Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor

Knowledge Discovery in Text applied to a Help Desk System

MSc. Marcelo Schiessl[1]

Dr.ª Marisa Bräscher[2]

Resumo

Analisa um Serviço de Atendimento ao Consumidor de instituição financeira que centraliza, em forma textual, os questionamentos, as reclamações, os elogios e as sugestões, verbais ou escritas, de clientes. Discute a complexidade da informação armazenada em linguagem natural e oferece alternativa para extração de conhecimento de bases textuais com a criação de agrupamentos e modelo de classificação automática de textos para agilizar a tarefa realizada atualmente por pessoas. Apresenta uma revisão de literatura que mostra a Descoberta de Conhecimento em Texto como uma extensão da Descoberta de Conhecimento em Dados que utiliza técnicas do Processamento de Linguagem Natural para adequar o texto ao formato apropriado para a mineração de dados e destaca a importância do processo na Ciência da Informação. Aplica a Descoberta de Conhecimento em Texto na base do Serviço de Atendimento ao Consumidor com objetivo de criar automaticamente agrupamentos de documentos para posterior criação de um modelo categorizador automático dos novos documentos recebidos diariamente. Essas etapas são validadas por especialistas de domínio que atestam a qualidade dos agrupamentos e do modelo. Finalmente, geram-se indicadores de desempenho do grau de satisfação do cliente referente a produtos e serviços oferecidos que subsidiam a gestão na política de atendimento.

Palavras-chave: descoberta de conhecimento em texto; mineração de textos; mineração de dados; descoberta de conhecimento em dados.

Abstract

It analyses a Help Desk System of a financial institution that centralizes customer answers, complains, compliments, and suggestions, spoken or written. It argues about information complexity stored in natural language. It intends to present an alternative for knowledge extraction from textual databases by creating clusters and automatic classification model of texts in order to improve the current tasks made by employees. It presents a literature revision that shows the Knowledge Discovery in Text as an extension of Knowledge Discovery in Data that utilizes the Natural Language Processing in order to adequate the text into an appropriated format to data mining and enhances the importance of the process in the Information Science field. It applies the Knowledge Discovery in Text techniques in the Help Desk Database in order to create cluster of documents and, after that, to build an automatic classification model to new documents received every day. These steps need to be validated by specialist in the area to verify the model and clusters quality. It creates performance indexes in order to measure the customer satisfaction related to products and services to provide information for decision makers.

Keywords: discovery in text; text mining; help desk system; data mining; knowledge discovery in data.

Resumen

Analiza un Servicio de Atención al Cliente de las instituciones financieras que centraliza, de forma textual, las preguntas, quejas, elogios y sugerencias, verbales o por escrito, de los clientes. Describe la

[1] Doutorando em Ciência da Informação. Mestre em Ciência da Informação. Especialista em Inteligência Competitiva. Área de Representação e organização da informação e do conhecimento E-mail: marcelo.schiessl@gmail.com

[2] Doutora em Ciência da Informação. Professora da Faculdade de Ciência da Informação da Universidade de Brasília. E-mail: marisa.brascher@gmail.com.br

complejidad de la información almacenada en lenguaje natural y ofrece una alternativa para la extracción de conocimiento a partir de bases de datos textuales con la creación de grupos y el modelo de clasificación automática de textos para abreviar la tarea realizada actualmente por las personas. Presenta una revisión de la literatura que muestra el descubrimiento del conocimiento en texto como una extensión del descubrimiento de conocimiento en datos, que emplea técnicas de procesamiento del lenguaje natural para ajustar el texto al formato adecuado para la minería de datos y pone de relieve la importancia del proceso en Ciencias de la Información. Aplica el descubrimiento del conocimiento en el texto, en la base de datos del Servicio de Atención al Cliente con el fin de crear automáticamente grupos de documentos para la posterior creación de un modelo de separador automático de los nuevos documentos recibidos diariamente. Estos pasos son validados por expertos de dominio que comprueban la calidad de las agrupaciones y el modelo. Por último se generan indicadores de desempeño del grado de satisfacción del cliente en lo que respecta a los productos y servicios que se ofrecen para subvencionar la gestión en la política de asistencia.

Palabras clave: descubrimiento de conocimiento en el texto, la minería de textos, la minería de datos, descubrimiento de conocimiento en los datos.

1.Introdução

Este trabalho relata o resultado de pesquisa que aplica técnicas de mineração de texto numa base de Serviço de Atendimento ao Consumidor (SAC), a fim de demonstrar a utilidade da Descoberta de Conhecimento em Textos (DCT) para a criação de agrupamentos de textos a partir de coleção de documentos existentes.

Grande parte da informação eletrônica está disponível em texto, cujo formato está adequado ao homem que, através da leitura, é capaz de decodificá-lo e apreendê-lo. Entretanto, para a máquina, a tarefa não é trivial. A DCT propõe soluções para tratar a informação eletrônica textual com o auxílio de máquinas para amenizar o impacto da sobrecarga de informação.

Um SAC estabelece um canal de comunicação entre cliente e empresa que visa identificar as necessidades dele em relação aos produtos e serviços oferecidos. Uma base de dados de SAC é exemplo de fonte rica em informações textuais passadas pelos clientes sem que se faça pesquisa de mercado ou similares que, em geral, são muito caras. Geralmente, essas bases de dados contêm o ponto de vista do cliente transcrito em linguagem natural.

Pragmaticamente, discute-se aqui a DCT em uma base de SAC de uma instituição financeira demonstrando que, com a utilização de ferramentas e metodologias adequadas, é possível maximizar a descoberta e a utilização de informações úteis ainda não identificadas.

Discute-se, ainda, a relação entre Descoberta de Conhecimento em Dados (DCD) e DCT. Para tanto, são apresentadas as visões de pesquisadores sobre a DCD e o processo de execução. Aborda-se o Processamento de Linguagem Natural (PLN) e sua ligação com a DCD e a DCT, seus aspectos relevantes e sua fundamentação em outras áreas do conhecimento, além de um esboço de suas aplicações.

2. Fundamentos

2.1 Serviço de Atendimento

A satisfação do cliente nem sempre foi o foco das empresas que ofereciam seus produtos conforme as suas necessidades e padrões de negócios. Kotler (1972) introduziu a satisfação do cliente como elemento fundamental, na teoria do Marketing, para a sustentabilidade das organizações.

No Brasil, se comparado aos países mais desenvolvidos, vive-se um período relativamente curto de proteção garantida pelo Estado aos clientes das empresas que se estabeleceram por todo país. Assim, apenas em 1991, o Código de Defesa do Consumidor foi promulgado para atender ao desejo da sociedade de garantir o seu direito no ato de consumir e poder contestar aquilo que lhe foi vendido, caso não satisfaça as suas expectativas.

De acordo com Chauvel (2000), o consenso entre empresas – de dar voz ao cliente – levou à criação dos Serviços de Atendimento ao Consumidor como canal de comunicação que auxilia na correção de produtos, de serviços e da própria estratégia dentro do mercado. Pois, o simples fato de existência do SAC requer mudança de postura, mais abertura e predisposição ao diálogo.

Por fim, em tempos de massificação, os consumidores querem que seus nomes sejam conhecidos e que suas diferenças sejam observadas pelas empresas, isto é, que o atendimento seja sob medida. Uma porta aberta para manifestação do consumidor é o SAC, porém há que se extrair o máximo que ele pode dar: informação na medida do cliente.

2.2 Descoberta de Conhecimento em Dados

Com o advento da digitalização de documentos e o desenvolvimento das redes, o volume de informação aumenta além da capacidade humana de apreensão e, dessa forma, existe um lapso crescente entre a criação de dados e a sua compreensão (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

A DCD[3] apresenta-se como uma opção para atender a essa necessidade. Para Fayyad, Piatetsky-Shapiro e Smyth (1996), em geral, o campo de pesquisa da DCD preocupa-se com o desenvolvimento de métodos e técnicas que buscam trazer sentido aos dados. Seu processo básico 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 métodos tradicionais de transformação de dados em informação situam-se na análise manual e na

[3] Do inglês: Knowledge Discovery in Database (KDD), existe uma discussão entre diversos autores da área a respeito da abrangência do termo KDD e Data Mining (DM), porém os termos são referidos em vários trabalhos indistintamente.

interpretação, porém, em contraste com a farta disponibilidade de bases de dados, tornam-se lentos, caros e altamente subjetivos.

Frawley, Piatetsky-Shapiro e Matheus (1992) afirmam que a Descoberta de Conhecimento é a extração não trivial da informação implícita, nos dados, previamente desconhecida e potencialmente útil e Fayyad, Piatetsky-Shapiro e Smyth (1996) complementam que a DCD é o processo de descoberta não trivial, em dados, de padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis. Dessas afirmações, entende-se que dado é um conjunto de fatos e padrão é a estrutura implícita que será encontrada. O termo processo envolve a preparação dos dados, a busca por padrões, a avaliação do conhecimento descoberto e os refinamentos necessários em repetidas iterações. Pelo termo não-trivial depreende-se que a busca ou inferência não seja uma operação direta de quantidades pré-definidas, como por exemplo, o cálculo de uma média. Além disso, que os padrões descobertos sejam válidos em novos dados com algum grau de confiabilidade. Deseja-se, ainda, que a descoberta seja uma novidade que agregue alguma utilidade e benefício ao usuário e, por último, que seja compreensível, mesmo que necessite de pós-processamento.

Segundo Berry e Linoff (1997), a DCD é a análise e exploração automáticas ou semiautomáticas de grandes quantidades de dados com o objetivo de descobrir regras e padrões significativos.

As definições acima exprimem visões com nuances sobre o mesmo tema. Enquanto a segunda privilegia os aspectos computacionais tais como algoritmos dedicados à mineração de dados e poder de processamento para execução de tarefas de manipulação de grandes volumes de dados e sua transformação em informação capaz de ser utilizada pelo homem, a primeira trata do processo de descoberta como um todo. Isto é, desde a aquisição do dado, seu armazenamento, a mineração que retira a informação codificada até a sua apresentação ao usuário final. Visto dessa forma, a técnica pode ser assemelhada à automatização do ciclo informacional da Ciência da Informação (CI).

O processo da DCD apresenta-se como uma atividade multidisciplinar que se apropria de técnicas que vão além do escopo de uma área em especial (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Do ponto de vista histórico, de acordo com Kodratoff (1999), a DCD integra várias abordagens de aquisição de conhecimento de diversos campos da ciência como o aprendizado de máquina, que inclui os tipos de aprendizado simbólico, estatístico, neural, bayesiano, a tecnologia de bancos de dados, a visualização de dados, a recuperação de informação, a interação homem-máquina e a Ciência Cognitiva (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992) (HAND; MANNILA; SMYTH, 2001).

Por fim, o termo em inglês *Kowledge Discovery in Data* (KDD) foi cunhado por Frawley, Piatetsky-Shapiro e Matheus (1992) e a palavra conhecimento, nesse contexto, não é definida por sua visão na filosofia[4] e, sim, por considerar que os padrões descobertos possam adicionar alguma informação nova ao usuário.

2.3 Processamento de Linguagem Natural

Com a evolução tecnológica, houve uma massificação de textos de toda ordem, seja em correspondências eletrônicas, em publicações científicas ou em sítios na Internet, com diversos propósitos.

Segundo Hearst (1999), o texto expressa uma fonte de informação tão vasta, quanto rica, porém codificada de maneira que é difícil de ser decifrada automaticamente. Assim, a ciência vem buscando soluções para simular a cognição humana que é capaz de processar o texto e de apreendê-lo de maneira satisfatória.

De acordo com Manning e Schütze (1999), o estudo da Linguística vem contribuir para resolver esse problema, pois busca caracterizar e explicar a diversidade de observações linguísticas que nos cerca, seja em diálogos, seja na escrita, seja em qualquer outro meio. Uma parte preocupa-se com o lado cognitivo de como o homem adquire, produz e entende a linguagem, outra parte, a compreensão da relação entre discurso linguístico e o mundo e, a terceira, com a compreensão das estruturas linguísticas por meio das quais o homem se comunica.

Paralelamente, o desenvolvimento da informática possibilita grandes avanços no estudo das línguas naturais. A área que examina as relações entre a Linguística e a Informática é a Linguística Computacional que objetiva a construção de sistemas especialistas em reconhecer e produzir informação em linguagem natural. Encontram-se nesse contexto os estudos de Processamento de Linguagem Natural que têm por objetivo a interpretação e geração de informação nos diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso (VIEIRA; LIMA, 2001).

Não se propõe aqui uma discussão detalhada do PLN, seus métodos e suas técnicas e, sim, a contextualização da relação entre a DCD e a DCT. Assim, durante várias décadas, inúmeras pesquisas têm provocado avanços no PNL e, atualmente, encontram-se procedimentos disponíveis capazes de realizar o tratamento do dado textual de maneira a possibilitar sua transformação e sua estruturação na forma adequada ao uso pela DCD.

[4] Ao leitor interessado na visão filosófica, sugere-se Hessen, J. Teoria do conhecimento: tradução João Vergílio Gallerani Cuter, 2ª ed., São Paulo, Martins Fontes, 2003.

2.4 Descoberta de Conhecimento em Textos

Tan (1999) e Feldman et al. (2001) declaram que 80% da informação de empresas está em documentos textuais e, de acordo com Dörre, Gerstl e Seiffert (1999), a informação textual não está prontamente acessível para ser usada por computadores, ou seja, ela é apropriada para que pessoas, através da leitura e dos processos cognitivos característicos dos humanos, manipulem e apreendam as informações contidas nesse formato.

Para Trybula (1999), a DCT assemelha-se à DCD com exceção ao foco em coleções de documento textuais. Essa visão está de acordo com Feldman et al. (2001) que afirma que a DCT é a área dentro da DCD que se concentra na descoberta de conhecimento em fontes de dados textuais.

Vários autores afirmam que as bases textuais apresentam-se de forma não-estruturada. Porém, possuem uma estrutura implícita que necessita de técnicas especializadas para ser reconhecida por sistemas automatizados. O PLN trata exatamente da descoberta destas estruturas implícitas, como por exemplo, a estrutura sintática (RAJMAN; BESANÇON, 1997).

A integração de técnicas de PLN e DCD constitui a Descoberta de Conhecimento em Texto que objetiva automatizar o processo de transformação de dados textuais em informação para possibilitar a aquisição do conhecimento.

A figura 1 apresenta o ciclo do processo da DCT e pode ser visto como uma consolidação das concepções de autores como Wives e Loh (1999), Dörre, Gerstl e Seiffert (1999) e Tan (1999) que equivale a uma adaptação ao modelo proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996).

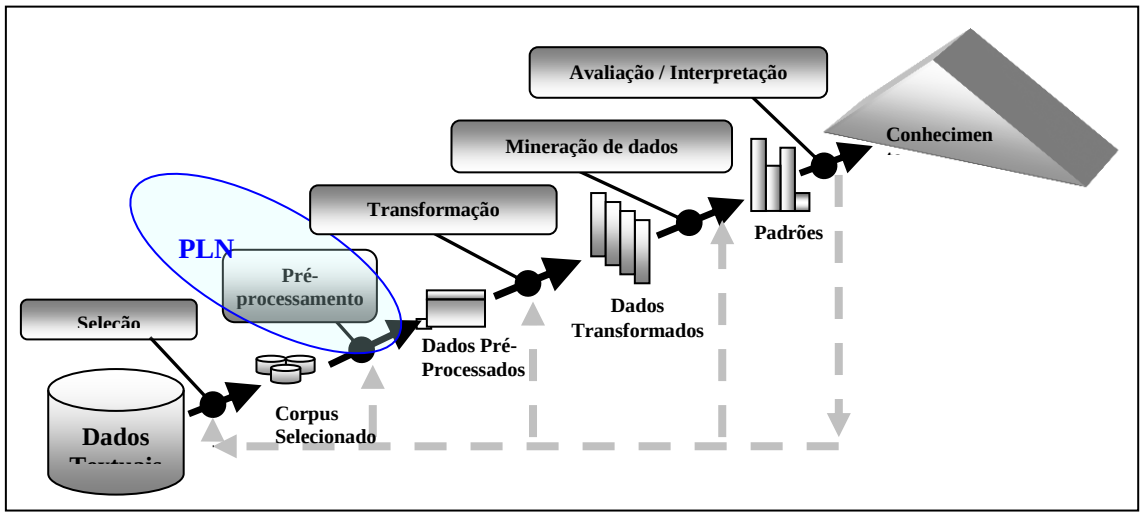

Figura 1: Processo de DCT.

De acordo com a figura 1, observa-se que o processo de DCT abrange a seleção do corpus, o pré-processamento, que envolve sua adequação aos algoritmos, a efetiva mineração de dados textuais, a validação dos resultados e, finalmente, a análise e interpretação dos resultados para a aquisição do conhecimento.

Vale a pena destacar que, embora a aderência entre o PLN e DCT seja muito forte, seus objetivos não se confundem, pois o PLN trata dos aspectos linguísticos relacionados a um texto específico, enquanto que a DCT busca as relações contidas entre os textos de uma coleção com o objetivo de apresentar a informação relacionada a um grupo ou grupos de textos. Ainda, o PLN analisa o conteúdo dos textos e, por outro lado, a DCT se utiliza dessas análises, em uma fase de pré-processamento, para transformá-las em dados apropriados para a descoberta de padrões e conhecimento entre os textos da coleção (KODRATOFF, 1999).

A DCT possibilita então o reconhecimento e a produção da informação apresentada em linguagem natural e, nesse sentido, vem contribuir enormemente com a Ciência da Informação no que tange ao tratamento e recuperação da informação.

3 Considerações sobre a DCT e a CI

A descoberta de conhecimento em textos é a conjunção de várias metodologias e conceitos, logo esse trabalho apresenta uma reprodução estática de seu desenvolvimento e implementações até esse momento. O desenvolvimento e aperfeiçoamento do processo são constantes, dada a natureza da língua e das ferramentas tecnológicas.

Konchady (2006) declara que a DCT é uma prática relativamente nova derivada da Recuperação da Informação (RI) e da PLN e essa afirmação estabelece um vínculo significativo, visto que a RI é uma das áreas principais de pesquisa da CI.

Segundo Bräscher (1999), os avanços tecnológicos influenciam a CI e favorecem o surgimento de novas técnicas de representação e recuperação de assunto considerando os aspectos cognitivos envolvidos no processo de comunicação homem-máquina que exigem modelos de representação do conhecimento capazes de contextualizar os significados expressos nos textos armazenados.

Lima (2003), em uma releitura de Saracevic (1995), expõe que a CI é uma área interdisciplinar que reúne a Biblioteconomia, a Ciência Cognitiva, a Ciência da Computação (CC) e a Comunicação, com forte associação dos processos humanos da comunicação e da tecnologia no seu contexto contemporâneo.

De fato, a CC trata de algoritmos relacionados à informação, enquanto a CI se dedica a compreensão da natureza da informação e de seu uso pelos humanos. A CI e a CC são áreas complementares que conduzem a aplicações diversas (SARACEVIC, 1995).

Robredo (2003) reafirma a interdisciplinaridade da CI orientando que não se pode restringir o escopo e a abrangência da informação ao campo exclusivo da biblioteconomia e da CI, pois variados estudiosos, pesquisadores e especialistas lidam com a informação de um ponto de vista cientifico e nas mais variadas abordagens e aplicações. Ainda, ensina que ela pode ser dividida, para fins de estudo e delimitação do(s) objeto(s), mas sem perder de vista o interesse comum de todos os seus domínios, a entidade informação.

Ainda, Lima (2003) aponta as possibilidades de interseção entre a CI e a CC que se concentram nos processos de categorização, indexação, recuperação da informação e interação homem-computador.

Nesse sentido, a DCT pode ser vista como a interposição da Estatística que utiliza métodos quantitativos para transformar dados em informação, da CC fornece suporte tecnológico para manipulação dela e da CI que concentra o foco de atuação na sua gestão.

Diante do exposto, a DCT encontra-se nesta área de intersecção da CI, da Estatística e da CC que utiliza métodos linguísticos para tratamento de textos. Essas áreas são mutuamente beneficiadas pelo aporte teórico de cada uma que favorece o desenvolvimento conceitual interdisciplinar. De tal modo, fica caracterizada a evidente contribuição do estudo da DCT no âmbito da CI.

4 Desenvolvimento da Pesquisa

O trabalho consiste na descoberta de conhecimento nos textos contidos na base do SAC de uma instituição bancária. Nesta base, o campo com a argumentação em linguagem natural fornecida por clientes é o texto a ser minerado.

A ferramenta utilizada para todo processo é o SAS Enterprise Miner que é capaz de fornecer uma interface amigável para a construção de modelos e, além disso, uma eficiente linguagem de programação para resolução de problemas específicos. O programa suporta a leitura de textos em vários formatos como, por exemplo, word, html, pdf, ASCII entre outros. O programa também possui ferramentas especializadas de modelagem [5] tais como Redes Neurais, Regressão Logística Árvore de Decisão e *Memory-based Reasoning* entre outras. Para validação desses modelos, existem procedimentos automatizados nessa ferramenta que quantificam os resultados encontrados de maneira a auxiliar o analista nas suas inferências.

A base explorada foi extraída de um sistema de ouvidoria que armazena todas as mensagens enviadas pelos clientes desde 2000. Essa base é atualizada diariamente, portanto, para a produção de material para esse estudo estabeleceu-se que o corte deveria ser até o

[5] Os aspectos teóricos na construção desses modelos e na aplicação destas técnicas no mundo dos negócios são discutidos em Berry e Linoff (1997).

mês de junho de 2006 inclusive. A organização dos campos analisados da base do SAC é apresentada na tabela 1.

Tabela 1: Descrição da Base

Grupo Assunto	Assunto a que se refere à mensagem;
Origem	O canal por onde se fez o contato (Internet, telefone,...);
Natureza	Se reclamação, sugestão ou elogio;
Produto	O produto do qual se está comentando;
Motivo	Pré-classificação do atendente;
Descrição da Ocorrência	Texto descrevendo a ocorrência;

Existem 499.102 registros. No último ano, 2006, o número de registros corresponde a 6 meses. Considerando o volume de dados e a relevância de informações atualizadas espera-se que as informações mais recentes retratem mais fidedignamente a realidade atual, portanto a base para o estudo se restringiu ao primeiro semestre de 2006, registros referentes à natureza Reclamação e originados via Telefone. O que caracterizou uma redução do universo para 52.646 mensagens textuais.

Dada a complexidade de processamento textual versus recursos tecnológicos, a amostragem é sempre uma boa alternativa. Dessa forma foi extraída amostra de 7895 registros. O critério escolhido foi a amostragem estratificada pelas variáveis “Código do Produto” e “Código do Motivo” com o objetivo de manter as amostras com proporções semelhantes à população.

Na figura 2, encontra-se ilustrado o processo global de desenvolvimento da pesquisa.

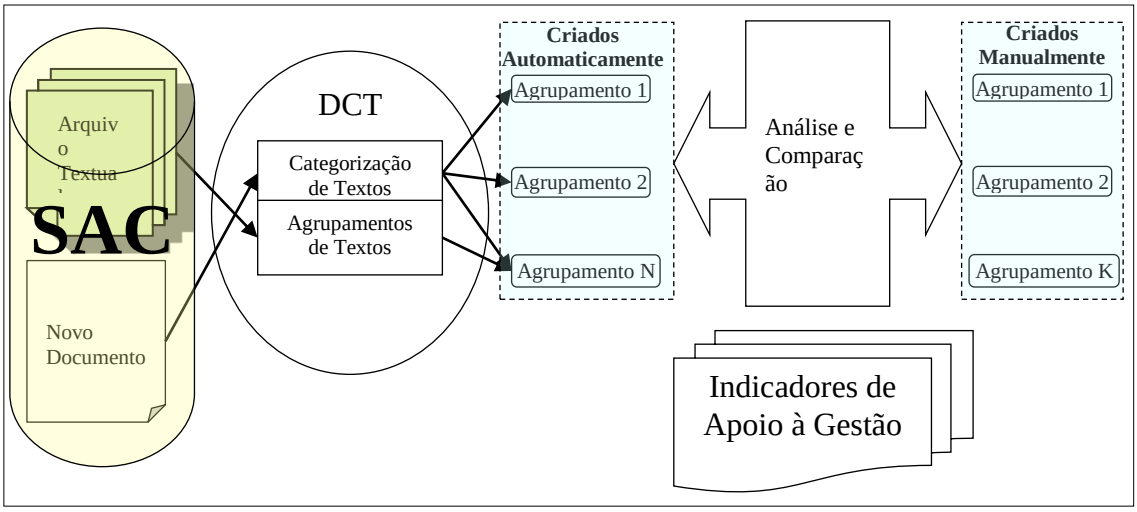

Figura 2: Processo de desenvolvimento da pesquisa

A base SAC contém os documentos necessários para o estudo. Os arquivos textuais históricos são o insumo para a mineração de texto que produze agrupamentos com o objetivo de reunir documentos similares quanto ao conteúdo. Após essa etapa, os grupos criados por processo automático são submetidos aos analistas de domínio para validação e interpretação de seu conteúdo. Em seguida, esses agrupamentos são comparados aos já existentes, criados manualmente.

Uma vez estabelecidos os grupos que representam de forma substantiva a categoria dos documentos, utiliza-se essa nova variável para a criação de um modelo categorizador de novos documentos que possibilita que o processo de categorização seja efetuado automaticamente.

Finalmente, tem-se a base textual segmentada e estruturada que possibilita a extração de relatórios que fornecem indicadores para apoiar a gestão.

Após a coleta e seleção dos documentos, um passo essencial e que consome a maior parte do tempo é a preparação dos dados. O processo envolve a extração de termos, entretanto deve-se ter o cuidado de não separar palavras compostas, como por exemplo, “mineração de textos”. Para isso, o uso de testes estatísticos tem papel relevante. Além disso, a eliminação de palavras frequentes e com baixa relevância é uma tarefa que deve ser executada. Seguindo o fluxo, a lematização, que visa diminuir a quantidade de termos utilizados no estudo, faz parte das tarefas a serem realizadas e é de grande utilidade na redução das dimensões. Outra tarefa importante é a criação da lista de sinônimos em parceria com analistas de domínio e uso de dicionário especializado. Finalmente, são efetuadas as transformações, utilizando ponderações, dos termos em notação numérica para, então, possibilitar transformações mais complexas como a decomposição de valores singulares (DVS) que também objetiva a redução da quantidade de termos.

O passo da preparação dos dados descrito no parágrafo anterior consiste na adequação da informação textual para o formato requerido pelo algoritmo de agrupamento. Essa etapa ocupa-se em rotular cada um dos documentos do corpus baseado no exame de seus termos e, dessa maneira, reuni-los em grupos menores que deverão conter documentos similares. Essa fase conta com o apoio do analista de domínio para identificação do tema de cada grupo.

Um dos objetivos específicos do projeto é verificar a aderência do processo de criação de agrupamentos automática e manual que já vem sendo executada por empregados que tratam as informações recebidas. Por meio dessa comparação propõe-se uma classificação e, a partir dela, a geração do indicador.

Uma vez criados os agrupamentos automaticamente, a tarefa a ser cumprida é a categorização de novos documentos que são recebidos diariamente.

Todo documento alocado em determinado grupo possui um rótulo que o identifica como membro portador das características que são específicas daquele grupo e não de outro. Esses documentos possuem padrões que os identificam e os diferenciam dos demais. A tarefa então é construir um categorizador automático baseado no conteúdo dos agrupamentos identificados.

Essa modelagem é baseada em processos já bem estabelecidos principalmente na Estatística e na Inteligência Artificial. Dessa forma, espera-se que os dados históricos possam indicar alguns critérios de decisão para categorização, isto é, que se aprenda, a partir dos dados, a estrutura implícita que caracteriza a alocação de um documento em um determinado grupo.

O algoritmo de categorização, chamado de categorizador, tem como entrada os novos documentos que passam pelos processos de transformação de texto em número. Assim, pela análise de peculiaridades do novo documento e comparação com sua aprendizagem na coleção para distinção de grupos, o categorizador indicará sua provável alocação.

Considerando que uma base SAC contempla a insatisfação dos clientes, é razoável supor que produtos ou serviços com grande número de queixas estejam com problemas e devem ser analisados a fim de resolver a insatisfação por parte da clientela.

O indicador foi formulado através de índices que destacam agrupamentos de produtos ou serviços de acordo com suas ocorrências. Esses índices podem servir de indicativos para uma verificação mais detalhada de pontos críticos.

O indicador para o acompanhamento é:

$$IS_i = \frac{G_i}{\sum_i G_i} x100 \qquad \textbf{(1)}$$

IS_i representa o índice de satisfação no agrupamento i

G_i é a soma de todas as ocorrências de reclamação no agrupamento i

$\sum_i G_i$ é a soma total das ocorrências de reclamação.

Os pontos críticos podem ser visualizados através de gráficos e tabelas que fornecem subsídios importantes para os gestores que prontamente podem tomar decisões na resolução do problema, como adequação de serviço ou produto, treinamento de pessoal, substituição de pessoas e assim por diante.

5 Resultados

Realizou-se a classificação automática e comparou-se com a classificação manual para aferir a qualidade da criação de agrupamentos automaticamente com base no conteúdo das mensagens. Na classificação manual foram criados 19 grupos, sendo que 3 deles não possuem denominação, portanto encaixam-se na categoria "outros". A classificação automática apresentou 12 grupos. A diferença entre as quantidades pode ser explicada pela subcategorização de grupos como Atendimento que, na classificação manual, são diferenciados em grupos distintos como atendimento pessoal e atendimento eletrônico. No quesito quantidade de grupos, a classificação automática apresentou uma quantidade menor, entretanto convincente em relação ao conteúdo dos textos alocados nos agrupamentos.

A figura 3 apresenta os agrupamentos finais com suas respectivas proporções.

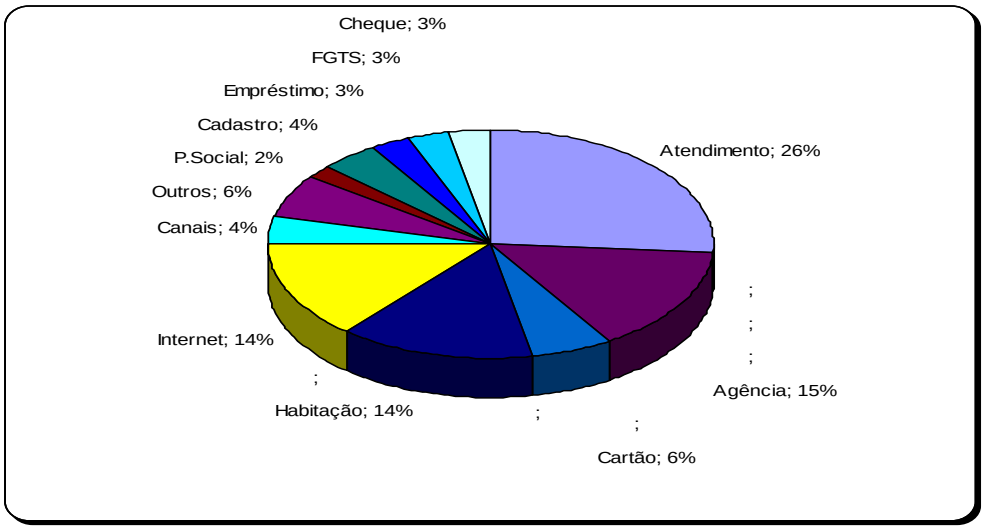

Figura 3: Representação da Proporção dos Agrupamentos

A identificação do tema dos agrupamentos é feita observando-se as palavras descritoras e, caso necessário, verifica-se o conteúdo de mensagens alocadas para o agrupamento em questão, visando assegurar o entendimento do tema que os textos abordam.

As proporções finais encontradas foram analisadas por especialistas de domínio que avaliaram satisfatoriamente, isto é, o resultado pode representar o dia-a-dia de reclamações que são encaminhadas para o SAC. Existem várias opções de ferramentas para construção de modelos preditivos no programa utilizado neste estudo. Alguns usam métodos estatísticos e outros não. O fato é que nenhuma técnica resolve todos os problemas na mineração de dados e cada uma delas possui o seu ponto forte e fraco. A escolha do melhor modelo é dependente da aplicação e deve ser feita baseada em medidas que validam a qualidade do modelo, isto é, o quão assertivo é o resultado produzido.

Para a construção do modelo de classificação neste trabalho, foram usadas 4 opções disponíveis no programa SAS: regressão logística, árvore de decisão, *memory-based reasoning* e rede neural. Ao leitor interessado nos aspectos teóricos na construção desses modelos e na aplicação destas técnicas no mundo dos negócios, sugere-se a leitura de Berry e Linoff (1997).

Para o desenvolvimento do modelo, uma prática comum é dividir os dados de modo que se possa sempre validar o modelo com dados que não foram utilizados para sua construção. Assim, a amostra foi dividida em três partes: para treinamento, para validação e para teste do modelo.

A primeira, treinamento, com 40% do total de documentos foi usada para construção do modelo inicial, a segunda, teste, 30%, foi utilizada para ajustar o modelo inicial e torná-lo mais geral, isto é, evitar que o modelo seja efetivo somente se aplicado à base de treinamento. A validação foi a terceira e última parte utilizada para medir a provável efetividade do modelo em dados novos, isto é, os 30% de dados restantes que não foram utilizados na construção dele. A divisão foi feita utilizando a estratificação pelas variáveis produto e grupo assunto.

Os resultados das taxas de erro na classificação de cada modelo estão na tabela 2 que se segue:

Tabela 2: Taxa de Erro na Classificação dos Modelos

Nome	Treinamento	Teste	Validação
Memory Based-reasoning	23%	25%	24%
Regressão Logística	5%	16%	14%
Árvore de Decisão	49%	53%	51%
Rede Neural	5%	15%	13%

A leitura dessa tabela diz que na fase de treinamento as melhores performances foram dos modelos de regressão logística e rede neural, 5% cada um. Isto quer dizer que os modelos erram em 5% dos casos de classificação, ou seja, o valor predito pelo modelo não confere com classificação realizada na criação dos agrupamentos. Na fase de teste, na qual o modelo é ajustado para generalizar a previsão, as taxas passam para 16% e 15% para Regressão Logística e Rede Neural, respectivamente. Nota-se que a diferença entre os modelos é muito pequena. Na última fase, aplicação em uma base de dados novos para validação o resultado é 14% para Regressão Logística e 13% para Rede Neural.

Considerando os resultados apresentados, o modelo de Rede Neural apresentou o melhor desempenho e, portanto, é o candidato a ser implementado para a classificação de novos registros que serão recebidos na base.

Vale ressaltar que todo modelo preditivo deve passar por uma reavaliação após um determinado período de uso, pois o seu poder de previsão pode deteriorar drasticamente, caso ocorram mudanças significativas nos dados.

A utilização da classificação automática de textos pode trazer mais velocidade ao tratamento das novas mensagens recebidas diariamente visto que as tarefas de recepção e classificação teriam menor intervenção humana do que a do modelo operacional adotado atualmente. O erro resultante da classificação automática sugere que uma parcela de empregados, hoje utilizados nos mutirões para reclassificação das mensagens, execute a tarefa de depuração destes agrupamentos e assinalem as inconsistências encontradas para que possam realimentar o modelo tornando sua depuração contínua e visando a redução do erro a cada iteração.

O objetivo da estruturação da informação textual contida no SAC culmina com a criação de indicadores de acompanhamento. Para que se faça uma gestão eficiente do desempenho dos produtos e serviços frente aos consumidores é necessária a elaboração de critérios que possibilitem a mensuração da satisfação do cliente.

Aplicando a equação 1 na base de dados tem-se o desempenho mensal por agrupamento.

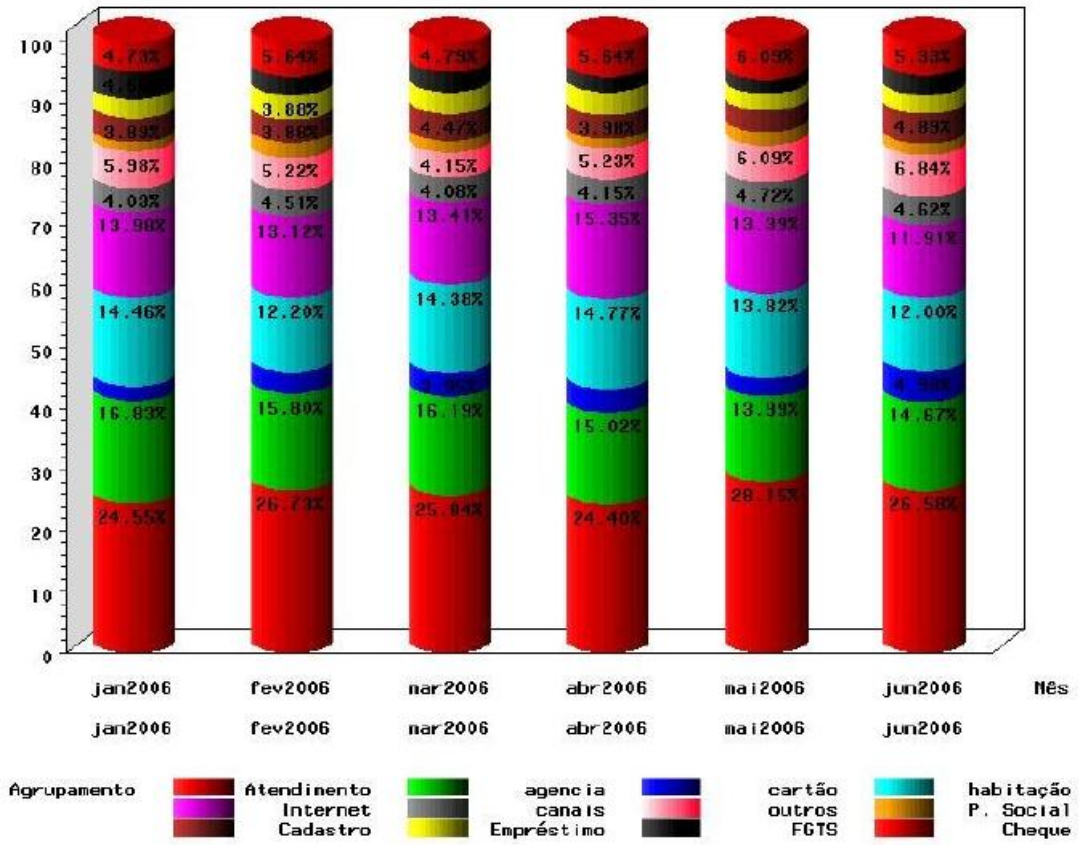

Figura 4: Acompanhamento Mensal

No gráfico da figura 4, pode ser observado mensalmente tanto o comportamento geral, quanto o dos agrupamentos separadamente. Nos agrupamentos pode-se perceber, por exemplo, a diminuição do agrupamento Internet que em janeiro estava com 13,98% de todas as ocorrências e em junho com 11,91%, isto representa um decréscimo de 2,07%. Por outro lado, no agrupamento Atendimento aponta uma leve tendência de crescimento das ocorrências de 24,55% em 26,58%.

Com esse tipo de indicador a Administração atenta pode reverter a insatisfação do cliente com intervenção pontual nos agrupamentos que forem se destacando no conjunto.

6 Conclusão

O trabalho desenvolvido aborda o tema de tratamento automático da informação textual que objetiva a menor intervenção humana possível. Almeja-se com isso a liberação de recursos humanos para atividades intelectuais que a máquina ainda não está apta a fazer.

Percebe-se que o desenvolvimento tecnológico auxilia na velocidade e no volume de tratamento de dados. Porém, a informação textual ainda carece de profissionais e ferramentas, utilizadas em larga escala, capazes de manuseá-la com a mesma destreza das

informações em formato de bancos de dados ou, comumente, chamadas informações estruturadas.

Para atingir os objetivos propostos utilizou-se a metodologia da DCT que vai desde a escolha da base de dados até a utilização efetiva da informação descoberta que se transforma em conhecimento diante das interpretações humanas para aplicação de forma prática.

Para tratamento dos textos, os erros de ortografia e os erros de pontuação comprometem o resultado final e, portanto, sua resolução constitui-se num ponto crítico do trabalho, no qual a intervenção humana interativa e iterativa é uma necessidade.

Durante a etapa de descrição da base, foi possível visualizar o potencial de negócio que a informação das necessidades do cliente, na forma de reclamação, suscita. Uma vertente é o canal direto do cliente com a Empresa que pode agir pontualmente baseada nas observações escritas e melhorar seus processos internos e oferecer melhores produtos e serviços. Se a Instituição atende a reclamação do cliente, ele se sente respeitado e estabelece uma relação duradoura que culmina na realização de muitos negócios vantajosos para ambos.

Considera-se que o objetivo dessa pesquisa foi alcançado, pois a proposta inicial foi concretizada: extrair conhecimento da base SAC, criar agrupamentos automáticos com utilização de ferramenta de mineração de texto, comparar os agrupamentos criados automaticamente e manualmente, criar modelo de classificação de automática das novas mensagens recebidas e propor indicador que reflete o grau de satisfação do cliente em relação aos produtos e serviços oferecidos.

A metodologia proposta foi útil e aplicável na transformação de textos em informações organizadas, na extração de conhecimento e na automatização de processos que dependem de leitura de pessoas dedicadas a essa tarefa.

No âmbito acadêmico, considera-se que a pesquisa obteve êxito, porém sua aplicação no âmbito profissional precisa de equipe dedicada. Neste contexto, o trabalho requer integração de profissionais que compõem este segmento, isto é, do atendente, dos especialistas de domínio, da equipe de tecnologia e dos pesquisadores dedicados à elaboração de modelos e de indicadores.

O sucesso nessa empreitada requer investimentos em treinamento específico: i) aos atendentes, técnicas para padronizar a entrada do texto visando minimizar o esforço de pré-processamento; ii) aos especialistas de domínio, insumos para criação de vocabulário especializado com o propósito de facilitar a identificação do tema da mensagem; iii) aos tecnólogos, atualizações para realização de limpeza e preparação de base de dados; iv) e aos pesquisadores, reciclagem para capacitá-los e atualizá-los em novas tecnologias, técnicas e metodologias.

A ciência é grande parceira da atividade empreendedora. No mundo empresarial, pesquisas têm sentido se promovem negócios de acordo com a missão de cada organização. Portanto, é fundamental promover a aproximação contínua entre equipe técnica e gestores para transformar ações de gestão em produtos e serviços. Melhor ainda se essas atividades são beneficiadas com técnica e ciência, daí o papel essencial da academia nas empresas.

Referências

BERRY, M. J. A.; LINOFF, G. **Data Mining Techniques**: For marketing, sales, and customer support. New York: Wiley, 1997.

BRÄSCHER, M. **Tratamento automático de ambiguidades na recuperação da informação**. 290 f. Tese (Doutorado em Ciência da Informação) — Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 1999.

CHAUVEL, M. A. **Consumidores insatisfeitos**: uma oportunidade para as empresas. Rio de Janeiro: Mauad, 2000.

DÖRRE, J.; GERSTL, P.; SEIFFERT, R. Text mining: finding nuggets in mountains of textual data. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 5., 1999, San Diego. **Proceedings...** New York: ACM Press, 1999. p. 398 – 401.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996. Disponível em:<http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131>. Acesso em: 27 jul. 2011.

FELDMAN, R. et al. A domain independent environment for creating information extraction modules. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 10., 2001, Atlanta. **Proceedings...** New York: ACM Press, 2001. p. 586–588.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview.**AI Magazine**, v. 13, n. 3, p. 57–70, 1992. Disponível em:<http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011>. Acesso em: 16 jul. 2011.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. Cambridge: MIT Press, 2001. 546 p.

HEARST, M. A. Untangling text data mining. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37., 1999, College Park.**Proceedings...** Stroudsburg: Association for Computational Linguistics, 1999. p. 3–10.

KODRATOFF, Y. Knowledge discovery in texts: A definition, and applications. In: INTERNATIONAL SYMPOSIUM ON FOUNDATIONS OF INTELLIGENT SYSTEMS, 11., 1999, Warsaw. **Proceedings...** Warsaw: Springer, 1999. p. 16–29.

KONCHADY, M. **Text Mining Application Programming**. Boston: Charles River Media, 2006. 412 p.

KOTLER, P. A generic concept of Marketing. **Journal of Marketing**, American Marketing Association, v. 36, n. 2, p. 46–54, April 1972.

LIMA, G. Â. B. Interfaces entre a ciência da informação e a ciência cognitiva. **Ciência da Informação**, Brasília, v. 32, n. 1, p. 77–87, jan./abr. 2003. Disponível em:

<http://revista.ibict.br/ciinf/index.php/ciinf/article/view/133/113>. Acesso em: 21 mai. 2011.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: The MIT Press, 1999. 680 p.

RAJMAN, M.; BESANÇON, R. Text mining: Natural language techniques and text mining applications. In: IFIP TC2/WG2.6 WORKING CONFERENCE ON DATABASE SEMANTICS (DS-7), 7., 1997, Leysin.**Proceedings...** Leysin: Chapman & Hall, 1997.

ROBREDO, J. **Da Ciência da informação revisitada aos sistemas humanos de informação**. Brasília: Thesaurus, 2003. 262 p.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, v. 24, n. 1, p. 36–41, 1995. Disponível em: <http://dici.ibict.br/archive/00000598/01/natureza\interdisciplinar.pdf>. Acesso em: 03 jun. 2011.

TAN, A.-h. Text mining: The state of the art and the challenges. In: WORKSHOP ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES, 1999, Beijing. **Proceedings...** 1999. p. 71 – 76. Disponível em:<http://www.ntu.edu.sg/home/asahtan/>. Acesso em: 11 mai. 20011.

TRYBULA, W. Text mining. **Annual Review of Information Science and Technology**, v. 34, n. Text mining: The state of the art and the challenges, p. 385–419, 1999.

VIEIRA, R.; LIMA, V. L. S. de. Linguística computacional: princípios e aplicações. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 21., 2001, Fortaleza. **Anais...** Fortaleza: SBC, 2001. v. 2, p. 47–88. Disponível em: <http://goo.gl/6HKaI>. Acesso em: 08 jun. 2011.

WIVES, L. K.; LOH, S. Tecnologias de descoberta de conhecimento em informações textuais (ênfase em agrupamento de informações). In: OFICINA DE INTELIGÊNCIA ARTIFICAL, III., 1999, Pelotas. **Proceedings...** Pelotas: EDUCAT, 1999. p. 28–48. Disponível em: <http://www.leandro.wives.nom.br/pt-br/publicacoes/OIA99.pdf>. Acesso em: 7 mai. 2011.