VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação
28 a 31 de outubro de 2007 • Salvador • Bahia • Brasil

GT 7 – Produção e Comunicação da Informação em CT&I
Pôster

EXPLORANDO O CONTEÚDO DE UM SISTEMA DE INFORMAÇÃO DESENHADO PARA MICROEMPRESAS: Aplicação da Descoberta de Conhecimento em Texto para apoio à geração de indicadores de CT&I

EPLORING THE CONTENT OF AN INFORMATION SYSTEM DESIGNED TO SMALL ENTERPRISES: Application of Knowledge Discovery in Text as a support to ST&I indicators production

Hélia de Sousa Chaves Ramos (PPGCInf/UnB, helia@ibict.br)
Marisa Bräscher (PPGCInf/UnB, marisab@unb.br)

Resumo: Esta pesquisa propõe a aplicação da Descoberta de Conhecimento em Texto (DCT) no Serviço Brasileiro de Respostas Técnicas (SBRT), um sistema de informação tecnológica na *Web* destinado ao setor produtivo brasileiro, notadamente empreendedores, micro e pequenas empresas, fruto de um esforço compartilhado entre governo, instituições de pesquisa, universidades e iniciativa privada. O objetivo central da pesquisa é extrair informações para apoiar a geração de indicadores e a tomada de decisão estratégica, assim como a definição de políticas públicas para o setor produtivo de pequeno porte. A metodologia adotada contempla a utilização da Descoberta de Conhecimento em Texto (DCT) em 6.014 registros extraídos do sistema de informação SBRT. Após a aplicação dessa ferramenta espera-se obter os indicadores necessários para a etapa de análise e conclusão da pesquisa.
Palavras-chave: Descoberta de Conhecimento em Texto (DCT), mineração de textos, indicadores de CT&I, serviços de informação, microempresa.

Abstract: *This research proposes the application of Knowledge Discovery in Textual Databases (KDDT) in the Brazilian Service for Technical Answers (Serviço Brasileiro de Respostas Técnicas – SBRT), a technological information service in the Web destined to the Brazilian production sector, specially micro and small enterprises or entrepreneurs. SBRT is an effort shared by government, research institutions, universities and the private sector. The main objective of the research is to mine information in this database in order to give support to the generation of ST&I indicators and to the strategic decision-making process, as well as the definition of public policies for the production sector composed by small enterprises. The methodology adopted encompasses the use of Knowledge Discovery in Text in 6.014 documents extracted from SBRT information system. After the application of this tool, it is expected to obtain indicators necessary for the analysis and conclusion of the research.*
Keywords: *Knowledge Discovery in Texts (KDT), text mining, ST&I indicators, information services, small enterprises.*

1. Introdução

Esta pesquisa tem por objetivo central explorar 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) –, com a finalidade de apoiar a geração de indicadores e a tomada de decisão estratégica, assim como a definição de políticas públicas para o setor produtivo de pequeno porte. A pesquisa será realizada com a aplicação da técnica de descoberta de conhecimento em texto nos documentos técnicos que compõem o conteúdo textual da base de dados.

Dentre as ações governamentais voltadas para o apoio à micro e pequena empresa brasileira, está o apoio à criação da Rede de Serviços de Informação Tecnológica, composta por instituições governamentais, universidades e iniciativa privada. O primeiro produto dessa rede é SBRT, concebido pelo Ministério da Ciência e Tecnologia (MCT) como um "Serviço Brasileiro que pudesse responder à demanda dos pequenos e médios empresários por informações de fácil acesso e que contribuísse para a melhoria de seus produtos ou processos, por meio da articulação das competências instaladas no País." (BRASIL, 2006). Responderam a esse anseio do MCT sete renomadas instituições brasileiras[1], atuantes nos campos de ensino, pesquisa e prestação de serviços tecnológicos, que se reuniram para a criação da Rede SBRT com o apoio do governo federal.

O sistema de informação SBRT contém um conjunto de informações que expressam a demanda do micro e pequeno empresário brasileiro, assim como a resposta a essa demanda, na forma de conhecimento produzido por instituições de ensino e pesquisa atuantes no setor de informação tecnológica.

Armazenado em forma textual, o conteúdo não está codificado para exploração por máquina e sua recuperação por meio de buscas com palavras-chave não reflete sua riqueza, dada a sua dispersão. A base de dados é composta por documentos técnicos (RT), cuja indexação é feita com o uso de uma tabela de assunto adaptada da CNAE (Classificação de Atividades Econômicas – Fiscal) e por palavras-chave extraídas do texto das RTs. Essa indexação é fundamental para a busca e recuperação de documentos específicos, para atendimento a demandas pontuais. É insuficiente, no entanto, para a realização de estudos do conteúdo da base de dados que permitam a extração de informações de interesse estratégico e gerencial, e para a identificação de tendências do setor e das relações existentes entre os documentos técnicos produzidos pela rede de instituições que compõem o SBRT.

O iminente crescimento da base de dados de respostas técnicas representará maior produção e disponibilização de conteúdos e a impossibilidade de realizar seu tratamento exclusivamente pelo homem.

Assim, torna-se necessária a aplicação de ferramentas e técnicas que possibilitem a exploração e sistematização desses conteúdos, de forma a melhor aproveitá-los, seja na associação dos conteúdos para geração de novos conhecimentos, seja na extração de informações estratégicas para gestão da rede SBRT e definição de políticas públicas voltadas ao setor produtivo brasileiro.

Acredita-se que a descoberta de conhecimento em texto aplicada no conteúdo da base de dados do SBRT poderá propiciar a extração de informação para apoio à geração de indicadores de Ciência e Tecnologia e à tomada de decisão, no que diz respeito ao direcionamento de ações, visando a obter o melhor resultado dos investimentos públicos no setor produtivo nacional de pequeno porte.

---

[1] São elas: o Centro de Desenvolvimento Científico e Tecnológico da Universidade de Brasília (**CDT/UnB**), o Disque-Tecnologia da USP (**CECAE/USP**), a Fundação Centro de Tecnologia de Minas Gerais (**CETEC/MG**), a Rede de Tecnologia do Rio de Janeiro (**REDETEC/RJ**), o Instituto de Tecnologia do Paraná (**TECPAR/PR**), o Instituto Euvaldo Lodi – Núcleo Regional da Bahia (**IEL/BA**), e o Serviço Nacional de Aprendizagem Industrial – Departamento Regional do Rio Grande do Sul (**SENAI/RS**). Somam-se a estas, na qualidade de parceiras, o Instituto Brasileiro de Informação em Ciência e Tecnologia (**Ibict**) e o Serviço de Apoio às Micro e Pequenas Empresas (**Sebrae Nacional**).

Vislumbra-se, ainda, a possibilidade de se caracterizarem as necessidades de informação tecnológica de empreendedores e da micro e pequena empresa brasileira, assim como mapear e organizar os conteúdos das Respostas Técnicas (RTs), visando identificar as relações entre conteúdos de RTs distintas e eliminar redundâncias, evitando-se, assim, o retrabalho.

Assim, propõe-se com este estudo investigar os conteúdos veiculados pelo sistema de informação do SBRT, com vistas a extrair informações para apoio à geração de indicadores de C&T, à tomada de decisão, seja em nível governamental, seja em nível institucional, ou ainda, extrapolando os limites organizacionais, contemplar a gestão da própria rede SBRT.

2. O contexto

Sabe-se das necessidades de acesso à informação pelas micro e pequenas empresas, para se manterem no mercado cada vez mais competitivo. Assim, serviços que disponibilizem informações e conhecimentos para a capacitação tecnológica de empresários e, por conseqüência, melhoria de suas empresas, se tornam indispensáveis. O SBRT não só atende a essa necessidade do ponto de vista de disponibilizar a resposta certa, personalizada, para a questão colocada pelo microempresário, mas também promove a aproximação dessa clientela do setor produtivo com as unidades de pesquisa que produzem o conhecimento técnico necessário para o atendimento a essa necessidade.

O banco de dados gerado pelo serviço é de acesso público, gratuito e constitui um dinâmico banco de conhecimento gerado a partir de demandas reais postadas por empreendedores de todo o território nacional.

O atendimento às necessidades de informação é potencializado pelo fato de a informação poder ser acessada no momento imediatamente após a sua geração e publicação. Some-se a isso a possibilidade de uma RT poder ser utilizada, na íntegra, por vários outros empreendedores, o que caracteriza um crescente potencial de auto-atendimento, podendo, ainda, servir de base orientadora para novas solicitações, mais pontuais ou complementares.

Visualiza-se o SBRT como indutor da utilização da informação como estratégia de desenvolvimento do setor produtivo de pequeno porte.

Assim, entende-se a importância do estudo desses conteúdos para extração de informações para apoio à geração de indicadores e à tomada de decisão, seja no campo da gestão interna da rede SBRT, seja no âmbito da política governamental de apoio à promoção da competitividade da micro e pequena empresa brasileira, buscando, assim, dar maior retorno à sociedade dos recursos públicos investidos na pesquisa tecnológica.

A relevância da pesquisa para a Ciência da Informação reside na abordagem de alguns temas ainda pouco explorados na literatura da área, notadamente a informação tecnológica e a questão do setor produtivo como usuário do conhecimento armazenado em sistemas de informação para a melhoria de seus produtos e processos.

Soares (2003) defende a expansão da Ciência da Informação no que diz respeito à investigação de outros ambientes nos quais a informação também se destaca como um dos principais elementos: o ambiente das empresas de base tecnológica. Segundo ele, essas empresas constituem um primeiro extrato de um novo modelo, tendo como principal insumo o conhecimento. Define informação tecnológica como a “aplicação de conhecimento no desenvolvimento de um processo/produto/serviço”.

3. Sobre o Serviço Brasileiro de Respostas Técnicas (SBRT)

Em novembro de 2004, foi lançado, em nível nacional, o Serviço Brasileiro de Respostas Técnicas (SBRT), um serviço de informação tecnológica na *Web* voltado para o registro e atendimento gratuito de demandas tecnológicas de baixa complexidade postadas por empreendedores e micro e pequenos empresários de todo o País.

A criação do SBRT teve como objetivos facilitar o rápido acesso a soluções tecnológicas de baixa complexidade e em áreas específicas, difundir e potencializar conhecimentos acumulados nas instituições de C,T & I e contribuir para com o processo de transferência de tecnologia (conexão entre as demandas e as competências), especialmente para as empresas de menor porte. Propicia, ainda, às instituições membros a rica experiência de atuarem em rede, unindo suas competências no fornecimento de soluções tecnológicas elaboradas sob medida. (SBRT, 2005)

O Serviço SBRT é materializado por meio de um sistema de informação *online*, que possibilita o cadastramento dos interessados em obter soluções personalizadas a suas questões de natureza tecnológica em forma de documentos elaborados com o apoio de especialistas das diversas instituições que compõem a rede SBRT.

Os interessados em ter uma questão solucionada pelo SBRT podem encaminhar suas perguntas por telefone, correspondência, pessoalmente, nos postos espalhados pelo País – as próprias instituições membros da rede SBRT e balcões Sebrae – ou via Internet (http://www.respostatecnica.org.br), por meio do cadastramento gratuito como cliente da rede SBRT e do posterior preenchimento de formulário eletrônico, registrando sua demanda.

As solicitações são encaminhadas automaticamente às instituições membros da rede SBRT e respondidas com o auxílio de especialistas, que elaboram respostas personalizadas na forma de documentos técnicos, as chamadas Respostas Técnicas (RTs). As RTs são validadas pela instituição responsável e enviadas aos respectivos clientes. Em seguida, são publicadas na base de dados do sistema de informação SBRT – sem a identificação dos solicitantes – para ampla divulgação e livre utilização pela população interessada, novos potenciais clientes do serviço.

As RTs têm por finalidade proporcionar aos microempresários e empreendedores condições de aplicar com facilidade as soluções apresentadas, de forma a propiciar a melhoria do processo ou produto e contribuir para a melhoria da sua competitividade, ou, ainda, para a formalização de novos empreendimentos e geração de emprego e renda.

3.1 O foco de atuação do SBRT

As respostas técnicas publicadas pelo SBRT são geradas para responder a demandas consideradas de baixa complexidade – a partir da pesquisa em fontes de informação de confiáveis, de domínio público – e de média complexidade, que envolvem análises mais específicas incluindo opiniões de especialistas.

As questões consideradas complexas, que exigem conhecimento altamente especializado e pesquisa mais aprofundada recebem outro tipo de tratamento, podendo gerar consultorias específicas, dependendo do interesse do cliente, com custos previamente negociados. Questões complexas, portanto, não fazem parte do escopo de Respostas Técnicas gratuitas veiculadas pelo serviço *online.*

3.2 A dimensão social do SBRT

Em pesquisa realizada em dezembro de 2005 sobre o uso do SBRT, chegou-se à conclusão de que 92% dos usuários do serviço são pessoas físicas, empreendedores em “busca de novos negócios”. (CHAVES, CORDEIRO e BASTOS, 2006). Mais recentemente, em março de 2007, verificou-se que essa tendência se manteve, quando apenas 9% das solicitações provinham de microempresas.

Há que se destacar, portanto, o papel social que o SBRT exerce na sociedade, na medida em que fomenta o empreendedorismo e, como conseqüência, apóia a formalização de pequenos negócios e contribui para a geração de emprego e renda. Além disso, contribui para a capacitação do empreendedor, que se utiliza das técnicas sugeridas para solução de seus

problemas, e, ainda, para o combate à conhecida alta mortalidade da microempresa brasileira, abordada no item 4.2 desta pesquisa.

Essa dimensão social do SBRT é reconhecida pelo MCT, que declara em seu Relatório de Gestão 2003-2006 que a "informação tecnológica, por permitir a consolidação de empreendimentos de pequeno porte, se presta ao combate à informalidade e, conseqüentemente, estimula a inclusão social." (BRASIL/MCT, 2006)

4. Fundamentação teórica

4.1 Gestão de C&T

As primeiras discussões sobre o papel e prioridades da política de Ciência e Tecnologia (C&T) ocorreram nas décadas de 1960 e 1970, no âmbito das discussões acerca da economia da inovação, e permeavam dois campos considerados opostos. O primeiro deles defendia os investimentos em P&D, principalmente na pesquisa básica, com a finalidade de promover os avanços científicos e tecnológicos, considerados os "principais alavancadores do progresso técnico (teorias classificadas como *science & technology-push*)". O segundo enfatizava as "forças do mercado e da demanda como o determinante primordial do progresso técnico (teorias classificadas como *demand-pull*)". Nesta visão, acreditava-se não serem os esforços de P&D os responsáveis pela maior parte das inovações, mas a atuação de outras partes da empresa (como as áreas de engenharia, produção e controle de qualidade), de outros elementos da cadeia produtiva (produtores de equipamentos, insumos e prestadores de serviços) ou dos próprios consumidores.". (LASTRES, 1995)

A evolução do setor de C&T no Brasil passou pela criação do Ministério da Ciência e Tecnologia (MCT), em 15 de março de 1985, pelo Decreto nº 91.146, como órgão central do sistema federal de Ciência e Tecnologia, com as seguintes responsabilidades inseridas na sua área de competência: "o patrimônio científico e tecnológico e seu desenvolvimento; a política de cooperação e intercâmbio concernente a esse patrimônio; a definição da Política Nacional de Ciência e Tecnologia; a coordenação de políticas setoriais; a política nacional de pesquisa, desenvolvimento, produção e aplicação de novos materiais e serviços de alta tecnologia". (MCT)

Observa-se, ao longo do tempo de atuação do MCT, uma crescente preocupação com a inserção do setor produtivo nas prioridades de investimento dos recursos destinados à Ciência e Tecnologia e nas iniciativas de estímulo ao crescimento desse setor como propulsor da economia nacional. A esse respeito, Helena Lastres (1995) cita o relatório de atividades do MCT relativo ao período de 1992-1994, cujos objetivos principais contemplavam a consolidação da base científica e tecnológica “de forma a permitir um desenvolvimento endógeno capaz de oferecer soluções criativas e duradouras aos principais problemas nacionais” e a mobilização do setor produtivo a uma maior participação neste esforço (uma vez que se reconhece que a participação das empresas nos gastos nacionais não passa de 10%).". (LASTRES, 1995)

4.2 A geração de indicadores em CT&I

Romão, Pacheco e Niederauer (2000) definem indicadores de C&T como "numéricos capazes de resumir informações generalizadas sobre investimentos, produção e tendências no campo da C&T.".

Segundo o MCT – órgão responsável pela formulação e implementação da Política Nacional de Ciência e Tecnologia – há uma série de fatores que tornam "extremamente complexa a seleção e construção de indicadores", que vão desde a heterogeneidade e amplidão de abrangência das atividades que envolvem o setor, passando pelo envolvimento de uma multiplicidade de agentes e instituições públicas e privadas, até a questão do tempo transcorrido entre as iniciativas e os resultados das ações. Outra questão levantada é o fato de os resultados

produzidos não serem "facilmente computáveis, como é o caso dos ativos intangíveis". Para vencer esses desafios e fazer face à responsabilidade de organizar e divulgar as informações de C&T no País, o MCT conta com a colaboração de instituições públicas, no âmbito federal e estadual, e de organizações privadas que produzem informações de interesse para a construção de indicadores de C&T e para o desenvolvimento de estudos sobre o tema. (MCT, 2002)

Dessa forma, é desejável que iniciativas envolvendo a aplicação de recursos públicos voltados para o desenvolvimento sejam acompanhadas e mensuradas sistematicamente como fontes de retroalimentação para o norteamento de ações futuras visando ao melhor aproveitamento dos recursos investidos.

Dentro da ótica de adaptação dos indicadores de C&T propostos pela Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) para melhor atender às necessidades dos países em desenvolvimento, Edson Kondo (1998) sugeriu expandir o foco da geração de indicadores de C&T, tradicionalmente voltados para a eficiência econômica, para abranger indicadores "vinculados ao bem-estar social".

Essa temática é abordada por Lea Velho (2001), quando levanta as questões relativas ao estabelecimento de um sistema de indicadores de C&T "útil e relevante para a tomada de decisão" e afirma: "No contexto atual, a ciência deixou de ser valorizada simplesmente por avançar o conhecimento e passou a ter sentido por seus resultados em termos de impacto na sociedade e na produção" e que 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.

Dessa forma, acredita-se que serviço SBRT está inserido nessa nova concepção de conteúdos para apoio à geração de indicadores, já que se trata de um estímulo à aplicação do conhecimento tecnológico gerado para melhoria da competitividade da microempresa brasileira e a conseqüente contribuição, tanto para a economia quanto para o bem-estar social citado por Kondo.

É sabido que as MPEs são, no Brasil, um dos pilares de sustentação da economia, em razão de seu número, abrangência, capilaridade e capacidade de geração de emprego. De acordo com o Sebrae, há mais de 15 milhões de empreendimentos informais no país, três vezes mais que o número de micro e pequenas empresas formalmente constituídas: 4,6 milhões. “As micro, pequenas e médias empresas oferecem a absoluta maior parte dos empregos no país. De sua modernização e expansão dependem a curto prazo o surgimento de mais trabalho e emprego para a população”. (TAKAHASHI, 2005)

No Relatório de Gestão do MCT para o período 2003-2006, o SBRT figura entre as iniciativas que compõem o chamado Eixo Política Industrial, Tecnológica e de Comércio Exterior, como uma das ações de apoio à inovação e à competitividade que "tem tido ótimos resultados". Cita o relatório que a "importância estratégica do projeto SBRT para o aumento da competitividade nacional é, ainda, reforçada por sua contribuição para o estabelecimento de uma cultura de geração e difusão da informação tecnológica e para o desenvolvimento de negócios no setor produtivo". (BRASIL/MCT, 2006)

4.3 A micro e pequena empresa

O Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (Sebrae) adota o critério número de empregados para a classificação das empresas brasileiras e expõe o seguinte conceito, adotado como referência aos tipos de empresa abordados neste trabalho:

> ...considera-se como microempresa aquela com até 19 empregados na indústria e até 09 no comércio e no setor de serviços; as pequenas empresas são as que possuem, na indústria, de 20 a 99 empregados e, no comércio e serviços, de 10 a 49 empregados; as médias empresas de 100 a 499 empregados na indústria e de 50 a 99 no comércio e serviços. (SEBRAE)

De acordo com Araújo (2005), há poucas pesquisas acadêmicas sobre as microempresas, sendo essas mais freqüentes sob o ponto de vista econômico (geração de emprego e crescimento nacional). A autora cita alguns estudos que abordaram a informação como estratégia para tomada de decisões no âmbito das micro, pequenas e médias empresas: Campos (1977), Borges (2002), Costa (2003) e Dias e Belluzzo (2003). Afirma que, ao longo das últimas três décadas, o governo Federal tem sido o maior investidor em serviços de informação voltados para as micro, pequenas e médias empresas.

Em suas reflexões sobre o setor empresarial na América Latina, Julio Cubillo (1997) afirma que a pequena e média empresa é considerada como um dos atores-chave do desenvolvimento, e isso se confirma quando se examinam os indicadores econômico-sociais, assim como na sua consolidação como tema relevante nas agendas do desenvolvimento.

O alto índice de mortalidade da microempresa brasileira é bastante conhecido. Em pesquisa realizada em 2004, o Sebrae noticia que foram extintas 59,9% das empresas com até 4 anos de existência, 56,4% daquelas que tinham até 3 anos, e 49,4% das que tinham apenas 2 anos de existência. Dentre as causas apontadas pelos entrevistados da pesquisa realizada em todo o território nacional, destacam-se, em primeiro lugar, as falhas gerenciais na condução dos negócios, seguida de causas econômicas conjunturais e tributação. Deduz-se, no relatório, que as falhas gerenciais podem estar relacionadas à "falta de planejamento na abertura do negócio". (SEBRAE, 2004).

Não é difícil associar essa carência à falta de preparo do microempresário, não só do ponto de vista de acesso a informações gerais de extrema relevância para a abertura do seu negócio, como também na manutenção do sucesso de seu empreendimento, ou de conhecimentos específicos sobre seu ramo de negócio.

Nesse sentido, Blaise Cronin, citado por Araújo, Freire e Mendes (1997), sustenta que, na sociedade atual, o sucesso do setor produtivo "tem sido caracterizado pela busca de informação, pela comunicação com fontes de conhecimento relevantes, pela capacidade de absorção de tecnologias nas unidades produtivas e, especialmente, pela capacidade para produção e avaliação de informações". As autoras abordam a demanda da indústria por informação que represente acesso a "conhecimento para ação" e afirmam ser a informação (e sua efetiva comunicação), cada vez mais, um dos recursos mais importantes para a produção de bens e serviços. (ARAÚJO, FREIRE e MENDES, 1997)

Entende-se que o serviço SBRT é um forte aliado à capacitação do microempresário na aplicação das soluções geradas pelas instituições membros da rede SBRT para melhoria de seus produtos e processos e, conseqüentemente, da competitividade do setor produtivo brasileiro de pequeno porte.

4.4 Gestão da Informação

4.4.1 A Informação Tecnológica

O tipo de informação veiculada pelo SBRT é a chamada informação tecnológica, grosso modo, considerada como a aplicação do conhecimento científico gerado em universidades e instituições de pesquisa, ou até mesmo em grandes empresas que investem em investigações científicas para consumo próprio.

Januzzi e Montalli (1999) abordam questões terminológicas na área de informação referente à indústria/empresa, que inclui a informação tecnológica e para negócios, chamando atenção para a importância da qualidade no uso da informação como insumo para a competitividade brasileira e na consolidação das redes de informação, fundamentadas na especialização: "Ciência, tecnologia e negócios são as palavras de ordem no mundo atual, que formam o tripé da competitividade global.".

Adota-se nesta pesquisa a definição de informação tecnológica publicada no Glossário de Informação Tecnológica, publicado pelo Senai, em 2001:

> Informação tecnológica é aquela relacionada com o modo de fazer um produto ou prestar um serviço para colocá-lo no mercado, servindo para difundir tecnologia de domínio público para possibilitar a melhoria da qualidade e da produtividade de empreendimentos existentes e construir insumo para o desenvolvimento de pesquisa tecnológica. (RODRIGUES, ABE, DIB, 2001)

4.4.2 A Descoberta de Conhecimento em Texto (DCT)

Muito se tem investido na organização e disponibilização de conteúdos na *Web* com a finalidade de ofertar o conhecimento produzido ao maior número possível de usuários interessados em utilizá-lo, no menor espaço de tempo desde a sua geração. Fatores como a rápida evolução das tecnologias de informação e as mudanças de paradigma no que diz respeito ao compartilhamento de informações permitiram que se somassem aos tradicionais bancos de dados referenciais um número cada vez maior de conteúdos textuais, com acesso direto à fonte primária do conhecimento, ou seja, ao próprio documento produzido. À medida que se avolumam esses conteúdos, mais se desenvolvem técnicas de tratamento automático da informação para possibilitar o manuseio dessas grandes massas de dados e seu melhor aproveitamento, seja para geração de novos conhecimentos, seja para extração de informações estratégicas para gestão.

Para melhor compreensão dos processos que envolvem a DCT, vale a pena abordar alguns conceitos básicos apresentados por Walter Trybula (1999) em sua revisão de literatura sobre a mineração de textos. Ele define base de dados como uma "coleção organizada de dados armazenados", os quais normalmente se referem a dados ativos e não a compilações de dados históricos e podem compor sistemas armazenados em locais distintos. As bases textuais, por sua vez, são repositórios de informações textuais, compilações históricas de informações eletrônicas que podem não ter sido previamente organizadas. (TRYBULA, 1999).

Essa definição pode ser complementada pela observação de repositórios atuais de informação textual, onde se observam grandes volumes de informação não estruturada inserida em sistemas de informação previamente organizada e estruturada. Exemplo disso é o sistema de informação SBRT, que possui uma estrutura de campos padronizados de informações associadas a textos elaborados na forma de documentos técnicos.

São muitas as discussões em torno das definições das técnicas de extração automática de conhecimento em grandes volumes de informação, onde conceitos e termos se misturam, por vezes sendo utilizados como sinônimos ou com pequenas nuances na diferenciação. Araújo Júnior (2005) enumera alguns termos, como prospecção, descoberta de conhecimento em banco de dados, mineração de dados, descoberta de conhecimento em textos, mineração de textos, que correspondem às siglas em inglês: *KDD* (*Knowledge Discovery in Databases*) e *KDT* (*Knowledge Discovery in Texts*). O autor afirma que os termos "mineração de textos", "descoberta de conhecimento em textos" e "mineração de dados" têm sido utilizados como sinônimos na literatura.

Quoniam (2005) define *Data Mining* como "todas as técnicas que permitem extrair conhecimento de uma massa de dados que, de outra maneira permaneceria escondido nas grandes bases". Afirma que sua aplicação "torna possível comprovar o pressuposto da transformação de dados em informação e posteriormente em conhecimento" e, por esta razão, ela se configura em uma técnica imprescindível para o processo de tomada de decisão.

Trybula (1999) aborda a terminologia da área como ainda incipiente, dada a juventude do tema, e apresenta sua definição em conjunto com a de "mineração de dados" e de "descoberta de conhecimento". Para ele, a mineração de textos (sigla em inglês *DM* para o termo *data mining)* é o processo básico empregado para analisar padrões em dados e extrair informações – processo esse que inclui a "limpeza" e a validação dos dados – e tem por objetivo gerar, além de verificar, uma hipótese sobre o dado selecionado. Costuma ser utilizada em gran-

des bases de dados organizacionais contendo informações sobre clientes para obtenção de informações sobre seu comportamento diante de incentivos do mercado.

Já a mineração de textos, na visão do autor, é o processo básico para analisar padrões em textos e apresentar informações, processo esse que inclui também a "limpeza" e validação dos dados com maior grau de dificuldade, dada a necessidade de isolamento de raízes de palavras e identificação de suas categorias gramaticais (substantivo, verbo, advérbio etc.). O autor define, ainda, a descoberta do conhecimento (sigla em inglês *KD* para o termo *knowledge discovery*), como o "processo de transformação de dados em relações previamente desconhecidas e insuspeitas, que podem ser empregadas como previsores de futuras ações". (TRYBULA, 1999)

Marti Hearst (2003) comenta o uso errôneo do termo *data mining* e a escassez de pesquisas sobre *text mining*. Considera a mineração de textos como uma ferramenta de suporte ao e valorização do conhecimento gerado, já que proporciona sua exploração e reutilização, e a define como

> Descoberta, por computador, de novas informações, previamente desconhecidas, pela extração automática de informações de diferentes recursos escritos, onde o elemento chave é a interligação das informações extraídas para formar novos fatos e novas hipóteses a serem posteriormente exploradas pelos meios de experimentação mais convencionais.

O autor afirma que a mineração de textos, ao contrário do que muitas pessoas pensam, não é uma forma de facilitar a busca de informação na *Web* e não deve, portanto, ser confundida com o processo de Recuperação da Informação (RI). Na RI, ocorre a descoberta de informação já conhecida, inserida em um documento pelo autor, uma forma de o usuário selecionar, em uma coleção de documentos, aqueles que lhe interessam e desprezar os demais. Enquanto que a mineração de textos tem como meta descobrir informação desconhecida, que não está escrita em nenhum documento individualmente, ou derivar novas informações a partir dos dados analisados, encontrar padrões em conjuntos de dados e/ou separar o signo do ruído. (HEARST, 2003)

Da mesma forma, Marty Lucas (2007) afirma que mineração não é RI; na mineração de dados textuais (*Text Data Mining – TDM*), os relacionamentos entre os documentos podem gerar novos fatos, não previamente conhecidos.

Esse diferencial em relação à RI é o que a torna especialmente adequada ao manuseio do crescente volume de informação contido em repositórios diversos de documentos textuais para extração de informações aparentemente inexistentes e a identificação de padrões e o relacionamento de conceitos nessas coleções.

De acordo com Hearst (2003), a mineração de textos, por facilitar a transferência de informação em conhecimento, propicia não somente o manuseio, mas a possibilidade de se manter atualizado no controle a vasta quantidade de informações relevantes para a organização.

Marcelo Scheissl (2007) realizou minucioso estudo sobre a literatura acerca da descoberta de conhecimento em dados e os diversos processos que a compõem, onde registra as particularidades apontadas por diversos autores. Segundo ele, o "processo de descoberta de conhecimento em dados compreende a seleção de dados, o pré-processamento que envolve sua adequação aos algoritmos, a efetiva mineração de dados, isto é, o uso de técnicas de mineração, a validação dos resultados e, finalmente, a análise e interpretação dos resultados para a aquisição do conhecimento." (SCHIESSL, 2007)

"A DCD vem sendo consolidada como um poderoso ferramental para auxiliar o homem na exploração da grande quantidade de informação disponível em formato eletrônico, dadas as limitações humanas no manuseio e interpretação dessa informação." (SCHIESSL, 2007)

Rogério Araújo Júnior (2005) afirma que a mineração de textos só terá sentido se for aplicada a uma situação concreta e sugere que seja utilizada, por exemplo, para enriquecer os instrumentos de apoio ao processo de indexação em um sistema de RI.

A partir da literatura analisada, observa-se a diversidade de termos utilizados para se definirem as várias técnicas utilizadas na extração de informações relevantes de grandes massas de dados. Esta pesquisa adota a DCT e explora as suas potencialidades para apoiar a geração de indicadores, a partir do tratamento automático e da análise de conteúdos armazenados em forma textual.

5. Procedimentos metodológicos

A pesquisa encontra-se em andamento, em fase de construção da base de dados de trabalho. A seguir, encontra-se a descrição das etapas concluídas até o presente.

5.1 Seleção da ferramenta e preparação do ambiente de trabalho

Há várias ferramentas disponíveis no mercado para a realização de estudos por meio da descoberta de conhecimento em conteúdos armazenados em formato textual. Optou-se, preliminarmente, pelo uso de um indexador textual de documentos, o BR/Search, para o tratamento dos dados e sua preparação para a pesquisa.

Diante da apresentação da proposta de pesquisa acadêmica, a empresa representante dessa ferramenta no Brasil, a Policentro, cedeu uma versão genérica do *software*, chamada BRS/Fácil.

Foi providenciado, no Ibict, um equipamento servidor (Athlon XP 1800Mhz, 512 MB RAM HD 40 GB), com sistema operacional Linux, onde foi instalado o BRS/Fácil, para a manipulação e preparação dos dados para a pesquisa.

5.2 Os dados para a pesquisa

Com autorização do Comitê Gestor da Rede SBRT, os dados para a pesquisa foram extraídos do sistema de informação SBRT, no dia 8 de agosto de 2007, e compõe-se de 6.014 registros.

Como o sistema SBRT é composto por um banco de dados estruturado em metadados que identificam o usuário e sua demanda, a classifica e a relaciona com a solução apresentada, a chamada Resposta Técnica, decidiu-se pela extração de alguns campos da base de dados que possam ser importantes também para as análises: i) Cliente Pessoa Física (gênero, cidade, estado (UF), escolaridade); ii) Cliente Pessoa Jurídica (nome da empresa, ou razão social, cidade, estado, natureza do vínculo do contato).

As demais informações fazem parte do corpo do texto da RT: título da RT, resumo, data de publicação, palavras-chave, assunto, demanda (a pergunta feita pelo cliente) e instituição respondente.

Como o BR/Search opera com unidades de parágrafos, decidiu-se pela adoção de um formato de extração de dados no qual cada documento a ser incluído na base de trabalho fosse composto de três parágrafos: i) campo de identificação do documento; ii) parágrafo contendo os metadados; iii) parágrafo de texto integral.

Originalmente, os dados extraídos do SBRT estavam parte em uma base de dados MySQL (metadados) e parte em arquivos PDF (textos completos) armazenados no servidor. Os procedimentos para sua extração foram realizados por meio de um *backup* da base MySQL, com o apoio da ferramenta phpMyAdmin, e posterior restauração desses dados em um PC contendo um ambiente de programação.

Em seguida, foi desenvolvida uma página PHP para “ler” os registros de Respostas Técnicas válidas da base MySQL e gerar um *Shell Script*, responsável pela conversão dos arquivos em PDF para o formato TXT, que é universal e se adapta perfeitamente ao exigido pe-

las ferramentas de mineração de dados. Finalmente, os dados foram mesclados em um único arquivo (BANCO_SBRT.txt).

Essas operações foram realizadas em servidor com sistema operacional Linux Ubuntu 7.04; servidor Web Apache versão 2.0, linguagem de Script PHP versão 5 e Shell Script, Servidor de Banco de Dados MySQL versão 5.0 e o conversor PDFTOTEXT.

Os dados em formato TXT foram gerados na estrutura de três "parágrafos", conforme definido anteriormente, sendo que o primeiro parágrafo contém o identificador do documento, o segundo os metadados relativos ao documento e o terceiro o texto do documento, conforme ilustrado a seguir:

..DOCN (Identificador do registro)

..METADATA (conjunto de metadados relativos às Respostas Técnicas)

..TXT (conteúdo da resposta técnica).

Procedeu-se, então, à criação da base de trabalho. O texto fornecido foi carregado no BR/Search, onde foi criada uma estrutura de parágrafos, e, iniciou-se a limpeza para eliminação dos caracteres de controle que foram inseridos automaticamente quando de sua importação para o formato TXT. Procedeu-se à eliminação das palavras não significativas do texto, como primeira ação de preparação dos dados para a mineração. Após a limpeza da base, serão necessários mais alguns procedimentos antes do início da mineração do texto.

5.3 A base de trabalho - próximos passos

O próximo passo será a geração de uma tabela contendo uma lista de todos os termos, com a contagem de sua freqüência e ocorrência na base de dados; em seguida, será realizada uma primeira análise e filtragem, para eliminação de termos pouco significativos remanescentes. Ao concluir essa tarefa, a base estará pronta para a aplicação da técnica de descoberta de conhecimento, por meio da extração de termos, geração de *clusters*, criação de tabelas, e realização das análises correspondentes.

Espera-se conseguir identificar, no relacionamento entre conteúdos, informações importantes que estejam ocultas nos textos dos documentos, conforme prevê a literatura, e que forneçam subsídios para a geração de indicadores sobre as necessidades de informação tecnológica do microempresário brasileiro usuário do sistema SBRT.

6. Considerações finais

A presente pesquisa encontra-se em andamento no âmbito do curso de Mestrado do Departamento de Ciência da Informação da Universidade de Brasília, com previsão de conclusão em dezembro de 2007.

Esta investigação é inédita, pois faz uso da DCT – assunto ainda muito pouco explorado no Brasil –, em conteúdos de informação tecnológica, contribuindo, assim, para a integração da investigação científica voltada para o setor produtivo.

A motivação para a realização do estudo se deveu à identificação da importância dessa ação inovadora que é o SBRT, que comporta um rico *corpus* de conteúdo sobre a necessidade de informação tecnológica dos micro e pequenos empresários brasileiros, e sua potencialidade para contribuir para a geração de indicadores de CT&I, visando nortear as ações voltadas para o setor de micro e pequenos empreendimentos.

Referências

ARAÚJO, Nelma Camêlo. **Análise do uso efetivo da informação por empresários de microempresas alimentícias do Estado de Minas Gerais**. Dissertação (Mestrado em Ciência da Informação), Universidade Federal de Santa Catarina, Florianópolis, 2005. 120f.

ARAUJO, Vania M. R. Hermes de, FREIRE, Isa Maria e MENDES, Teresa Cristina M. Demanda de informação pelo setor industrial: dois estudos no intervalo de 25 anos. **Ciência da Informação**, Brasília, v.26, n.3, p. 283-289, set/dez. 1997. (Cap Inf. Tec.)

ARAUJO JUNIOR, Rogerio Henrique de. **Precisão no processo de busca e recuperação da informação**. Orientador: Tarapanoff, Kira Maria Antônia. 2005. xiv, 223 p. Tese(Doutorado em Ciência da Informação)-Universidade de Brasília. Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação. Departamento de Ciência da Informação e Documentação. Inclui bibliografia e anexos.

BRASIL. Ministério da Ciência e Tecnologia. **Relatório de Gestão** Janeiro de 2003 a Dezembro de 2006. Brasília, MCT, 140 p.

CHAVES, H., CORDEIRO, F. e BASTOS, M. Avaliação do uso do Serviço Brasileiro de Respostas Técnicas: um serviço de informação destinada à microempresa brasileira. **Ciência da Informação**, Brasília, v. 35, n. 3, p. 255-269, set./dez. 2006.

CUBILLO, Julio. La inteligencia empresarial en las pequeñas y medianas empresas competitivas de América Latina : algunas reflexiones. **Ciência da Informação**, Brasília, v.26, n.3, p. 260-267, set/dez 1997.

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, acessado em 15 novembro de 2005.

HEARST, Marti. **What is Text Mining?** Disponível em: <http://www.ischool.berkeley.edu/~hearst/text-mining.html>. Acesso em: 31 jul. 2007.

JANUZZI, Celeste Aída Sirotheau Corrêa e MONTALLI, Katia Maria Lemos. **Informação tecnológica e para negócios no Brasil: introdução a uma discussão conceitual**. Ciência da Informação, IBICT: Brasília, v. 28, n. 1, p. 28-36 1999.

LASTRES, H. M. M. . Dilemas da política de desenvolvimento cientifico e tecnológico. **Ciência da Informação**, v. 24, n. 2, p. 189-193, 1995.

LUCAS, Marty. Mining in textual mountains. **Mapa Mundi Magazine**, disponível em http://mappa.mundi.net/trip-m/hearst/, acessado em 31 jul. 2007.

MCT – Ministério da Ciência e Tecnologia – Brasil. Disponível em: <http://www.mct.gov.br/index.php/content/view/105.html>. Acesso em: 12 ago. 2007.

MCT – Ministério da Ciência e Tecnologia - Brasil. Indicadores nacionais de ciência e tecnologia – 2002. Brasília: MCT, 2004, 140 p. ISSN 1413-3148. Disponível em: <http://www.mct.gov.br/index.php/content/view/2042.html>. Acesso em: 13 ago. 2007.

QUONIAM, Luc et al. Intelligence obtained with the application of data mining analysing the French DocThéses on subjects about Brazil. **Ciência da Informação**, Brasília, v. 30, n. 2, 2001.Disponível em:<http://www.ibict.br/cionline/viewarticle.php?id=216&layout=abstract>. Acesso em: 26 Mar 2007.

ROMÃO, W. ; PACHECO, R. C. S. ; NIEDERAUER, C. A. P. . Planejamento em C&T: uma abordagem para descoberta de conhecimento relevante em banco de dados de grupos de pesquisa. **Revista Tecnológica**, Maringá, v. 9, p. 139-152, 2000.

RODRIGUES, Margarete de Luna; ABE, Naguiça; DIB, Simone Faury (Org.) **Glossário de informação tecnológica – GLIT**. Brasília, SENAI/DN, 2001. 51 p.

SBRT. Plano de negócios do Serviço Brasileiro de Respostas Técnicas. Rio de Janeiro, 2005. 31 p.

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.

SEBRAE. **Serviço Brasileiro de Apoio às Micro e Pequenas Empresas**. *Site*: http://www.sebrae.com.br.

SEBRAE. **Fatores condicionantes e taxa de mortalidade de empresas no Brasil** : relatório de pesquisa. Brasília, Agosto/2004. 56 p.

SOARES, Bruno Jorge. **Comportamento de gestores de empresas de base tecnológica na busca e uso de informações.** Orientador: Costa, Sely Maria de Souza. Brasília, 2003, 163 p. Dissertação(Mestrado em Ciência da Informação)-Uiversidade de Brasília. Departamento de Ciência da Informação e Documentação.

TAKAHASHI, Tadao. Inclusão social e TICs. **Inclusão Social**, v. 1, n. 1, p. 56-59, out./mar., 2005.

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.