UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO

ALEXANDRE CHOW

AVALIAÇÃO DA INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS ELETRÔNICOS NO SOFTWARE *ADOBE ACROBAT* VERSÃO 5.0

Porto Alegre

2004

ALEXANDRE CHOW

AVALIAÇÃO DA INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS ELETRÔNICOS NO SOFTWARE *ADOBE ACROBAT* VERSÃO 5.0

Monografia elaborada como requisito para conclusão da disciplina BIB 03037 – Trabalho de Conclusão de Curso, do Departamento de Ciências da Informação, do Curso de Biblioteconomia, da Faculdade de Biblioteconomia e Comunicação, da Universidade Federal do Rio Grande do Sul

Orientadora: Profa. Ms. Glória Isabel Sattamini Ferreira – CRB10/176

Porto Alegre

2004

025.4:004.6
C459a Chow, Alexandre
Avaliação da indexação automática de documentos eletrônicos no software Adobe Acrobat versão 5.0 / Alexandre Chow; orientação de Glória Isabel Sattamini Ferreira

53 f. ; 30 cm.

Trabalho de Conclusão de Curso

1. Indexação automática. 2. Portable document format. 3. Adobe Acrobat. I. Título.

Departamento de Ciências da Informação
Rua Ramiro Barcelos, 2705 – 5° andar
Bairro Santana
Porto Alegre (RS)
CEP 90035-007
Telefone: (51) 3316-5146
Fax: (51) 3316-5435
E-mail: fabico@ufrgs.br

ALEXANDRE CHOW

AVALIAÇÃO DA INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS ELETRÔNICOS NO SOFTWARE *ADOBE ACROBAT* VERSÃO 5.0

Monografia elaborada como requisito para conclusão da disciplina BIB03037 – Trabalho de Conclusão de Curso, do Departamento de Ciências da Informação, do Curso de Biblioteconomia, da Faculdade de Biblioteconomia e Comunicação, da Universidade Federal do Rio Grande do Sul

Aprovada em 16/12/2004

BANCA EXAMINADORA

Glória Isabel Sattamini Ferreira (Orientadora)
Ms. em Educação pela PUC/RS
UFRGS

Regina Helena van der Laan
Dra. em Lingüística pela UFRGS
UFRGS

Marcia Raymundo Bernardes
Esp. em Automação de Bibliotecas pela UFPE
UFRGS

AGRADECIMENTOS

Agradeço à Profa. Ms. Glória Isabel Sattamini Ferreira não só pela orientação recebida, mas também pelo estímulo e apoio durante o desenvolvimento deste estudo.

Aos colegas, demais professores, bibliotecários e funcionários da Faculdade agradeço pela oportunidade de troca de experiências e pela hospitalidade durante minha (longa) estada na FABICO.

Aos profissionais bibliotecários Raquel da Rocha Schimitt, Simone Augustinho Rocha, Denise Nunes Pithan, Marcia Raymundo Bernardes e Márcio Rohan da Silva pelas oportunidades de aprendizado durante os estágios realizados.

A minha família pelo apoio e compreensão durante a concretização deste trabalho.

E, finalmente, todos aqueles que, direta ou indiretamente, me auxiliaram na superação de mais esta etapa.

RESUMO

Avaliação da indexação automática realizada pelo *Adobe Acrobat*. Este trabalho teve como objetivo a avaliação da indexação automática em documentos eletrônicos em formato PDF realizada pelo recurso *Adobe Catalog*, disponível na versão 5.0 do *Adobe Acrobat*, quanto à recuperação de informações, levando-se em conta o nível de especificidade dos termos, gênero (masculino/feminino), número (singular/plural), equivalência e o comportamento do software na recuperação de sintagmas nominais. O universo estudado foi composto dos 20 Trabalhos de Conclusão de Curso (TCC) do curso de Biblioteconomia da Universidade Federal do Rio Grande do Sul no segundo semestre de 2003. Após a construção automática do índice pelo software, cada uma das palavras-chave atribuídas nos trabalhos de conclusão foram utilizadas como termos de busca e o conjunto de documentos recuperados em cada uma das buscas foi anotado na ficha de observação. Os resultados mostraram que 77% das palavras-chave presentes nos TCC recuperaram seus documentos de origem. As palavras-chave que não trouxeram resultados nas buscas eram sintagmas de segundo e terceiro níveis. Três das cinco variações de gênero não produziram resultados nas buscas. Das 36 palavras flexionadas por número, seis não recuperaram documentos. Não foram detectados problemas na especificidade do vocabulário que interferissem na recuperação de documentos. Os termos equivalentes não apresentaram problemas, exceto quando localizados em trabalhos distintos gerando uma dispersão nos resultados das buscas. Apesar de suas limitações, o *Adobe Catalog* conseguiu recuperar de maneira satisfatória os documentos indexados em um acervo especializado na área de Ciência da

Informação. O aprimoramento desta ferramenta poderá contribuir para o aumento de sua eficácia, justificando também a sua utilização em acervos de cunho geral.

Palavras-Chave: Indexação automática. *Portable Document Format*. *Adobe Acrobat*.

ABSTRACT

Evaluation of automatic indexing performed by *Adobe Acrobat*. This work aimed to evaluate the automatic indexing of electronic documents in PDF format performed by *Adobe Catalog* feature available in *Adobe Acrobat* version 5.0. The specificity level of indexing, gender (masculine/feminine), number (singular/plural), equivalence issues and nominal sintagma retrieval were studied. The total of 20 Final Library Science Courseworks presented in the second term of 2003 at the Federal University of Rio Grande do Sul will be automatically indexed. After the indexing process, each of the keywords presented in those Final Courseworks will be used as search terms and the documents retrieved during those searches will be listed in the observation card. The results showed that 77% of the keywords appearing in those Final Courseworks retrieved its origin documents. The keywords that didn't retrieve any documents belonged to second and third grade nominal sintagma. Three out of five keywords with gender variations didn't retrieve any document. Among 36 keywords with number variations, six produced no results during searches. No major problems were detected due to specificity issues. Equivalence issues were not detected, except when those equivalents were present at different documents. Although its limitations, Adobe Catalog feature fairly retrieved documents in a Library Science digital collection. The improvement of this tool will contribute to achieve a better level of efficacy, justifying its use in general subject collections.

Keywords: Automatic indexing. Portable Document Format. Adobe Acrobat.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Atualmente, com o uso da Internet e das novas tecnologias, a informação supera barreiras de tempo e espaço. Acervos em bibliotecas *on-line* podem ser consultados sem a necessidade do deslocamento físico de seus usuários, além de poderem ser acessados por mais de uma pessoa simultaneamente. A comunicação científica ficou mais ágil: o tempo necessário para se disponibilizar um fascículo de periódico na Internet é bem menor do que aquele que seria gasto para impressão e distribuição do mesmo da maneira tradicional.

Boa parte dos periódicos científicos *online* disponibilizam seu conteúdo em formato PDF (*Portable Document Format*), gerado pelo programa *Adobe Acrobat*. Entre as principais razões de sua ampla utilização, estão a fidelidade ao formato original (impresso), a redução da chance de alteração das informações nele contidas e o seu grande poder de compactação.

Com o aumento do intercâmbio de documentos nesse formato de arquivo e a formação de acervos eletrônicos em formato PDF, é necessário o tratamento desses documentos, possibilitando sua recuperação de maneira rápida e eficaz. O software *Adobe Acrobat*, responsável pela geração de arquivos em formato PDF, oferece um módulo de indexação automática de acervos eletrônicos, denominado *Adobe Catalog*. Com recursos humanos cada vez mais escassos, a possibilidade de indexar acervos de forma automática parece ser uma saída para o descongestionamento do setor de processamento técnico das unidades de informação, principalmente em áreas especializadas.

2 JUSTIFICATIVA

Através do presente estudo, pretende-se apresentar as características e limitações do processo de indexação automática realizado pelo *Adobe Acrobat* versão 5.0, através da função *Adobe Catalog*, auxiliando na decisão de adoção desta ferramenta em unidades de informação que possuam acervos em formato PDF.

A utilização da versão 5.0 para a realização deste estudo deve-se à sua compatibilidade com o computador disponibilizado para a bateria de testes: a versão mais recente do software *Adobe Acrobat* (6.0) na sua versão *Professional* é incompatível com o sistema operacional *Windows* 98.

3 OBJETIVO

Este trabalho visou analisar a indexação automática de acervos eletrônicos compostos de arquivos PDF realizada pelo *Adobe Catalog*, disponível na versão 5.0 do *Adobe Acrobat*, quanto à recuperação de informações, levando-se em conta o nível de especificidade dos termos, gênero (masculino/feminino), número (singular/plural), equivalência e o comportamento do software na recuperação de sintagmas nominais. Serão mostradas as características e limitações desta modalidade de indexação, possibilitando a sua adoção em unidades de informação.

4 REVISÃO DA LITERATURA

A presente revisão de literatura versará sobre as modalidades de indexação (automática e manual), as linguagens documentárias e os sintagmas nominais.

4.1 O Processo de Indexação

A indexação é o "[. . .] processo intelectual que envolve atividades cognitivas na compreensão do texto e a composição da representação do documento." (LIMA, 2003, p. 83). A indexação não é uma atividade que tem um fim em si mesma: ela deve levar em consideração as necessidades dos usuários do sistema de informação onde é executado, possibilitando uma melhor representação temática do item para a comunidade atendida pelo sistema.

O processo de indexação pode ser realizado pelo homem (indexação manual) ou por programas de computador (indexação automática). Cada uma dessas modalidades de indexação será tratada nas seções seguintes.

4.1.1 Indexação Manual

O processo de indexação manual é decorrente de uma análise intelectual que é, basicamente, dividida em três etapas distintas (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 1992):

a) estabelecimento do assunto através do exame do documento;
b) identificação dos conceitos presentes;
c) tradução dos conceitos para uma linguagem de indexação.

Por ser uma atividade intelectual e, conseqüentemente, subjetiva, a indexação deve ser executada de acordo com políticas bem claras e definidas para que a subjetividade dos indexadores seja minimizada (LANCASTER, 1993), contribuindo para a eficiência do sistema de recuperação de informação.

4.1.2 Indexação Automática

A indexação automática consiste na análise de conteúdo de documentos por programas de computador. Diferentemente do processo manual de indexação, nessa modalidade os assuntos do documento são extraídos do documento analisado. As primeiras iniciativas do uso da indexação automática datam do final da década de 1950, quando Luhn desenvolveu o índice KWIC (*Key Word in Context*) em que as palavras

significativas dos títulos dos documentos sofriam uma rotação automática. Nas décadas seguintes, vários estudos foram realizados para comprovar a validade do uso do título como fonte de indexação. Os autores de artigos científicos, cientes da utilização do KWIC na indexação de documentos, passaram a compor os títulos mais relevantes e precisos para os seus trabalhos, na tentativa de se sobressaírem em relação à explosão bibliográfica, facilitando a recuperação das mesmas (VIEIRA, 1988).

Com o desenvolvimento da informática e o conseqüente aumento no poder de processamento e armazenagem de dados, não somente os títulos puderam ser objeto de indexação, mas também o texto integral do documento, como ocorre na indexação feita pelo *Adobe Catalog*.

4.2 Linguagens Documentárias

As linguagens documentárias são utilizadas para representar o conteúdo dos documentos. Elas classificam-se em dois grandes grupos: linguagem natural e linguagem controlada.

4.2.1 Linguagem Natural

Na linguagem natural, as palavras usadas pelo autor são utilizadas para a representação do assunto de um documento. Segundo Lopes (2002), entre as principais vantagens da adoção da linguagem natural, estão:

a) registro imediato da informação sem prévia consulta a uma ferramenta;
b) dispensa do uso de treinamentos para uso de uma linguagem de controle;
c) eliminação de conflitos entre indexadores e usuários, pois os mesmos terão acesso aos mesmos termos.

A principal desvantagem da adoção da linguagem natural reside na necessidade de se prever, no ato da busca, todas as variações (sinônimos, grafias alternativas, etc.) do assunto, aumentando o tempo despendido para a procura de documentos no sistema.

4.2.2 Linguagem Controlada

A linguagem controlada utiliza a noção de conceitos e a representação dos assuntos de um documento é extraída do instrumento de indexação (tesauro, lista de cabeçalhos de assunto, etc.) utilizado pelo sistema.

De acordo com Lopes (2002), as principais vantagens do uso da linguagem controlada são:

a) diminuição dos problemas de comunicação entre indexadores e usuários, através do controle total do vocabulário de indexação;
b) melhor atribuição de descritores com o uso de um tesauro e de suas respectivas notas de escopo;

c) buscas mais elaboradas e melhor identificação de conceitos relacionados, através das relações hierárquicas e remissivas do instrumento de vocabulário controlado.

Entre as principais desvantagens da utilização do vocabulário controlado, estão:

a) alto custo de produção e manutenção/atualização da ferramenta;
b) necessidade de treinamento para o uso da ferramenta tanto para os intermediários (indexadores), quando para os usuários finais.

4.3 Vocabulário Livre ou Controlado?

Buscar a escolha mais adequada sobre qual tipo de vocabulário adotar para a indexação num sistema de recuperação de informação parece ser uma tarefa impossível. E na verdade tanto uma opção quanto a outra oferecem facilidades e limitações que podem influir no bom desempenho de um sistema de informação.

O uso da linguagem controlada (através de listas de cabeçalho de assunto, tesauros, códigos de classificação) permite uma padronização de termos utilizados entre diferentes indexadores, pois por mais diferentes termos que um conceito possa ser expresso, ele, teoricamente, receberá o mesmo rótulo no sistema.

Por outro lado, a adoção de ferramentas que controlem o vocabulário exige um treinamento tanto de indexadores quanto de usuários, além de ser exigida uma

atualização constante da ferramenta, principalmente nas áreas do conhecimento em que há rápido surgimento de novos conceitos.

A linguagem natural, por sua vez, dispensa a prévia consulta a uma ferramenta (e o conseqüente treinamento para o seu uso) tanto para a execução da indexação quanto da busca. O problema mais saliente na adoção de tal vocabulário é a necessidade de se prever todas as expressões possíveis para um determinado conceito para que a busca seja a mais abrangente possível.

Uma tendência descrita em artigos indica uma utilização conjunta dos dois tipos de linguagem: "A LC e a LN não podem mais ser tratadas como técnicas de busca separadas, mas devem sempre ser tratadas em conjunto, como uma combinação ideal para ampliar os resultados das buscas de informação." (MUDDAMALLE, 1998, p. 887)

4.4 Os Sintagmas Nominais

A representação dos assuntos dos documentos em grande parte dos sistemas de informação é feita através de descritores ou palavras-chave. Segundo Pinto (2001), o grande problema nesta sistemática de indexar é que as palavras são retiradas de seu contexto, perdendo a significação determinada pelo mesmo.

Os sintagmas nominais são, de acordo com Kuramoto (1996, p. 184), "[. . .] a menor parte do discurso portadora de informação". A utilização dos sintagmas nominais para a indexação de documentos permite que a representação do conteúdo do documento não seja desvinculada de seu contexto, tornando-se uma alternativa aos sistemas existentes de recuperação da informação.

Os sintagmas nominais podem ser classificados de acordo com o seu grau de complexidade. Quanto maior for o grau do sintagma (ou, em outras palavras, quanto maior for o seu nível), mais delimitada é a informação nela contida:

Sintagma nominal de primeiro nível: as bibliotecas

Sintagma nominal de segundo nível: as bibliotecas universitárias

Sintagma nominal de terceiro nível: sistema de bibliotecas universitárias

No processo de indexação tradicional, considera-se somente o núcleo dos sintagmas nominais para a representação e recuperação dos documentos. Uma abordagem alternativa foi proposta por Kuramoto (1996) ao desenvolver um estudo utilizando os sintagmas nominais no tratamento e na recuperação de informações.

4.5 O Formato PDF

O PDF (*Portable Document Format*) é um formato de arquivo para distribuição e troca de documentos eletrônicos. Seu formato preserva as fontes, figuras, gráficos e o *layout* do documento original, independentemente do aplicativo e plataforma usados para criá-lo. É possível também controlar o acesso ao documento (através de senhas), impedir a impressão e/ou extração de texto. Os arquivos PDF são compactos e podem ser compartilhados, exibidos e impressos por qualquer usuário do software gratuito *Adobe Reader*. De acordo com a *Adobe Systems* (2004), empresa que comercializa o programa, mais de 500 milhões de cópias da versão gratuita do software (*Adobe Reader*) já foram distribuídas.

O formato PDF é originário de aprimoramentos ao padrão *postscript* (utilizados em arquivos de impressão). Por ser um formato de arquivo proprietário, a *Adobe Systems* define suas especificações e as publica a cada nova versão do software.

4.5.1 O Programa Adobe Acrobat

O programa *Adobe Acrobat* é o software que possibilita a criação de arquivos PDF. Ao contrário do *Adobe Reader*, é um software pago (sua licença custa entre R$ 900,00 e R$ 1.500,00) e, atualmente, encontra-se na versão 6.0 disponível em três modalidades: *Standard*, *Professional* e *Elements*. Além de converter documentos para o formato PDF para distribuição eletrônica, o software também permite a criação de formulários eletrônicos e inserção de arquivos multimídia em arquivos PDF (ADOBE SYSTEMS, 2004).

Além do *Adobe Acrobat*, há também outros softwares disponíveis gratuitamente na Internet para geração de arquivos PDF como, por exemplo, o PDF Livre disponível em português para ambiente Windows.

4.5.2 O Adobe Catalog

O *Adobe Catalog* é uma função disponível a partir da versão 4.0 do *Adobe Acrobat* que possibilita a geração de índices para um conjunto de documentos em

formato PDF. A seguir, serão listadas as configurações disponíveis para a geração do índice através desta funcionalidade oferecida pelo software.

4.5.3 Configurações do Adobe Catalog *Quanto ao Acervo a Ser Indexado*

O *Adobe Catalog* possibilita a indexação de uma ou mais pastas de documentos eletrônicos e suas respectivas subpastas. Na caixa de diálogo de definição do índice (*Index Definition*) pode-se definir o(s) diretório(s) cujo(s) documento(s) fará(ão) parte do índice, adicionando-os na seção "*Include These Directories*". Caso haja alguma pasta com documentos que não necessitam ou devem fazer parte do índice, ela deverá ser incluída na seção "*Exclude These Directories*" para que seus documentos não sejam indexados.

4.5.4 Configurações do Adobe Catalog Quanto ao Índice a Ser Gerado

De acordo com a *Adobe Systems* (2001), o *Adobe Catalog* oferece as seguintes funcionalidades para a definição do índice:

a) Palavras para não serem incluídas no índice (*Words not to include in the index*): o *Adobe Catalog* possibilita a definição de uma lista de palavras que não farão parte do índice por serem semanticamente vazias (por exemplo: artigos, conjunções, preposições, etc.), reduzindo o tamanho do conjunto dos

arquivos de índice. O programa possibilita a utilização de uma lista de até 500 palavras de até 24 caracteres cada;

b) Não incluir números (*Do not include numbers*): caso esta opção seja marcada, o índice gerado não possibilitará a busca de números nos documentos;

Em *Word Options* (opções de palavra) são disponibilizadas três configurações referentes às palavras que farão parte do índice:

a) *case sensitive*: caso esta opção seja marcada, o *Adobe Catalog* possibilitará recuperação do texto ou expressão exatamente como foi entrada no campo de busca quanto ao uso de maiúsculas e minúsculas;
b) *sounds like*: caso esta opção seja marcada, a opção "*Sounds Like*" estará disponível na interface de busca, possibilitando a expansão das buscas por nomes próprios;
c) *word stemming*: esta opção, quando selecionada, ativa a visualização "*Word Assistant*" no momento da busca, possibilitando o acesso ao índice para seleção do termo para a busca.

Finalmente, na parte inferior da janela de configuração do índice, encontram-se as seguintes configurações:

a) *Optimize for CD-ROM*: caso esta opção seja ativada, os arquivos do índice serão dispostos de maneira a agilizar o processo de busca, quando gravados em CD-ROM;

b) *Add IDs to Acrobat 1.0 PDF files*: ativando esta opção, o programa gera uma etiqueta eletrônica nos arquivos PDF gerados na versão 1.0, tornando-os compatíveis para o processo de indexação e recuperação de documentos.

5 METODOLOGIA

O objeto do presente estudo é a análise da indexação automática realizada pela função *Adobe Catalog*, disponível na versão 5.0 do software *Adobe Acrobat*, através do desempenho do índice durante as buscas realizadas através das palavras-chave contidas nos documentos indexados eletronicamente.

O *corpus* de pesquisa são os vinte Trabalhos de Conclusão de Curso (TCC) do segundo semestre de 2003 do Curso de Biblioteconomia da Universidade Federal do Rio Grande do Sul, disponibilizados pela coordenação da disciplina BIB 03037 – Trabalho de Conclusão de Curso em formato eletrônico (CD-ROM).

Para que o trabalho de análise fosse realizado, foram necessários alguns procedimentos prévios descritos a seguir:

a) escolha da versão do software a ser utilizada: embora a versão 6.0 do *Adobe Acrobat Professional* seja a mais recente (e atualmente disponível no mercado), optou-se por utilizar a versão anterior do software (5.0) pelo fato de o computador disponibilizado para a bateria de testes de indexação automática ter o sistema operacional *Windows* 98, incompatível com a versão mais recente do *Adobe Acrobat* na versão *Professional*.
b) preparo dos documentos para indexação automática: de um total de 20 trabalhos disponíveis no CD-ROM, quatro deles estavam em formato *Word* (.DOC) e precisaram ser convertidos para o formato PDF, para que pudessem ser indexados pelo software. A seguir, cada um dos trabalhos recebeu uma codificação alfabética para a sua identificação.

c) configuração dos parâmetros da indexação automática: após a etapa de preparo dos documentos, o recurso *Adobe Catalog*, disponível na versão 5.0 do software *Adobe Acrobat* foi configurado com os seguintes parâmetros:

- Utilização de *stopwords* (palavras que não entram no processo de indexação): por não serem semanticamente válidas e, conseqüentemente, não auxiliarem na recuperação de documentos, as preposições, os artigos e suas variações foram incluídos na lista de *stopwords* configurada pelo usuário que gera o índice automatizado. As seguintes palavras foram desconsideradas na geração do índice avaliado neste estudo: o, os, a, as, um, uns, uma, umas, algum, alguns, alguma, algumas, ante, após, até, com, contra, de, desde, em, entre, para, per, perante, por, sem, e, ou, O, Os, A, As, Um, Uns, Uma, Umas, Algum, Alguns, Alguma, Algumas, Ante, Após, Até, Com, Contra, De, Desde, Em, Entre, Para, Per, Perante, Por, Sem, E, Ou;
- Inclusão de números no índice: como o conjunto de documentos a ser indexado não apresentava números que deveriam fazer parte do índice, optou-se por desconsiderá-los para a geração do mesmo, deixando a opção *do not include numbers* (não incluir números) ativada;
- Sensibilidade a maiúsculas/minúsculas: visando facilitar o acesso aos documentos, a opção *case sensitive* foi desativada, permitindo que uma palavra ou expressão consiga recuperar documentos, independente do uso de maiúsculas ou minúsculas que o usuário tenha feito;

– Similaridade (*sounds like*): de acordo com a documentação que acompanha o software, a ativação desta opção possibilita a recuperação de grafias diferentes para nomes próprios, justificando a sua utilização;
– Palavras de mesmo radical: com o objetivo de facilitar a recuperação de documentos, optou-se por deixar a opção *word stemming* ativada, possibilitando a recuperação de documentos que compartilham o mesmo radical da palavra usada como termo de busca.

d) geração do índice automático: após o processo de configuração dos parâmetros do índice a ser gerado, o processo de indexação automática foi iniciado. A Figura 1 ilustra o processo de acompanhamento da indexação automática realizada pelo *Adobe Acrobat*:

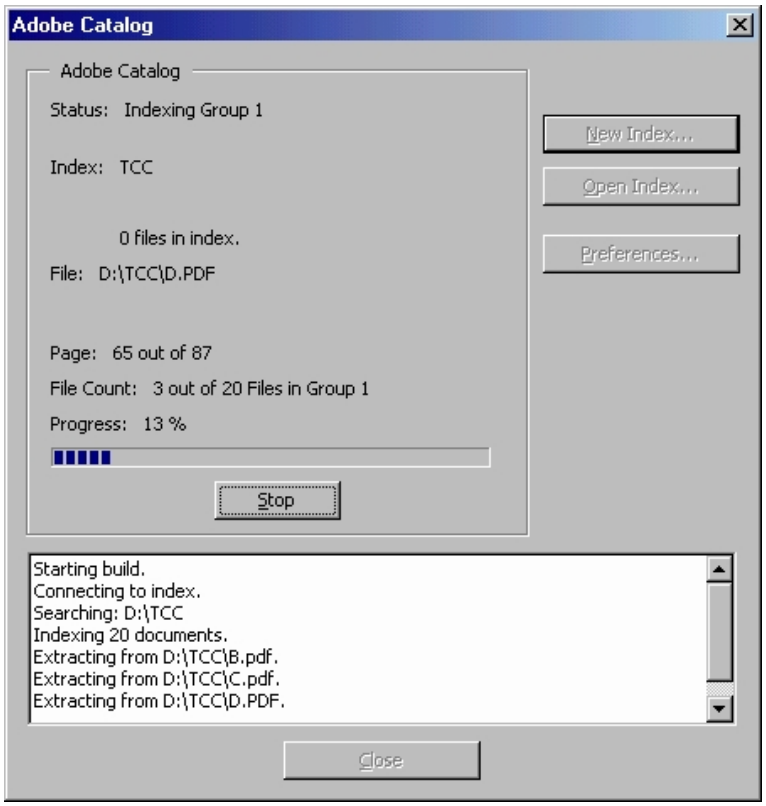

Figura 1 – Geração do Índice Automático

Durante o processo de geração do índice, o programa cria um arquivo de registro (*log*) contendo data, hora e ação executada, permitindo a análise posterior caso seja necessário. O arquivo de registro gerado durante a realização deste estudo encontra-se no ANEXO.

Depois de configurado, o Adobe Catalog segue as atvidades representadas no fluxograma a seguir para a geração do índice automático:

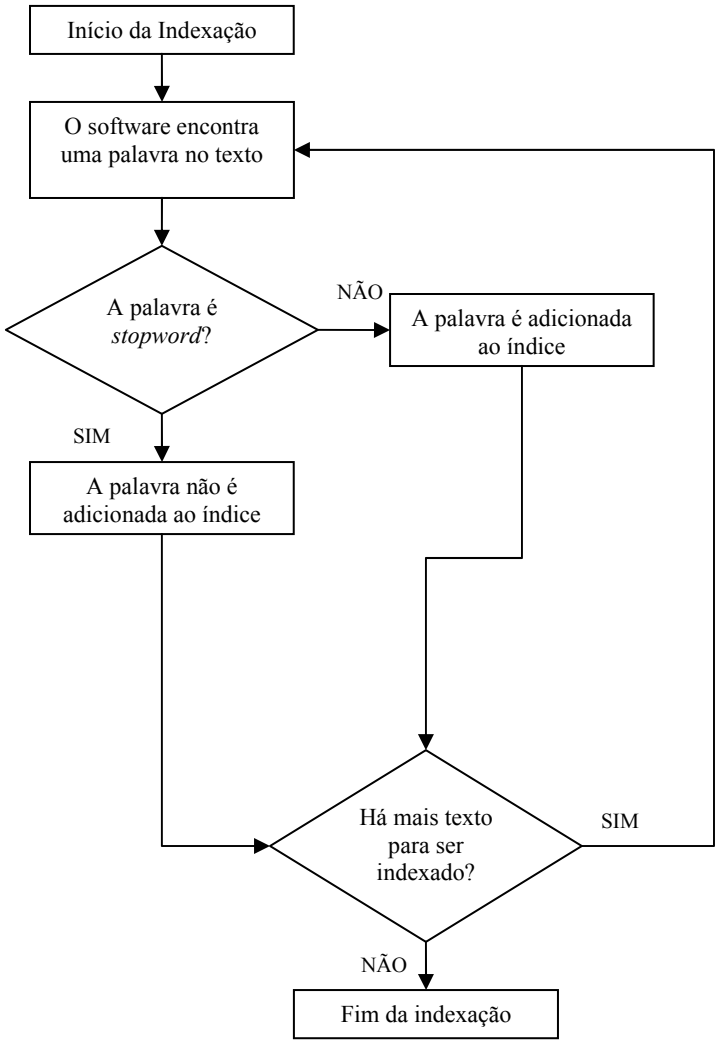

Figura 2 – Fluxograma do Processo de Indexação Automática no *Adobe Acrobat* Versão 5.0

O *Adobe Catalog* possibilita o acesso ao seu índice gerado automaticamente. Para tanto, é necessário que, durante o processo de configuração do índice (anterior a sua geração), a opção "Word Stemming" esteja ativada (vide 5.3 item "e"). O acesso ao índice é feito a partir do *Word Assistant*, opção disponível no menu *File* > *Search* > *Word Assistant*. A seguir, encontra-se a Figura 3 que ilustra o índice gerado eletronicamente.

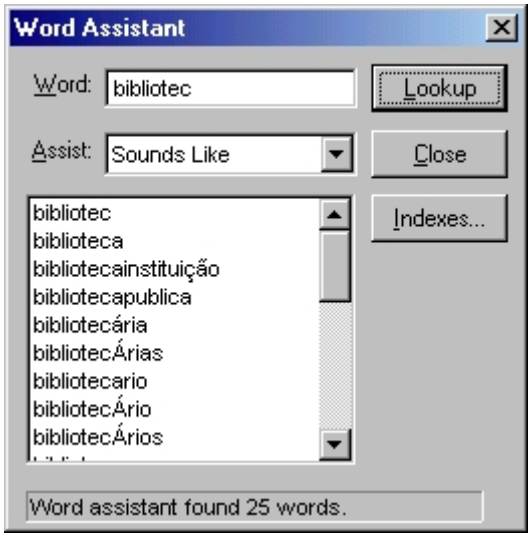

Figura 3 – Índice Gerado pelo *Adobe Acrobat* Versão 5.0

O índice gerado pelo *Adobe Acrobat* é composto de dois elementos: o arquivo de definição do índice com a extensão PDX e sua pasta auxiliar que é criada automaticamente durante o processo da geração do índice. Esses arquivos devem estar contidos em uma única pasta para que o programa funcione corretamente. O índice dos 20 TCC gerado neste trabalho (excluindo os trabalhos propriamente ditos) ocupou 1,69Mb de espaço em disco.

A Figura 4 mostra a estruturação das pastas e seus arquivos:

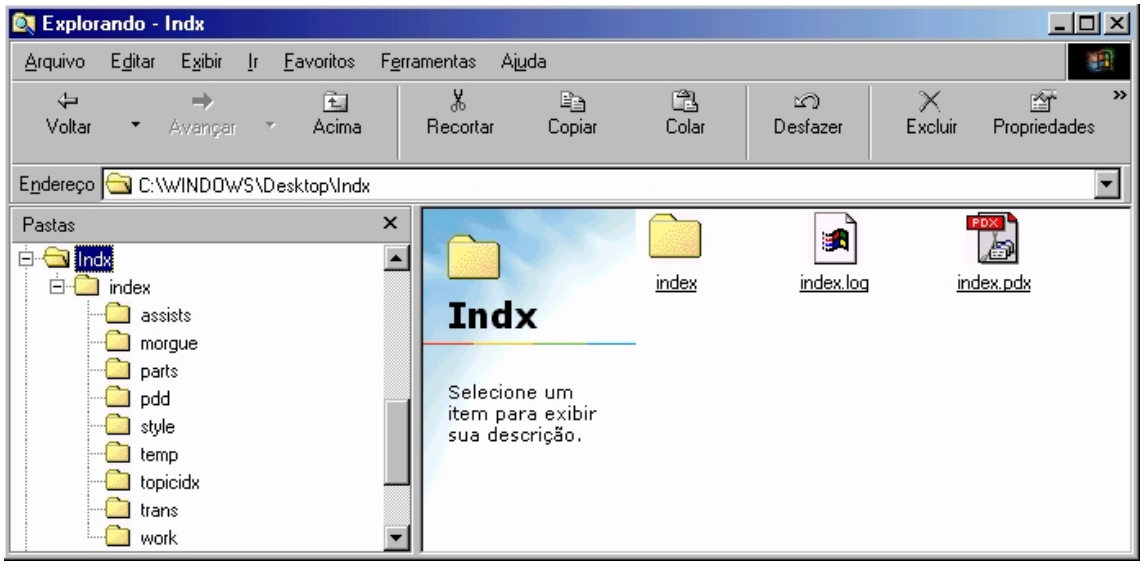

Figura 4 – Estrutura dos Arquivos de Índice Gerados pelo *Adobe Acrobat* Versão 5.0

Terminado o processo de indexação do conjunto de documentos, cada uma das palavras-chave atribuídas nos trabalhos de conclusão foi utilizada como termo de busca. O conjunto de documentos recuperados em cada uma das buscas foi anotado na ficha de coleta de dados (APÊNDICE A). O mesmo procedimento foi adotado para as variações (singular/plural, masculino/feminino, etc.) das palavras-chave. A metodologia detalhada para avaliação de cada uma das variáveis envolvidas na indexação está descrita a seguir.

5.1 Variações de Gênero

As palavras-chave apresentadas nos TCC que eram passíveis de flexão de gênero (como, por exemplo, usuário e bibliotecário gestor) foram utilizadas como termos de

busca e os documentos recuperados foram registrados no formulário de observação. Uma amostra do conjunto de palavras-chave que sofreram variação de gênero encontra-se no APÊNDICE B.

5.2 Variações de Número

A metodologia utilizada para analisar as variações de gênero também foi adotada para analisar as variações de número: as palavras-chave expressas nos TCC que eram passíveis de flexão de número (como, por exemplo, biblioteca universitária e aluno de biblioteconomia) foram usadas como termos de busca e os documentos recuperados foram registrados no formulário de observação. Uma amostra da listagem de palavras-chave que sofreram variação de número encontra-se no APÊNDICE C.

5.3 Equivalência

A análise da equivalência foi realizada através da leitura dos trabalhos de conclusão. As diversas expressões utilizadas pelos autores do conjunto de trabalhos de conclusão para expressar o mesmo conceito de cada uma das palavras-chave dos TCC foram anotadas e utilizadas como termos de busca no índice gerado automaticamente pelo software. Uma amostra da lista de palavras-chave equivalentes encontra-se no APÊNDICE D.

5.4 Especificidade

A análise da especificidade da indexação foi realizada de três maneiras distintas:

a) através da leitura dos TCC: os trabalhos foram lidos e as palavras-chave foram analisadas quanto à sua especificidade;
b) através da busca por proximidade: a utilização deste recurso para a análise da especificidade não se mostrou preciso, pois a proximidade é considerada válida pelo software quando os termos de busca estão até três páginas distantes um do outro;
c) com o uso de operadores booleanos. A busca pela palavra-chave "Fontes de Informação Jurídica" não produziu resultados, porém a busca feita pela expressão fontes and informação jurídica, recuperou três documentos, entre eles o que continha a palavra-chave em questão (H).

Uma amostra da lista de palavras-chave e seus termos específicos encontra-se no APÊNDICE E.

5.5 Sintagmas Nominais

Para realizar a análise da recuperação dos sintagmas nominais, cada uma das palavras-chave foi classificada de acordo com o seu grau de complexidade, de acordo

com os princípios definidos por Kuramoto (1996). Após esta etapa, cada uma das palavras-chave foi utilizada como termo de busca no índice gerado automaticamente pelo *Adobe Acrobat* e os resultados das buscas foram anotados na ficha de observação. Uma amostra da listagem de palavras-chave e seus respectivos graus de complexidade de sintagmas nominais encontra-se no APÊNDICE F.

6 INSTRUMENTOS METODOLÓGICOS

Para auxiliar a realização deste estudo, foi necessário utilizar os seguintes instrumentos para coleta, armazenamento e análise dos dados.

6.1 Ficha de coleta

Para realizar a coleta de dados, foi elaborada uma ficha de observação (APÊNDICE A) contendo campos para o registro da palavra-chave, suas variantes (plural, singular, masculino, feminino, etc.) para que os documentos recuperados em cada uma das buscas feitas fossem anotados.

6.2 Base de Dados

Para fazer o arranjo das palavras-chave e possibilitar uma melhor manipulação dos registros feitos durante a coleta de dados, foi desenvolvida uma base de dados em CDS/ISIS para Windows, popularmente conhecido como Winisis, utilizando sua versão mais recente (1.5 build 3) disponível para download no site da UNESCO.

7 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A seguir, encontram-se a análise e discussão dos resultados do presente estudo.

7.1 Questões gerais de busca

Das 81 palavras-chave encontradas ao longo dos 20 TCC do curso de Biblioteconomia da Universidade Federal do Rio Grande do Sul do segundo semestre de 2003, 62 delas (77%) recuperaram seus documentos de origem. O Gráfico 1 mostra esta proporção:

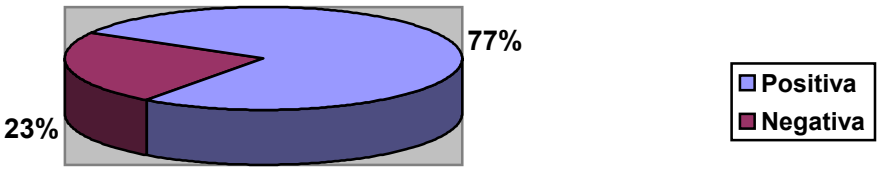

Gráfico 1 – Recuperação dos Trabalhos através das Palavras-Chave

De acordo com Lancaster (1993), o percentual mencionado acima se encontra dentro dos padrões razoáveis de recuperabilidade de documentos, tendo em vista que os termos utilizados para as buscas eram as próprias palavras-chave.

7.2 Limitações do Sistema

Os termos de busca devem ser, necessariamente, acentuados corretamente (como figuram nos documentos, partindo do princípio de que estejam redigidos corretamente quanto à grafia) para que os resultados das buscas realizadas não sofram interferências.

Outro problema verificado durante o desenvolvimento do trabalho foi a não recuperação de documentos.

Do conjunto de 20 trabalhos analisados, treze deles apresentaram uma ou mais palavras-chave (de um total de 18) que, ao serem utilizadas como termos de busca, produziram um resultado de 'falso negativo', não recuperando nenhum documento. O Gráfico 2 mostra os percentuais de trabalhos que apresentaram problemas na recuperação.

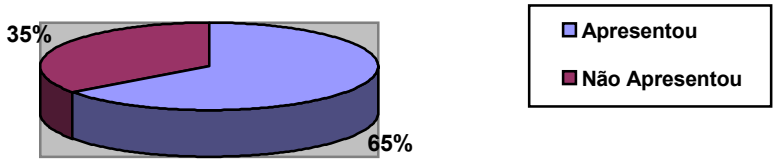

Gráfico 2 – Percentual de Trabalhos que Apresentaram Problemas na Recuperação

A palavra-chave "Comportamento de Busca", presente em três trabalhos recuperou apenas um deles quando utilizada como termo de busca no índice produzido automaticamente pelo *Adobe Catalog*.

7.3 Questões de Gênero

Dentre as 81 palavras-chave presentes nos TCC, apenas quatro delas (5%) eram passíveis de flexão de gênero e uma delas (1,23%) era comum de dois gêneros (profissional da informação) cuja flexão dependia do artigo que a precedia. O Gráfico 3 apresenta os percentuais das palavras-chave que foram flexionadas por gênero.

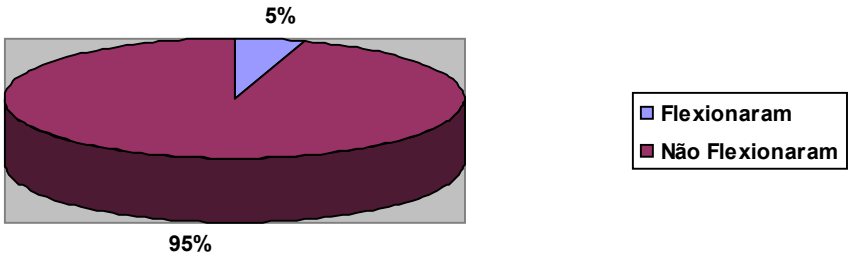

Gráfico 3 - Percentual de Palavras-Chave Flexionadas por Gênero

Três das cinco variantes de gênero das palavras-chave não produziram resultados nas buscas realizadas. Já a busca pela forma feminina da palavra-chave "usuário" recuperou 5 documentos, sendo que em dois deles (os de códigos C e T, onde é recuperada a palavra "usuária" na expressão "comunidade usuária") a palavra "usuária" não era o núcleo do sintagma nominal, aumentando a revocação de documentos, dispendendo um maior tempo para a localização dos resultados.

Outra possibilidade de busca foi a utilização de "caracteres-máscara". O ponto de interrogação (?) foi utilizado para substituir um caractere, possibilitando a recuperação de palavras independente do seu gênero no teor dos TCC (por exemplo: a

expressão usuári? recuperaria os termos usuário e usuária). O asterisco (*), utilizado para substituir mais de um caractere, não foi utilizado.

7.4 Questões de Número

Do conjunto de 81 palavras-chave coletadas nos 20 TCC de 2003/2, 36 delas (44%) foram flexionadas quanto ao número. As palavras-chave que estavam no singular foram pluralizadas e vice-versa. O Gráfico 4 mostra a proporção de palavras-chave flexionadas quanto ao número.

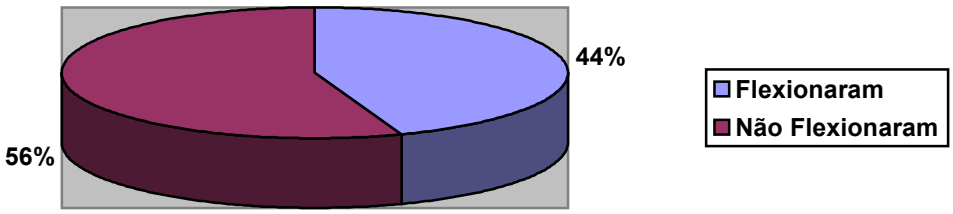

Gráfico 4 – Percentual de Palavras-Chave Flexionadas por Número

Do conjunto de 36 palavras que sofreram flexão de número, 6 apresentaram divergências na recuperação pois uma de suas formas (singular ou plural) não recuperou nenhum documento. O Gráfico 5, situado na próxima página, mostra o percentual de palavras-chave flexionadas por número que apresentou problemas na recuperação de documentos.

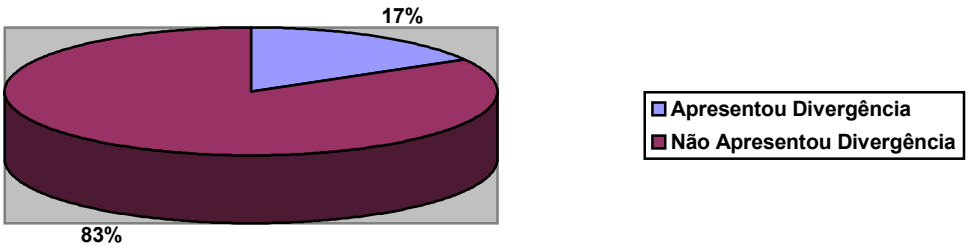

Gráfico 5 – Percentual de Flexões de Número que Apresentou Divergência na Recuperação

7.5 Questões de Especificidade

Durante a leitura dos TCC analisados, apenas os trabalhos P e S apresentaram problemas de especificidade.

No trabalho P, havia a presença de um termo genérico (Bibliotecas) e de seus termos específicos (Bibliotecas Escolares, Bibliotecas Públicas e Bibliotecas Comunitárias).

Já no trabalho S, os termos genéricos (Universidade e Biblioteca) foram atribuídos pelo autor, bem como termos específicos (Universidade Virtual, Universidade Corporativa, Universidade Corporativa CAIXA, Biblioteca Virtual, Biblioteca Especializada).

Nos demais trabalhos, o vocabulário dos autores para expressar os assuntos de seus trabalhos apresentaram um nível de especificidade adequado e não apresentaram conceitos implícitos que poderiam causar problemas na recuperação dos documentos.

7.6 Questões de Equivalência

A análise da equivalência foi dividida em dois grandes grupos:

a) equivalentes presentes em um mesmo documento: nesta circunstância, a recuperação de documentos através de expressões equivalentes às palavras-chave encontradas no texto de cada um dos TCC dependeu, predominantemente, do uso das mesmas pelos autores ao longo de seus trabalhos. Por exemplo: a palavra-chave "Programa Sociedade da Informação no Brasil" e sua denominação abreviada "SocInfo", presentes em um dos TCC analisados, não apresentaram problemas na recuperação;

b) equivalentes presentes em trabalhos distintos: quando os termos equivalentes estavam presentes em documentos diferentes, os problemas de recuperação se tornavam mais visíveis. A palavra-chave "Educação à Distância", presente no trabalho I, não apresentou em seu teor a sua expressão equivalente (Ensino à Distância). Esta última, por sua vez, recuperou apenas um documento (o trabalho S). Pode-se concluir, portanto, que, nesta situação, o software não está preparado para lidar corretamente com questões de equivalência.

7.7 Recuperação dos Sintagmas Nominais

O comportamento do software em relação à recuperação de sintagmas nominais foi satisfatório. A Tabela 1 apresenta a distribuição do conjunto de 81 palavras-chave analisadas, de acordo com os níveis de sintagma nominal:

Tabela 1 – Distribuição das Palavras-Chave dos TCC de Acordo com os Níveis de Sintagma Nominal

Tipo	*f*	*%*
Sintagma Nominal de Primeiro nível	21	25,9
Sintagma Nominal de Segundo nível	45	55,5
Sintagma Nominal de Terceiro nível	12	14,8
Sintagma Nominal de Quarto nível	3	3,7
TOTAL	81	100

A Tabela 2 mostra a distribuição do conjunto das palavras-chave que não recuperaram documentos, de acordo com os níveis de sintagma nominal:

Tabela 2 – Distribuição das Palavras-Chave que Não Recuperaram Documentos de Acordo com os Níveis de Sintagma Nominal

Tipo	*f*	*%*
Sintagma Nominal de Primeiro nível	-	-
Sintagma Nominal de Segundo nível	10	55,5
Sintagma Nominal de Terceiro nível	8	44,5
Sintagma Nominal de Quarto nível	-	-
TOTAL	18	100

8 CONCLUSÕES E RECOMENDAÇÕES

O presente trabalho não teve a intenção de encerrar os estudos sobre a indexação automática realizada através da função *Adobe Catalog*, mas sim propiciar uma discussão e subsidiar futuros estudos utilizando versões mais recentes do software, bem como outras ferramentas de indexação automática. As principais conclusões decorrentes do presente estudo encontram-se a seguir.

A interface de busca do índice não mostra as últimas pesquisas realizadas e, conseqüentemente, não permite o cruzamento entre elas. A possibilidade de combinação de pesquisas é um recurso interessante para usuários que não estão familiarizados com operadores booleanos, pois se evita a estruturação de expressões de busca muito complexas que correm o risco de serem elaboradas erroneamente.

Os termos de busca poderiam ser considerados válidos com ou sem acentuação, facilitando a recuperação de informações. De acordo com a maneira que foi concebido, o software considera os termos usuário (com acento) e usuario (sem acento) como termos de busca distintos, produzindo diferentes resultados com a primeira grafia e não trazendo registros com o uso da segunda.

A recuperação de sintagmas nominais foi considerada satisfatória: das 81 palavras-chave encontradas nos TCC de 2003/2, a sua grande maioria (63 delas, ou 77,7%) apresentou resultado positivo quando utilizada como termos de busca, recuperando seus documentos de origem.

Notou-se, também, que todas as palavras-chave que apresentaram problemas de recuperação eram sintagmas nominais de níveis 2 e 3. Estas palavras-chave que apresentaram problemas de recuperação foram partidas e seus elementos foram ligados

pelo operador booleano AND para a realização de novas buscas. Utilizando-se esta sistemática, apenas uma das 18 palavras-chave não recuperou o seu documento de origem, apontando para outro trabalho de conclusão do acervo. Não foi possível detectar um padrão de palavra-chave que gerasse o silêncio durante a busca, mas foi constatado que a utilização de sintagmas menos complexos combinados por operadores booleanos produzem melhores resultados.

Durante a configuração do índice a ser gerado, o software oferece a possibilidade de exclusão de palavras não significativas (chamadas de *stopwords*). Seria interessante que também fosse oferecida a opção para o usuário definir quais palavras deveriam ser incluídas na elaboração do mesmo, ou seja, a estruturação de vocabulário para a geração do índice, possibilitando um direcionamento no processo de indexação.

Além disso, o software poderia permitir ao usuário que irá configurar os parâmetros da indexação automática quais os sufixos mais comuns em outras línguas, além do inglês.

A possibilidade de busca por proximidade oferecida pela indexação realizada no *Adobe Acrobat* é muito ampla (o software considera uma proximidade válida caso os termos de busca estejam até três páginas distantes um do outro). A recuperação por proximidade de termos poderia ser mais eficaz se o software permitisse, durante a elaboração da estratégia de busca, a seleção do número de palavras que poderiam existir entre os termos de pesquisa digitados, possibilitando uma melhor recuperação de sintagmas nominais utilizados no teor dos documentos.

Apesar de suas limitações, algumas decorrentes do processo de análise do software (que tem como base cada uma das palavras presentes nos textos) outras, oriundas do baixo nível de customização dos parâmetros de indexação, o *Adobe Catalog* conseguiu, de maneira satisfatória, recuperar os documentos indexados em um acervo

especializado na área de Ciência da Informação. Espera-se que o aprimoramento desta ferramenta possa contribuir para o aumento de sua eficácia, justificando sua utilização não só em acervos de uma área específica do conhecimento, bem como acervos de cunho geral.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 12676*: métodos para análise de documentos – seleção de termos de indexação. Rio de Janeiro, 1992.

ADOBE SYSTEMS INCORPORATED. *Acrobat 5.0*: guia autorizado Adobe. Rio de Janeiro: Campus, 2001.

_____. *O Que é Adobe PDF?* Disponível em: <http://www.adobe.com.br/products/acrobat/adobepdf.html>. Acesso em: 20 jun. 2004.

KURAMOTO, H. Uma Abordagem Alternativa para o Tratamento e a Recuperação de Informação Textual: os sintagmas nominais. *Ciência da Informação*, Brasília, v. 25, n. 2, p. 182-192, maio/ago. 1996.

LANCASTER, F. W. *Indexação e Resumos*: teoria e prática. Brasília: Briquet de Lemos/Livros, 1993.

LIMA, G. A. Interfaces entre a Ciência da Informação e a Ciência Cognitiva. *Ciência da Informação*, Brasília, v. 32, n. 1, p. 77-87, jan./abr. 2003.

LOPES, I. L. Uso das Linguagens Controlada e Natural em Bases de Dados: revisão da literatura. *Ciência da Informação*, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

MUDDAMALLE, M. R. Natural Language versus Controlled Vocabulary in Information Retrieval: a case study in soil mechanics. *Journal of the American Society for Information Science*, Silver Spring, v. 49, n. 10, p. 881-887, Oct. 1998.

PINTO, V. B. Indexação Documentária: uma forma de representação do conhecimento registrado. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 6, n. 2, p. 223-234, jul./dez. 2001.

VIEIRA, S. B. Indexação Automática e Manual: revisão de literatura. *Ciência da Informação*, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988.

APÊNDICE A – Modelo de Ficha de Coleta de Dados

Palavra-chave:			
Variante 1	Variante 2	Variante 3	Variante 4
Documentos recuperados			
1)	1)	1)	1)
2)	2)	2)	2)
3)	3)	3)	3)
4)	4)	4)	4)
5)	5)	5)	5)
6)	6)	6)	6)
7)	7)	7)	7)
8)	8)	8)	8)
9)	9)	9)	9)
10)	10)	10)	10)
11)	11)	11)	11)
12)	12)	12)	12)
13)	13)	13)	13)
14)	14)	14)	14)
15)	15)	15)	15)
16)	16)	16)	16)
17)	17)	17)	17)
18)	18)	18)	18)
19)	19)	19)	19)
20)	20)	20)	20)

APÊNDICE B – Palavras-Chave Flexionadas por Gênero

Palavra-Chave (10): Alunos da Faculdade de Direito da UFRGS
Trabalhos (20): H
Flex. Gênero (50): Sim

Palavra-Chave (10): Alunos de Biblioteconomia
Trabalhos (20): D
Flex. Gênero (50): Sim

Palavra-Chave (10): Bibliotecário gestor
Trabalhos (20): E
Flex. Gênero (50): Sim

Palavra-Chave (10): Profissional da informação
Trabalhos (20): S
Flex. Gênero (50): Sim

Palavra-Chave (10): Usuário
Trabalhos (20): A
Flex. Gênero (50): Sim

APÊNDICE C – Amostra de Palavras-Chave Flexionadas por Número

Palavra-Chave (10): Biblioteca universitária
Trabalhos (20): C
Trabalhos (20): F
Trabalhos (20): T
Flex. Número (60): Sim

Palavra-Chave (10): Registros bibliográficos
Trabalhos (20): G
Flex. Número (60): Sim

Palavra-Chave (10): Telecentros comunitários
Trabalhos (20): L
Flex. Número (60): Sim

Palavra-Chave (10): Bibliotecas públicas
Trabalhos (20): P
Trabalhos (20): R
Flex. Número (60): Sim

Palavra-Chave (10): Estudos de comunidade
Trabalhos (20): R
Flex. Número (60): Sim

Palavra-Chave (10): Produtos de informação
Trabalhos (20): S
Flex. Número (60): Sim

APÊNDICE D – Amostra de Palavras-Chave Equivalentes

Palavra-Chave (10): Educação à Distância
Trabalhos (20): I

Palavra-Chave (10): Ensino à Distância
Trabalhos (20): S

Palavra-Chave (10): Serviço de Referência
Trabalhos (20): F

Palavra-Chave (10): Serviço de Referência e Informação
Trabalhos (20): K

Palavra-Chave (10): Programa Sociedade da Informação no Brasil
Trabalhos (20): N

Palavra-Chave (10): SocInfo
Trabalhos (20): N

APÊNDICE E – Amostra de Palavras-Chave e Seus Termos Específicos

Palavra-Chave (10): Bibliotecas
Trabalhos (20): P
Trabalhos (20): S

Palavra-Chave (10): Bibliotecas Públicas
Trabalhos (20): P
Trabalhos (20): R

Palavra-Chave (10): Informação
Trabalhos (20): J

Palavra-Chave (10): Informação Especializada
Trabalhos (20): A
Trabalhos (20): K

Palavra-Chave (10): Universidade
Trabalhos (20): S

Palavra-Chave (10): Universidade Corporativa
Trabalhos (20): S

APÊNDICE F – Amostra de Palavras-Chave e Seus Respectivos Níveis Quanto ao Sintagma Nominal

Palavra-Chave (10):Internet
Trabalhos (20):D
SN (70): 1

Palavra-Chave (10): Impactos
Trabalhos (20): L
SN (70): 1

Palavra-Chave (10): Alunos de Biblioteconomia
Trabalhos (20): D
SN (70): 2

Palavra-Chave (10): Seleção de informações
Trabalhos (20): H
SN (70): 2

Palavra-Chave (10): Gestão de unidades de informação
Trabalhos (20): E
SN (70): 3

Palavra-Chave (10): Serviço de referência e informação
Trabalhos (20): K
SN (70): 3

Palavra-Chave (10): Alunos da Faculdade de Direito da UFRGS
Trabalhos (20): H
SN (70): 4

ANEXO – Arquivo de Registro (*log*) Gerado pelo *Adobe Acrobat*

```
09/16/2004 13:39:04: Starting build.

09/16/2004 13:39:05: Connecting to index.

09/16/2004 13:39:07: Searching: C:\WINDOWS\Desktop\TCC Etapa 2

09/16/2004 13:39:07: Indexing 20 documents.

09/16/2004 13:39:07: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\B.pdf.

09/16/2004 13:39:11: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\C.pdf.

09/16/2004 13:39:17: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\D.PDF.

09/16/2004 13:39:20: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\E.pdf.

09/16/2004 13:39:24: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\F.pdf.

09/16/2004 13:39:30: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\G.pdf.

09/16/2004 13:39:35: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\A.pdf.

09/16/2004 13:39:38: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\H.pdf.

09/16/2004 13:39:45: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\I.pdf.

09/16/2004 13:39:48: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\J.PDF.

09/16/2004 13:39:49: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\K.pdf.

09/16/2004 13:39:54: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\L.pdf.

09/16/2004 13:39:59: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\M.pdf.

09/16/2004 13:40:01: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\N.pdf.

09/16/2004 13:40:15: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\O.pdf.

09/16/2004 13:40:21: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\P.pdf.

09/16/2004 13:40:26: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\Q.PDF.

09/16/2004 13:40:29: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\R.pdf.

09/16/2004 13:40:35: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\S.PDF.

09/16/2004 13:40:38: Extracting from C:\WINDOWS\Desktop\TCC Etapa 2\T.pdf.

09/16/2004 13:40:55: Search engine is reextracting text from C:\WINDOWS\Desktop\TCC
Etapa 2\T.pdf after writing a temporary index.

09/16/2004 13:41:11: Removing index entries for changed or deleted documents.

09/16/2004 13:41:11: Removing unnecessary index entries.

09/16/2004 13:41:11: Building Word Assist List.

09/16/2004 13:41:14: Waiting for about 20 seconds to perform routine index update.

09/16/2004 13:41:35: TesteCatalog - Index Build Successful.

09/16/2004 13:41:35:   Total Acrobat files in all directories: 20

09/16/2004 13:41:35:   Total new files: 20

09/16/2004 13:41:35:   Total pages indexed: 1739

09/16/2004 13:41:35:   Total number of files skipped: 0

09/16/2004 13:41:35:   Total deleted files: 0

09/16/2004 13:41:35: C:\WINDOWS\Desktop\TCC Etapa3\index.pdx - Index Build Successful.
```

10/11/2004 19:23:05: Purging index.

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1300 (Io): Couldn't delete C:\WINDOWS\Desktop\TCC Etapa3\index\parts\00000001.ddd delete error 13 (Permission denied)

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1313 (Io): Error deleting file C:/WINDOWS/Desktop/TCC Etapa3/index/parts/00000001.ddd

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1300 (Io): Couldn't delete C:\WINDOWS\Desktop\TCC Etapa3\index\parts\00000001.did delete error 13 (Permission denied)

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1313 (Io): Error deleting file C:/WINDOWS/Desktop/TCC Etapa3/index/parts/00000001.did

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1306 (Io): Couldn't delete directory C:\WINDOWS\Desktop\TCC Etapa3\index\parts rmdir error 13 (Permission denied)

10/11/2004 19:38:12: Search Engine Message: (0) Error E0-1311 (Io): Error removing directory C:/WINDOWS/Desktop/TCC Etapa3/index/parts

10/11/2004 19:40:07: Resetting document count to reflect actual number of documents in the index.

10/11/2004 19:40:29: Waiting for about 20 seconds to perform routine index update.

10/11/2004 20:30:23: Starting build.

10/11/2004 20:30:23: Connecting to index.

10/11/2004 20:30:23: Searching: D:\TCC

10/11/2004 20:30:24: Indexing 20 documents.

10/11/2004 20:30:24: Extracting from D:\TCC\B.pdf.

10/11/2004 20:30:29: Extracting from D:\TCC\C.pdf.

10/11/2004 20:30:35: Extracting from D:\TCC\D.PDF.

10/11/2004 20:30:39: Extracting from D:\TCC\E.pdf.

10/11/2004 20:30:42: Extracting from D:\TCC\F.pdf.

10/11/2004 20:30:49: Extracting from D:\TCC\G.pdf.

10/11/2004 20:30:54: Extracting from D:\TCC\A.pdf.

10/11/2004 20:30:58: Extracting from D:\TCC\H.pdf.

10/11/2004 20:31:06: Extracting from D:\TCC\I.pdf.

10/11/2004 20:31:09: Extracting from D:\TCC\J.PDF.

10/11/2004 20:31:10: Extracting from D:\TCC\K.pdf.

10/11/2004 20:31:16: Extracting from D:\TCC\L.pdf.

10/11/2004 20:31:21: Extracting from D:\TCC\M.pdf.

10/11/2004 20:31:25: Extracting from D:\TCC\N.pdf.

10/11/2004 20:31:43: Extracting from D:\TCC\O.pdf.

10/11/2004 20:31:53: Extracting from D:\TCC\P.pdf.

10/11/2004 20:32:05: Extracting from D:\TCC\Q.PDF.

10/11/2004 20:32:09: Extracting from D:\TCC\R.pdf.

10/11/2004 20:32:17: Extracting from D:\TCC\S.PDF.

10/11/2004 20:32:20: Extracting from D:\TCC\T.pdf.

```
10/11/2004 20:32:42: Search engine is reextracting text from D:\TCC\T.pdf after writing
a temporary index.

10/11/2004 20:33:03: Removing index entries for changed or deleted documents.

10/11/2004 20:33:03: Removing unnecessary index entries.

10/11/2004 20:33:03: Building Word Assist List.

10/11/2004 20:33:06: Waiting for about 20 seconds to perform routine index update.

10/11/2004 20:33:27: TesteCatalog - Index Build Successful.

10/11/2004 20:33:27:   Total Acrobat files in all directories: 20

10/11/2004 20:33:27:   Total new files: 20

10/11/2004 20:33:27:   Total pages indexed: 1739

10/11/2004 20:33:27:   Total number of files skipped: 0

10/11/2004 20:33:27:   Total deleted files: 0

10/11/2004 20:33:27: C:\WINDOWS\Desktop\Indx\index.pdx - Index Build Successful.
```