Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T

Hélia de Sousa Chaves Ramos
Mestre em ciência da informação pela Universidade de Brasília (PPGCInf/UnB)
E-mail: helia@ibict.br

Marisa Bräscher
Professora doutora do Departamento de Ciência da Informação e Documentação da Universidade de Brasília (PPGCInf/UnB)
E-mail: marisab@unb.br

Resumo

Relata resultados de pesquisa aplicando a descoberta de conhecimento em texto (DCT) em conteúdos textuais, importantes fontes de informação para tomada de decisão. O objetivo central da pesquisa foi verificar a eficácia da DCT na descoberta de informações para apoio à construção de indicadores e definição de políticas públicas. O estudo de caso foi o Serviço Brasileiro de Respostas Técnicas (SBRT) e a técnica aplicada a de agrupamento de documentos a partir dos termos minerados na base de dados. Comprovou-se a aplicabilidade da DCT na extração de informações ocultas em documentos textuais para subsidiar a tomada de decisão e a construção de indicadores, informações essas que não poderiam ser visualizadas utilizando-se recursos tradicionais de recuperação da informação. Observou-se a preocupação com o meio ambiente nas demandas feitas pelos usuários do SBRT e a aplicabilidade da DCT para orientação de políticas internas à rede SBRT.

Palavras-chave

Descoberta de conhecimento em texto (DCT). Mineração de textos. Indicadores de C&T. Serviços de informação tecnológica. Micro e pequenas empresas (MPEs). Empreendedores.

Applying knowledge discovery in texts (KDT) to support the construction of S&T infometric indicators

Abstract

This article describes the results of a research applying Knowledge Discovery in Texts (KDT) in textual contents, which are important sources of information for decision-making purposes. The main objective of the research is to verify the effectiveness of KDT for discovering information that may support the construction of ST&I indicators and for the definition of public policies. The case study of the research was the textual content of the Brazilian Service for Technical Answers (Serviço Brasileiro de Respostas Técnicas – SBRT) and the technique adopted was document clustering from terms mined in the database. The use of DCT for extracting hidden information – that could not be found by using the traditional information retrieval – from textual documents proved to be efficient. The presence of environmental concerns in the demand posted by SBRT's users and the applicability of DCT to orient internal policies for SBRT network were also evidenced by the research results.

Keywords

Knowledge Discovery in Texts (KDT). Text mining. S&T indicators. Information services. Micro-enterprises. Entrepreneurs.

INTRODUÇÃO

A velocidade e a amplitude com que o conhecimento gerado passou a ser compartilhado provocaram o surgimento de uma dinâmica de reaproveitamento e produção de novos conhecimentos, bem como o aparecimento de novas necessidades de tratar a informação. Para suprir essas necessidades, ferramentas e técnicas para tratamento de grandes massas de dados foram criadas e aperfeiçoadas, tratamentos estatísticos aplicados no processamento e análise de dados e informações, em busca de retratar o que não seria possível com a limitada capacidade humana de leitura e registro. As informações estruturadas em bases de dados – organizadas, indexadas e dotadas de ferramentas cada vez mais sofisticadas e velozes para busca e recuperação da informação – têm sido objeto de estudo com a finalidade de extrair conhecimento para apoio à tomada de decisão.

Nos últimos anos, os focos da pesquisa têm se voltado para aqueles conteúdos armazenados em meio digital sem a preocupação com o rigor da estruturação – os documentos textuais – comumente chamados de "informação não estruturada". Esses conteúdos se revelaram portadores de informações valiosas, camufladas em grandes volumes textuais, que passaram a ser explorados em busca de padrões de conhecimento até então desconhecidos, para tomada de decisão e geração de novos conhecimentos.

Segundo Tan (1999), mais de 80% das informações de uma organização estão contidos em documentos textuais, que são a forma mais natural de armazenamento de informações. Não resta dúvida de que esse tipo de documento adquire importância fundamental para a descoberta do conhecimento gerado dentro das organizações.

Este artigo apresenta os resultados de uma pesquisa de mestrado que teve como objetivo testar a eficácia da descoberta de conhecimento em textos (DCT) na descoberta de informações para apoiar a construção de indicadores úteis à tomada de decisão estratégica, assim como a definição de políticas públicas para o setor produtivo de pequeno porte. O estudo de caso foi o conteúdo textual de um sistema de informação criado para prover soluções a questões de natureza tecnológica apresentadas por empreendedores e microempresários brasileiros – o Serviço Brasileiro de Respostas Técnicas (SBRT)[1]. Os textos analisados contêm soluções elaboradas por especialistas em atendimento a questões de natureza tecnológica levantadas por microempreendedores de todo o país. Constitui-se, assim, em rica fonte de conhecimento tecnológico que pode se revelar importante origem de indicadores sobre as necessidades dos pequenos negócios e nortear investimentos para solucioná-las.

Buscou-se, com o estudo, comprovar a aplicabilidade da DCT no apoio à construção de indicadores em ciência e tecnologia, motivação para a descoberta de associações entre os documentos e identificação de tendências que apóiem a tomada de decisão governamental com relação ao setor empresarial de pequeno porte.

A construção de indicadores

Em minucioso estudo sobre indicadores de CT&I, Sartori e Pacheco (2007) afirmam que há, entre os estudiosos do assunto, o reconhecimento de que os indicadores em CT&I são de fundamental importância para "nortear a formulação e a avaliação de políticas e, principalmente, para permitir à sociedade acompanhar e avaliar os esforços dirigidos a tais atividades e os resultados obtidos" e que o atual conjunto de indicadores é insuficiente para

[1] O SBRT constitui uma iniciativa governamental encampada por instituições de ensino e pesquisa atuantes na prestação de serviços de informação tecnológica. Trata-se de uma ação inovadora criada por iniciativa do Ministério da Ciência e Tecnologia (MCT), que reúne universidades, iniciativa privada e governo. São membros do SBRT: CDT/UnB, Disque-Tecnologia da USP (Cecae/USP), Cetec/MG, Redetec/RJ, Tecpar/PR, IEL/BA, Senai/RS; e parceiros: Ibict e Sebrae Nacional. Destina-se a micro e pequenas empresas e empreendedores e oferece um serviço de informação gratuito na *Web* (http://sbrt.ibict.br).

atender a essa questão. Os autores salientam que a pesquisa científica e tecnológica adquire cada vez mais importância e impacto perante a sociedade ao mesmo tempo em que se verifica a existência de grandes lacunas no conjunto de indicadores de CT&I brasileiros.

Na visão do Ministério da Ciência e Tecnologia (MCT) – órgão responsável pela formulação e implementação da Política Nacional de Ciência e Tecnologia –

> "o conjunto de indicadores de C&T hoje disponível para o Brasil será continuamente enriquecido, na medida em que as dificuldades metodológicas e de acesso aos dados forem sendo superadas e novos indicadores produzidos."

Inicialmente, os indicadores limitavam-se ao dimensionamento dos recursos financeiros e humanos investidos em ciência e tecnologia – os chamados "indicadores de insumo". Em seguida, foram criados os "indicadores de resultados", contendo o registro da produção científica, a produção de patentes e a transferência de tecnologia entre países. Há, mais recentemente, a preocupação em se mensurar os indicadores de impacto, aqueles que procuram avaliar

> "como determinado resultado científico ou tecnológico afeta as várias dimensões das condições de existência dos indivíduos, seja no próprio campo científico e tecnológico, seja na dimensão econômica, seja na dimensão social (BRASIL, 2004).

A preocupação com os indicadores de impacto acompanha a tendência em se buscar o melhor conhecimento entre a relação das atividades de C&T e as atividades inovativas no Brasil, onde a soma de investimentos e a aplicação dos conhecimentos gerados possam promover reais impactos na economia e no bem-estar social.

É nessa natureza de indicadores, ou seja, no campo dos indicadores de impacto para a área de C&T que se insere esta pesquisa, onde os resultados dos investimentos públicos em um sistema de informação voltado a empreendedores e microempresários possam reverter em informações úteis à tomada de decisão nesse campo. Esses indicadores se inserem no campo da Infometria, considerada por Le Coadic (2005) como um novo eixo de pesquisa e desenvolvimento na Ciência da Informação, onde ocorre a aplicação da matemática e da estatística ao estudo dos fenômenos informacionais. Segundo o autor, "uma boa gestão de serviços públicos necessita cada vez mais da utilização de uma larga gama de ferramentas de gestão adaptadas aos contextos culturais, educativos, científicos e também às dimensões e características do serviço. São ferramentas de análise de necessidades de informação da comunidade atendida, ferramentas de acompanhamento e de avaliação e ferramentas de medição de performance. Estas ferramentas possibilitam à organização dispor de um conjunto de indicadores de desempenho."

Dentro da ótica de adaptação dos indicadores de C&T propostos pela Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) com vistas a melhor atender às necessidades dos países em desenvolvimento, Kondo (1998) sugeriu expandir o foco da construção de indicadores de C&T, tradicionalmente voltados para a eficiência econômica, para abranger indicadores "vinculados ao bem-estar social". Essa temática é abordada por Velho (2001), quando levanta questões relativas ao estabelecimento de um sistema de indicadores de C&T "útil e relevante para a tomada de decisão" e chama atenção para a importância do uso do conhecimento científico na produção, com a finalidade de propiciar melhoria da qualidade de vida da sociedade. Em sua opinião, os indicadores tradicionais passaram a ser questionados para se considerar a mudança técnica, o conceito de sistema nacional de inovação. De acordo com a autora, a inovação tem uma dimensão local e contingente.

O MCT chama atenção para as "reconhecidas e marcantes especificidades nacionais" relativas à base

técnico-científica, as quais evidenciam a necessidade de

> associar à produção de informação quantitativa o desenvolvimento de estudos mais aprofun-dados para validar ou redefinir os pressupostos sobre os quais se apóiam os indicadores. (BRASIL, 2004)

Nesse sentido, acredita-se que o SBRT esteja inserido na nova concepção de conteúdos adequados ao apoio à construção de indicadores, vez que se trata de um estímulo à aplicação do conhecimento tecnológico gerado pelas instituições de ensino e pesquisa para melhoria da competitividade da microempresa e a consequente contribuição tanto para a economia brasileira quanto para o bem-estar social.

A descoberta de conhecimento

São muitas as discussões em torno das definições das técnicas de extração automática de informações relevantes em grandes massas de dados, nas quais conceitos e termos se misturam, por vezes sendo utilizados como sinônimos: prospecção de conhecimento, descoberta de conhecimento em bases de dados, mineração de dados, descoberta de conhecimento em textos, mineração de textos. De forma abrangente, utiliza-se o termo "descoberta de conhecimento", passando-se a qualificá-lo a partir do conteúdo a ser analisado: se este foi previamente organizado e estruturado (descoberta de conhecimento em dados – DCD) ou se se encontra disperso em documentos textuais dos mais diversos formatos e tamanhos (descoberta de conhecimento em textos – DCT).

Grandes repositórios textuais contêm informações adormecidas, camufladas, até que o minerador as encontre e as transforme em informações preciosas para a organização. Descobrir conhecimento significa identificar, receber informações relevantes e poder computá-las e agregá-las ao seu conhecimento prévio, mudando o estado de conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido (WIVES, 2004).

Fayyad et al. (1996) consideram a DCT uma etapa da DCD, a qual se preocupa com o desenvolvimento de métodos e técnicas que buscam trazer sentido aos dados. Na visão dos autores, o processo básico da DCD é traduzir a informação do seu nível mais elementar, o dado, geralmente armazenado em grandes volumes, em formas mais compactas, mais resumidas e mais úteis. Os autores afirmam que a DCD tem sido cada vez mais empregada para a solução de problemas do mundo real, tanto no campo das ciências como nos negócios.

Weiss at al. (2005) compartilham da ideia da DCT como etapa da DCD e afirmam que para a conclusão do processo de mineração de textos, estes serão processados e transformados em representação numérica, distinção inicial entre elas.

Na visão de Trybula (1999), a descoberta do conhecimento é o "processo de transformação de dados em relações previamente desconhecidas e insuspeitas, que podem ser empregadas como previsores de futuras ações".

Minucioso estudo sobre a literatura acerca da descoberta de conhecimento em dados e os diversos processos que a compõem foi realizado por Schiessl (2007), que registra as particularidades apontadas por vários autores e chama atenção para a necessidade de adaptação da DCD "para que a linguagem natural seja passível de processamento automático visando à extração de conhecimento".

Embora a conceituação das técnicas de tratamento automático de grandes volumes de dados ainda se encontre de certa forma difusa, é possível identificar pontos convergentes fundamentais para a compreensão do seu funcionamento. Fica clara a forte evolução dessas técnicas com a possibilidade de tratamento de conteúdos em linguagem natural, que representa, de modo geral, a maioria dos conteúdos gerados por uma organização.

Esta pesquisa explora as potencialidades da DCT e faz uso da mineração de textos como uma das etapas de todo o processo, coerente com a concepção defendida por Hearst (1999), que a

define como "descoberta, por computador, de novas informações, previamente desconhecidas, pela extração automática de informações de diferentes recursos-chave da mineração de textos a interligação das informações extraídas para "formar novos fatos e novas hipóteses a serem posteriormente exploradas pelos meios de experimentação mais convencionais". Apoiado por Weiss at al. (2005), o autor trata dessas relações.

Hearst (1999), Aires (2005) e Lucas (2007) alertam para os perigos de se confundir mineração de textos com recuperação da informação e consideram como diferenciais da mineração o relacionamento entre documentos e a possibilidade de se extrair deles informação previamente desconhecida.

METODOLOGIA APLICADA NO ESTUDO

O universo de estudo da pesquisa foi o conteúdo textual da base de dados de Respostas Técnicas (RTs)[2]. O conteúdo estudado foi extraído do sistema de Informação SBRT no dia 8 de agosto de 2007, mediante autorização do Comitê Gestor da Rede SBRT. Os dados representavam, naquela data, a totalidade das RTs enviadas aos clientes e publicadas no *site*: 6.041 documentos.

As informações constantes do corpo do texto da RT são título da RT, resumo, data de publicação, palavras-chave, assunto, demanda (a pergunta feita pelo cliente) e instituição respondente (responsável pela elaboração da RT). Todos esses campos foram considerados na extração dos termos para análise.

Aplicação da DCT

A aplicação da DCT se dá por meio de técnicas diversas, sendo as mais conhecidas, segundo Wives (2004): análise de conglomerados (*clustering*), classificação, extração de informações, sumarização, análise qualitativa e quantitativa e identificação de regras de associação. Ainda nesse aspecto, o autor afirma que o processo de DCT é "interativo e iterativo, correspondendo à aplicação repetida de métodos de mineração e interpretação dos resultados pelo usuário." A pesquisa concentrou-se na técnica de análise de "conglomerados", isto é, do agrupamento de documentos textuais do sistema de informação SBRT.

A pesquisa foi realizada utilizando-se o pacote SAS Data Mining Solutions, composto de dois aplicativos – o SAS Enterprise Miner e o SAS Text Miner for Portuguese – desenvolvidos para revelar padrões e relações ocultos em dados, objetivando contribuir para o entendimento de tendências históricas e a previsão de oportunidades futuras.

O desenvolvimento da pesquisa pode ser sintetizado nas seguintes etapas: seleção do conteúdo dentre as bases de dados do sistema de informação do SBRT, extração dos dados, conversão dos dados para o formato legível pela ferramenta de mineração, preparação dos dados (limpeza e padronização), construção da base de trabalho, mineração do texto, agrupamento dos documentos e análise dos agrupamentos.

Conforme amplamente discutido na literatura, para que seja possível realizar qualquer tratamento automático de uma coleção de documentos escritos em linguagem natural em busca do conhecimento nela embutido, torna-se necessária a limpeza e padronização do texto. Sob essa ótica, Tan (1999) considera dois componentes estruturais da técnica de mineração de textos: o refinamento do texto, que transforma os documentos com textos não estruturados para o que ele chamou de "formato intermediário", e a destilação do conhecimento, que deduz padrões ou conhecimento a partir desse formato intermediário.

A experiência mostra que se gasta muito tempo na remoção de ruídos com o intuito de padronizar os dados, de forma a possibilitar maior precisão e acurácia no processo de mineração textual. De acordo com Quoniam et al. (2005), a etapa de preparação dos dados é crucial para a qualidade final dos resultados e corresponde a 60% de todo

[2] Respostas Técnicas (RTs) são as soluções elaboradas por especialistas da Rede SBRT em resposta às questões postadas pelos microempresários no sistema de informação do SBRT.

o processo de mineração. Nesta pesquisa, isso ficou bastante evidenciado, pois diversos processamentos se fizeram necessários até que se pudesse considerar que o conteúdo estava pronto para análise.

PREPARAÇÃO DOS DADOS (PRÉ-PROCESSAMENTO, LIMPEZA E TRANSFORMAÇÃO)

A etapa de preparação dos dados comprovou ser a mais trabalhosa na mineração de textos e, portanto, merece destaque para melhor compreensão do processo como um todo. Duas operações foram essenciais nessa etapa: a remoção de palavras não significativas e a lematização.

Remoção de palavras não significativas

As palavras não significativas – as chamadas *stopwords*, em inglês – são palavras comuns, encontradas em grande quantidade em um arquivo textual e não carregam significado em si próprias. Em geral, pertencem às seguintes classes gramaticais: artigos, conjunções, preposições, pronomes e advérbios. Wives (2004) apresenta três outras possibilidades de denominação do termo em língua portuguesa: "palavras negativas", "palavras-ferramenta" ou "palavras-vazias".

Em contraponto à lista de palavras não significativas (*stopwords*), o programa SAS oferece uma lista de *startwords*, termos que caracterizam o domínio do assunto a ser pesquisado, ou seja, todos os termos a serem considerados no processo de mineração. Após a extração de 416 palavras não significativas da base de trabalho do SBRT, restaram 43.271 termos na lista de *startwords*, que representam o *corpus* da base de dados utilizado nos processamentos.

Lematização

Em sequência à remoção das palavras não significativas, foi realizada outra etapa de preparação de dados para a mineração. Trata-se da lematização, ou extração de inflexões de termos, reduzindo-os a seus radicais, na intenção de se criarem padrões para proporcionar maior confiabilidade ao processo de mineração. Obviamente, há que se cuidar para que os radicais expressem um conceito comum. Essa tarefa é possível por meio da criação de um dicionário de termos com suas respectivas categorias gramaticais.

Segundo Schiessl (2007), a lematização é utilizada também para reduzir a quantidade de termos com a finalidade de facilitar a análise e reduzir o custo computacional, visto que restringe a quantidade de termos que serão processados.

Após as etapas de limpeza e padronização, os dados textuais ficaram prontos para a aplicação da técnica de mineração.

A mineração do texto – idas e vindas

Após as idas e vindas do processo de limpeza e padronização das palavras significativas do texto, foi iniciada a mineração propriamente dita, adotando-se a técnica de agrupamento dos documentos da base de trabalho. Utilizou-se, então, o recurso do SAS para agrupamento dos textos com base nas semelhanças entre eles. Esse agrupamento é referenciado com frequência na literatura a partir do seu termo em inglês, *clustering*, ou "geração de *clusters*" ou de conglomerados, na linguagem de alguns autores.

Agrupamento dos documentos

O agrupamento de documentos constitui o grande diferencial da técnica de mineração de textos, visto que identifica associações entre documentos aparentemente sem nenhuma relação. Ou seja, são apresentadas possibilidades de extração de conhecimentos totalmente novos e imprevistos.

O agrupamento permite que novas classes sejam descobertas, já que consegue agrupar documentos mesmo que estes não pertençam a assuntos conhecidos. Isso porque não há necessidade de conhecimento prévio sobre os assuntos (ou os possíveis assuntos) dos documentos. Os assuntos ou as classes dos documentos são descobertos após o agrupamento, durante o processo de análise dos grupos obtidos (WIVES, 2004)

A despeito do esforço inicial empregado no processo de limpeza e padronização dos dados, as primeiras análises trouxeram à tona alguns dos problemas que provocaram a necessidade de se realizarem diversos processamentos. A título de exemplo, são apresentados a seguir os erros detectados no tocante à lematização dos termos.

Erros de lematização

Ao se analisarem os primeiros agrupamentos gerados, verificou-se que alguns dos termos resultantes estariam invalidados, dada a sua frequência em praticamente todos os documentos da base de trabalho, como por exemplo: "data", "*site*", "palavras-chave". Outra observação foi a forte presença dos termos "podar" e "parir". Passou-se, então, a analisar a planilha gerada pelo SAS contendo todos os termos da base – 43.271 – e suas respectivas lematizações. Confirmou-se, por exemplo, que o termo "podar" não era um "lema perfeito", pois representava palavras de diferentes categorias, assim como o termo "parir", que incorporava entre suas representações o termo "para", que, embora fizesse parte da lista de palavras não significativas, não havia sido completamente eliminada durante o processo de limpeza e padronização dos dados.

O quadro 1, a seguir, traz um detalhamento desses erros, assim como os termos inadequados, com a respectiva explanação sobre a necessidade de eliminá-los e realizar novo processamento.

Cabe aqui fazer uma observação sobre a importância da lematização na língua portuguesa. Como observado em pesquisa de Bräscher (1999), com a lematização obtêm-se todos os contextos em que o lema foi empregado no *corpus*, independentemente de sua forma, o que enriquece as análises linguísticas que são objeto da pesquisa e evita a dispersão das frequências. No entanto, a autora alerta para o fato de a homografia provocar erros de lematização, uma vez que o sistema não tem como determinar, *a priori*, o lema correto para uma forma homógrafa, segundo seu emprego no *corpus*. Os exemplos do quadro 1 ilustram esse problema e ressaltam a necessidade da análise humana para solucioná-lo.

O *software* SAS utilizado já possui uma adaptação para a língua portuguesa, o que facilitou o processo de lematização. Ainda assim, erros como os descritos não puderam ser evitados. Acredita-se que poderá ocorrer melhora sensível nos resultados dos agrupamentos, se a ferramenta utilizada incorporar técnicas de processamento automático da linguagem natural para o tratamento de homografias.

Produto final da mineração: dados prontos para análise

Após as correções dos erros detectados, realizou-se novo processamento, que deu origem a 12 agrupamentos, os quais foram considerados representativos da realidade da base de trabalho e adequados para a realização das análises subsequentes. Eles representam o resultado da iteratividade típica da operação de mineração de textos.

O quadro 2, a seguir, mostra o detalhamento desses agrupamentos, isto é, os termos que provocaram a união dos documentos que os compõem, assim como a representatividade percentual de cada um, a quantidade de documentos em que os termos aparecem e a variabilidade deles no agrupamento.

ANÁLISES DOS RESULTADOS DA MINERAÇÃO

Logo na primeira leitura dos termos apresentados no quadro 2, pôde-se observar a presença da natureza tecnológica do conteúdo da base de dados SBRT, a partir dos verbos característicos de orientações para se realizar a aplicação do conhecimento contido na solução tecnológica fornecida, como por exemplo: alimentar, comer, cultivar, dever, elaborar, formar, plantar, processar, produzir, reciclar, resumir, usar, utilizar.

Da mesma forma, é possível inferir os principais temas de que trata a base, por meio da leitura dos substantivos mais freqüentes na lista de termos representativos dos agrupamentos: agricultura, animal, criação, equipamento, espécie, máquina, óleo, plástico, produção, produto, químico, resíduo, técnico.

QUADRO 1
Lista de erros de lematização

Termos representantes	Termos representados	Por que há necessidade de reprocessamento
+ podar	poda podar podadas podas podado pode podados pode-se podá-la podem podam podem-se podar podemos podam	Para o sistema, o infinitivo "podar" não diz respeito somente ao verbo podar, mas "poder" = corrigir essa relação.
+ fonte	fontes consultadas fonte:	Termos comuns a todos os documentos textuais da base = extrair Manter as demais formas do termo
+ parir	para	Palavra não significativa = extrair
end	end end.: end:	Termos comuns na maioria dos documentos textuais da base = extrair
bahia	Aparece em nomes próprios, exemplo: Rede de Tecnologia da Bahia - RETEC/BA	Nome de uma das instituições do SBRT, constante em todas as RTs respondidas por ela = extrair
site	site sites	Termos comuns a todos os documentos textuais da base = extrair
+ datar	data data de finalização	Termos comuns a todos os documentos textuais da base = extrair
	date	Termo presente em endereço de *site* referenciado = extrair
+ palavra-chave	palavra-chave palavras-chave	Termos comuns a todos os documentos textuais da base = extrair

Vale observar que quanto menor for o índice de variabilidade dos termos no agrupamento, maior é a sua precisão na representatividade e coesão de conteúdo. Identificou-se, portanto, o agrupamento de número 10 como sendo o mais coeso, por ter apresentado o menor coeficiente (0,0876160473). Some-se a este fato o de que os termos que formam o agrupamento (material, + reciclar, + resíduo, + plástico, + processar)[3] sugerem que o seu tema central está voltado para a preocupação com o meio ambiente, estímulo interessante para o objetivo geral desta pesquisa. Assim, o agrupamento 10 foi considerado o ideal para aprofundamento das análises em busca de comprovação da questão central da pesquisa: a possibilidade de uso da mineração de textos para extração de informações para apoiar a construção de indicadores de ciência e tecnologia.

O agrupamento 10 foi, portanto, o selecionado para o aprofundamento das análises em busca de informações para apoiar a construção de indicadores. Esse agrupamento foi analisado sob dois aspectos: a) classificação dos documentos e b) termos minerados, conforme detalhado a seguir.

a) Classificação dos documentos

A primeira análise que se fez no agrupamento 10 foi realizada considerando-se o assunto, campo

[3] O sinal "+" que antecede a maioria dos termos indica que se trata de um termo que representa uma classe de termos, como por exemplo: +reciclar representa: recicla, reciclada, reciclado, reciclador, recicladora, reciclando, reciclagem, entre outros termos.

QUADRO 2
Lista de agrupamentos e termos

Agrupamento	Peso (presença na base)	Frequência (quantidade documentos)	Variabilidade dos termos no agrupamento	Termos de Agrupamento
1	5%	120	0,1032064562	+ óleo, + químico, + usar, + processar, + utilizar
2	19%	464	0,1139042878	+ produto, + alimentar, + processar, + dever, + comer
3	5%	122	0,1075986813	+ químico, + produto, + produzir, + técnico, + resumir
4	20%	485	0,1144369151	+ material, + utilizar, + fonte, + usar, + processar
5	5%	115	0,1017080841	+ animal, + alimentar, + agricultura, + dever, + produção
6	7%	172	0,0941117077	+ cultivar, + solar, + plantar, + dever, + apresentar
7	8%	193	0,1034777388	+ máquina, + fornecedor, + utilizar, + elaborar, + usar
8	6%	136	0,099722373	+ identificação, responsável, + necessário, + dever, + informação
9	2%	37	0,093348348	+ espécie, + criação, + animal, + alimentar, + formar
10	3%	84	0,0876160473	+ material, + reciclar, + resíduo, + plástico, + processar
11	10%	231	0,0976318382	fax, + fornecedor, + equipamento, + máquina, + indústria
12	10%	242	0,1193337586	+ químico, + utilizar, + usar, + resumir, + processar
Total de documentos: [4]		2.401		

[4] Vale observar que, para efeito do tratamento estatístico, o *corpus* da base de trabalho, 6.041 documentos, foi automaticamente subdividido pela ferramenta utilizada em três arquivos, sendo um com 40% dos documentos, os quais são utilizados nos processamentos, e dois contendo, cada um, 30% dos documentos, que são reservados para validação. Portanto, os 2.401 documentos minerados representam estatisticamente o *corpus* total da base de trabalho.

da base que apresenta a classificação da Resposta Técnica (RT), de acordo com uma tabela de assuntos adaptada da Tabela CNAE (Classificação Nacional de Atividades Econômicas).

Os 84 documentos do agrupamento trazem uma diversidade de assuntos característica da base de RTs. Mesmo agrupados, dada a sua inter-relação em torno de um tema central, os documentos obtiveram 44 diferentes classificações, 34 das quais ocorrendo apenas uma vez, conforme detalhado no quadro 3, a seguir..

À primeira leitura dos assuntos, já se observa a coerência entre os documentos, que, de maneira geral, sugerem tratar de temas diversos, contudo, com uma visão central: o de tratamento e/ou reaproveitamento de materiais de diferentes naturezas para fins diversos.

b) Análise dos termos minerados

A partir da extração dos termos, foi possível aprofundar um pouco mais as análises, ficando evidenciado o diferencial de se utilizar a DCT para exploração de informações ocultas em documentos textuais. A seguir, estão descritas algumas particularidades da análise que fundamentam essa afirmativa.

Tome-se como base, por exemplo, o termo "resíduo", presente em 70% dos documentos do agrupamento 10, sob vários aspectos – coleta, estocagem, embalagem, acondicionamento, tratamento, disposição, transporte, descarte, incineração, aproveitamento. O termo está relacionado a vários tipos de materiais de diversas origens (alimentos, plásticos, vidros, *nylon*, gesso, madeira, borracha, sucatas metálicas, couro, dejetos humanos, óleos, materiais de construção, entulhos, embalagens de remédios, lixo hospitalar, lixo biológico, aterro sanitário, materiais eletroeletrônicos, resinas) e de diferentes naturezas: química, industrial, biológica, urbana, farmacêutica.

Pode-se observar que todos os documentos do agrupamento trazem consigo, implícita ou explicitamente, uma preocupação, e até mesmo

QUADRO 3
Lista geral de assuntos do agrupamento 10

Nº	Assunto	Ocorrências	Porcentagem %
1	Alimentos e bebidas	1	1,19
2	Borracha e plástico	2	2,38
3	Brinquedos e jogos recreativos	1	1,19
4	Cerâmica	1	1,19
5	Coleta de entulhos e refugos de obras e demolições	1	1,19
6	Coleta de resíduos não-perigosos	1	1,19
7	Coleta de resíduos perigosos (2)	5	5,95
8	Descontaminação e outros serviços de gestão de resíduos	1	1,19
9	Extração de minério de níquel	1	1,19
10	Fabricação de brinquedos e jogos recreativos	1	1,19
11	Fabricação de aditivos de uso industrial	1	1,19
12	Fabricação de artefatos de material plástico para outros usos não especificados anteriormente	1	1,19
13	Fabricação de bolsas de plástico	1	1,19
14	Fabricação de embalagens de material plástico	2	2,38
15	Fabricação de móveis de outros materiais	1	1,19
16	Fabricação de peças e acessórios para o sistema motor de veículos automotores	1	1,19
17	Fabricação de produtos de material plástico	1	1,19
18	Fabricação de produtos químicos orgânicos não especificados anteriormente	1	1,19
19	Fabricação de resinas de poliuretano	1	1,19
20	Fabricação de resinas termoplásticas	1	1,19
21	Fabricação de sacolas de material plástico	1	1,19
22	Fabricação de tecidos especiais	1	1,19
23	Fabricação de vassouras	1	1,19
24	Medição da poluição	1	1,19
25	Meio ambiente	18	21,43
26	Meio ambiente, reciclagem e tratamento de resíduos	5	5,95
27	Minerais não metálicos	1	1,19
28	Mobiliário	1	1,19
29	Reciclagem	1	1,19
30	Reciclagem de sucatas não-metálicas	1	1,19
31	Recuperação de materiais metálicos produto reciclado	1	1,19
32	Recuperação de materiais não especificados anteriormente	3	3,57
33	Recuperação de materiais plásticos	5	5,95
34	Recuperação de resíduos contendo produtos químicos	1	1,19
35	Recuperação de sucatas de alumínio	1	1,19
36	Reforma de pneumáticos usados	1	1,19
37	Reparação e manutenção de computadores e de equipamentos periféricos	1	1,19
38	Seleção latas de alumínio usadas	1	1,19
39	Serviço de coleta acondicionamento e transporte de lixo hospitalar	1	1,19
40	Serviços industriais	5	5,95
41	Tecelagem de fios de fibras artificiais e sintéticas	1	1,19
42	Tratamento e disposição de resíduos	1	1,19
43	Tratamento e disposição de resíduos não-perigosos	2	2,38
44	Tratamento e disposição de resíduos perigosos	3	3,57
Total		**84**	**99,97%**

certo comprometimento, com o meio ambiente. Isso fica claro quando se analisam os termos resultantes da mineração nos documentos, embora eles não tenham sido classificados sob a categoria "meio ambiente".

Em suma, a aplicação da DCT proporcionou a visualização das fortes relações existentes entre documentos que aparentemente não teriam interação entre si, caso fossem consideradas apenas as classificações e palavras-chave que lhes foram atribuídas.

Outro exemplo interessante de ser citado é o fato de algumas RTs do agrupamento 10 tratarem de assuntos tão diversificados e aparentemente distanciados do tema central identificado, ou seja, o meio ambiente. Eis alguns exemplos: fabricação de brinquedos (jogos recreativos, bolinhas de gude); fabricação de vassouras, produção de energia elétrica; fabricação de sofás. No entanto, um olhar mais aproximado revelou que as RTs se enquadram com perfeição no tema, dada a natureza das matérias-primas utilizadas. A fabricação de vassouras e sofás utiliza como matéria-prima garrafas PET recicladas, assim como a produção de energia elétrica, fruto do calor emanado pela queima desse produto. Os jogos recreativos são feitos de papel reciclado e as bolinhas de gude são produzidas a partir de sucatas de vidro.

Detectou-se, ainda, outra natureza de envolvimento com o meio ambiente em RTs do agrupamento, pela presença de expressões como "padrões de emissão de efluentes", "normas técnicas", "legalização", "legislação", "procedi-mentos legais", "licenciamento ambiental", as quais sugerem uma preocupação dos microempresários em exercer suas atividades de forma regulamentar, dentro de padrões que preservem o meio ambiente.

Em todo o agrupamento, visualiza-se a presença marcante de termos como "aproveitamento", "reaproveitamento", "reciclagem", "resíduos", "recuperação", "sucatas", "descontaminação", "produção limpa", os quais evidenciam processos condizentes com questões ambientais.

Esse pode ser visto como um dos temas de destaque do SBRT que merece ser explorado para a extração de informações úteis à construção de indicadores, vez que se trata de um assunto de extrema importância para a sociedade. Em sendo o meio ambiente uma preocupação nacional, a análise aprofundada dessas soluções tecnológicas poderia facilmente nortear ações governamentais de incentivo a iniciativas que de fato contribuíssem para o bem-estar social.

Entende-se que uma forma de melhor explorar esse conteúdo seria o cruzamento dos dados obtidos a partir da DCT com os demais dados disponíveis no sistema de informação. Isso possibilitaria a identificação, por exemplo, de regiões onde estariam sendo empreendidos determinados esforços, que tipo de empresa ou o perfil do empreendedor busca essas orientações, assim como os tipos de iniciativas que poderiam ser associadas umas às outras em busca de melhor aproveitamento dos esforços empreendidos.

Acredita-se que, a partir de dados como esses, que expõem a preocupação do microempresário e do empreendedor brasileiro em buscar orientações para atuar no setor produtivo em consonância com os anseios do bem-estar público, poder-se-á chegar a níveis de decisões estratégicas que facilitem essas ações e produzam melhores resultados a menores custos.

CONSIDERAÇÕES FINAIS

A aplicação da DCT é ainda uma atividade muito pouco explorada no Brasil, conforme foi possível detectar pela escassez de literatura em língua portuguesa sobre o tema. Muitos dos documentos localizados estão inseridos em áreas como informática e estatística, nas quais se pesquisam temas relacionados à descrição de metodologias, funcionalidades de ferramentas de mineração, ou tratamentos estatísticos, lingüísticos, indexação automática e bases de dados. Não foram localizados muitos documentos em língua portuguesa tratando de análise de conteúdos utilizando a DCT. Entende-se, portanto, que esta pesquisa apresentou um grau de

ineditismo, por ser pioneira no estudo desta técnica aplicada a conteúdos de informação tecnológica, voltada à aplicação dos conhecimentos gerados em prol do setor produtivo de pequeno porte.

A aplicação da DCT nos conteúdos textuais do Serviço Brasileiro de Respostas Técnicas trouxe à tona uma diversidade de informações agrupadas que não poderiam ter sido visualizadas sem o uso dessa técnica, cuja capacidade de extrair informações ocultas em acervos textuais os transforma em preciosas fontes de novos conhecimentos. O estudo demonstrou que a DCT pode aproximar textos de temas aparentemente díspares e, assim, proporcionar um mergulho diferenciado no conteúdo existente visando seu melhor aproveitamento. Além disso, pode propiciar a identificação de informações inesperadas, como, por exemplo, a preocupação dos microempresários com os aspectos regulatórios e legais e com as questões ambientais. O conhecimento desses aspectos é valioso para a condução de políticas públicas que visem explorar esse potencial identificado.

As informações obtidas por meio da aplicação da DCT podem se configurar em importantes fontes para a construção de indicadores, onde será possível identificar os impactos sociais de um serviço criado pela aplicação de recursos públicos destinados a Ciência e Tecnologia. A partir do cruzamento dos dados extraídos na mineração de conteúdos textuais (Respostas Técnicas) com os metadados disponíveis na base de dados de Respostas Técnicas, será possível extrair os mais diversos tipos de indicadores que possam nortear ações futuras, como, por exemplo, as regiões do país em que determinados temas estão sendo mais explorados; o tipo de cliente que lida com determinado assunto e com que finalidades; ou que tipo de técnica está sendo utilizada para a produção determinados produtos.

Extrapolando os limites do estudo de caso da pesquisa, entende-se que os resultados alcançados podem ser vistos como teste de utilização da DCT, uma prática aplicável a outros conteúdos informacionais com características semelhantes ao analisado. A despeito da energia despendida nos repetidos processamentos e análises aprofundadas que se fazem necessários à aplicação da DCT, os resultados comprovam ser compensador lançar mão desse recurso para a extração de conhecimento de conteúdos textuais anteriormente desconhecidos e pouco valorizados.

AGRADECIMENTO

Agradecemos à Coordenação da rede SBRT, pela liberação dos dados, e ao SAS Institute Brasil Ltda., pela cessão das ferramentas, elementos essenciais à realização do estudo que deu origem a este artigo, assim como à Diretoria do Ibict, pelo apoio incondicional ao desenvolvimento da pesquisa.

Artigo submetido em 19/01/2009 e aceito em 06/02/2009.

REFERÊNCIAS

AIRES, Rachel Virgínia Xavier. *Uso de marcadores estilísticos para a busca na Web em português*. Orientadora: Profa. Dra. Sandra Maria Aluísio, Co-orientadora: Dra. Diana Santos. 2005, 202 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional)-USP-São Carlos- Instituto de Ciências Matemáticas e de Computação - ICMC-USP.

BRÄSCHER, M. *Tratamento automático de ambigüidades na recuperação da informação*. 290 f. Tese (Doutorado em Ciência da Informação) - Curso de Pós-graduação em Ciência da Informação, Universidade de Brasília, Brasília, 1999.

BRASIL. Ministério da Ciência e Tecnologia. *Indicadores de Ciência & Tecnologia – 2002*. Brasília: MCT, 2004, 140 p. ISSN 1413-3148. Disponível em: < http://www.mct.gov.br/index.php/content/view/3770.html>. Acesso em: 18 jan. 2009.

FAYYAD, U., *et al. From Data Mining to Knowledge Discovery in Databases.* AI Magazine, Fall 1996, p. 37-53. Disponível em: < http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf>. Acesso em 10 jan. 2009.

HEARST, Marti. *Untangling Text Data Mining*. In: *Proceedings of* ACL'99: *the 37th Annual Meeting of the* Association for Computational Linguistics, *University of Maryland, June 20-26, 1999 (invited paper*). Disponível em: <http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>. Acesso em: 12 jan. 2009.

KONDO, Edson Kenji. *Desenvolvendo indicadores estratégicos em ciência e tecnologia*: as principais questões. Ciência da Informação, 1998, vol.27, no.2, p. 128-133.

LUCAS, Marty. Mining in textual mountains. Mapa Mundi Magazine, disponível em <http://mappa.mundi.net/trip-m/hearst/>. Acessado em 18 jan. 2009.

LE COADIC, Yves F. Mathématique et statistique en science de l'information et en science de la communication: Infométrie mathématique et infométrie statistique des revues scientifiques. *Ciência da Informação*. Brasília, DF. v. 34, n. 3, p.15-22, set./dez. 2005. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/818/0>. Acesso em: 18 jul 2009.

QUONIAM, Luc, TARAPANOFF, Kira, ARAÚJO JÚNIOR, Rogério Henrique, ALVARES, Lillian. I*nteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. Ciência da Informação*, Brasília, v. 30, n. 2, p. 20-28, maio/ago. 2001. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/viewFile/183/162>. Acesso em 15 jan. 2009.

SCHIESSL, José Marcelo. *Descoberta de Conhecimento em Texto aplicada a um sistema de atendimento ao consumidor*. Orientador: Profa. Dra. Marisa Bräscher, 2007. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília. Disponível em: < http://bdtd.bce.unb.br/tedesimplificado/tde_busca/arquivo.php?codArquivo=1538>. Acesso em: 18 jan. 2009.

SARTORI, R. ; PACHECO, R. C. dos S. . Indicadores de Ciência, Tecnologia e Inovação: a interação humana nos grupos de pesquisa. In: VII Congreso Iberoamericano de Indicadores de Ciencia y Tecnología, 2007, São Paulo. Annales del VII Congreso Iberoamericano de Indicadores de Ciencia y Tecnología - Nuevos Indicadores para Nuevas Demandas de Información. Buenos Aires : RICYT, 2007, 2007.

TAN, A.-H. Text mining*:* The state of the art and the challenges. *In: Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining – PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases*, Beijing, p. 65–70, 1999. Disponível em: < http://www.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf>. Acesso em: 18 jan. 2009.

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

VELHO, L. Estratégias para um sistema de indicadores de C&T no Brasil. Parcerias estratégicas, Brasília, Brasil, v. 13, n. -, p. 109-121, 2001. Disponível em: < http://www.cgee.org.br/arquivos/pe_13.pdf>. Acesso em: 25 jun. 2009.

WEISS, Sholom, INDURKHYA, Nitin, ZHANG, Tong e DAMERAU, Fred. *Text Mining*: Predictive Methods for Analyzing Unstructured Information. Springer, New York, 2005.237 p.

WIVES, Leandro Krug. *Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos*. Orientador: Oliveira, José Palazzo Moreira de, 2004. 126f : il. Tese(Doutorado)-Universidade Federal do Rio Grande do Sul. Porto Alegre: Programa de Pós-Graduação em Computação. Disponível em: <http://www.lume.ufrgs.br/handle/10183/4576>. Acesso em: 16 jan. 2009.