METODOLOGIA PARA DESCOBERTA DE CONHECIMENTO EM TEXTOS (TEXT MINING)

EDILBERTO M. SILVA[1, 2], , EDILSON FERNEDA[1], HÉRCULES. A. DO PRADO [1, 3]

[1] *Universidade Católica de Brasília (UCB/DF) Campus Universitário II – Pró-Reitoria de Pós-Graduação e Pesquisa – Brasília/DF*

[2] *Uniplac - União Educacional do Planalto Central, Faciplac – Faculdade de Ciência e Tecnologia do Planalto Central – Campus II - Gama/DF*

[3] *Embrapa Agroindústria de Alimentos*

Av. das Américas, 29501 – Guaratiba - 23020-470 Rio de Janeiro, RJ

E-mails: edilms@yahoo.com, hercules@cpac.embrapa.br, eferneda@pos.ucb.br

Abstract— In this paper a methodological proposal for knowledge acquisition using Text Mining is presented, independent of the type of technique used for the creation of models. Studying new approaches for knowledge acquisition from data in textual format is justified since a great amount of information existing in the organizations come in this format. It is well-know that this huge amount of texts can hide useful and non-trivial knowledge that could subsidize a well founded decision take. Departing from well-know approaches for developing Text Mining applications, an unified methodology were elaborated that aggregates the most interesting characteristics from those approaches. The method CRISP-DM (*Cross Industry Standard Process for Data Mining*) served as inspiration to defining the pipeline for this proposal.

Keywords— CRISP-DM, Text Mining, KDT, Decision Support Systems

Resumo— Apresenta-se neste artigo uma proposta metodológica para aquisição de conhecimento usando mineração de textos, independente do tipo de técnica (e.g., simbólica, conexionista ou evolutiva) utilizada para a criação de modelos. O estudo de novas formas de se obter conhecimento a partir de informações em formato textual justifica-se pelo fato de que a grande maioria das informações existentes atualmente nas empresas encontra-se neste formato. É reconhecido que esta enorme quantidade de textos pode gerar conhecimento útil e não-trivial e, assim, subsidiar os gestores na tomada de decisão bem fundamentada. Partindo de várias abordagens já consagradas para o desenvolvimento de aplicações de Mineração de Textos, procurou-se elaborar uma metodologia unificada que agregasse as características mais interessantes daquelas. A proposta foi elaborada a partir do modelo proposto no método CRISP-DM (*Cross Industry Standard Process for Data Mining*).

Palavras-chave— CRISP-DM, Mineração de Texto, KDT, Sistemas de Suporte a Decisão

1 Introdução

O crescimento contínuo do volume de dados eletrônicos disponíveis (e-mails, imprensa, transações, websites, newsgroups, fóruns, listas de correspondência etc.) e úteis para tomar decisões cruciais de negócio tornou as técnicas de extração de conhecimento automáticas cada vez mais necessárias, com vistas a valorizar a gigantesca quantidade de dados armazenada nos sistemas de informação.

As metodologias para mineração de dados foram desenvolvidas inicialmente para dados estruturados. Técnicas específicas para Mineração de Textos têm sido desenvolvidas com objetivo de processar uma parte importante da informação disponível na forma de dados não-estruturados.

As técnicas para mineração de texto evoluíram ao longo do tempo e os avanços nesta área se refletem nos diversos softwares de *text mining* (mineração de texto) disponíveis, baseadas em modelos bem conhecidos (e.g. simbólicos, conexionistas ou evolutivos). Silva (2002b) mostra esta diversidade de técnicas para sumarização, recuperação por termos gramaticais, radicais, semântica, comparação, classificação, entre outras.

A metodologia proposta nesta pesquisa é o resultado da análise de várias abordagens de renomados especialistas da área de Descoberta de Conhecimento, aplicada e validada por meio de estudo de caso. O método proposto detalha o procedimento de descoberta de conhecimento em dados no formato textual e não-estruturados a partir das ferramentas metodológicas e computacionais citadas no tópico de abordagens para Descoberta de Conhecimento em Texto (DCT) fundamentada e testada em um problema real: obtenção de padrões úteis de notícias publicadas por uma agência de notícias. (Silva 2002, 2002a, 2002b; Prado 2004, 2005).

2 Abordagens para DCT

A metodologia proposta aqui é baseada em abordagens recomendadas por pesquisadores da área de DCT. A metodologia base é a CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) (Chapman, 2001), desenvolvida com intuito de promover a padronização de conceitos e técnicas na busca de informações específicas para tomada de decisões.

Segundo Loh (2000a), DCT inicia-se com a seleção dos textos, análise da coleção (toda ou partes), análise de grupos de textos (todo ou partes), comparação de subcoleções e validação de hipóteses.

Conforme Tan (1999), o processo de mineração de texto consiste em duas etapas: o refinamento do texto e o extrator de conhecimento. O refinamento consiste na transformação do texto de forma livre para uma forma intermediária. O extrator do conhe-

cimento faz o reconhecimento de padrões, baseado nessa forma intermediária.

Para Halliman (2001), o processo de mineração de texto inicia-se com a compreensão do domínio da empresa, por meio do entendimento dos competidores e das forças que estes exercem sobre suas atividades que poderão melhorar as táticas e estratégias desenvolvidas.

3 Metodologia Proposta para DCT

A proposta metodológica descrita aqui é neutra, independente de produtos e com a proposição de etapas padronizadas e não-rígidas para DCT. Visa transformar necessidades de negócio em objetivos factíveis de serem resolvidos através de transformações nos dados e técnicas e algoritmos adequados. É constituída de 11 etapas mostradas na figura 1.

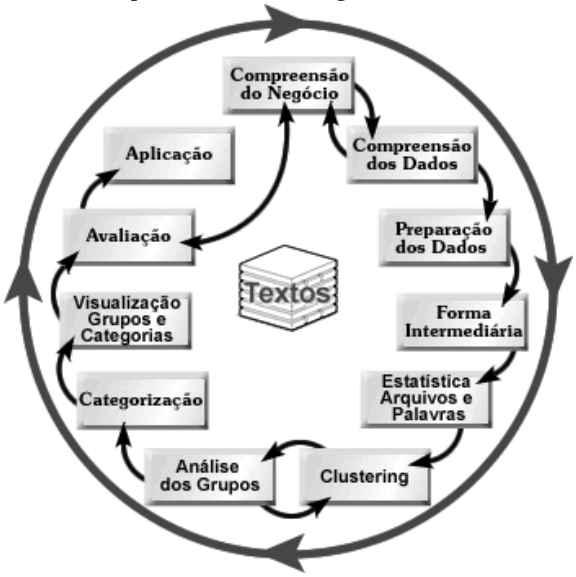

Figura 1. Metodologia Proposta para DCT

3.1 Compreensão do negócio

Ao se iniciar um projeto de mineração de dados deve estar bem claro para a organização e para os responsáveis pelo projeto de mineração de conhecimento. As técnicas a serem usadas para alcançar os resultados desejados (objetivos da mineração) devem estar alinhadas aos objetivos traçados pela organização. As tarefas a serem executadas nesta fase são:

(i) Identificar os Objetivos do Negócio – entender de forma unívoca a missão da empresa e relacionar, na visão do negócio, os anseios do cliente, determinando o que pontualmente a empresa deseja ter como ganho;

(ii) Avaliação da Situação – A partir do momento que se têm os objetivos definidos e a identificação inicial dos recursos, faz-se necessário realizar um levantamento apurado destes recursos. Isto envolve a relação minuciosa de recursos pessoais (gerentes, administradores, suporte técnico, especialistas, engenheiro do conhecimento...), de dados (fontes operacionais, documentação, formas de acesso, formato...) e computacionais (hardware existente, ferramentas de mineração de dados...);

(iii) Objetivos da Mineração de Dados – os objetivos da organização (ex.: "Melhorar relacionamento com meus clientes internautas.") devem ser traduzidos para problemas típicos de Data Mining tais como agrupamento, classificação, predição ou descrição (ex.: "Quais são as categorias de problemas mais reclamados por tipo de cliente?");

(iv) Plano do Projeto – nesta tarefa, é traçado o cronograma do projeto contendo os objetivos da mineração, das datas limites para execução das etapas do projeto bem como os recursos responsáveis por cada uma das etapas.

3.2 Compreensão dos dados

A fase de compreensão de dados visa avaliar o ambiente de mineração em relação aos insumos existentes na organização e/ou fora dela de modo a permitir que se encontrem respostas para os objetivos traçados. Nesta fase, uma amostra de dados é avaliada e são identificados a qualidade, relevância, formatos e conteúdos das fontes úteis de dados. São executadas as seguintes tarefas:

(i) Relatório inicial da coleção – relação de todos os dados necessários ao projeto de mineração de dados com ênfase aos detalhes de atributos e fontes, buscando classificá-los por importância e relevância para o projeto;

(ii) Descrição dos dados – etapa inicial de entendimento dos dados donde são extraídas as primeiras estatísticas da fonte de dados incluindo os formatos, quantidade de linhas, palavras, registros ou campos;

(iii) Exploração dos dados – envolve todos os passos para se obter conhecimento apurado dos dados tal como visualização e refino na descrição de dados. É detalhada a qualidade dos dados e descritas todas as transformações necessárias para análise dos dados e validação de hipóteses iniciais.

(iv) Qualidade dos dados – a partir do momento que se tem o levantamento dos dados, faz-se necessário a elaboração do relatório de qualidade dos dados. Este relatório descreve em detalhes aspectos relevantes quanto a sua completude e grau de confiabilidade.

3.3 Preparação dos dados

A Preparação dos Dados consiste numa série de atividades destinadas a obter o conjunto final de dados, a partir do qual será criado e validado o modelo. Nessa fase, são utilizados programas de extração, limpeza e transformação dos dados. Compreende a junção de tabelas e a agregação de valores, modificando seu formato, sem mudar seu significado a fim de que reflitam as necessidades dos algoritmos de aprendizagem. São executadas as seguintes tarefas:

(i) Seleção dos Dados – não é necessária a utilização de todos os dados encontrados na organização para o sucesso do projeto. Aliado a isto, há possibilidade de se ter diversas fontes de dados. Por isso, é cogente descrever os critérios de seleção/exclusão dos arquivos, textos ou dados a serem utilizados na mineração de dados;

(ii) Limpeza dos Dados – Esta tarefa aprimora a seleção dos dados elevando a qualidade dos textos a serem minerados pela retirada de textos ou partes não relevantes para técnicas a serem utilizadas. Isto inclui a execução de rotinas precisas para controle dos textos a fim de identificar arquivos corrompidos, nulos, inválidos ou fora do escopo do projeto (textos em outro idioma, padrões de acentuação inválidos);

(iii) Engenharia nos dados – os textos, após serem selecionados e limpos, devem ser preparados de fato para a mineração dados. Isto inclui a sua formatação/construção em formato compatível com a(s) ferramenta(s) a serem usadas como, por exemplo, separar textos que se encontre em um só arquivo para arquivos independentes ou mudança nos nomes dos arquivos textos.

(iv) Integração de dados – dependendo da disposição que os textos se encontram, é necessário combiná-los para se obter êxito com a(s) técnica(s) escolhida(s). Esta fusão refere-se à combinação de dois ou mais textos que tenham diferentes informações sobre o mesmo objeto como, por exemplo, e-mails, suas respostas e encaminhamentos.

(v) Formatação dos dados – A última etapa na preparação dos dados é a formatação no modelo final exigido pela(s) ferramenta(s) a ser utilizada. Algumas ferramentas exigem que a primeira linha dos textos contenha metadados (informações sobre o texto: tamanho, nome, etc...), outras exigem que todos os textos tenham a mesma extensão (.txt, .rtf, .csv...) ou ainda que tenha o mesmo padrão de codificação (ASC II, UTF-8...). Há ainda algoritmos que exigem a classificação dos textos por tamanho, ordem alfabética ou separados em diretórios.

3.4 Forma Intermediária

Após o pré-processamento, os textos encontram-se aptos a serem submetidos para criação e validação dos padrões obtidos na mineração de dados de fato. A forma intermediária inicia a fase de seleção e aplicação de técnicas para descoberta de conhecimento aliadas aos objetivos do projeto. Aqui, se seleciona ferramentas e algoritmos a serem utilizados e o efetivo processamento do modelo.

3.5 Elaboração de Estatísticas

Corresponde à transformação dos dados em informação e conhecimento. Nesta fase, podem ser utilizadas diversas ferramentas computacionais e métodos de análise para elaboração de estatística de quantidade de textos, palavras e tamanho de arquivos. Pode ainda contemplar o levantamento da limpeza dos dados, de arquivos corrompidos, nulos, inválidos ou fora do escopo do projeto, palavras com maior o menor freqüência de uso.

3.6 Stopwords

A técnica de *stopwords* (agrupamento) permite associar objetos similares em categorias distintas, onde os objetos desse grupo detêm características comuns. No caso de DCT, separam-se os textos tomando por base a verificação das palavras existentes nestes textos, onde em cada conjunto formado (grupo) determinadas palavras são compartilhadas.

Desse modo, é possível separar logicamente uma grande coleção de documentos isolando aqueles pertencentes a um mesmo assunto, facilitando a identificação de documentos relevantes para o usuário. Aplicando-se técnicas adicionais, é possível destacar o assunto ou conhecimento específico de cada grupo, facilitando o processo de recuperação de informações ou descoberta de conhecimento.

O agrupamento de texto é a tarefa de maior esforço computacional nesta abordagem de DCT. O processamento completo compreende uma série de subtarefas, onde algumas exigem intervenção do usuário, tais como: ajuste da lista de *stopwords*, legitimação dos centróides encontrados, definição do grau de similaridade e escolha do(s) algoritmo(s) e parâmetros a serem usados.

Vários softwares podem ser usados para obtenção de grupos de objetos textuais. No estudo de caso que valida a metodologia aqui apresentada, foi utilizada a ferramenta Eurekha®, desenvolvida por Wives (1999) e baseada nas análises e comparações de algoritmos de *stopwords*, permitido obtenção de conhecimento (padrões, relacionamentos) a partir de textos de forma interativa.

3.7 Análise dos Grupos

Com a análise dos agrupamentos, pode-se criar e explicar hipóteses. Baseado no conhecimento já apropriado pelo especialista, adquirido com a experiência nos dados da empresa ou ainda pelas informações fornecidas pela descrição e sumarização dos dados, obtêm-se relações relevantes ao negócio da empresa. Com esse conhecimento, o especialista, pela observação, formula pressupostos, cria critérios de classificação, estabelecendo hierarquias e assim detecta conceitos e categorias presentes nos grupos.

3.8 Categorização

Depois de formados os grupos e respectivos centróides, com vistas a extrair os conhecimentos a partir das informações encontradas, é realizada a categorização. Esta análise deve ser realizada com apoio de

um especialista da área, que usa, além das informações dos grupos, seu conhecimento do negócio.

Utilizando-se o percentual de distribuição de textos nos grupos, o especialista categoriza por assunto quantificando as categorias com base na distribuição (%) dos centróides. Com isto, é possível relacionar os principais assuntos tratados pelos grupos analisando as respectivas palavras chave que os compõem. Por exemplo: poder-se-ia assumir que o grupo que contenha como centróide as palavras: "ministro", "presidente", "desenvolvimento" e "governo" refere-se à Presidência.

A categorização assemelha-se a construção de um *thesaurus* utilizado para determinar as relações significativas entre termos próximos, em associações diretas, indiretas (transitivas só em primeiro grau) e categorias comuns de termos. Essa abordagem vai fazer com que o conteúdo seja organizado de acordo como a informação em si deve ser classificada.

3.9 Visualização de Grupos e Categorias

Sugere-se, para uma análise mais acurada das categorias, a elaboração de diversos gráficos com objetivos de apurar melhor a interpretação realizada pelo especialista. A visualização dos grupos e categorias é muito importante para avaliação, pois aproxima o entendimento de todos os envolvidos no projeto, incluindo o cliente. Pode-se considerar a apresentação da evolução das palavras com maior destaque da coleção, a evolução das categorias por determinado período e ainda como cada uma das categorias se comporta em um instante determinado.

3.10 Avaliação

Na avaliação do modelo, é realizada uma revisão dos passos seguidos, verificando se os resultados obtidos vão ao encontro dos objetivos, previamente, determinados na Compreensão do Negócio. São executadas as seguintes tarefas:

(i) Avaliação dos resultados – estima se os resultados obtidos estão alinhados com os objetivos do negócio e com os objetivos iniciais da mineração de dados. Isto se dá a partir do completo entendimento dos padrões encontrados e sua correta interpretação em relação ao negócio. Os impactos futuros e iniciais dos resultados encontrados também devem ser avaliados, juntamente com mensuração banco de conhecimento adquirido;

(ii) Revisão dos processos – Verificado que o modelo resultante é satisfatório para suprir as necessidades da organização, deve-se realizar uma completa revisão nos passos efetuados. A documentação destes passos deve seguir a ordem adotada no projeto tendo como principal meta relatar fatos ou tarefas importantes que possam ser esquecidos. Este passo é muito importante para o exame de qualidade do produto, antes de se efetivamente terminar o projeto.

(iii) Próximos passos – de acordo com a avaliação dos resultados e com a revisão dos processos, deve-se decidir como proceder com o projeto. Neste estágio, é verificada a continuação da avaliação estratégica do projeto, novas interações ou ainda modificar a abrangência do projeto.

3.11 Aplicação

O conhecimento em suas várias definições pode ser considerado como a efetiva aplicação da informação. Assim, nesta fase, é realizado um conjunto de ações de forma que possa ser utilizado eficientemente pelo cliente. Nessa fase, gera-se um relatório final para explicar os resultados e as experiências, procurando utilizá-los no negócio. É dividida nas seguintes tarefas:

(i) Plano estratégico de aplicação – sumariza todos os passos necessários e como aplicá-los para permitir que a organização obtenha vantagem competitiva no uso do conhecimento descoberto. Neste plano, devem ser relacionadas todas as decisões para propagação do conhecimento na empresa entre seus colaboradores, incluindo estratégias de aplicação, softwares e benefícios esperados para cada área atingida;

(ii) Plano de acompanhamento e manutenção – a partir do momento que se aplica o conhecimento em prol do negócio, o acompanhamento e manutenção devem se tornar parte do dia-a-dia do projeto do ambiente organizacional. Esta manutenção deve ser permanente com vistas a evitar usos errôneos dos conhecimentos adquiridos, sempre levando em consideração o tipo de abordagem e propagação a ser usado. O plano contém ainda como os procedimentos podem mudar o ambiente e como os padrões encontrados resolverão os problemas;

(iii) Relatório final – nesta etapa, os conhecimentos adquiridos são relatados baseando-se em como resultados foram conduzidos à organização. Gera-se um relatório final para explicar os resultados e as experiências, procurando utilizá-los no negócio. Na descrição dos resultados obtidos, é importante ressaltar os processos realizados incluindo custos, dificuldades e desvios que ora tenham ocorridos no plano original. No relatório, deve constar ainda o resumo de custo X benefício do projeto e recomendações para trabalhos futuros.

(iv) Revisão do projeto – na revisão geral do projeto, avalia-se o que foi feito de certo e errado, bem como o que necessita ser melhorado. A revisão do projeto completa o relatório final com um resumo dos fatos mais relevantes acontecido durante o projeto.

4 Conclusão

Considerando que a tomada de decisão é um processo de investigação, reflexão e análise, justifica-se a necessidade da alta administração obter informação qualitativa que contenha alto valor agregado. A apropriação do conhecimento tem sido também fator imprescindível para a sobrevivência das empresas.

Neste sentido, observa-se que grande parte das informações dentro e fora das organizações encontra-se em forma não-estruturada (texto, por exemplo). No entanto, poucas aplicações são dirigidas a tal tipo de informação. Considera-se, assim, relevante a proposta de uma nova abordagem para adquirir conhecimento neste tipo de dados, onde há carência metodológica e baixo nível de aplicação em mineração de texto.

A metodologia proposta tem base na CRISP-DM, metodologia mundialmente consagrada, e é comprovada por meio de estudo de caso onde foi possível reverter conhecimentos adquiridos com o DCT em benefício do melhoramento da eficiência organizacional em uma empresa pública brasileira de jornalismo.

É verossímil adquirir conhecimentos úteis à organização a partir da análise de textos, onde as técnicas envolvidas permitem que a empresa alcance vantagens competitivas e com isto os gestores tem um importante ferramenta na ajuda da gestão organizacional.

Referências Bibliográficas

Chapman, P., Kerber R., Clinton J., Khabaza T., Reinartz t., Wirth R., 2000. The CRISP-DM Process Model. http://www.crisp–dm.org. (07/07/07).

Halliman, C., 2001. "*Business intelligence using smart techniques: environmental scanning using text mining and competitor analysis using scenarios and manual simulation. Information Uncover"*, Houston.

Loh, S.; Wives, L.K.; Palazzo, J.M.O. Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva. In *Proceedings of the*: *International Symposium on Knowledge Management / Document Management* (ISKM/DM), III. Curitiba, PR, 2000a. p.125-147.

Prado, H. A., Oliveira, J. P. M., Ferneda, E., Wives, L. K., Silva, E. M., Loh, S. "*Transforming Textual Patterns in Knowledge*", In: Mahesh S. Raisinghani (Org.) Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks, pp. 207-227. Idea Group Publishing, Hershey(EUA), 2004.

Prado, H. A., Oliveira, J. P. M., Ferneda, E., Wives, L. K., Silva, E. M., Loh, S. "*Text Mining in the Context of Business Intelligence*", In: Mehdi Khosrow-Pour. (Org.), Encyclopedia of Information Science and Technology, Vol. V, 2793-2798. Idea Group Reference, Hershey (EUA), 2005.

Silva, E. M.; Prado, H.P.; Ferneda, Edilson – "*Text mining: crossing the chasm between the academy and the industry"*. In: DM2002 - Third International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields - Data Mining 2002. Bologna, WIT Press, 2002, p.351-361.

Silva, E. M.; Prado, H.P.; Ferneda, E. "*Text Mining for Organizational Intelligence: A Case Study*" On A Public News Agency. In: 5th International Conference on Enterprise Information Systems - ICEIS, Angers, França, 2002a.

Silva, E. M, Descoberta de conhecimento com o uso de text mining: cruzando o abismo de Moore, Dissertação de mestrado, UCB, Brasília, DF, 2002b.

Tan, A.-H. "*Text mining: The state of the art and the challenges"*, Kent Ridge Digital Labs, 1999. http://textmining.krdl.org.sg. (08/23/01).

Wives, L. K., 1999. Um Estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Stopwords. Disertação de Mestrado, PPGC/UFRGS, Porto Alegre (Brasil).