Precisão no processo de busca e recuperação da informação: uso da mineração de textos

Rogério Henrique de Araújo Júnior
Doutor em ciência da informação pela Universidade de Brasília.
E-mail: araujojr@unb.br

Kira Tarapanoff
Doutora em ciência da informação pela Universidade de Sheffield – Inglaterra.
E-mail: kat309@unb.br

Resumo

Trata da comparação entre a indexação manual e a ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. O estudo de caso escolhido para o desenvolvimento da pesquisa foi o Centro de Referência e Informação em Habitação (Infohab), cuja base de dados sobre habitação, saneamento e urbanização foi indexada de forma manual por bibliotecários da Caixa Econômica Federal, com base em uma lista de palavras-chave. Houve o desenvolvimento de um protótipo cujos itens bibliográficos correspondem às teses e dissertações contidas no Infohab, o que permitiu a aplicação do *software* BR/Search para a execução da mineração de textos. As pesquisas no Infohab e no protótipo foram realizadas a partir da demanda de especialistas da Caixa nos assuntos contidos na base. Conclui que não há ganhos significativos na precisão ao se aplicar a ferramenta de mineração de textos em relação à indexação manual.

Palavras-chave

Processo de recuperação da informação. Índice de precisão. Processo de indexação. Mineração de textos.

Precision rate in the information retrieval process: the use of text mining

Abstract

This research deals with the comparison between manual indexing and the text mining tool, using the analysis of reply precision rate in the information retrieval process. The case study selected for this research was the Centro de Referência e Informação em Habitação – Infohab. The center which contains the database on habitation, sanitation and urbanization was manually indexed by the librarians of Caixa Econômica Federal, using a list of key words. A prototype was developed, containing bibliographic references that corresponded to the theses and dissertations of Infohab, which allowed the application of BR/Search software of text mining. The researches performed on the prototype and in Infohab were demanded by specialists of Caixa in database subjects. The research gave evidenced that there are no significant profits in the precision rate with the applications of text mining tool in relation to the manual indexing.

Keywords

Information retrieval process. Precision rate. Indexing process. Text mining.

INTRODUÇÃO

Partindo da premissa de que a análise do documento é uma significativa contribuição para a comunicação e o fluxo da informação em qualquer organização e para qualquer sistema de recuperação da informação, o presente artigo, baseado em tese de doutorado defendida no Departamento de Ciência da Informação e Documentação da Universidade de Brasília, trata da comparação entre indexação manual e ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação.

O estudo de caso escolhido para o desenvolvimento da pesquisa foi o Centro de Referência e Informação em Habitação (Infohab) liderado pela Associação Nacional de Tecnologia do Ambiente Construído (Antac) que tem por finalidade a captação, seleção e divulgação de toda a informação acerca da tecnologia do ambiente construído, sobretudo sobre a habitação de interesse social, englobando a sua produção, manutenção e uso.

O ambiente construído envolve todas as atividades, recursos, conhecimento, *expertises*, experiências, tecnologia, equipamentos, instrumentos, mão-de-obra e mercado relacionados à habitação e às políticas públicas sobre saneamento, urbanização, além das questões técnicas que envolvem o estatuto das cidades.

O foco do estudo de caso incidiu sobre uma das competências do Infohab, a manutenção de uma base de dados atualizada com referências dos resultados de pesquisas, legislação federal, estadual/municipal, normas pertinentes, levantamentos governamentais e demais tipologias de documentação.

O universo da pesquisa foi representado por 1.520 documentos da Caixa Econômica Federal (Caixa) inseridos na base de dados do Infohab e, como amostra, o acervo de teses e dissertações inseridas no Infohab pela Centralizadora de Documentação e Informação (Cedin), vinculada à Gerência Nacional de *Marketing* Interno da Caixa (Gemac) (tabela 1, a seguir). A amostra contou com 56 itens bibliográficos, e a sua escolha deveu-se a três fatores:

TABELA 1
Estatística da base de dados do Infohab

TIPO DE ITEM BIBLIOGRÁFICO	QUANTIDADE
Anais de congresso	25
Artigo de congresso	15
Artigo de periódico	39
Dicionário	4
Documento sonoro	1
Especificação técnica	2
Folheto	528
Imagem em movimento	3
Legislação	3
Livro	795
Manual	42
Periódico	7
Tese e dissertação	56
Total de itens na base de dados	**1.520**

Fonte: Cedin/Caixa.

• todas as teses e dissertações que compõem a amostra são o resultado final de pesquisas realizadas pelo pessoal da Caixa nas áreas de saneamento, desenvolvimento urbano e habitação, sendo, portanto, uma amostra fiel dos assuntos de especialização do universo do trabalho: o ambiente construído. Correspondem também à documentação do capital intelectual* de uma das mais importantes áreas-fim da Caixa, o Desenvolvimento Urbano;

• a amostra corresponde ao terceiro tipo de bibliografia mais numeroso na base de dados do acervo Caixa no Infohab, conforme a tabela 1;

• universo relativamente "fácil", em comparação a outros tipos de documentos disponíveis na organização.

O PROBLEMA

O estudo consistiu em avaliar se no processo de busca e recuperação da informação a mineração de textos trouxe ganho no índice de precisão em relação à lista de palavras-chave utilizadas na indexação manual por bibliotecários da Caixa. Assim, o trabalho concentrou-se em comparar a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação.

Além disso, interessou ao escopo da investigação verificar a viabilidade de propor uma sistemática de uso dos termos gerados a partir de mineração de textos para auxiliar o processo de indexação manual no aumento do índice de precisão de resposta na recuperação da informação.

Como forma de operacionalização da metodologia da pesquisa, foi construído um protótipo cujos itens bibliográficos corresponderam às teses e dissertações contidas no Infohab, o que possibilitou o emprego do *software* BR/Search para a execução da mineração de textos.

A mineração de textos consiste na extração de informações sobre tendências ou padrões em grandes volumes de documentos textuais, em que uma amostra significativa de informações é avaliada em textos contidos em bases textuais e em fontes de informação em linha (POLANCO; FRANÇOIS, 2000).

As bases textuais são coleções de documentos em linguagem natural, sem formato predefinido para seus conteúdos, como acontece com as bases de dados. Dividem-se em bases textuais cujo conteúdo é estruturado de acordo com a sua localização no documento. Como exemplos, têm-se relatórios policiais, relatórios de instituições financeiras, ou seja, o conteúdo pode variar, mas a estrutura do documento é predefinida, e informações não estruturadas, em que se têm como exemplos os relatórios, publicações e a maioria dos documentos textuais (TRYBULA, 1999).

Na figura 1, a seguir, são apresentados os fatores de influência nos resultados de busca e recuperação da informação em uma base de dados.

Neste modelo adaptado de Lancaster (1998), visualizamos as peças básicas que compõem a problemática apresentada, em que o usuário, a partir de suas necessidades de informação, dá início ao processo de busca e recuperação da informação.

O especialista em informação, ao qual cabe a responsabilidade de indexador, vale-se da compreensão da demanda para trazer da base de dados, na forma do pedido, a resposta o mais adequada possível às necessidades do usuário. Além disso, Lancaster (1998) ressalta que a qualidade da estratégia de busca e o vocabulário são fatores importantes para a atividade.

Todavia, temos também as questões da precisão e da qualidade da própria base de dados, sem contar que o indexador (especialista em informação) depende dos

* **Capital intelectual** – compreende o conhecimento que é de valor para uma organização construída de capital humano, capital estrutural e capital-cliente. Acredita-se que este fator possa ser analisado para permitir classificar a organização como rica ou pobre em informação (Halal & Kull, 2000 *apud* Tarapanoff, 2001).

termos autorizados no vocabulário, a fim de lograr êxito na indexação que deve alimentar e impactar a base.

O PROPÓSITO DA PESQUISA

A comparação do resultado da aplicação da mineração de textos com o resultado da aplicação da recuperação de itens (recuperação informacional de textos manualmente indexados), em função do índice de precisão de resposta no processo de busca e recuperação da informação, foi o propósito do trabalho.

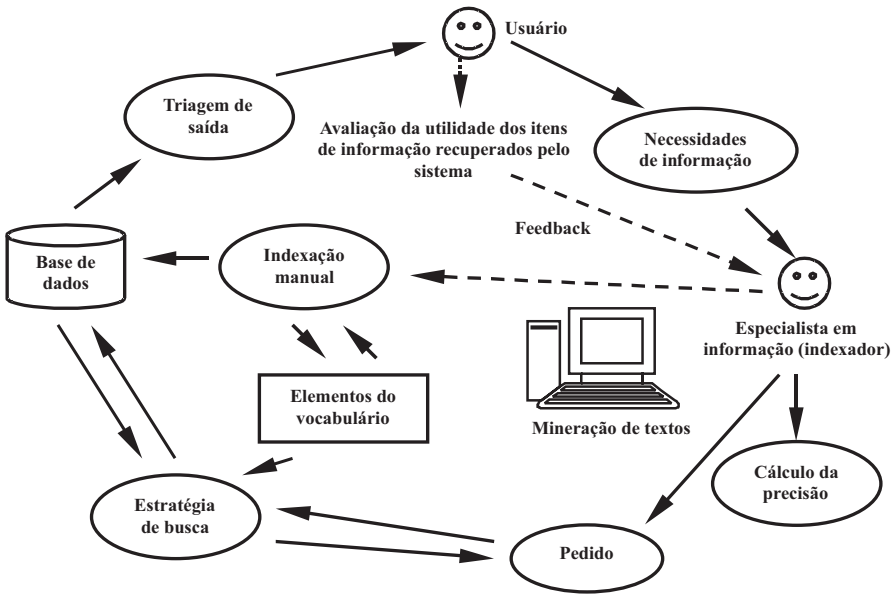
Desta forma, pretendeu-se com o estudo:

I. verificar se, na indexação manual, o índice de precisão resultante do processo de busca e recuperação da informação foi superado com o uso de ferramenta de mineração de textos;

II. verificar se a ferramenta de mineração de textos poderia ser convertida em ferramenta de indexação a partir da extração automática de termos, com o auxílio do julgamento dos indexadores na seleção de termos a serem utilizados na representação do conteúdo dos documentos em futuras pesquisas;

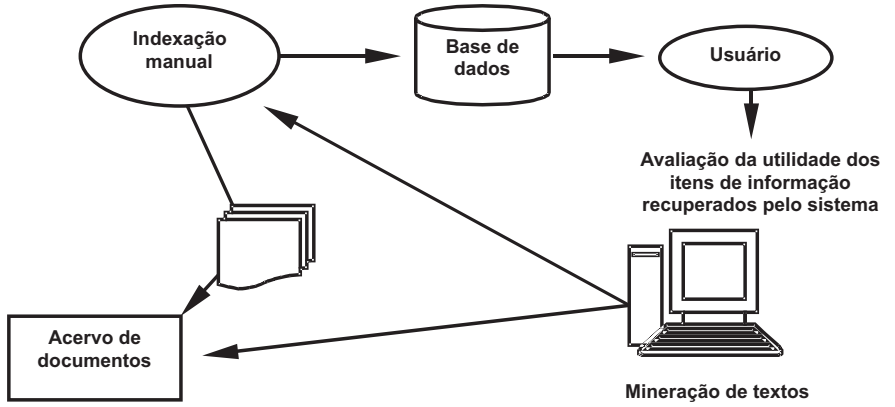
III. se a mineração de textos poderia apoiar a construção e/ou manutenção do tesauro, que é gerado e usado na indexação manual.

FIGURA 1
Fatores de influência nos resultados de busca em uma base de dados



Fonte: adaptado de Lancaster (1998).

FIGURA 2
Posição da mineração de textos no contexto da indexação manual



Como forma de viabilizar o propósito do trabalho, três questões foram formuladas:

a) A mineração de textos aplicada ao processo de busca e recuperação da informação traz ganhos de precisão se comparada à indexação manual?

b) A mineração de textos pode ser empregada como ferramenta complementar no processo de indexação visando ao aumento do índice de precisão na recuperação da informação?

c) É possível a construção de uma sistemática de uso de mineração de textos para complementar e aperfeiçoar o processo de indexação visando ao aumento do índice de precisão na recuperação da informação?

O trabalho pretendeu ainda apresentar uma sistematização dos termos gerados a partir da mineração de textos para complementar o processo de indexação manual, com vistas ao aumento do índice de precisão de resposta no processo de busca e recuperação da informação, conforme ilustra a figura 2.

A IMPORTÂNCIA DA PESQUISA

A partir do objetivo da pesquisa, ou seja, comparar a utilidade da ferramenta de mineração de textos com a lista de palavras-chave utilizadas na indexação manual por bibliotecários da Caixa, verificando a variação no índice de precisão no processo de busca e recuperação da informação na base de dados do Infohab, as tarefas fundamentais foram as seguintes:

- avaliar se a recuperação da informação pela aplicação da ferramenta de mineração de textos traz ganho no índice de precisão, se comparada com a lista de palavras-chave utilizadas na indexação manual na base de dados;
- verificar a viabilidade de utilizar os termos resultantes do emprego de ferramenta de mineração de textos, para enriquecer a lista de palavras-chave usada no Infohab, objetivando aprimorar o trabalho de indexação manual em relação ao índice de precisão de resposta no processo de busca e recuperação da informação;
- propor uma sistemática de uso dos termos gerados a partir da mineração de textos que apóie o processo de indexação manual, visando ao aumento do índice de precisão de resposta no processo de busca e recuperação da informação.

O cumprimento de cada um destes objetivos foi calcado em um processo de investigação por comparação entre um processo tradicional e consagrado na representação do conteúdo dos documentos, como é o caso do processo de indexação, com a mineração de textos. Para viabilizar esta comparação, optou-se pelo uso do índice de precisão, uma medida objetiva e consagrada desde a sua proposta por Cleverdon (1962), descrito na seção de metodologia empregada, e que também será capaz de mostrar as diferenças de desempenho entre a indexação e a mineração de textos na recuperação da informação.

Sobre a factibilidade de incluir na problemática da recuperação da informação a mineração de textos, pode-se citar, dentre os autores estudados, a percepção de uma lacuna: a utilização da mineração de textos voltada para a resolução das falhas no processo de busca e recuperação da informação. Nesta associação, há de fato uma possibilidade clara de uso da ferramenta na melhoria de desempenho destes sistemas; todavia, a cautela aponta na direção da utilização da mineração de textos para auxiliar o processo de indexação, já que a grande potencialidade da ferramenta está na sua capacidade de sumarizar grandes conjuntos de documentos em agrupamentos, apresentando-os sob a forma de listas de palavras que mais ocorrem por documento, ou por resultado de pesquisas (conjuntos de documentos), e em alguns casos com gráficos indicativos das relações semânticas entre os termos.

Assim, a possibilidade de extrair, de uma montanha de textos, informação útil às demandas torna-se um efetivo instrumento de gestão em bases textuais. Contudo, a informação útil não é dada de forma automática com a mineração de textos, mas por meio da interpretação que for dada aos resultados obtidos.

Deste modo, o aumento do índice de precisão da resposta obtida do processo de busca e recuperação da informação poderá ser alcançado a partir de dois fatores críticos de sucesso (FCS)*:

I) aplicação da mineração de textos;

II) integração dos resultados obtidos com a mineração de textos à indexação.

O quadro 1 associa os fatores críticos de sucesso aos seus objetivos-chave, a fim de torná-los mais claros.

QUADRO 1
FCS na gestão da precisão no processo de busca e recuperação da informação

FATORES CRÍTICOS DE SUCESSO (FCS)	OBJETIVOS – CHAVE
Aplicação da mineração de textos	Auxiliar o processo de indexação
Integração dos resultados obtidos com a mineração de textos à indexação	Montagem de uma sistemática de uso da mineração de textos no processo de indexação, com vistas ao aumento do índice de precisão no processo de busca e recuperação da informação

A proposição do estudo comparado entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação, pretendeu contribuir efetivamente no âmbito da ciência da informação para o estabelecimento de estratégias de uso da mineração de textos na melhoria contínua da resposta nestes sistemas,

* **Fatores Críticos de Sucesso (FCS) –** características, condições ou variáveis críticas para o sucesso (atingimento dos objetivos) em um dado processo ou até mesmo em uma organização (ROCKART, 1979).

além de verificar, com clareza, os ganhos que a mineração de textos pode trazer em relação ao processo de indexação na recuperação da informação.

METODOLOGIA

O desenvolvimento da metodologia está subdividido em etapas que englobam procedimentos tais como o desenvolvimento de um protótipo com aplicação da ferramenta de mineração de textos.

Etapa 1 – Definição da amostra: consistiu na escolha do item "tese e dissertação" retirado do universo de 1.520 documentos da Caixa Econômica Federal inseridos na base de dados do Infohab, conforme os fatores elencados na introdução deste artigo.

Etapa 2 – Extração da amostra: ocorreu em dois momentos: 1º) os "metadados" de todas as teses e dissertações, bem como os mecanismos de recuperação por palavras-chave, foram extraídos do universo do Infohab por meio de senha do administrador da base, o que possibilitou a recuperação por palavras-chave em uma pesquisa apenas das teses e dissertações que correspondem à amostra selecionada; 2º) extração dos textos completos de cada metadado que compõe a amostra em formato Adobe Portable Document Format (PDF) do *software* Acrobat Reader 4.0 da empresa Adobe Systems Incorporated para um CD-ROM.

Etapa 3 – Construção do protótipo: iniciou-se com a separação dos metadados de cada item bibliográfico da amostra e depois com a submissão e captura dos 56 arquivos PDF, pelo *software* de mineração de textos BR/*Search*. Com a base pronta, passou-se a refinar os instrumentos de mineração de texto conforme a necessidade de utilização na pesquisa. Esta etapa correspondeu às partes 2 (extração) e 3 (mineração) propostas por Trybula (1999), conforme a figura 3, a seguir. Na construção do protótipo um fator foi decisivo, ou seja, a adoção de uma série de procedimentos para adaptar o banco de dados gerado para que o protótipo fosse o espelho da base de dados Infohab, condição básica para a realização dos testes de precisão.

Etapa 4 – O *software* de mineração de textos utilizado: o BR/Search foi desenvolvido em ferramenta OLAP* e é um produto voltado para o gerenciamento de informações não estruturadas. Trata-se de uma tecnologia de gerenciamento eletrônico de documentos (GED) na qual o tipo de dado "texto corrido" é tratado com o mesmo nível de importância que qualquer outro campo-chave, tradicionalmente utilizado nos bancos de dados que suportam o modelo relacional, embora não siga tal modelo. Isto quer dizer, fundamentalmente, que palavras, expressões etc. podem ser encontradas, mesmo que estejam explicitadas apenas em pequeno trecho de um texto. Esta potencialidade é característica dos *softwares* de mineração de textos.

Etapa 5 – A coleta de dados para os testes de precisão: foram elaborados dois formulários de pesquisa aplicados junto aos usuários do Infohab, a fim de realizar testes de precisão com a amostra selecionada e o protótipo. O formulário foi estruturado para coletar dados relativos à pesquisa descrita em texto livre, palavras-chave tal como submetidas às duas bases de dados, inclusive com os operadores e campos pesquisados, o resultado das pesquisas e a posterior validação do usuário com a anotação de "útil" ou "inútil" para cada item bibliográfico recuperado.

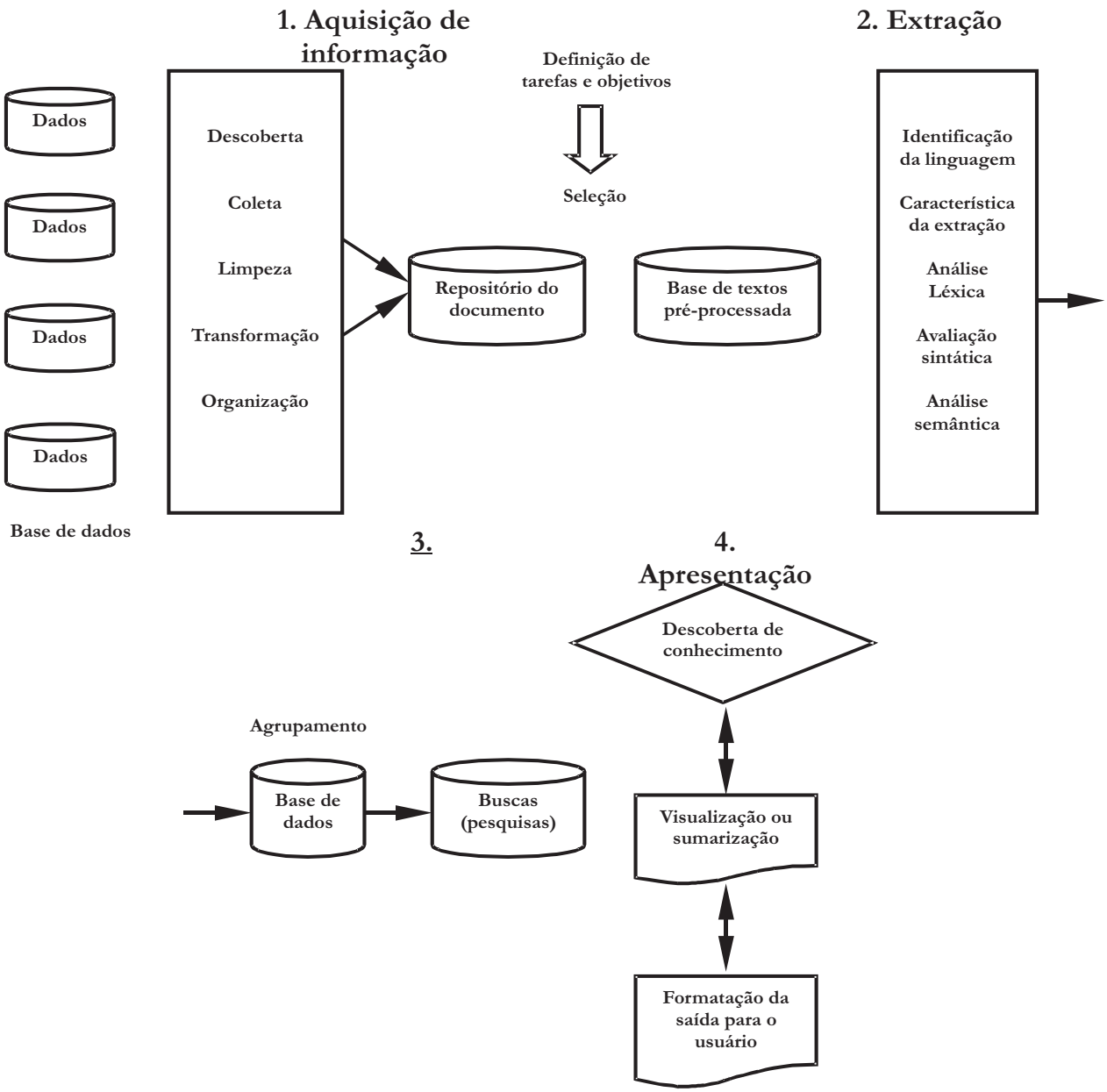
Etapa 6 – Seleção dos usuários para a coleta de dados: a seleção dos usuários seguiu o critério de gerência e consultoria técnica das áreas usuárias estratégicas do Infohab na Caixa: Vice-Presidência de Desenvolvimento Urbano e Governo (Viurb/Diretoria de Parcerias e Apoio ao Desenvolvimento Urbano (Dipup) e a Gerência Nacional de Normas e Padrões de Engenharia e Trabalho Social (Gepad). Entre os 22 usuários consultados, três eram bibliotecários que, além de indexar a base, eram seus usuários.

A questão estratégica que incidiu na escolha dos usuários está no fato de que tais gerentes utilizam as informações da base em decisões relativas aos programas de habitação prioritários da Caixa e, no caso dos bibliotecários, a manutenção da base de dados em condições adequadas para a perfeita disseminação da informação.

Etapa 7 – Testes de precisão e validação dos usuários: os testes de precisão para a geração dos resultados de pesquisa foram realizados com 44 formulários, sendo dois para cada entrevistado. Neste teste, as mesmas palavras-chave com os respectivos operadores no campo palavra-chave foram aplicadas de forma idêntica ao protótipo e à lista de palavras-chave utilizadas na indexação manual. Em seguida, o resultado das pesquisas nas bases foi encaminhado para a apreciação e validação dos usuários

* **Ferramenta OLAP** – programa que possibilita ao usuário a obtenção de informações armazenadas nas bases de dados dos *data warehouses*. Entre as suas principais funcionalidades estão o *drilling*, ou seja, detalhamento, e o *slice* & *dice* ou seleção e visualização de porções de base de dados (TARAPANOFF, 2001).

FIGURA 3
Processo de mineração de textos



Fonte: Trybula (1999).

com a seguinte anotação para cada item bibliográfico recuperado: documento útil (U) ou documento inútil (I). Os resultados foram apresentados em duas listas distintas, uma com o resultado da pesquisa na base do Infohab e a outra com o resultado da pesquisa no protótipo. Este procedimento visou a garantir a correta aplicação da fórmula do cálculo do índice de precisão para cada uma das bases.

Etapa 8 - Tratamento dos dados: o tratamento dos dados deu-se em dois momentos:

a) *Cálculo do índice de precisão* – com os resultados validados por 22 usuários dos itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual e do uso da ferramenta de mineração de textos, foi realizado o cálculo do índice de precisão. Para tanto, a fórmula usada foi a seguinte:

$$\textit{Precisão} = \frac{\textit{Número de } \textbf{\textit{documentos úteis}} \textit{ recuperados pelo sistema}}{\textit{Número } \textbf{\textit{total de documentos encontrados}} \textit{ pelo sistema}} X\ 100$$

b) *Tabulação dos dados* – com os resultados dos índices de precisão apurados, foi possível calcular e inferir se, no estudo de caso do Infohab, a lista de palavras-chave utilizadas na indexação manual trouxe um índice de precisão maior na busca e recuperação da informação do que com o uso da ferramenta de mineração de textos, além das inferências que complementam a tabulação dos dados na comprovação ou refutação dos pressupostos da pesquisa.

RESULTADOS

A análise dos resultados obtidos procurou responder às três questões formuladas na seção propósito da pesquisa.

O primeiro questionamento indagou se a mineração de textos aplicada ao processo de busca e recuperação da informação traz ganhos de precisão, se comparada à indexação manual. Os dados obtidos revelaram que não há ganho significativo no índice de precisão no processo de busca e recuperação da informação.

A segunda questão indagou se a mineração de textos pode ser empregada como ferramenta complementar no processo de indexação, visando ao aumento do índice de precisão na recuperação da informação. Os dados analisados comprovam que o uso da ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual. Já a submissão de termos específicos da base de dados do Infohab ao protótipo com aplicação da mineração de textos sempre trará itens bibliográficos. Tais conclusões abrem caminho para uma resposta positiva à segunda indagação e, por extensão, à terceira, que questiona se é possível a construção de uma sistemática de uso de mineração de textos para complementar e aperfeiçoar o processo de indexação, visando ao aumento do índice de precisão na recuperação da informação.

Considerando que há factibilidade para responder positivamente às duas últimas questões, podemos afirmar que o uso da mineração de textos pode, em associação com o processo de indexação manual, trazer ganhos ao índice de precisão no processo de busca e recuperação da informação, desde que o julgamento do indexador seja considerado indispensável na montagem de uma sistemática de uso dos termos gerados a partir da mineração de textos. A habilidade do indexador de contextualizar, relacionar palavras, usar a abstração, bem como decidir quais termos serão usados para identificar o conteúdo dos documentos é fator preponderante para a indexação de documentos e a sua posterior recuperação em uma base de dados.

Enfim, considerando os argumentos apresentados, conclui-se que a mineração de textos (utilizando-se do *software* BR/Search na criação do protótipo) trará, como resposta a uma consulta, uma quantidade de itens bibliográficos sempre maior do que na resposta obtida com a lista de palavras-chave utilizadas na indexação manual.

Esta possibilidade abre caminho para o emprego da mineração de textos como instrumento de enriquecimento da lista de palavras-chave e/ou construção de um vocabulário controlado, utilizando, para tanto, a lista de palavras mais freqüentes em cada documento recuperada em pesquisas realizadas no protótipo, além da lista de palavras mais freqüentes do resultado total da pesquisa, também realizada no protótipo. Faz-se necessário ressaltar, neste contexto, que a grande potencialidade da ferramenta é a captura de qualquer termo em qualquer parte do texto completo armazenado no protótipo. Isto se configura como um instrumento útil no aprimoramento contínuo do processo de indexação, já que a mineração de textos pode extrair automaticamente termos relacionados à pesquisa. A partir daí, o julgamento dos indexadores que deverão selecionar os termos a serem usados na representação do conteúdo dos documentos poderá enriquecer e/ou apoiar a construção de um tesauro, por exemplo.

Os resultados obtidos com a pesquisa são determinantes na formulação do argumento de que a mineração de textos pode ser utilizada também como instrumento de enriquecimento da lista de palavras-chave e/ou construção de um vocabulário controlado, utilizando a lista de palavras mais freqüentes em cada documento.

Em resumo, duas são as possibilidades concretas da mineração de textos por meio do *software* BR/Search:

a) a geração e utilização da lista de palavras mais freqüentes em cada item bibliográfico (texto completo) recuperada na pesquisa realizada no protótipo;

b) a geração e utilização da lista de palavras mais freqüentes no resultado total (textos completos agrupados) da pesquisa realizada no protótipo com uso de mineração de textos.

Apresentamos, a seguir, dois exemplos de pesquisa e geração de lista de palavras mais freqüentes em cada documento recuperadas na pesquisa realizada no protótipo e listas de palavras mais freqüentes do resultado total da pesquisa realizada no protótipo:

Exemplo 1:

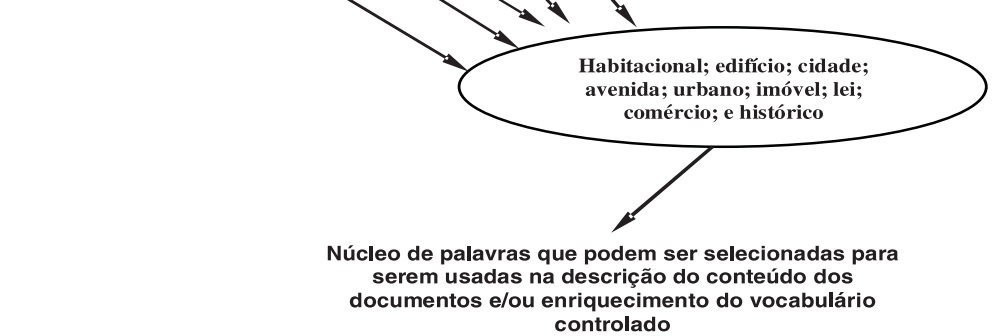
• *Termo pesquisado:* revitalização COM* urbana com dois itens bibliográficos recuperados.

• *Resultado:* a tabela 2 apresenta a lista das 20 (vinte) palavras mais freqüentes da pesquisa.

TABELA 2

Lista de palavras mais freqüentes do resultado total da pesquisa realizada no protótipo com o termo de busca: revitalização COM urbana

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
Recife	74	308	24,02	2	10	20
Área	69	1082	6,37	2	46	4,34
Uso	66	624	10,57	2	49	4,08
centro	61	259	23,55	2	37	5,40
áreas	47	813	5,78	2	45	4,44
habitacional	**44**	**707**	**6,22**	**2**	**27**	**7,4**
edifícios	**42**	**121**	**34,71**	**1**	**12**	**8,33**
programa	41	777	5,27	2	44	4,54
cidade	**38**	**817**	**4,65**	**2**	**44**	**4,54**
avenida	**37**	**50**	**74**	**1**	**9**	**11,11**
cidades	**37**	**336**	**11,01**	**2**	**39**	**5,12**
urbana	**36**	**640**	**5,62**	**2**	**41**	**4,87**
edifício	**33**	**157**	**21,01**	**2**	**8**	**12,5**
imóveis	**32**	**522**	**6,13**	**2**	**24**	**8,33**
onde	32	542	5,87	2	51	3,92
Pernambuco	32	54	59,25	1	3	33,33
Lei	**31**	**491**	**6,31**	**2**	**33**	**6,06**
Tem	31	865	3,58	2	53	3,77
Caixa	30	669	4,48	2	36	5,55
Comércio	**30**	**86**	**34,88**	**2**	**19**	**10,52**
Foi	30	995	3,01	2	53	3,77
Históricos	**30**	**78**	**38,46**	**2**	**12**	**16,66**
Urbano	**30**	**664**	**4,51**	**2**	**44**	**4,54**
Guararapes	29	34	85,29	1	2	50
Estudo	26	445	5,84	2	45	4,44



Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*

* **COM** – Operador booleano que combina dois ou mais termos em uma expressão de busca. Na recuperação da informação no protótipo, o operador booleano "com" foi empregado com a mesma função do operador "E", que executa a operação de intersecção entre conjuntos.

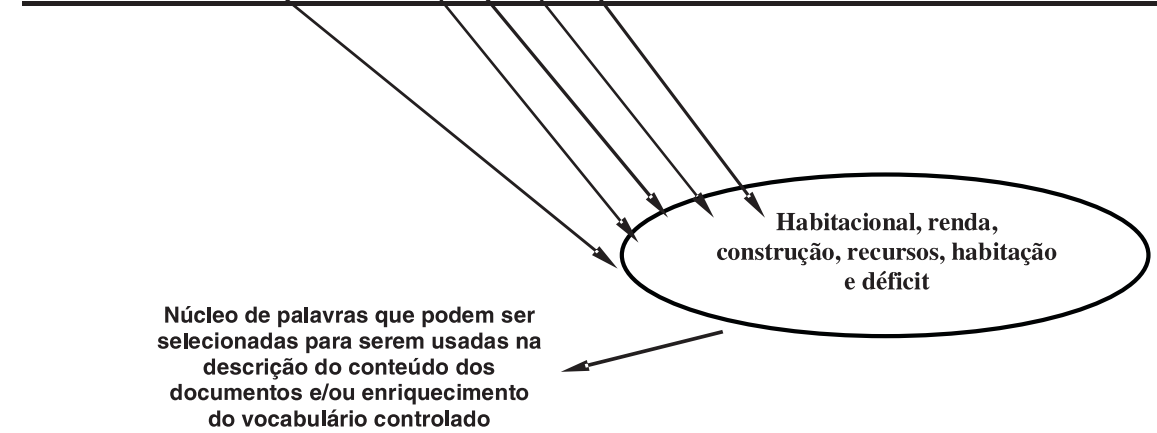
Exemplo 2:

• *Termo pesquisado:* arrendamento COM urbano três itens bibliográficos recuperados.

• *Resultado:* a tabela 3 apresenta a lista das 20 (vinte) palavras mais freqüentes do resultado total da pesquisa.

TABELA 3
Lista de palavras mais freqüentes do resultado total da pesquisa realizada no protótipo com o termo de busca: arrendamento COM urbano

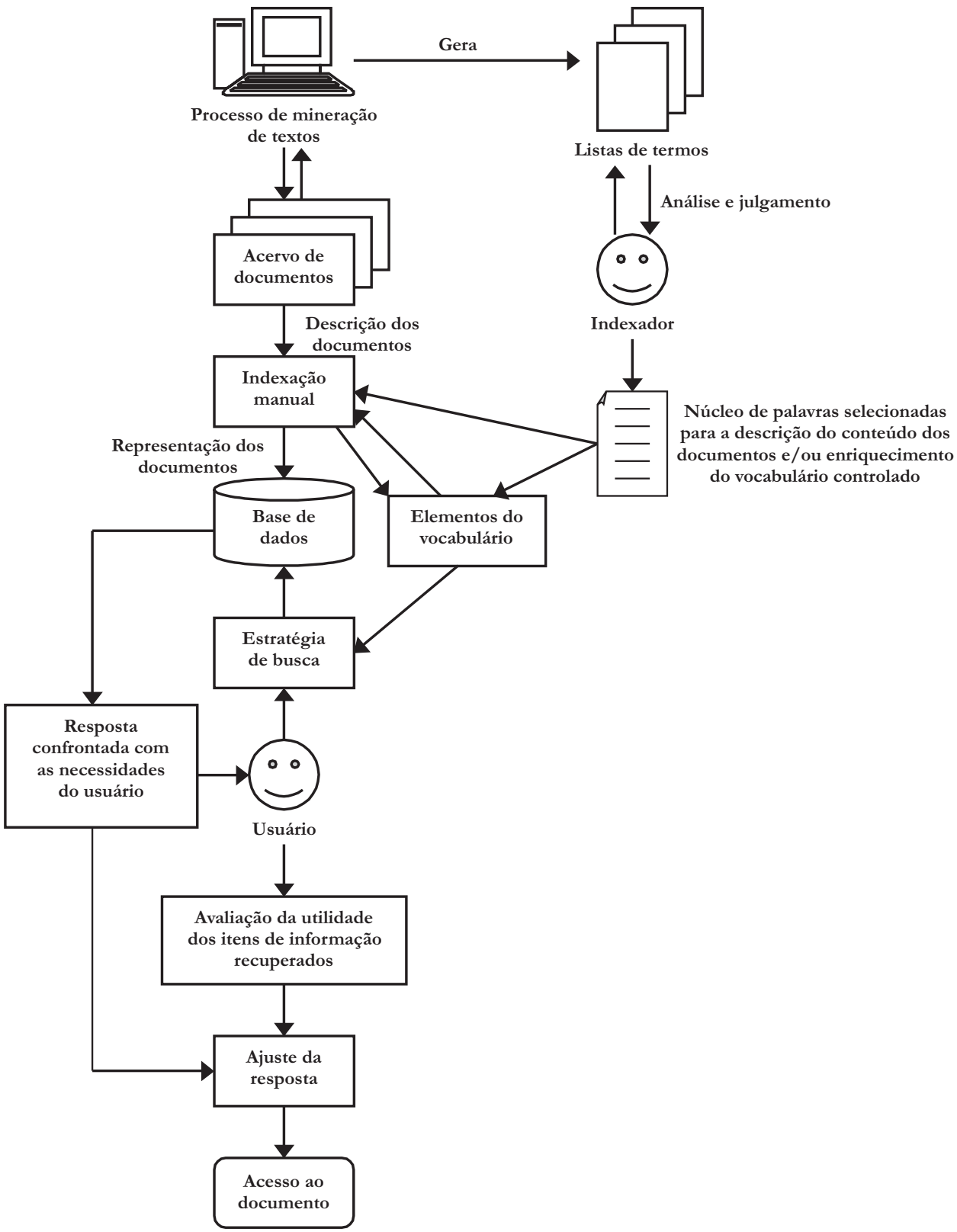
Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
to	307	948	32,38	2	31	6,45
link	305	803	37,98	2	15	13,33
page	305	803	37,98	2	15	13,33
habitacional	**274**	**707**	**38,75**	**3**	**27**	**11,11**
renda	**202**	**974**	**20,73**	**3**	**39**	**7,69**
construção	**181**	**547**	**33,08**	**3**	**46**	**6,52**
programa	179	777	23,03	3	44	6,81
recursos	**177**	**1013**	**17,47**	**3**	**45**	**6,66**
habitação	**157**	**702**	**22,36**	**3**	**33**	**9,09**
"1"	157	6953	2,25	3	55	5,45
par	156	169	92,30	3	13	23,07
população	**148**	**1054**	**14,04**	**3**	**50**	**6**
pela	144	1339	10,75	3	52	5,76
através	143	880	16,25	3	48	6,25
déficit	**137**	**274**	**50**	**3**	**18**	**16,66**
R	136	1501	9,06	3	41	7,31
"3"	134	4612	2,90	3	55	5,45
área	130	1082	12,01	3	46	6,52
"2"	124	5693	2,17	3	55	5,45
alianças	123	126	97,61	1	3	33,33
foi	123	995	12,36	3	53	5,66
"5"	120	4048	2,96	3	53	5,66
habitacionais	**119**	**295**	**40,33**	**3**	**28**	**10,71**
pesquisa	118	535	22,05	3	43	6,97
tem	117	865	13,52	3	53	5,66



Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*

Com base na análise de todos os resultados obtidos, foi proposto um fluxograma (figura 4, a seguir), a seguir, que representa a sistemática de uso da mineração de textos, a partir das listas de palavras mais freqüentes para apoiar o processo de indexação manual, com vistas ao aumento do índice de precisão de resposta no processo de recuperação da informação:

FIGURA 4
Fluxograma da sistemática de uso da mineração de textos na indexação manual



CONCLUSÕES

Durante a elaboração da pesquisa, algumas conclusões preliminares apoiaram o seu desenvolvimento, contribuindo, assim, para o alcance dos objetivos propostos:

• o aperfeiçoamento contínuo do processo de indexação deve passar pelo conhecimento proativo das necessidades dos usuários, o que deve proporcionar subsídios para a determinação dos requisitos a serem utilizados no âmbito do gerenciamento estratégico da informação;

• em termos práticos, a mineração de textos, no âmbito do processo de busca e recuperação da informação, poderá ser pensada como uma ferramenta a ser empregada na busca da melhoria das respostas nestes sistemas. A avaliação desta possibilidade poderá ser materializada a partir da utilização do índice de precisão para aferir o desempenho da ferramenta nesta tarefa, bem como compará-la aos instrumentos tradicionais, utilizados hoje;

• a utilização dos índices de precisão deverá sempre propiciar parâmetros para a melhoria contínua da resposta obtida dos sistemas de recuperação da informação. O conhecimento das necessidades de informação dos usuários é ponto de partida para a concretização desta meta. Tais necessidades, entretanto, acabam por gerar também graus de imprecisão, tal como observa Foskett (1996), ou seja, incapacidade de um sistema de informação recuperar documentos úteis ante a solicitação do usuário, sobretudo se envolver a negociação da questão;

• os autores estudados no âmbito da pesquisa concordam com o papel preponderante que cabe ao julgamento dos usuários para o cálculo do índice de precisão. Com este dado, torna-se factível pensar na precisão como um elemento importante de análise e decisão em busca da melhor resposta nos sistemas de busca e recuperação da informação. Não é possível calcular o número de documentos úteis encontrados pelo sistema sobre o total de documentos relevantes contidos no sistema, sem o julgamento do usuário que demanda tais documentos, ou seja, a precisão se consubstancia por meio de um julgamento externo ao sistema de busca e recuperação da informação, o que vai determinar, também, a sua capacidade de atendimento ou o seu desempenho;

• a precisão não se dá *per se*, mas no contexto em que operam a revocação, a exaustividade a especificidade e, sobretudo, tendo como ponto de equilíbrio o usuário que vai definir, em nome da sua necessidade de informação, o que é útil ou inútil dentre toda a informação recuperada.

Com base na proposta do estudo comparado entre a mineração de textos e a indexação manual, a investigação concentrou-se na avaliação da resposta obtida no processo de busca e recuperação da informação, por meio de uma medida objetiva, o índice de precisão. Desde a década de 70, a questão da precisão foi amplamente discutida em associação com a análise de desempenho de um sistema de recuperação da informação. Para Lancaster e Fayen (1973), ao considerar os fatores que interferem no desempenho destes sistemas, será necessário conhecer anteriormente os pré-requisitos do usuário em relação aos resultados de busca e recuperação da informação.

A escolha do índice de precisão permitiu avaliar, em termos percentuais, o desempenho de um protótipo com aplicação de mineração de textos, confeccionado para ser o espelho da amostra selecionada da base do Infohab, na qual os documentos são indexados manualmente. O resultado da avaliação dos desempenhos do protótipo e da base do Infohab atingiu um dos objetivos do estudo, qual seja, avaliar se a recuperação da informação por meio de ferramenta de mineração de textos traz ganho no índice de precisão, se comparada à lista de palavras-chave utilizadas na indexação manual na base de dados do Infohab.

A avaliação do desempenho do protótipo permitiu também a verificação da viabilidade de utilizar os termos resultantes do emprego da ferramenta de mineração de textos no enriquecimento da lista de palavras-chave utilizadas na indexação manual. Apesar de a ferramenta constituir importante instrumento na identificação de palavras-chave, o indexador continua um dos principais artífices no processo de indexação, dada a sua competência na escolha de termos a serem usados para identificar o conteúdo dos documentos.

Este debate, apesar de já antigo, sempre retorna como nova proposição a cada desenvolvimento de novas tecnologias, trazendo expectativas para a resolução dos problemas relacionados ao processo de busca e recuperação da informação. Para Lancaster (1993), entretanto, apesar de terem ocorrido muitos avanços no processamento da linguagem natural por computador, é mister admitir que a "compreensão" de textos pelo computador ainda se acha muito limitada. Ou seja, é possível construir instrumentos auxiliares morfológicos, sintáticos e semânticos que ajudem o computador a

interpretar textos, mas isto ainda está muito longe do que acontece quando um ser humano lê um texto e compreende o que o autor quer dizer.

A mineração de textos apresenta-se, neste contexto, como uma ferramenta de ponta, cujo objetivo, segundo Feldman e Hirsh, citados porWives (1999), é constituir-se em um meio efetivo de recuperação, filtragem, manipulação e resumo do conhecimento contido em grandes volumes de informações textuais, para apresentá-lo em forma de gráficos, listas ou tabelas.

Esta possibilidade da ferramenta de mineração de textos abriu caminho para a proposição de uma sistemática (disposta na figura 4) de utilização dos termos gerados a partir da mineração de textos que apóia o processo de indexação manual visando ao aumento do índice de precisão de resposta no processo de busca e recuperação da informação, alcançando, assim, o terceiro e último objetivo proposto.

De acordo com McGarry (1999), os sistemas de computador obedecem a algoritmos, mas o conteúdo semântico dos textos está além de sua compreensão. As mentes humanas têm conteúdos semânticos, significados e ressonâncias, o que enseja a discussão fomentada pelo autor a respeito do futuro da ciência da informação: poderá esta ciência ser conduzida dentro de um sistema fechado de raciocínio algorítmico? Poderá crescer e desenvolver-se no vazio cultural?

Apesar de toda a discussão em torno da utilização de um ou outro método, ou seja, a indexação manual ou a mineração de textos, os resultados, na verdade, só poderão ser validados por intermédio da avaliação dos usuários. Isto significa dizer que os sistemas de recuperação da informação, além de buscar atender às demandas informacionais dos usuários, dependem destes para que a qualidade dos seus serviços seja reconhecida.

Artigo submetido em 16/03/2006 e aceito em 23/02/2007.

REFERÊNCIAS

CLEVERDON, C. W. *Report on testing and analysis of investigation into comparative efficiency of indexing systems*. Cranfield: Aslib, 1962.

FELDMAN, R.; HIRSH, H. Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, v. 9, n. 1, p. 83-97, July/Aug. 1997.

FOSKETT, A. C. *The subject approach to information*. 5th ed. London : Unipub, 1996.

HALAL, E.; KULL, M. *Measuring organizational intelligence*. Disponível em: <http://www.auburn.edu/administration/horizon/measuring.html>. Acesso em: 12 fev. 2005.

LANCASTER, F. W. *Indexação e resumos*: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos, 1993.

_____. *Indexing and abstracting in theory and practice*. 2nd ed. London: Library Association, 1998.

_____; FAYEN, E. G. *Information retrieval on-line*. Los Angeles: Melville, 1973.

MCGARRY, K. *O contexto dinâmico da informação*: uma análise introdutória. Tradução de Helena Vilar de Lemos. [Brasília]: Briquet de Lemos, 1999.

POLANCO, X.; FRANÇOIS, C. Data clustering and cluster mapping or visualization in text processing and mining. In: INTERNATIONAL ISKO CONFERENCE, 6., 2000, Toronto. *Proceedings*... Toronto: Ergon Verlag: Würzburg, 2000. p. 359-365.

ROCKART, J. F. Chief executives define their own data needs. *Harvard Business Review*, v. 57, n. 2, p. 81-93, Mar./Apr. 1979.

TARAPANOFF, K. (Org.). *Inteligência organizacional e competitiva*. Brasília: Editora Universidade de Brasília, 2001.

TRYBULA, W. J. Text mining. *Annual Review of Information Science and Technology*, v. 34, p. 385-419, 1999.

TRYBULA, W. J. *Text mining and knowledge discernment*: an exploratory investigation. 1999. Tese (Doutorado em Ciência da Computação)- University of Texas, Austin, 1999.

WIVES, L. K. *Estudo sobre agrupamento de documentos textuais em processamento de informação não estruturadas usando técnicas de clustering*. 1999. Dissertação (Mestrado em Ciência da Computação)- Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999.