SANTAROSA, L. C. **Criação de ambientes de aprendizagem colaborativa**. X SBIE, 1999.

[77] OLDROYD, B. K; CITROEN, C. L. **A study of strategies used in online searching**. Online Review, 1977.

[78] PASSOS, E. L; GOLDSCHMIDT, R. **Data Mining: Um Guia Prático**. Editora Campus, Rio de Janeiro, 2005.

[79] PEAT, H. J; WILLETT, P. **The limitations of term co-occurrence data for query expansion in document retrieval systems**. Journal of the American Society for Information Science, 42:378–383, 1991.

[80] QIU, Y; FREI, H.-P. **Concept based query expansion**. In: SIGIR 93: PROCEEDINGS OF THE 16TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 160–169, New York, NY, USA, 1993. ACM.

[81] QUINT, B. **Inside a searcher's mind: the seven stages of an online search—part 1**. Online, 15(3):13–18, 1991.

[82] QUINT, B. **Inside a searcher's mind: the seven stages of an online search—part 2**. Online, 15(4):28–35, 1991.

[83] RADEV, D. R; HOVY, E; MCKEOWN, K. **Introduction to the special issue on summarization**. Comput. Linguist., 28(4):399–408, 2002.

[84] REZENDE, D. A. **Tecnologia da Informação Integrada à Inteligência Empresarial: Alinhamento Estratégico e Análise da Prática nas Organizações**. Editora Atlas, 2002.

[85] REZENDE, S. O. **Sistemas Inteligentes - Fundamentos e Aplicações**. Editora Manole, 2002.

[86] RIJSBERGEN, C. J. V. **Information Retrieval**. Butterworth-Heinemann, Newton, MA, USA, 1979.

[87] ROBERTSON, S. E. **On relevance weight estimation and query expansion**. Journal of Documentatzon, 42(3), 1986.

[88] ROBERTSON, S. E; JONES, K. S. **Relevance weighting of search terms**. Taylor Graham Publishing, London, UK, UK, 1988.

[89] ROSEMBERG, M. J. **E-Learning**. Editora Makron, São Paulo, 2002.

[90] RÖÜLING, G; TROMPLER, C; MÜHLHÄUSER, M; KÖBLER, S; WOLF, S. **Enhancing classroom lectures with digital sliding blackboards**. SIGCSE Bull., 36(3):218–222, 2004.

[91] SALTON, G. **Automatic text processing**. Addison-Wesley, Boston, MA, USA, 1988.

[92] SALTON, G. **The SMART Retrieval System—Experiments in Automatic Document Processing**. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[93] SALTON, G. **Dynamic information and library processing**. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1975.

[94] SCHIESSL, J. M. **Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor**. Master's thesis, Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação- Universidade de Brasília, Brasília, 2007.

[95] SCHNEIDER, J. W. **Verification of bibliometric methods' applicability for thesaurus construction**. SIGIR Forum, 39(1):63–64, 2005.

[96] SILVA, I; RIBEIRO-NETO, B; CALADO, P; MOURA, E; ZIVIANI, N. **Link-based and content-based evidential information in a belief network model**. In: SIGIR '00: PROCEEDINGS OF THE 23RD ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 96–103, New York, NY, USA, 2000. ACM.

[97] SILVA, K. X. S. **Webquest: uma metodologia para a pesquisa escolar por meio da internet**. Master's thesis, Universidade de Brasília, 2006.

[98] SILVERMAN, B. G. **Computer supported collaborative learning (cscl)**. Computers Education, 25(3), 1995.

[99] **Só portugues**. http://www.soportugues.com.br, 2009.

[100] TAKAO, E. L. **Análise comparativa dos modelos e sistemas probabilísticos em recuperação de informação em bases textuais**. Master's thesis, UFSC, 2001.

[101] TRAJBER, R; DA COSTA, L. B. **Avaliando a Educação Ambiental no Brasil**. Peirópoles, São Paulo, 1996.

[102] VALENTE, J. A; FREIRE, F. M. P; DA ROCHA, H. V; D'ABREU, J. V; BARANAUSKAS, M. C. C; MARTINS, M. C; PRADO, M. E. B. B. **O computador na sociedade do conhecimento**. In: COLEçãO INFORMáTICA PARA A MUDANçA NA EDUCAçãO. Ministério da Educação, 2002.

[103] VANTI, N. A. P. **Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento**. volume 31, p. 152–162. Ciência da Informação, Brasília, 2002.

[104] VIERA, A. F. G; VIRGIL, J. **Uma revisão dos algoritmos de radicalização em língua portuguesa**. http://informationr.net/ir/12-3/paper315.html, 2007.

[105] VILLARDI, R. **Ensinando a gostar de ler e formando leitores para a vida inteira**. Qualitymark Dunya, Rio de Janeiro, 1999.

[106] VOORHEES1994, E. M. **Query expansion using lexical-semantic relations**. In: SIGIR '94: PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, p. 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[107] VYGOTSKY, L. S. **A formação social da mente**. Editora Martins Fontes, São Paulo, 1987.

[108] WEISS, S; INDURKHYA, N; ZHANG, T; DAMERAU, F. **Text Mining: Predictive Methods for Analyzing Unstructured Information**. Springer, 2004.

[109] WILKS, Y. **Information extraction as a core language technology**. In: SCIE '97: INTERNATIONAL SUMMER SCHOOL ON INFORMATION EXTRACTION, p. 1–9, London, UK, 1997. Springer-Verlag.

[110] WIVES, L. K. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. PhD thesis, Instituto de Informática, UFRGS, Porto Alegre, 2001.

[111] YANG, Y; PEDERSEN, J. O. **A comparative study on feature selection in text categorization**. In: ICML '97: PROCEEDINGS OF THE FOURTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, p. 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[112] ZHONG, S; GHOSH, J. **A unified framework for model-based clustering**. J. Mach. Learn. Res., 4:1001–1037, 2003.

[113] ZIPF, G. K. **Human behavior and the principle of the least effort: an introduction to human ecology**. Addison-Wesley, Cambridge, MA, USA, 1949.

APÊNDICE A

Formulário de Pesquisa

Universidade Federal de Goiás
Instituto de Informática
Mestrado em Ciência da Computação
Mestranda: Luciana Oliveira e Silva
Orientadora: Prof. Dra. Ana Paula L. Ambrósio

Pesquisa Acadêmica

Nome Completo:	
Instituição:	
Curso:	
Série:	

A presente pesquisa tem como objetivo identificar as maiores dificuldades enfrentadas pelos alunos quando, no momento de estudo individual ou na resolução de uma tarefa, procuram por material de apoio para consulta.

Para responder o questionário de avaliação, marque o item que mais se aproxima da sua opinião **quanto à busca e recuperação de material didático**.

Dados Gerais

1.O que você utiliza, com maior freqüência, como material didático na hora de estudar?
( ) Livros
( ) Revistas especializadas
( ) Busca na internet
( ) Material indicado pelo professor
( ) Outro (qual): _____

2.Em sua opinião, que tipo de material possui conteúdo mais adequado para o seu aprendizado?
( ) Livros
( ) Material indicado pelo professor
( ) Revistas especializadas
( ) Internet
( ) Outro (qual): _____

3.Quando você utiliza material impresso (livros, entre outros), como você avalia o tempo gasto para encontrar o assunto que procura?
( ) Tenho muita dificuldade para localizar as informações que procuro
( ) Gasto um tempo considerável para encontrar a informação que procuro
( ) Não tenho dificuldades em localizar a informação procurada
( ) Encontro com facilidade a informação que procuro

4.Em média, qual a quantidade de material impresso (livros, por exemplo) você precisa consultar até encontrar as informações que procura?
( ) 1 ( ) 2 ( ) 3 ( ) 4 ou mais

5.Quando utiliza a internet para buscar material didático de apoio, você encontra o conteúdo desejado entre os 10 primeiros resultados mostrados pelo buscador?
( ) Sim ( ) Não ( ) Não soube informar

6.Em sua opinião, quando utiliza a internet para buscar material de apoio, como você avalia o seu desempenho?
( ) Tenho muita dificuldade para localizar as informações que procuro
( ) Gasto um tempo considerável para encontrar a informação que procuro
( ) Não tenho dificuldades em localizar a informação procurada
( ) Encontro com facilidade a informação que procuro

Figura A.1: *Formulário de pesquisa aplicado - página 1/2*

7.Quando utiliza o computador, em média, quantas interações você precisa executar para encontrar material que considere adequado (quantas vezes precisa refinar os termos da busca ou quantos documentos são abertos até localizar algum que seja indicado para o que procura)?
( ) Somente 1 ( ) 2 a 6 ( ) 7 a 11 ( ) 12 ou mais

Uso da Ferramenta

1.Em sua opinião, ao utilizar a ferramenta indicada, os 10 primeiros ARQUIVOS recuperados eram referentes ao assunto que você procurava?
( ) Sim ( ) Não ( ) Não sei informar

2.Em sua opinião, ao utilizar a ferramenta indicada, as 10 primeiras PÁGINAS DA INTERNET recuperadas eram referentes ao assunto que você procurava?
( ) Sim ( ) Não ( ) Não sei informar

3.Ao utilizar a ferramenta, quantas interações você precisou executar até encontrar um material com conteúdo adequado à informação que procurava?
( ) Somente 1 ( ) 2 a 6 ( ) 7 a 11 ( ) 12 ou mais

4.Os ARQUIVOS recuperados pela ferramenta eram relevantes?
( ) Sim ( ) Não ( ) Não sei informar

5.As PÁGINAS DA INTERNET recuperadas pela ferramenta eram relevantes?
( ) Sim ( ) Não ( ) Não sei informar

6.Como você avalia a facilidade de utilização da ferramenta para auxiliá-lo na localização de material didático?
( ) Muito Difícil
( ) Difícil
( ) Regular
( ) Fácil
( ) Muito Fácil

6.Como você avalia a utilização dessa ferramenta em sala de aula.
( ) Inadequada
( ) Pouco adequada
( ) Adequada
( ) Muito adequada

7.Você utilizaria a ferramenta indicada em sala de aula ou em casa?
( ) Sim ( ) Não ( ) Não sei informar

8. Espaço aberto para críticas e sugestões:

Figura A.2: *Formulário de pesquisa aplicado - página 2/2*