UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO – BACHARELADO

DESCOBERTA DE CONHECIMENTO COM O USO DE TEXT MINING APLICADA AO SAC

JOSÉ LINO UBER

BLUMENAU
2004

2004/2-27

JOSÉ LINO UBER

DESCOBERTA DE CONHECIMENTO COM O USO DE TEXT MINING APLICADA AO SAC

Trabalho de Conclusão de Curso submetido à Universidade Regional de Blumenau para a obtenção dos créditos na disciplina Trabalho de Conclusão de Curso II do curso de Ciência da Computação — Bacharelado.

Prof. Paulo Roberto Dias

BLUMENAU
2004

2004/2-27

DESCOBERTA DE CONHECIMENTO COM O USO DE TEXT MINING APLICADA AO SAC

Por

JOSÉ LINO UBER

Trabalho aprovado para obtenção dos créditos na disciplina de Trabalho de Conclusão de Curso II, pela banca examinadora formada por:

Presidente:	Prof. Paulo Roberto Dias, FURB
Membro:	Prof. Jomi Fred Hubner, FURB
Membro:	Prof. Roberto Heinzle, FURB

Blumenau, 16 de dezembro de 2004

Dedico este trabalho a minha mãe Olga Odorizzi exemplo pessoal e profissional, sabedoria e generosidade, ao meu padrasto Anselmo Benazzi, ambos por toda formação, educação, amor e carinho que me deram, e a todos aqueles que acreditam no ideal acadêmico.

CERTEZA

De tudo, ficaram três coisas:

A certeza de que estamos sempre começando...

A certeza de que precisamos continuar...

A certeza de que seremos interrompidos antes de terminar...

Portanto devemos:

Fazer da interrupção um caminho novo...

Da queda um passo de dança...

Do medo, uma escada...

Do sonho, uma ponte...

Da procura, um encontro...

(Fernando Pessoa)

AGRADECIMENTOS

À Deus, pelo seu imenso amor, graça e presença em minha vida.

À minha família, que sempre esteve presente, me incentivando e apoiando em minhas decisões.

À minha irmã Jacqueline Uber Silva por toda dedicação, paciência e orientação no desenvolvimento deste trabalho.

À minha namorada Lílian Katiane Hellmann e família, pela atenção e carinho.

Aos meus amigos, pelos empurrões e cobranças.

Ao meu orientador, Paulo Roberto Dias, por ter acreditado na conclusão deste trabalho.

RESUMO

A partir do final dos anos 80, vem sendo desenvolvidas pesquisas com o intuito de extrair padrões úteis (até então desconhecidos) a partir de grande volume de dados existentes nas organizações. Atualmente, as organizações começam a buscar informações em suas bases de dados, objetivando um melhor atendimento e disponibilidade de seus produtos no mercado. Mas essas informações, além de formarem uma massa de dados grande, encontra-se de forma desorganizada e não padronizada, o que dificulta a sua localização e acesso. Esses fatores acabam dificultando o tratamento e o entendimento das informações. O objetivo deste trabalho é apresentar a área denominada de Descoberta de Conhecimento em Textos e Descoberta de Conhecimento em Base de dados, que trata os problemas relacionados ao entendimento, classificação e tratamento de informações. Aborda uma visão geral dos conceitos e técnicas envolvidas.

Palavras chave: Text Mining; Cross-Industry Standard Process for Data Mining; Descoberta do conhecimento em base de dados; Descoberta de Conhecimento em Textos; Extração de Informação; Recuperação da Informação.

ABSTRACT

Since the end of the 1980's, research has been developed with the intention of establishing useful standards (unknown as of then), beginning with the large volume of existent data in organizations. Currently, organizations begin searching for information in their own data banks, with the objective of arriving at better servicing and availability of its products in the market. But this information, besides forming a great mass of data, is neither organized nor standardized, which makes it difficult to both locate and access it. These factors contribute to difficulties in dealing with and understanding information. The objective of this work is to present the area known as the Discovery of Knowledge in texts and the Discovery of Knowledge in Data Bases, which deal with problems related to the understanding, classification and treatment of information. It approaches a general vision of the concepts and techniques involved.

Key-Words: Text Mining; Cross-Industry Standard Process for Data Mining; Knowledge Discovery in Databases; Knowledge Discovery from Texts; Information Extraction; Information Retrieval.

LISTA DE ILUSTRAÇÕES

LISTA DE SIGLAS

AG – Algoritmos Genéticos

CRISP-DM – Cross-Industry Standard Process for Data Mining

DCBD – Descoberta do Conhecimento em Base de Dados

DCT – Descoberta de Conhecimento em Textos

EI – Extração de Informação

FA – Ficha de Atendimento

NCR – National Cash Register Company

PLI – Programação Lógica Indutiva

RI – Recuperação da Informação

SAC – Serviço de Atendimento ao Cliente

SPSS – Statistical Package for the Social Sciences

SUMÁRIO

1 INTRODUÇÃO

Em um mundo moderno, o conhecimento tem sido imprescindível para a sobrevivência das empresas. Nas últimas décadas, as empresas estão cada vez mais aumentando suas bases de dados. Vários fatores têm contribuído para este comportamento. A queda nos custos de armazenamento pode ser vista como a principal causa do surgimento dessas enormes bases. Outro fator é a disponibilidade de computadores de alto desempenho a baixo custo.

Tan (1999) afirma que 80% das informações de uma empresa estão em formato textual. Entretanto, as organizações e as pessoas têm dificuldade para tratar adequadamente este tipo de informação por não estar estruturada. A área de *Text Mining* consiste em duas formas de mineração em textos: a Descoberta de Conhecimento em Textos (DCT) e Descoberta do conhecimento em base de dados (DCBD). A DCT surgiu para minimizar este problema, ajudando a explorar conhecimento armazenado em meios textuais, e pode ser definida como sendo o processo de extrair padrões ou conhecimento, interessantes e não-triviais. A DCBD consiste na mineração em base de dados estruturados que contenham textos.

Por tratar-se de um assunto novo, Nuggest (2001) afirma que *Text Mining* representa atualmente apenas 4% das técnicas usadas regularmente para mineração de dados, conforme figura 1. A utilização dessa técnica vem aumentando ao passar dos anos, devido a novas pesquisas na área e pela necessidade de resultados rápidos e eficientes no processamento de um grande volume de informação.

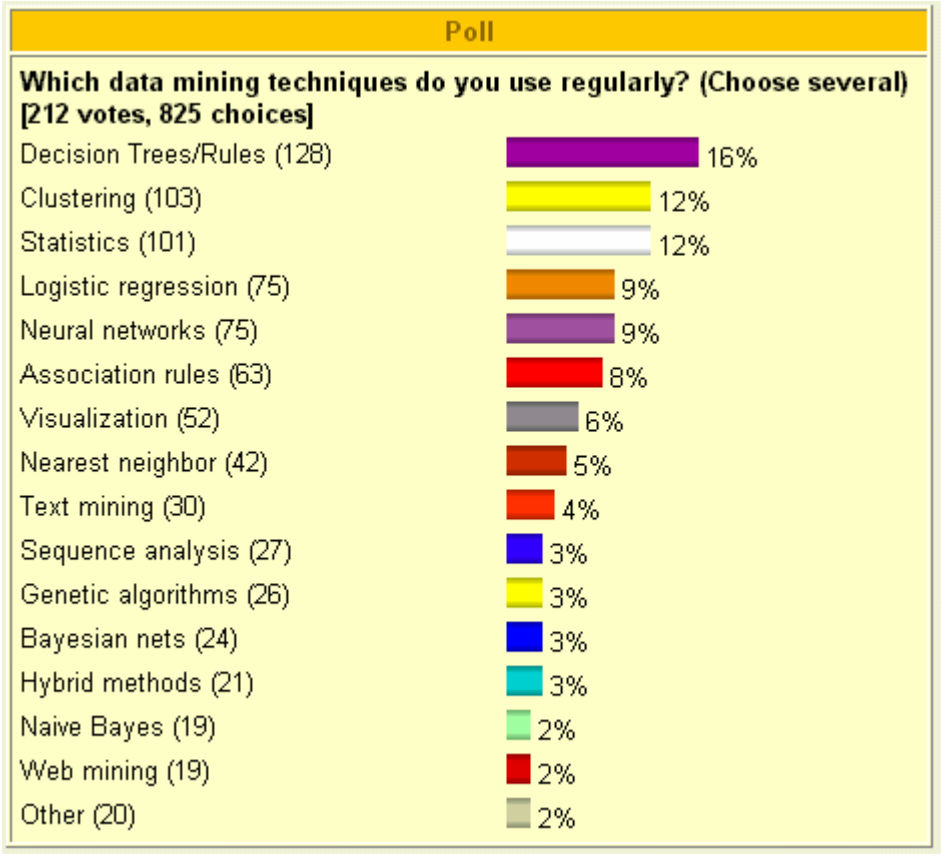
Uma forma de realizar a mineração consiste em utilizar a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM, 2000). Essa metodologia foi criada em 1996 pelo grupo composto pelas empresas DaimlerChrysler, a Statistical Package for the Social Sciences (SPSS) e National Cash Register Company (NCR). É constituída de seis etapas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e aplicação.

Os dados estudados neste trabalho são os chamados telefônicos que estão armazenados em uma base de dados. Esses dados são de propriedade da Operacional Têxtil Ltda[1].

[1] Operacional Têxtil Ltda é uma empresa de desenvolvimento de sistemas para área têxtil.

Foi desenvolvido um software para minerar os chamados telefônicos. Os chamados telefônicos são compostos por: data de abertura, software utilizado, versão do software, descrição do problema, situação da ficha de atendimento, dentre outros, sendo que a descrição do problema é um texto livre e esta foi a variável analisada. Para isto verificou-se a necessidade da implementação de uma ferramenta para automatizar e auxiliar na análise do elevado número de registros de chamados acumulados, atualmente em torno de 8.000. Pretende-se com isto a identificação do motivo do chamado dos clientes, prevendo situações ou diminuindo novos chamados.

Este trabalho também serve como fonte para novos pesquisadores, já que não existe ou não foi encontrada nenhuma pesquisa em chamados telefônicos de uma software house, portanto torna-se inédito.



Fonte: Nuggest (2001)

Figura 1 – Representação das técnicas utilizadas para *Text Mining*

1.1 OBJETIVOS DO TRABALHO

O objetivo deste trabalho é desenvolver um software para descobrir novos conhecimentos em textos armazenados em um banco de dados (descrição do problema), utilizando para isso técnicas de mineração em texto, para automatizar a análise dos chamados telefônicos, identificar automaticamente quais os chamados mais freqüentes, identificando o que ocasionou o chamado e fornecer indicadores em nível gerencial.

1.2 ESTRUTURA DO TRABALHO

Os capítulos iniciais deste trabalho apresentam o problema a ser resolvido em seguida será visto a parte de fundamentação teórica, conceitos, técnicas e métodos comuns para *Text Mining*. Tratando-se da fundamentação teórica, explanou-se uma visão geral dos conceitos da metodologia CRISP-DM, focando o processo de classificação ou categorização das palavras.

Em seguida, pode-se obter conhecimento das metodologias de Descoberta do Conhecimento em Base de Dados e Descoberta de Conhecimento em Textos. O capitulo seguinte refere-se ao desenvolvimento do trabalho, onde constam os requisitos, especificação (diagramas de caso de uso, diagrama de classes), implementação, documentação e operacionalidade do software desenvolvido. E por fim, os resultados, conclusão e extensões para este trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Tan (1999) considera o armazenamento através de texto a forma mais natural de armazenamento de informação.

Segundo (FAYYAD, 1996) a descoberta de conhecimento ocorre por meio de complexas interações realizadas entre homem e uma base de dados, geralmente por meio do uso de uma série heterogênea de ferramentas.

Existem duas formas de descoberta de conhecimentos em textos: a DCT e a DCBD. A única diferença entre a DCT e a DCBD consiste em que, na DCBD os dados encontram-se parcialmente estruturados, e na DCT estão em forma de texto.

Segundo Stanley Loh (apud SILVA, 2002, p. 25), existem três grandes áreas que lidam com informações em grandes bases de dados: *Data Mining* (Mineração de Dados), Recuperação da Informação (RI) e Extração de Informação (EI). No software implementado neste trabalho foram utilizadas técnicas de RI e EI.

2.1 CONCEITOS BÁSICOS

Neste tópico serão vistos alguns conceitos referente a: *Stopwords*, *Keywords*, *Collocations*, *Stemming* e *Corpus*.

2.1.1 Stopwords

Dentro de uma verificação enumerável do que está sendo estudado, têm-se as formas e palavras que não demonstram a mínima relevância. Analisando-se as formas mais freqüentes de se escrever tem-se algumas palavras que não possuem representatividade alguma. Pode-se citar como exemplo as vogais e as preposições de e para.

Estes termos formam a maior parte dos textos da língua portuguesa, e dificultam a busca quando são solicitados, pois estão contidos em quase todos os títulos.

Gonçalves (2003) salienta que por conta disso, a eliminação de tais palavras no processo de indexação salva uma enorme quantidade de espaços em índices, e não prejudica a eficácia da recuperação. A lista das palavras filtradas durante a indexação automática, em virtude de essas palavras produzirem índices pobres, é chamada de *stoplist* ou *dicionário negativo*. Uma maneira de melhorar a performance do sistema de recuperação de informação,

então, é eliminar as *stopword* – palavras que fazem parte da *stoplist* – durante o processo de indexação automática.

Palavras desse tipo são chamadas de palavras negativas, e devem ser retiradas na etapa de preparação dos dados, que será vista no capítulo 3, conhecida como Remoção de *Stopwords.* Na etapa de remoção das *stopwords* essas palavras são eliminadas. No quadro 1, estão relacionadas as *stopwords* mais comuns.

TIPO	PALAVRAS
CONSOANTES	B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y, Z
PREPOSIÇÕES	A, À, ANTE, AO, APOS, ATE, ATÉ, COM, CONTRA, DA, DE, DESTE, DO, NA, NO, PARA, PERANTE, POR, SEM, SOB, SOBRE, TRAS, DURANTE, COMO, CONFORME, EXCETO, MEDIANTE, AFORA, ENTRE, COMO, PER
VOGAIS	A, E, I, O, U
ARTIGOS	A, AS, O, OS, UM, UMA, UMAS, UNS
PRONOMES	COMIGO, CONTIGO, DELE, DELES, ELE, ELES, EU, ME, MEU, MI, NOS, NÓS, NOSSAS, NOSSOS, SEU, TEU, TU, VOS, VÓS, ELA, ELAS, ME, TE, O, A, SE, LHE, MIM, TI, SI, NOS, OS, AS, LHES, COMIGO, CONTIGO, CONSIGO, CONOSCO, CONVOSCO, VOCE, VOCÊ, VOCES, VOCÊS, SENHOR, SENHORA, VOSSA, MEU, SEU, NOSSO, VOSSO, SEU, MINHA, TUA, SUA, NOSSA, VOSSA, SUA, MEUS, TEUS, SEUS, NOSSOS, VOSSOS, SEUS, MINHAS, TUAS, SUAS, NOSSAS, VOSSAS, SUAS, ESTE, ESTA, ISTO, ESSE, ESSA, ISSO, AQUELE, AQUELA, AQUILO, MESMO, PROPRIO, PRÓPRIO, SEMELHANTE, TAL, ALGUÉM, ALGUEM, NINGUÉM, NINGUEM, TUDO, NADA, ALGO, OUTREM, NENHUM, OUTRO, UM, CERTO, QUALQUER, QUAISQUER, ALGUM, CADA, QUEM, QUAL, ESTE, QUANTOS, QUE, CUJO, QUAIS, CUJA, CUJOS, CUJAS, QUANTO, QUANTA, QUANTOS, QUANTAS, ONDE
CONJUNÇÕES	E, MAS, OU, QUE, QUANDO, PORQUE, OU, NEM, NÃO, MAS, PORÉM, POREM, CONTUDO, TODAVIA, ENTRETANTO, SENÃO, SENAO, LOGO, POIS, PORTANTO, COMO, QUANTO, EMBORA, CONQUANTO, APESAR, AINDA, CONFORME, SEGUNDO, TAL, TÃO, TAO, TANTO, QUANDO, DEPOIS, ANTES, ENQUANTO

Quadro 1 – Palavras negativas

2.1.2 Keywords

São as palavras importantes do texto, ignorando-se símbolos e caracteres de controle de arquivo de formatação. Para uma correta determinação das *keywords* (palavras-chave) é imprescindível que sejam removidas as *stopwords*. Um dos recursos utilizados para descobrir a importância dessas palavras é calcular a freqüência com que elas aparecem no texto. Gerard Salton (apud WIVES, 1999, p. 23) determina essa importância de peso que indica o grau de relação entre a palavra e o documento no qual ela aparece e que pode ser calculada pela freqüência absoluta ou pela freqüência relativa.

Segundo Santos (apud WIVES, 1999, p.23) a técnica mais comum de identificação de atributos (palavras) marcantes é a freqüência relativa, que indica o quanto determinada palavra é importante para um documento de acordo com o número de ocorrências desta palavra no documento. Segue fórmula da freqüência relativa conforme quadro 2.

$$F_{rel}x = \frac{F_{abs}x}{N}$$

Fonte: Wives (1999, p. 23)

Quadro 2 – Fórmula da freqüência relativa

Onde:

a) F_{rel} freqüência relativa de uma palavra x em um documento;
b) F_{abs} número de vezes que a palavra aparece no documento;
c) N número total de palavras no texto (N).

2.1.3 Collocations

Pode-se definir as *collocations* ou expressões compostas, como agrupamentos de palavras onde o significado é composto pela soma dos significados das partes mais algum componente semântico adicional.

Segundo Santos (2002, p. 10) pode-se citar como exemplo: cabelo branco, pele branco e vinho branco onde o branco do cabelo é cinza, o branco da pele é rosado e o branco do vinho é amarelado.

2.1.4 Stemming

Segundo Chaves (2003, p. 2), *stemming* consiste em reduzir todas as palavras ao mesmo *stem*[2], por meio da retirada dos afixos da palavra, permanecendo apenas a raiz dela.

O propósito, segundo Chaves (2003, p. 2) é chegar a um *stem* que captura uma palavra com generalidade suficiente para permitir um sucesso na combinação de caracteres, mas sem perder muito detalhe e precisão. Um exemplo típico de um *stem* é “conect” que é o *stem* de

[2] Conjunto de caracteres resultante de um procedimento de *stemming*

“conectar”, “conectado” e “conectando”. Dois erros típicos que costumam ocorrer durante o processo de *stemming* são *overstemming* e *understemming. Overstemming* se dá quando a cadeia de caracteres removida não é um sufixo, mas parte do *stem*. Por exemplo, a palavra gramática, após ser processada por um *stemmer*, é transformada no *stem grama*. Neste caso, a cadeia de caracteres removida eliminou parte do *stem* correto, a saber “gramát”. Já *understemming* ocorre quando um sufixo não é removido completamente. Por exemplo, quando a palavra “referência” é transformada no *stem* “referênc”, ao invés do *stem* considerado correto “refer”.

2.1.5 Corpus

> É um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SANCHEZ, 1995, p. 8-9).

2.2 MÉTODOS MAIS COMUNS DE TEXT MINING

Dentre os métodos mais comuns de *Text Mining* pode-se destacar: recuperação da informação, extração da informação e indexação automática.

2.2.1 Recuperação da informação

Dixon (1997, p. 2) define a RI como sendo “localização e recuperação dos textos relevantes para o quê o usuário necessita descobrir”. A RI tem como objetivo localizar os documentos que contém informações definidas pelo usuário em uma consulta. Para agilizar, utiliza-se a indexação, extraindo assim os termos mais significativos e excluindo os que não tem importância.

A RI vem sendo estudada por muitos anos dentro da área de banco de dados. Ao contrário dos sistemas de banco de dados que focalizam a questão de transação, a RI está preocupada com a organização e recuperação da informação em grande número de documentos textuais. O problema da RI consiste em localizar documentos pertinentes com as palavras chaves que o usuário deseja encontrar. Para localizar essas informações, faz-se uso da indexação.

A indexação é considerada como um tipo de filtro capaz de selecionar e identificar as características de um documento, extraindo termos mais significativos e excluindo aqueles que não são importantes. Segundo Yates (apud SILVA, 2000, p. 29), de três formas:

a) tradicional – os termos descritivos dos documentos são selecionados manualmente, especificando quais farão parte do índice;
b) *full-text* – os termos que compõem o documento são usados como parte do índice;
c) por parte do texto (*tags*) – a seleção dos termos é feita de forma automática.

2.2.2 Indexação automática

Segundo Riloff (apud WIVES, 2000), a indexação automática consiste de quatro etapas: identificação de termos, remoção de *stopwords*, normalização e padronização do vocabulário e seleção de termos relevantes.

A identificação de termos identifica as palavras importantes do texto, ignorando símbolos e caracteres de controle de arquivo ou de formatação. Segundo Wives (2000), essa etapa deve considerar e tratar termos compostos (por exemplo: processo judicial, processo computacional) por considerar que pode fazer parte do mesmo índice para não perderem o significado que as palavras expressam quando estão juntas.

A seleção de termos relevantes consiste em descobrir a importância das palavras em um texto. Para isso utiliza-se da freqüência com que elas aparecem. Gerard Salton (apud WIVES, 2000) denomina essa importância de peso que indica o grau de relação entre a palavras e o documento.

2.2.3 Extração da informação

Dixon (1997) define a EI como a identificação de itens (características, palavras), relevantes nos documentos (geralmente o usuário deve, de alguma forma estabelecer quais são estes itens e como eles podem ser identificados). Esses itens devem ser extraídos e convertidos em dados.

A EI consiste de algumas técnicas, dentre elas: Sumarização, *Clustering* e Filtragem de Informação e Classificação ou categorização.

2.2.3.1 Sumarização

A sumarização é a abstração das partes mais importantes do conteúdo do texto, assim fazendo a produção de um resumo do texto original. Ao observar os procedimento de busca através de palavras-chave, verificar-se que a aplicabilidade e utilização do quesito como instrumento na área comercial, pode ser utilizado o *WordSmith Tools* sistema de *Text Mining* para que é denominado sumarização e extração para o português denominado por Larocca Neto et al. (2000). Destacando que as mudanças referem-se a procedimentos da língua natal, ou seja, a que é disposta a atualizar. Os procedimentos restantes são individuais da língua, e são reproduzidos de acordo com a língua ao qual estão instalados. Verifica-se dentro de um sistema de sumarização que no período de pré-processamento, é eliminada as *stopwords* afiliando-na a partir da lista para o português pré-determinado. Sendo que não se é utilizada algoritmos de *stemming* já que existem grandes recursos de aplicabilidade na língua portuguesa.

Os metadados ou dicionário de dados são as informações sobre os dados mantidos pela empresa. São definidos como “dados dos dados”, “informações das informações”. Dada a complexidade das informações de um *Text Mining*, a documentação dos sistemas e dos bancos de dados tornou-se de vital importância, pois, sendo um projeto gigantesco, se não houver documentação não será possível analisá-lo. Devem ser constituídos para o projeto de *Text Mining*, segundo Carvalho (2001) metadados de negócio e técnico.

2.2.3.2 Clustering

Observando algumas denominações realizadas por Carvalho (2001) observa-se que *clustering* (ou *agrupamento*) é o sistema que foi utilizado para traçar paralelos e associar entre formas, objetos, dando uma maior identificação às formas. O *clustering* tem como função essencial alocar documentos que possuam assuntos similares, desta forma criando novos grupos para cada elemento distinto. O *clustering* é utilizado no processo de classificação, facilita a definição das classes, possibilitando ao operador observar e co-relacionar os elementos e diversidade de documentos.

2.2.3.3 Classificação ou categorização

Classificação é uma das tarefas mais referenciadas na literatura. É também denominada de aprendizado supervisionado, pois a entrada e a saída desejadas são fornecidas

previamente por um supervisor externo (FAUSETT, 1994). Por exemplo, pessoas podem ser previamente grupadas nas classificações de bebês, crianças, adolescentes, adultos, e idosos. Dois anos ou menos pode ser mapeado para a categoria bebê.

Segundo Yang (apud UBER, 2004), classificação “é uma técnica empregada para identificar qual classe ou categoria determinado documento pertence, utilizando como base o seu conteúdo. Para tanto, as classes devem ter sido previamente modeladas ou descritas através de suas características, atributos ou fórmula matemática”.

As principais técnicas de classificação são três: estatística, aprendizagem de máquina simbólica e redes neurais. Nesta monografia, o interesse é investigar principalmente o campo chamado da estatística, focalizando a tarefa de classificação, e muita atenção tem sido dada a técnicas baseadas em árvores de decisão. Outras técnicas, tais como Algoritmos Genéticos (AG) e Programação Lógica Indutiva (PLI) têm sido alvo de mais interesse por parte de pesquisadores recentemente. As técnicas de aprendizagem de máquina simbólica para a tarefa de classificação possuem a vantagem de gerar expressões simples o suficiente para a compreensão humana. Cabe ressaltar que nenhum algoritmo é o melhor em todas as aplicações. A *performance* de um algoritmo de classificação depende muito do domínio da aplicação.

Nesta monografia, assume-se que o problema é projetar um algoritmo para ser aplicado em um banco de dados onde as classes são pré-definidas e cada novo dado deve ser associado a uma destas classes. Este processo é conhecido como reconhecimento de padrões, discriminação, aprendizagem supervisionada ou classificação. Na literatura de estatística, a aprendizagem supervisionada usualmente é referenciada como discriminação.

Segundo Romão (1995, p. 77), o fato das regras de classificação serem obtidas utilizando apenas os dados de treinamento, não garante que as regras terão boa exatidão diante dos dados de teste, os quais não foram utilizados durante o treinamento. A expressão “não determinística” se refere ao fato das regras serem obtidas com base em dados passados para prever o futuro, caracterizando uma forma de indução.

Muitos algoritmos necessitam a experiência de um especialista para usá-lo adequadamente, isto é, para ajustar os parâmetros. Idealmente, todos algoritmos deveriam ser

automáticos no ajuste destes parâmetros. O usuário deveria apenas alimentar os dados e executar o algoritmo. Entretanto, nem sempre este é o caso.

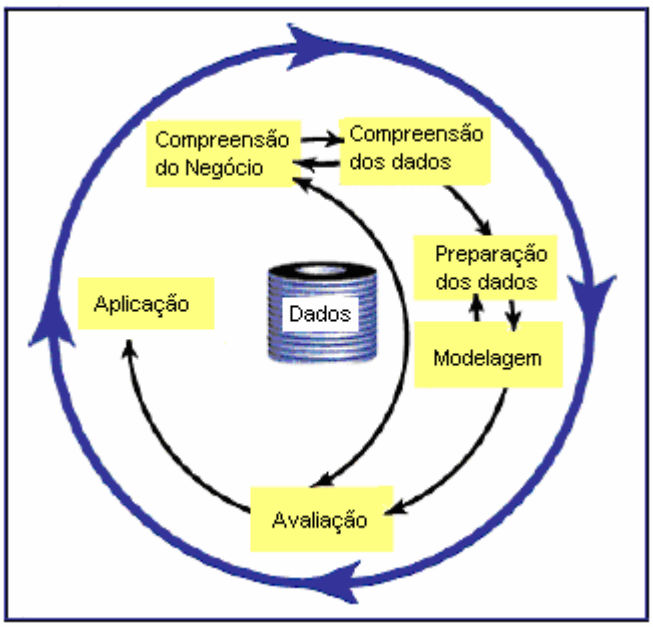
3 METODOLOGIAS PARA DESCOBERTA DE CONHECIMENTO EM TEXTO

Com a finalidade de assegurar a compreensão do estudo, descreve-se neste capítulo a utilização de métodos e técnicas de procedimento cientifico, bem como as metodologias de Palazzo, Ah-Hwee Tan, Halliman e CRISP-DM. Essas metolodogias visam padronizar os procedimentos de mineração em texto. Nessa monografia será utilizada a metodologia CRISP-DM.

3.1 METODOLOGIA PARA DCBD

Segundo CRISP-DM (2000), em 1996 foi criado o grupo de trabalho CRISP-DM (Cross-Industry Standard Process for Data Mining), com o intuito de promover a padronização de conceitos e técnicas na busca de informações específicas para tomada de decisões. Esse grupo propôs uma metodologia como o mesmo nome, destinada a auxiliar administradores e responsáveis no processo geral de planejar e executar a mineração de dados, englobando a especificação do processo até a apresentação dos resultados. Esse grupo era composto por três empresas pioneiras no setor: a Daimlerchrysler, a SPSS (Data Mining) e a NCR (Data Warehouse).

O modelo CRISP-DM segue um ciclo de vida conforme figura 2.



Fonte: SILVA (2002, p. 44)

Figura 2 – Ciclo de vida de DCBD segundo a CRISP-DM

Essa metodologia é constituída de 6 etapas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e aplicação. Essas etapas são executadas de forma interativa. O encadeamento das ações, dependendo do objetivo e de como as informações encontram-se, pode ocasionar o retorno a etapas já realizadas.

A compreensão do negócio procura identificar as necessidades e objetivos do cliente, convertendo esse conhecimento em uma tarefa de mineração de dados.

A compreensão dos dados procura identificar informações que possam ser relevantes para o estudo e uma primeira familiarização com seu conteúdo, descrição, qualidade e utilidade. A coleção inicial dos dados procura adquirir a informação com a qual se irá trabalhar, relacionando suas fontes, o procedimento de leitura e os problemas detectados. Nessa tarefa, descreve-se ainda a forma como os dados foram adquiridos, listando seu formato, volume, significado e toda informação relevante. Durante essa etapa, são realizadas as primeiras descobertas.

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.

Na modelagem, são selecionados e aplicadas técnicas de mineração de dados mais apropriadas, dependendo dos objetivos pretendidos. A criação de um conjunto de dados para teste permite construir um mecanismo para comprovar a qualidade e validar os modelos que serão obtidos. A modelagem representa a fase central da mineração, incluindo escolha, parametrização e execução de técnicas sobre o conjunto de dados visando à criação de um ou vários modelos.

A avaliação do modelo consiste na revisão dos passos seguidos, verificando-se os resultados obtidos vão ao encontro dos objetivos, previamente, determinados na compreensão do negócio, como também tarefas a serem executadas. De acordo com os resultados alcançados, na revisão do processo, decide-se pela sua continuidade ou se deverão ser efetuadas correções, voltando a fases anteriores ou ainda, iniciando novo processo.

A aplicação é o conjunto de ações à organização do conhecimento obtido e à sua disponibilização 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.

3.2 METODOLOGIAS PARA DCT

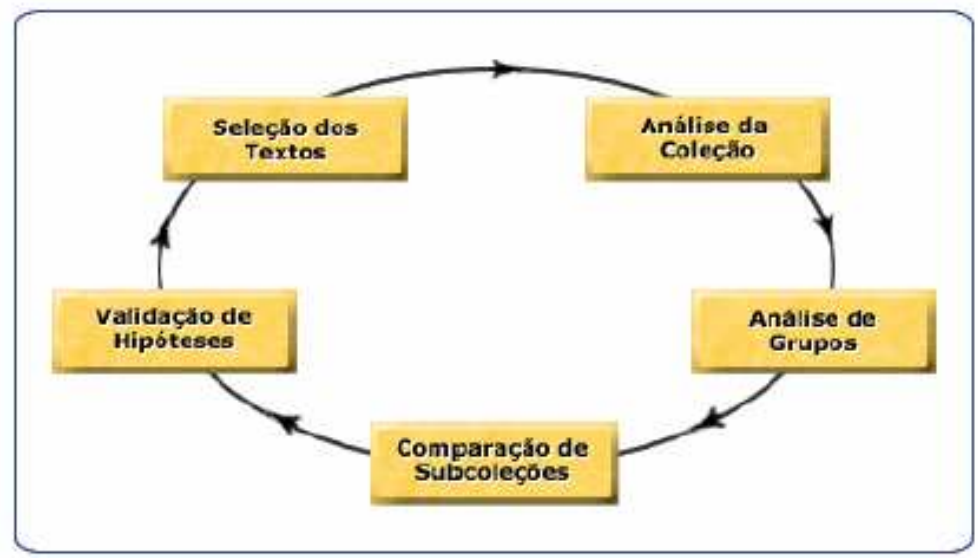
Segue algumas metodologias que podem ser aplicadas para a descoberta de conhecimento em texto.

3.2.1 Abordagem de Palazzo

Segundo Loh (2000), a descoberta de conhecimento em texto divide-se de acordo com o tipo de ação, proativa ou reativa. No modo reativo, o objetivo é direcionado para a solução especificada pelo usuário que, nesse caso, sabe como solucionar o problema. O usuário segue utilizando pistas que deseja provar, direcionando o processo de descoberta. Ele sabe o que quer e tem idéia de onde achar a resposta. Nesse modo, o usuário deve definir, da forma a mais precisa possível, sua necessidade, o que muitas vezes contradiz o processo de descoberta. Na maioria das vezes, o que acontece é o usuário não saber especificar as necessidades para resolução dos seus problemas.

No modo proativo, ao contrário, sem que haja uma intervenção inicial do usuário, as informações úteis para resolução do problema são encontradas automaticamente. Dessa maneira, o problema é definido pelo usuário, mas a descoberta ocorre de modo não-supervisionado.

Segundo Wives (2000), de maneira geral, o processo de descoberta do conhecimento é realizado de forma proativa, que é apresentada na figura 3:



Fonte: SILVA (2002, p. 52)

Figura 3 – Abordagem de Palazzo

Seleção de textos: aplicação de técnicas automáticas como a recuperação de informação (que encontra textos por palavras-chave ou termos presentes nos textos) e a classificação (que separa textos por assunto) ou selecionando manualmente.

Análise da coleção (toda ou partes): aplicação de técnicas de descoberta sobre todos os textos ou de partes da coleção. A separação em sub-coleções pode ser feita de forma automática com a técnica de agrupamento ou por algum critério estabelecido pelo usuário.

Análise de grupos de textos (todo ou parte): extração de uma lista de termos comuns a todos os textos ou que aparecem em mais de um (técnica de listagem de conceitos-chave ou centróide).

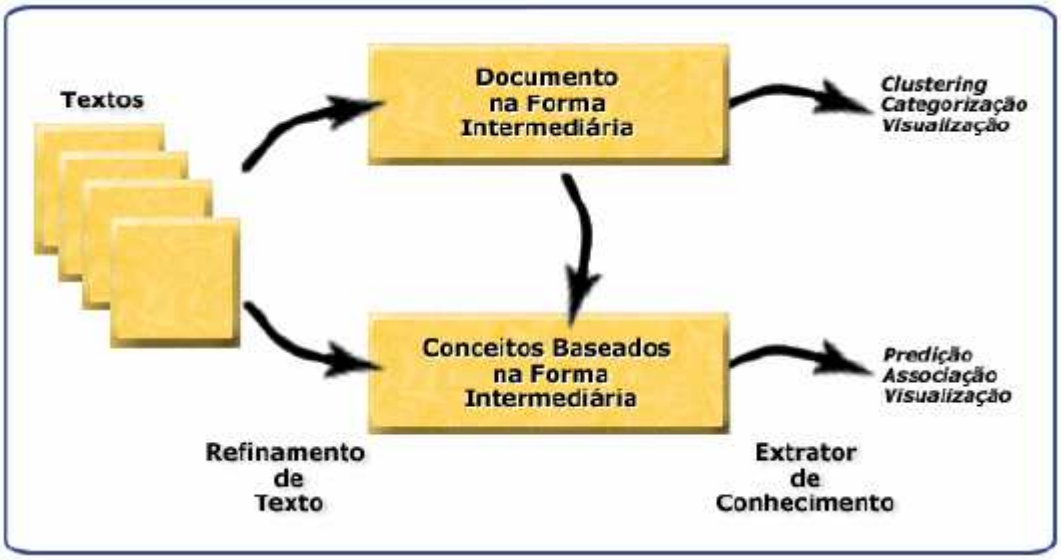
Comparar sub-coleções entre si ou em relação à coleção toda: comparação entre os resultados dos subgrupos e os obtidos da coleção toda.

Validar hipóteses: por meio de técnica de resumos, interpretar os resultados.

Retroalimentação: refazer e realimentar o processo até atingir o objetivo esperado.

3.2.2 Abordagem de Ah-Hwee Tan

Conforme Tan (1999), o processo de mineração em texto, consiste em duas etapas: o refinamento do texto e o extrator do conhecimento, conforme figura 4:



Fonte: SILVA (2002, p. 53)

Figura 4 – Abordagem da mineração de texto segundo Ah-Hwee Tan

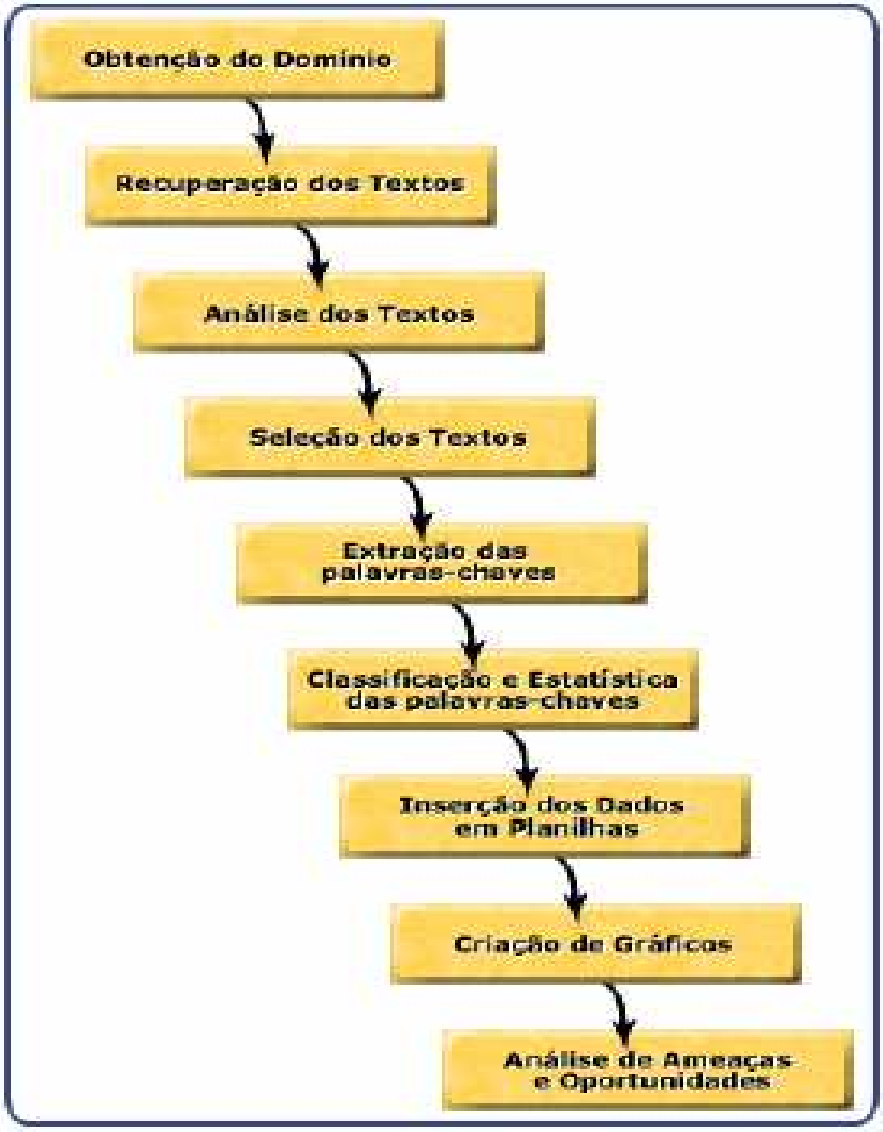
O refinamento consiste na transformação de texto de forma livre para uma forma intermediária. Essa forma intermediária pode ser semi-estruturada (gráficos) ou estruturada (tal como um banco de dados relacional). Por exemplo, dado um conjunto de artigos novos, o refinamento de texto (*text refining*) converte cada documento para forma intermediária. O objetivo dessa ação é organizar os artigos de acordo com seu conteúdo para visualização e navegação.

O extrator do conhecimento faz o reconhecimento de padrões, baseado nessa forma intermediária. Essa etapa consiste na própria mineração, podendo utilizar técnicas como o agrupamento e a classificação ou modelos de predição e de associação.

3.2.3 Abordagem de Halliman

Silva (2002) apresenta um estudo de caso em que são analisadas informações textuais externas à empresa. Neste estudo, a análise do ambiente é dividida em partes, conceituadas pelo autor como forças do ambiente.

Com base nessas tendências e na distribuição dessas forças, o autor mostra como detectar ameaças e fortalecer oportunidades para a empresa tendo como base a mineração de texto. O fluxo dessa metodologia é mostrado na figura 5.



Fonte: SILVA (2002, p. 53)

Figura 5 – Abordagem de Halliman

O processo de mineração de texto como base e 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.

Para cada força do ambiente, são associadas palavras-chave. Por exemplo: força marketing: mercado, marketing.

Tendo o domínio e as palavras, o processo seguinte é a recuperação dos textos, nos quais são verificados e analisados seus conteúdos. É realizada uma seleção por meio de pesquisa no texto com as palavras-chave.

São identificados e excluídos os textos não pertencentes ao domínio. Os arquivos restantes são analisados, tendo suas palavras-chave extraídas para compor uma planilha. O resultado é então classificado pela quantidade de palavras-chave encontrada em cada texto.

Com os dados da planilha, são elaborados gráficos a fim de facilitar o processo de análise das informações obtidas. Na pesquisa, o autor mostra a análise por meio de gráficos de distribuição, palavras mais usadas, gráficos de tendências entre outros. O processo usado reduz o tempo de aquisição de informação relevante, levando-se em conta que, na análise, a obtenção de vantagens dependerá das habilidades e do conhecimento do analista.

4 DESENVOLVIMENTO DO TRABALHO

Partindo-se de um software já desenvolvido, que serve para cadastramento da ficha de atendimento, foi estudada a descrição do problema informada pelos Atendentes do Suporte e pela equipe técnica da Operacional Têxtil. Nessa descrição estão relatados os problemas enfrentados pelos clientes e os erros encontrados pela própria equipe técnica.

4.1 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

Os requisitos principais do software implementado neste trabalho estão divididos em duas etapas:

a) requisitos funcionais: especificação das funcionalidades principais da implementação do software;
b) requisitos não funcionais: ao contrário dos requisitos funcionais os requisitos não funcionais não expressam nenhuma função a ser implementada em um sistema de informação. Eles expressam condições de comportamento e restrições que devem prevalecer.

4.1.1 Requisitos funcionais

Segue lista de requisitos funcionais:

a) lista de palavras: é gerada uma lista com todas as palavras, totalizando a freqüência em que essas palavras aparecem. Pode ser ordenada pela maior freqüência ou pela ordenação alfabética das palavras;
b) lista de palavras excluídas (*stopwords*): o usuário poderá determinar qual as palavras não lhe interessam;
c) criação e remoção de categorias;
d) lista de palavras chaves (*keywords*): o usuário poderá determinar qual as palavras que mais lhe interessam;
e) lista de freqüência: o usuário poderá visualizar a freqüência das palavras chaves (*keywords*), palavras excluídas (*stopwords*) e palavras dos registros lidos do bando de dados;
f) criação de gráfico: gráfico estatístico para facilitar a visualização das informações;
g) busca de registros por palavras: a busca poderá ser pelas palavras já encontradas ou por uma palavra escolhida pelo usuário;

h) busca de frases: pode-se pesquisar pelas frases mais freqüentes, determinando a quantidade de palavras.

4.1.2 Requisitos não funcionais

Segue lista de requisitos não funcionais:

a) desempenho: o software gera os resultados das pesquisas solicitadas pelo usuário em um tempo aceitável (em torno de 6 segundos);
b) banco de dados: o software utiliza o banco de dados Oracle, o mesmo utilizado pelo sistema de cadastro de chamados;
c) visualização: o sistema será simples e de fácil visualização, pois terá uma ajuda explicando as principais funções do software.

4.2 ESPECIFICAÇÃO

Seguindo a metodologia CRISP-DM descrita no Capítulo 3, inicializou-se compreendendo o funcionamento das Fichas de Atendimento (FA) para interar-se das funcionalidades da mesma. Para compreender o armazenamento dos dados verificou-se todas as tabelas envolvidas na (FA). Os dados encontravam-se limpos e organizados, sem erros de grafia, portanto a etapa de preparação dos dados não foi realizada.

A modelagem foi realizada através de dois diagramas:

a) diagrama de caso de uso;
b) diagrama de classes.

4.2.1 Diagrama de Caso de Uso

Para realizar a especificação do diagrama de caso de uso, foi utilizada a ferramenta de ajuda Jude Community conforme figura 6. Na especificação do diagrama fez-se a utilização da figura de um operador de computador. Este operador por sua vez irá ter a interatividade com o software desenvolvido da seguinte maneira:

a) carregar dados: o operador após clicar nesta opção fará a carga dos registros armazenados no base de dados do software utilizado pela Operacional Têxtil Ltda, desta carga poderá ser executado o passo de classificação dos registros e poderá obter-se informações sobre os dados armazenados através de pesquisas e gráficos;

b) definir stopwords: neste processo o operador irá informar através de um cadastro as palavras que não são relevantes na busca dos registros;
c) definir keywords: neste processo o operador irá informar através de um cadastro as palavras que são relevantes na busca dos registros, estas palavras serão cadastradas por categoria, a fim de efetuar a classificação dos registros;
d) definir categorias: neste processo o operador irá informar através de um cadastro as categorias que os registros armazenados na carga de dados irão ocupar (classificação).

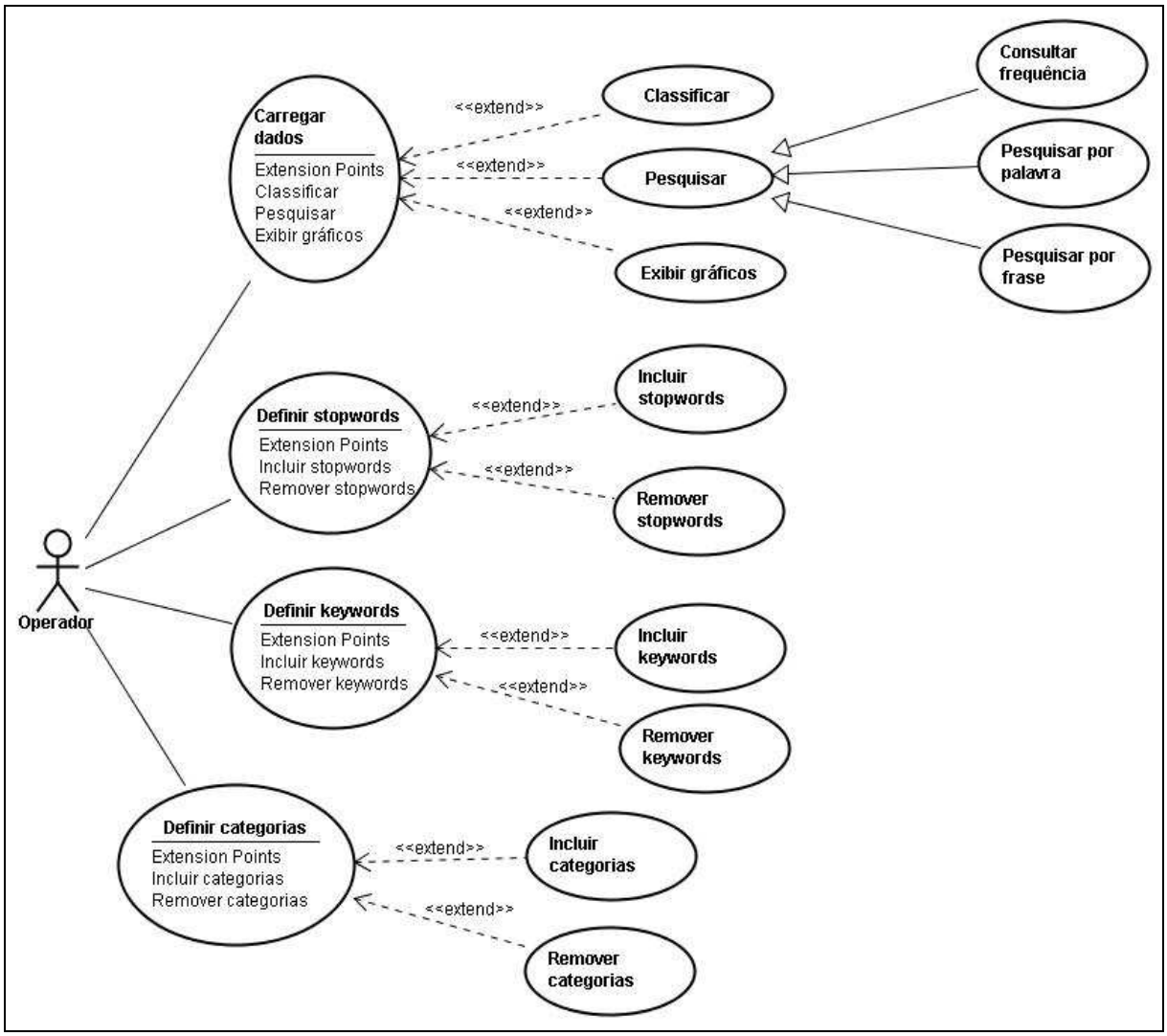


Figura 6 – Diagrama de caso de uso

4.2.2 Diagrama de Classes

Para realizar a especificação do diagrama de classes, foi utilizada a ferramenta Rational Rose conforme figura 7. A seguir são descritas as funcionalidades das classes modeladas no software desenvolvido, as quais são: TFichas_Texto, TStopWords, TCategorias, TListas_Processamento, TPalavras, TKeyWords, TObservacoes e TFicha_Atendimento.

A classe TListas_Processamento é a classe principal do sistema e agrega as classes TFichas_Texto, TStopWords, TCategorias, TListas_Processamento, TPalavras, TKeyWords, TObservacoes e TFicha_Atendimento. Esta classe é responsável por passar as mensagens para as outras classes executarem e interpretarem o código da linguagem.

A classe TFichas_Texto irá armazenar os registros das Fichas de Atendimento.

A classe TStpoWords irá armazenar as palavras negativas, sem relevância para classificação.

A classe TCategorias irá armazenar as categorias das palavras chave (keywords).

A classe TPalavras irá armazenar todas as palavras lidas feitas na carga dos registros.

A classe TKeyWords irá armazenar todas as palavras chaves, relevantes para a classificação.

Para especificação do software foram modeladas 8 classes citadas acima, as quais podem ser vistas na figura 7.

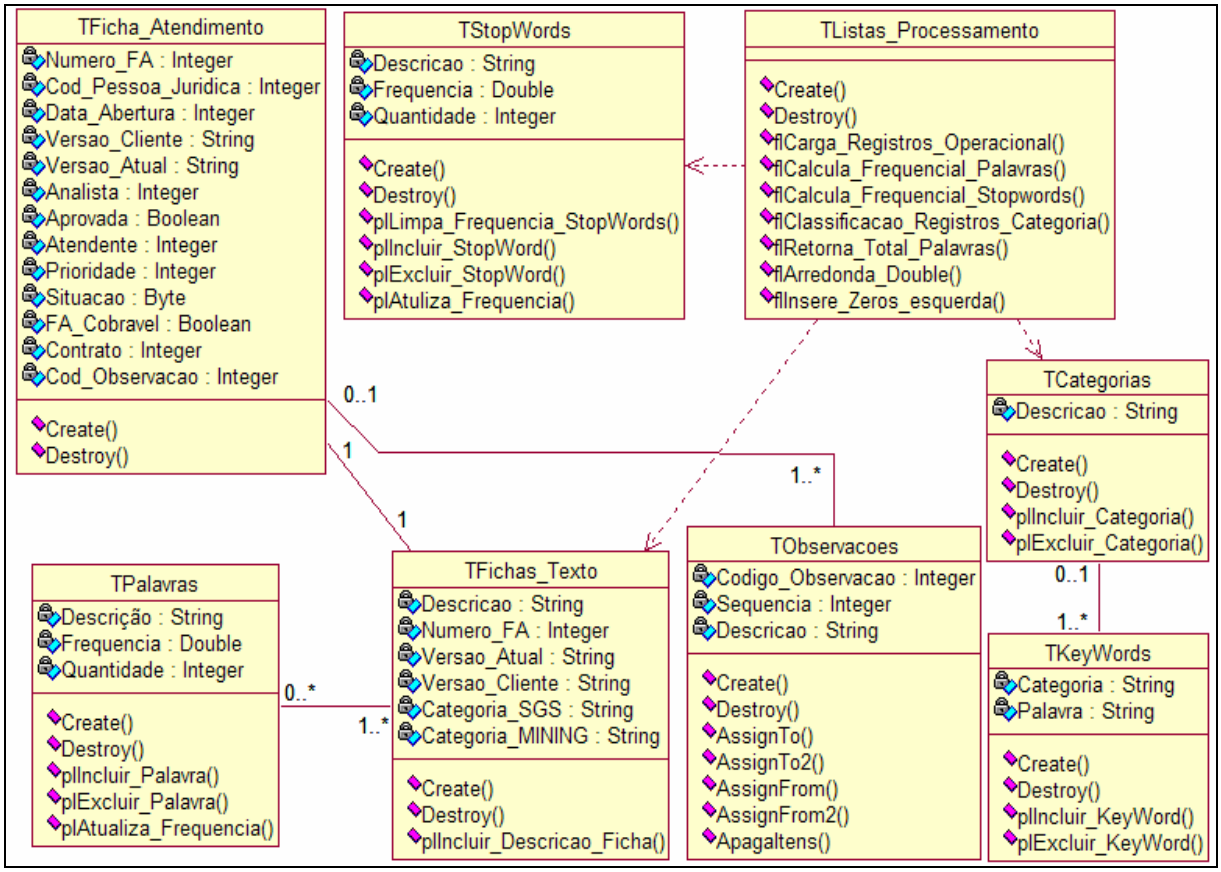


Figura 7 – Diagrama de classes

4.3 IMPLEMENTAÇÃO

Para fins de análise dos registros cadastrados pelos atendentes do suporte e pela equipe técnica da Operacional Têxtil, fez-se necessário à implementação de um software. Este software apresenta de algumas formas a análise dos resultados obtidos dos registros cadastrados. Porém, possui algumas limitações que devem ser ajustadas e refinadas em algum trabalho futuro.

Para implementação do software foi utilizada a metodologia da CRISP-DM, constituída de 6 etapas, e a técnica de classificação e categorização dos registros.

Na primeira etapa que é a *compreensão do negócio*, fez-se necessário um estudo das necessidades da empresa, ou eventuais sugestões de implementação pelo autor, estas sugestões são os requisitos funcionais, referenciados no capítulo 4 item 4.1.1.

Na etapa de *compreensão dos dados*, fez-se necessário uma analise, estudo das “tabelas”, classes e campos existentes, volume de dados do software da empresa, a fim de

buscar um melhor aproveitamento e desempenho dos algoritmos gerados para buscar os registros dos chamados dos clientes.

Já na etapa de *preparação dos dados*, após o conhecimento dos dados a serem buscados, verificou-se que não havia necessidade da limpeza dos dados, pois os registros estavam sem muitos erros de grafia, então esta etapa apenas foi verificada e não foi utilizada.

Na etapa de *modelagem*, a principio, a mais importante, foi onde o algoritmo, técnica de EI utilizando a classificação e categorização dos registros já carregados foi implementada.

Na etapa de *avaliação*, foi feita a validação das etapas aqui descriminadas, verificando se os resultados obtidos vão de encontro com os objetivos do trabalho. Constatou-se que a modelagem mostrou-se eficiente, não precisando voltar às etapas anteriores ou voltar a buscar alguma técnica nova, visto que a agilidade do algoritmo de classificação na busca do conhecimento são rápidos e podem ser analisados pelo software.

E por fim a etapa de *aplicação*, que consiste em mostrar através de relatórios, gráficos e consultas, a fim de explicar os resultados obtidos e uma melhor compreensão do negocio.

O software foi batizado de *MINING OF INFORMATION* conforme figura 8, que representa a tela principal do software desenvolvido. Permite que o usuário obtenha conhecimento dos textos de forma interativa. Implementado na linguagem de programação DELPHI, com algumas características de orientação à objetos. O ambiente de programação adotado foi o Borland Delphi 6.0, devido às facilidades de construção de interfaces. Este software possui algumas limitações que devem ser ajustadas e refinadas em algum trabalho futuro.

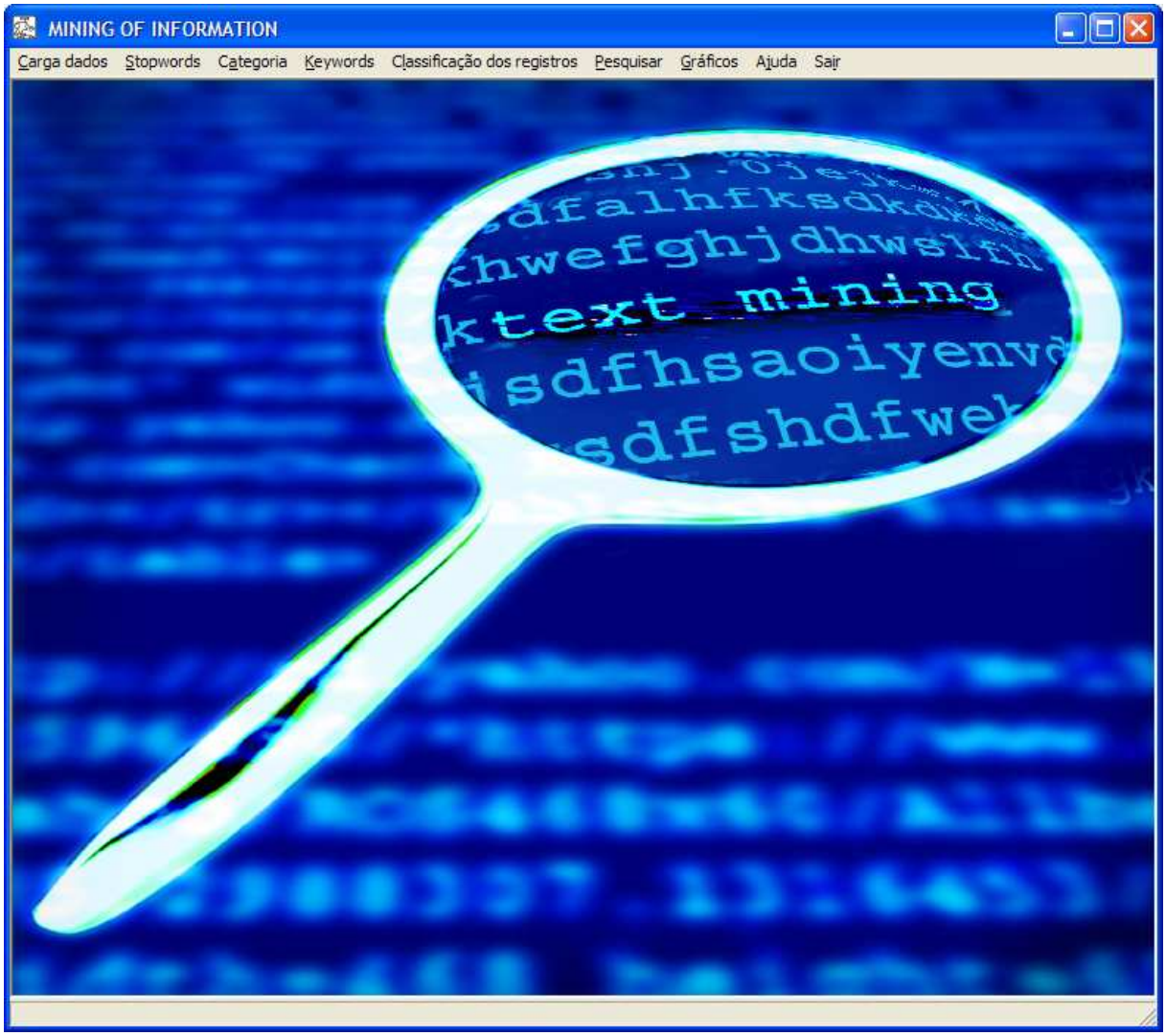


Figura 8 – Tela principal do software

4.3.1 Requisitos do sistema

Para o software desenvolvido nesta monografia, foi adotado o sistema operacional Windows XP como plataforma alvo da implementação. Os requisitos mínimos para o sistema são:

a) um ambiente ou sistema operacional gráfico padrão Windows 32 bits;
b) um processador Pentium™ ou compatível;
c) 128 Megabytes de memória RAM;
d) 6 Megabytes disponíveis no disco rígido.

Mas com objetivos de aumentar o tempo de respostas para busca do conhecimento, recomenda-se 256Mb de memória RAM disponível, visto que os dados lidos do bando de dados e ficam armazenados temporariamente em memória.

4.3.2 Técnicas e ferramentas utilizadas

Para a realização deste trabalho foram utilizadas algumas ferramentas e tecnologias a fim de especificar a análise e desenvolver o aplicativo. Além da linguagem de programação Delphi 6.0 que dá suporte tanto a orientação a objetivos quanto a programação procedural conforme figura 9, que representa a principal classe do software desenvolvido. Foram utilizadas mais duas ferramentas para modelagem das entidades envolvidas. Uma delas é denominada de Jude Community para modelar o diagrama de caso de uso e a outra Rational Rose para fazer o diagrama de classes. São ferramentas fáceis de serem utilizadas e foram muito úteis.

Com o auxilio das técnicas de extração da informação e recuperação da informação, foi possível chegar aos objetivos do trabalho. Estas técnicas foram implementadas respectivamente na carga de dados e classificação dos registros. Os registros foram classificados pela maior quantidade total de palavras chave (*keywords*) das categorias.

```
TListas_Processamento = Class
private
  tbObservacao       : TOBSERVACOES;
  //Abrir classes
  Procedure plAbreArquivos;
  //Finalizar classes
  Procedure plFechaArquivos;
protected

public
  tbFicha_Atendimento: TFICHA_ATENDIMENTO;
  giTotal_Palavras       : Integer;
  goLista_StopWords  : TStopWords;
  goLista_Categorias : TCategorias;
  goLista_Palavras   : TPalavras;
  goLista_Fichas     : TFichas_Texto;
  goLista_KeyWords   : TKeyWords;
  //Criação da classe
  constructor Create; Virtual;
  //Finalização da classe
  destructor  Destroy; Virtual;
  //Carga dos registros para a listas em memória
  function flCarga_Registros_Operacional: Boolean;
  //Calcula a freqüência das palavras
  function flCalcula_Frequencial_Palavras: Boolean;
  //Calcula a freqüência das stopwords
  function flCalcula_Frequencial_Stopwords: Boolean;
  //Classificação dos registros pela quantidade total de palavras chave
  //encontrada no registro, por categoria
  function flClassificacao_Registros_Categoria: Boolean;
  //Retorna o total de palavras
  function flRetorna_Total_Palavras: Integer;
  //Arredonda valores do tipo double
  function flArredonda_Double(vdNum   : Double;
                              Const vbNovDec: Byte): Double;
  //Insere zeros a esquerda
  function flInsere_Zeros_esquerda(lsCampo:String; Const lbTamanhoChave: Byte): String;
published

end;
```

Figura 9 – Classe principal do software

4.3.3 Operacionalidade da implementação

O software foi todo construído baseado em técnicas de ergonomia. Cuidou-se do tamanho das telas, cores utilizadas e respostas claras na interação com o usuário. A ajuda delimitou-se em esclarecimentos do funcionamento dos formulários de modo geral.

Preocupou-se em projetar um sistema que informe e conduza o usuário, fornecendo *feedback*[3] imediato e de qualidade às suas ações. A distribuição espacial dos itens nas telas na

[3] feedback são comentários e informações.

condução do usuário nas opções disponíveis. No próximo tópico será visto a interface do software com o usuário.

4.3.3.1 Interface

A interface do software é constituída de 8 etapas:

a) carga dados – Nesta etapa será feita à leitura dos registros da base de dados para listas em memória;
b) *stopwords* – Definição e manipulação das palavras não relevantes;
c) categorias – Definição de grupos de palavras de *keywords*;
d) *keywords* – Definição e manipulação das palavras chaves (mais relevantes) e manipulação das categorias e classificação das *keywords*;
e) classificação dos registros – Definir os registros lidos do banco de dados qual é de cada categoria respectivamente;
f) pesquisar – Opções para extração do conhecimento, como busca por frases, palavras e lista de freqüência das palavras;
g) gráficos – Gráficos estatísticos para orientação nas tomadas de decisões;
h) ajuda – auxilio nas principais funções do software, visualização das informações da autoria do software e hardware.

4.3.3.1.1 Carga dados

Nesta etapa, conforme a figura 10, são lidas todas as fichas de atendimento (FA), que são armazenadas em memória. Neste mesmo processo, é gerada uma lista em memória, contendo as palavras e a freqüência que elas aparecem nas descrições. Convém salientar que as palavras são agrupadas distintamente, excluindo caracteres numéricos e caracteres de especiais.

Para fazer a carga dos dados, bastar clicar no botão identificado pelo número um (1).

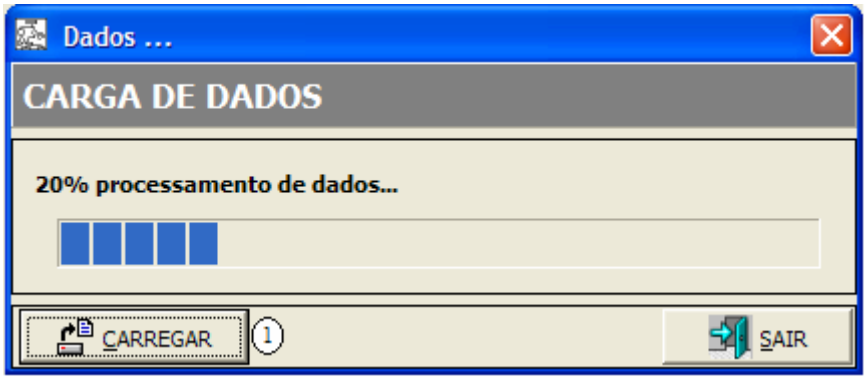


Figura 10 – Carga de dados

4.3.3.1.2 Lista de stopwords

As palavras negativas ou *stopwords* como foi visto no item 2.1.1, devem ser incluídas ou excluídas através deste item do menu, conforme figura 11.

Para incluir uma nova palavra, digite a mesma no campo *Stopword*, identificado pelo número um (1), em seguida clique no botão pesquisar (2). Se a mesma já estiver sida incluída anteriormente o sistema exibe uma mensagem informando ao usuário que a mesma já se encontra cadastrada, desta forma o usuário terá a opção de excluir caso assim o desejar.

Se a palavra não constar na lista de *stopwords*, o sistema vai incluí-la e posteriormente irá eliminá-la da lista de palavras gerada na carga de dados.

A forma de indexação das palavras (3) permite visualizar as palavras em ordem alfabética ou pela sua freqüência.

Os botões (4) são utilizados para mover as palavras da lista de palavras para a lista de *stopwords* e vice-versa.

A barra (5) serve para movimentar-se entre as palavras cadastradas, sendo que o botão (6) é utilizado para excluir uma *stopword* da lista.

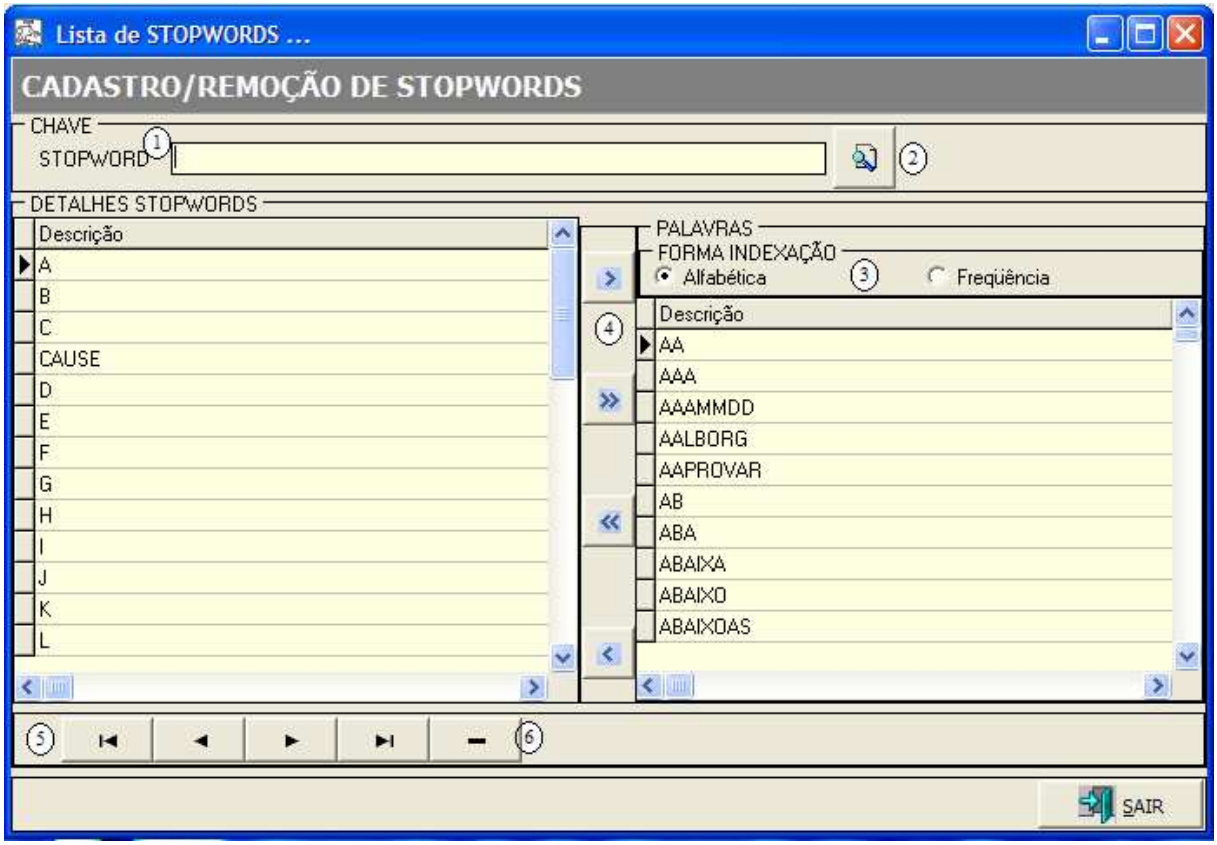


Figura 11 – Cadastro/Remoção de *stopwords*

4.3.3.1.3 Lista de categorias

As categorias são incluídas ou excluídas através deste item do menu, conforme figura 12.

Para incluir uma nova categoria, digite a mesma no campo Categoria, identificado pelo número um (1), em seguida clique no botão pesquisar (2). Se a mesma já estiver sida incluída anteriormente o sistema exibe uma mensagem informando ao usuário que a mesma já se encontra cadastrada, desta forma o usuário terá a opção de excluir caso assim o desejar.

Se a categoria não constar na lista de categorias, o sistema vai incluí-la.

A barra (3) serve para movimentar-se entre as palavras cadastradas, sendo que o botão (4) é utilizado para excluir uma categoria da lista.

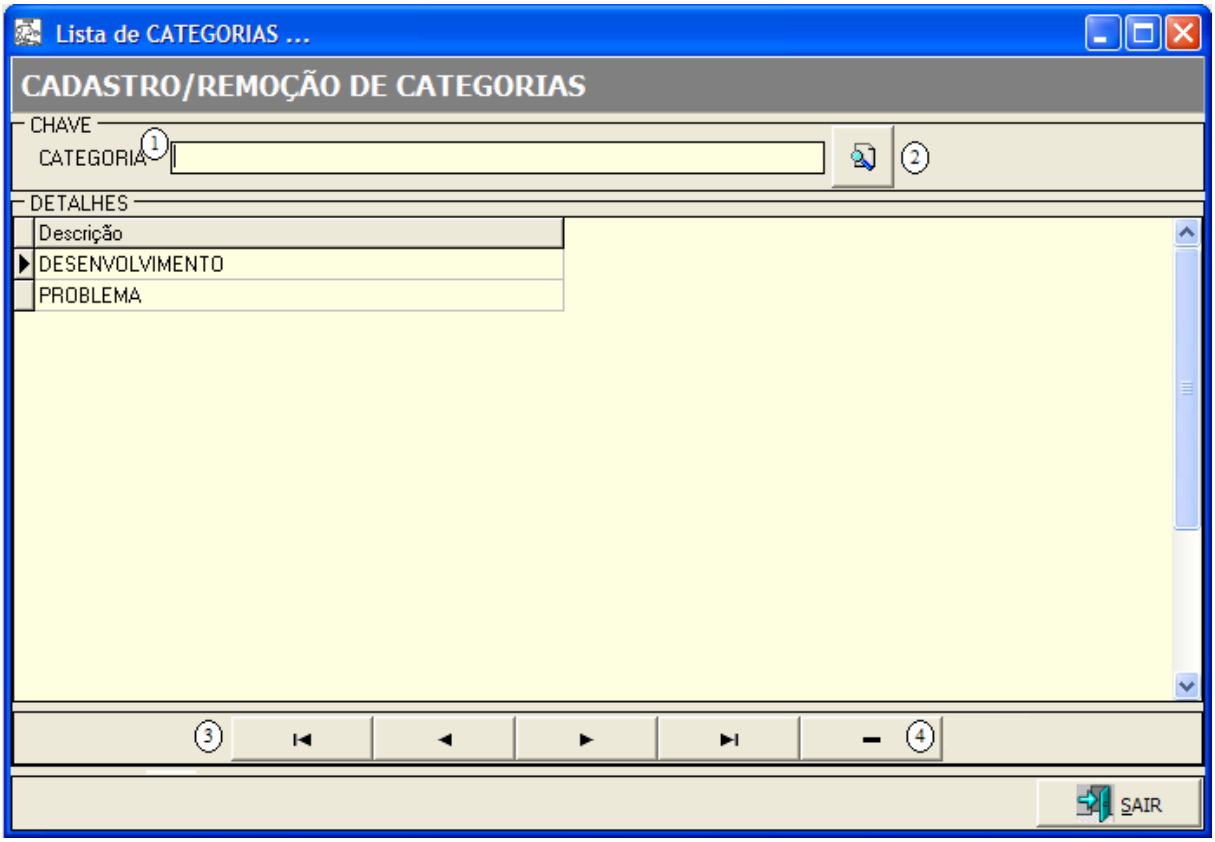


Figura 12 – Cadastro/Remoção de categorias

4.3.3.1.4 Lista de keywords

As palavras chaves ou *keywords* como foi visto no item 2.1.2, devem ser incluídas ou excluídas através deste item do menu, conforme figura 13.

Para incluir uma nova *keyword*, seleciona a categoria identificada pelo número um (1), digite a mesma no campo *keyword* (2), em seguida clique no botão incluir (3). Se a mesma já estiver sida incluída anteriormente o sistema exibe uma mensagem informando ao usuário que a mesma já se encontra cadastrada.

Para excluir uma *keyword*, o usuário deve clicar sobre a palavra desejada (4) ou digitá-la (2), e no botão excluir (5).

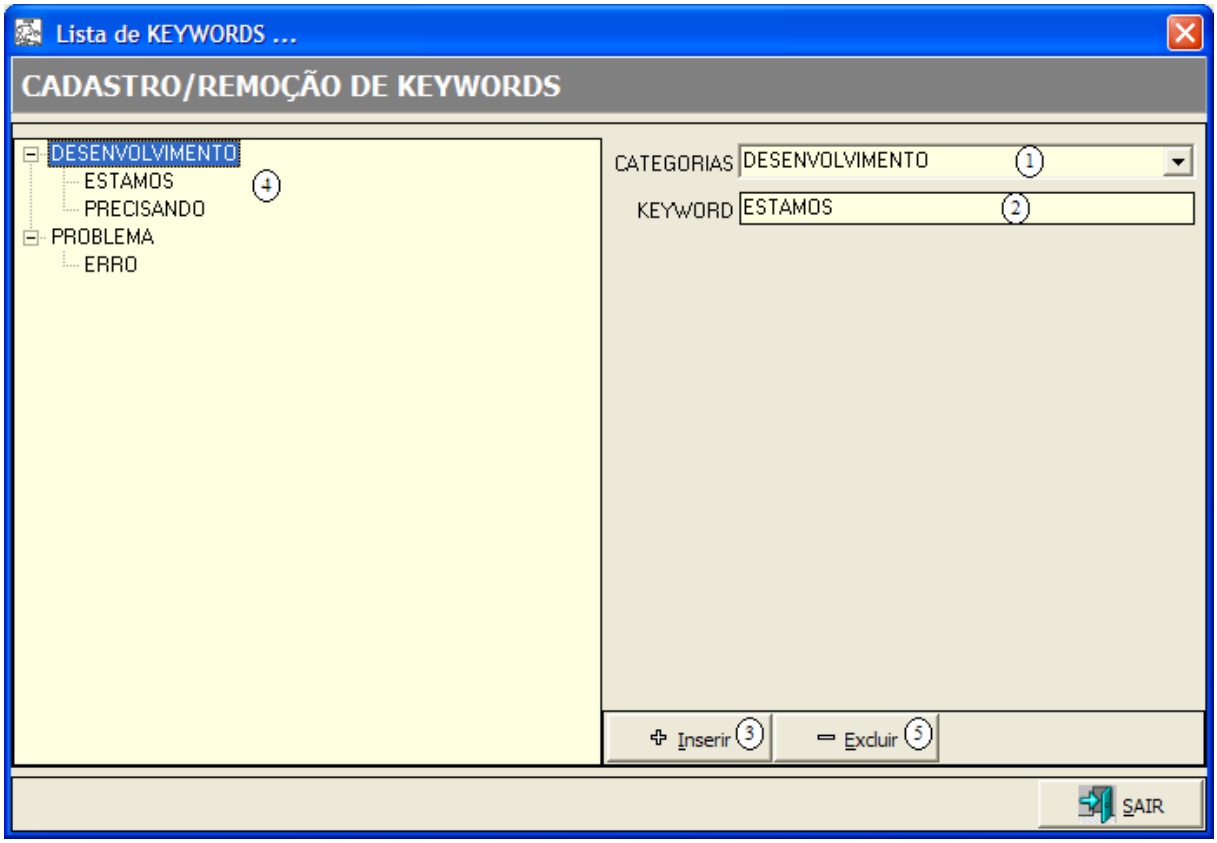


Figura 13 – Cadastro remoção de *keywords*

4.3.3.1.5 Classificação dos registros

Nesta etapa, conforme figura 14, são classificados todos os registros lidos das fichas de atendimento (FA), que estão armazenados em memória após a carga de dados e estes são classificados conforme as palavras chaves da cada categoria.

Para fazer a classificação, bastar clicar no botão identificado pelo número um (1).

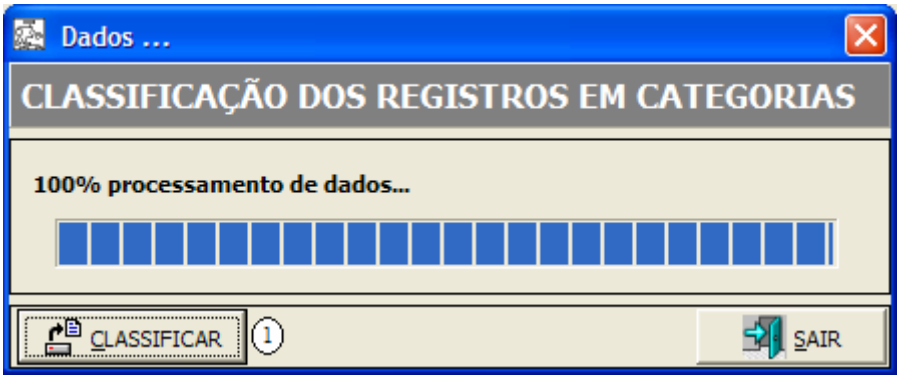


Figura 14 – Classificação dos registros em categorias

4.3.3.1.6 Pesquisar

O menu Pesquisar esta dividido em três opções para o usuário:

a) pesquisa por frase: o usuário pode pesquisar uma frase que estará contida nos registros lidos a partir da carga de dados, conforme figura 15. O usuário poderá digitar uma frase no item identificado por um (1), em seguida irá clicar no botão pesquisar (2), após isto o software irá buscar todos os registros com a frase selecionada;
b) pesquisa por palavra: o usuário pode pesquisar uma palavra chave (*keyword*) contida nos registros lidos a partir da carga de dados, conforme figura 16. O usuário poderá selecionar uma palavra no item identificado por um (1), em seguida irá clicar no botão pesquisar (2), após isto o software irá buscar todos os registros com a palavra selecionada;
c) lista de freqüência: o usuário poderá ver a freqüência com que as palavras aparecem nos registros lidos, conforme figura 17. A freqüência poderá ser visualizada pelas *stopwords*, *keywords* e pelas palavras lidas conforme identificado pelo número um (1).

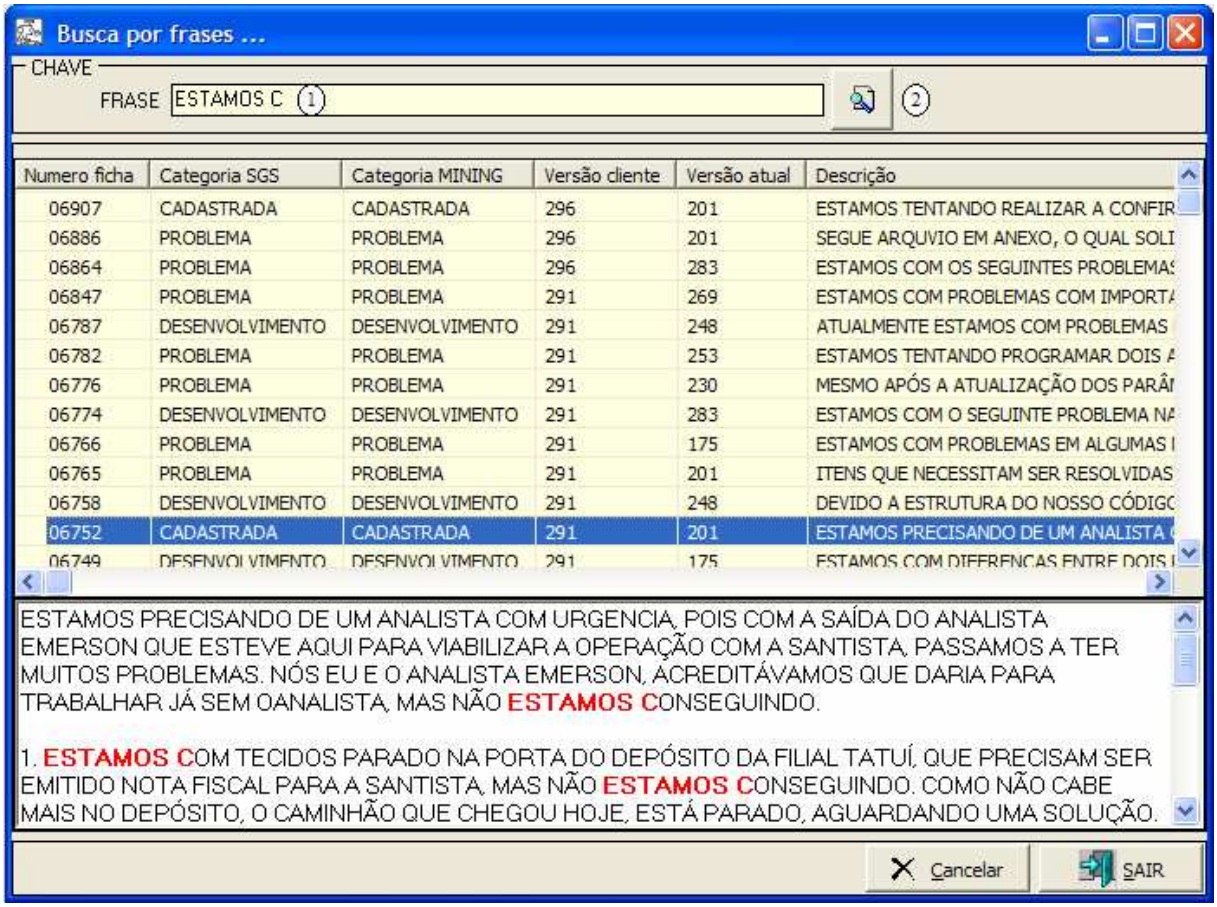


Figura 15 – Busca por frase

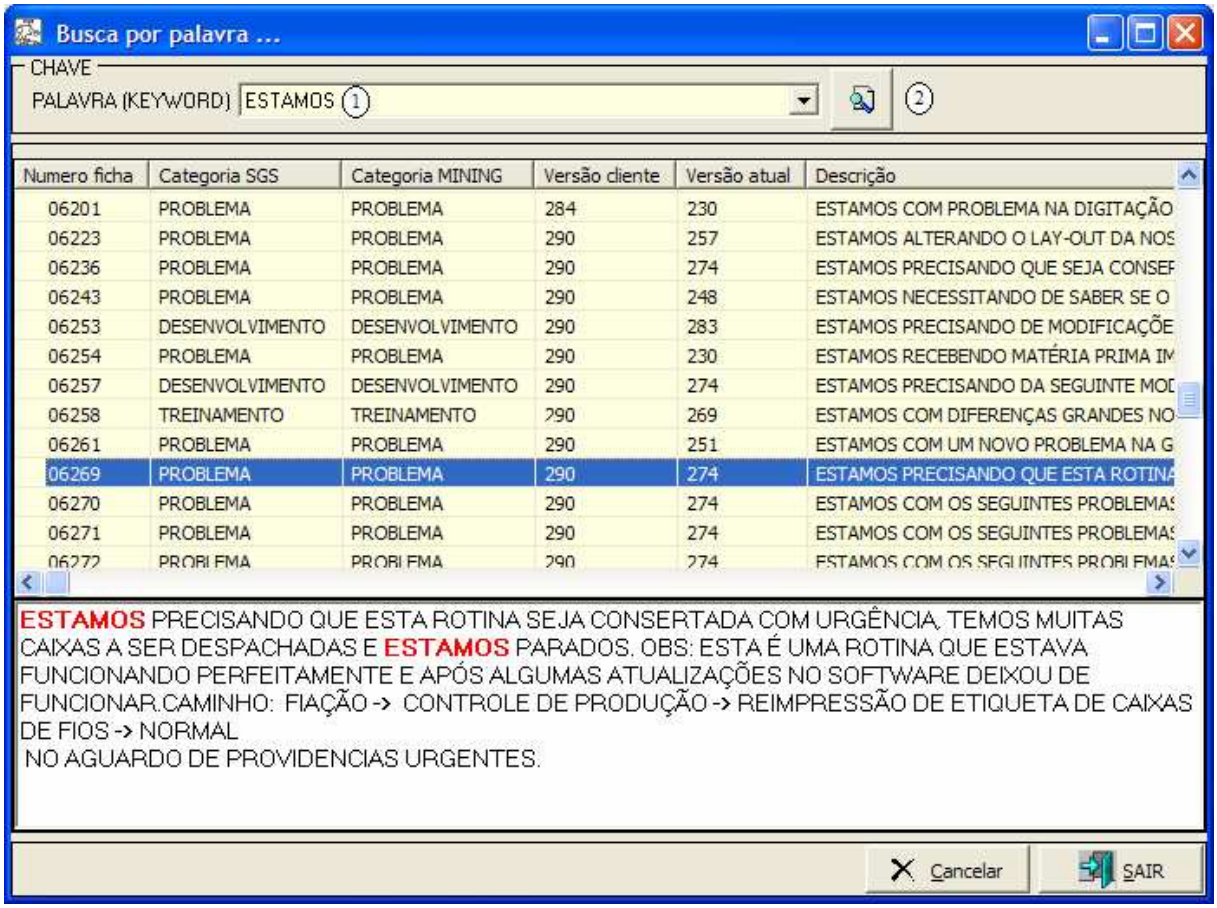


Figura 16 – Busca por palavra

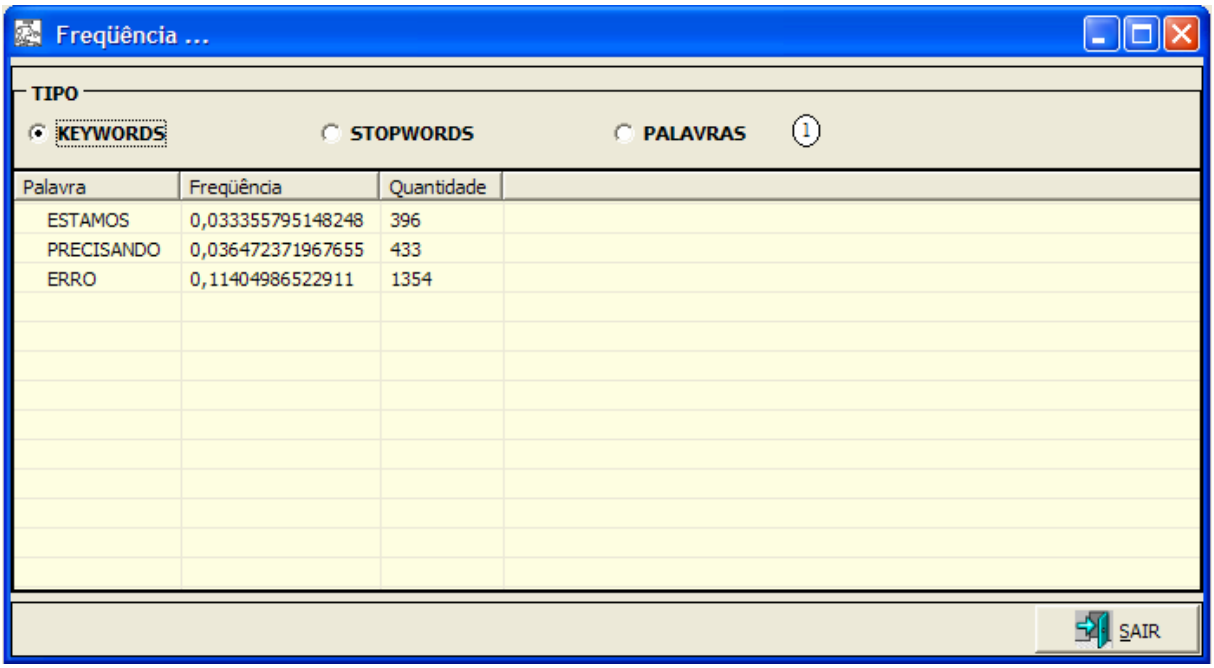


Figura 17 – Consulta da freqüência

4.3.3.1.7 Gráficos

Nesta etapa conforme figura 18, o usuário poderá visualizar os registros classificados no item 4.3.2.1.5 de acordo com as palavras chaves (*keywords*) de cada categoria.

Para gerar a visualização gráfica, o usuário poderá escolher o gráfico pela categoria ou pelas palavras chaves de cada categoria, identificado pelo número um (1).

Caso selecionado o tipo da consulta por categoria, o mesmo deverá clicar no botão pesquisar (4). Se for selecionado o tipo de consulta por *keywords*, o usuário deverá escolher a categoria (2), depois a quantidade de palavras da categoria o mesmo irá querer visualizar no gráfico (3) e clicar no botão pesquisar (4).

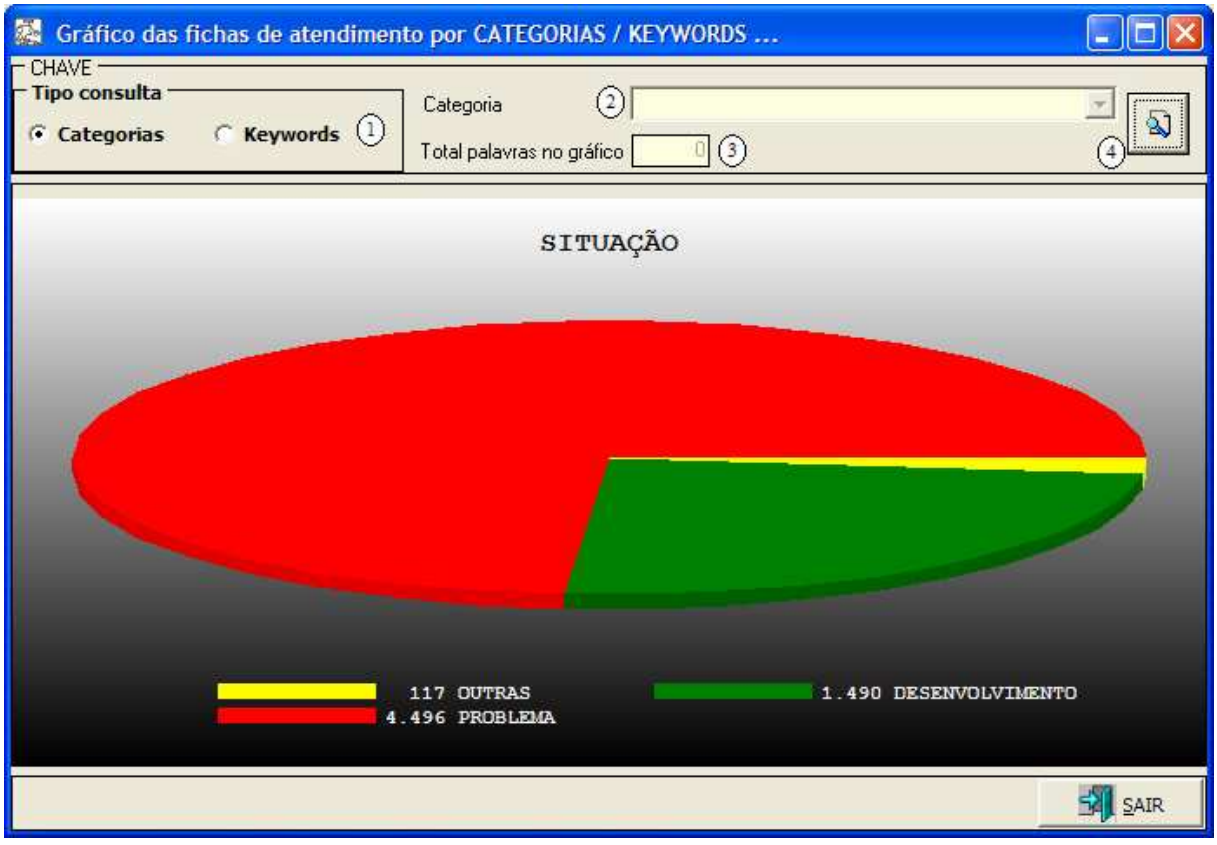


Figura 18 – Gráfico das categorias/*keywords*

4.3.3.1.8 Ajuda

Auxilia nas dúvidas dos usuários explicando o funcionamento do sistema, conforme figura 19.

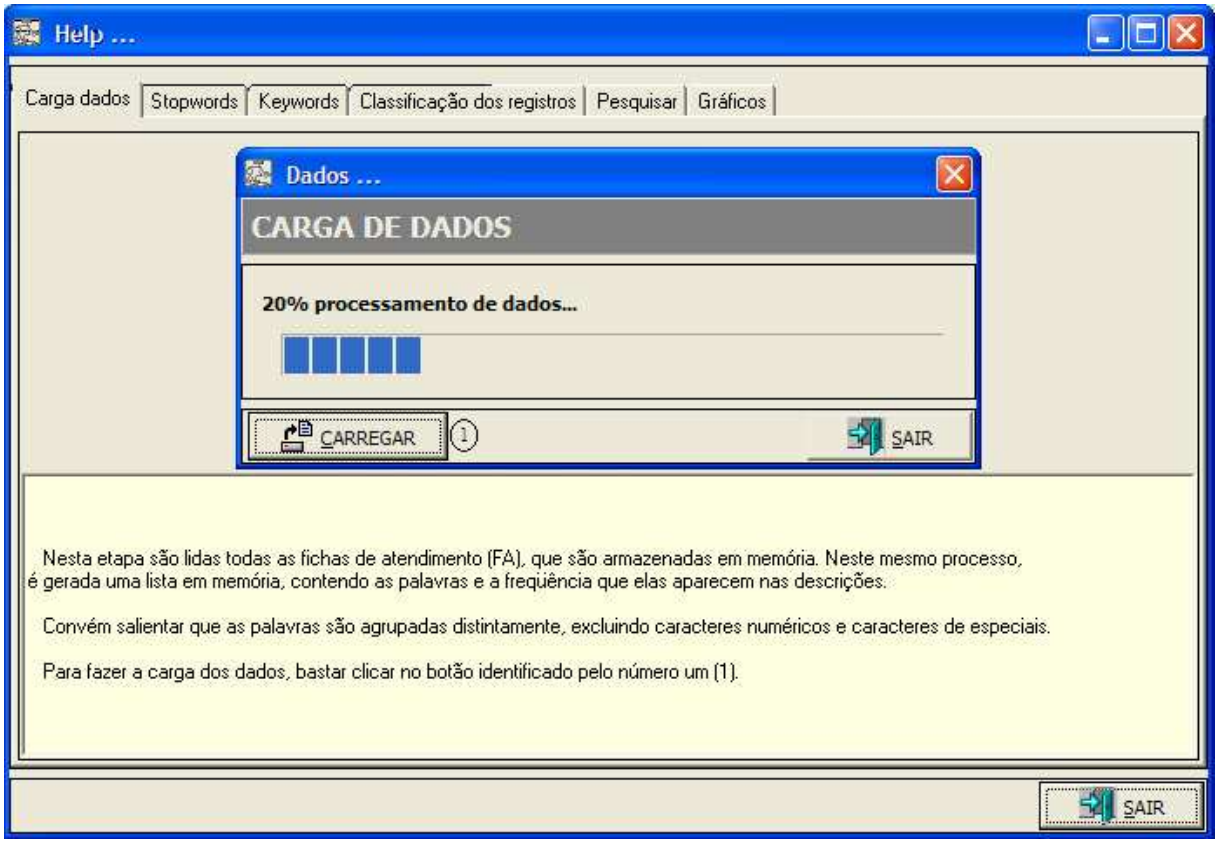


Figura 19 – Ajuda do software

4.4 RESULTADOS E DISCUSSÃO

Os resultados foram satisfatórios em relação ao que foi estudado e implementado. Confrontando as classificações geradas pelo software, pode-se observa que 85 % dos registros foram classificados corretamente.

Conforme figura 20, pode-se observar que ocorreram dois casos, identificados pelo Número da ficha 6.700 e 6.752 onde os atendentes do suporte, cadastraram as duas como sendo um problema de software. Após a classificação dos registros, seguindo os passos da metodologia CRISP-DM aplicados no desenvolvimento deste trabalho, o MINING OF INFORMATION classificou a ficha 6.700 como sendo um desenvolvimento e a ficha 6.752 como sendo uma consultoria de sistema, conforme figura 21.

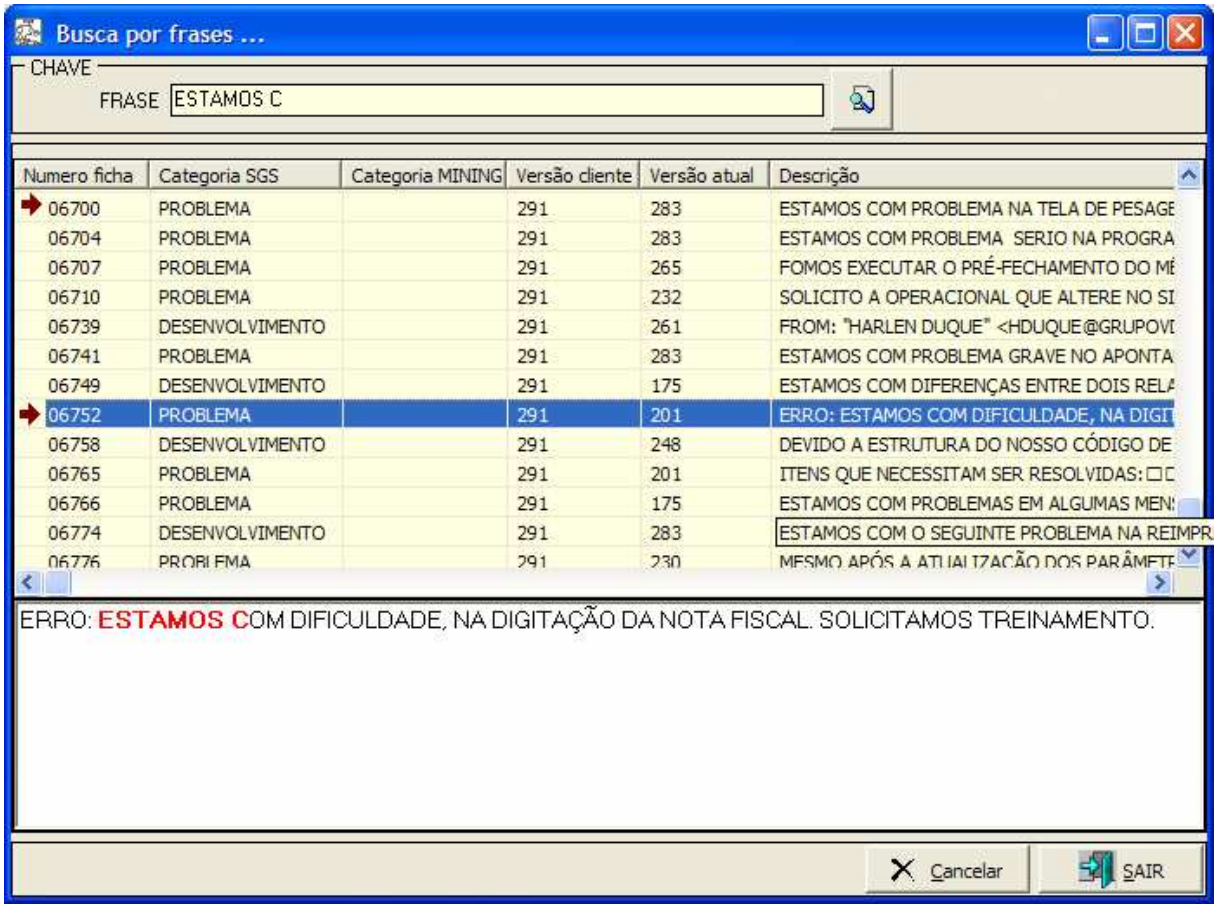


Figura 20 – Problema proposto

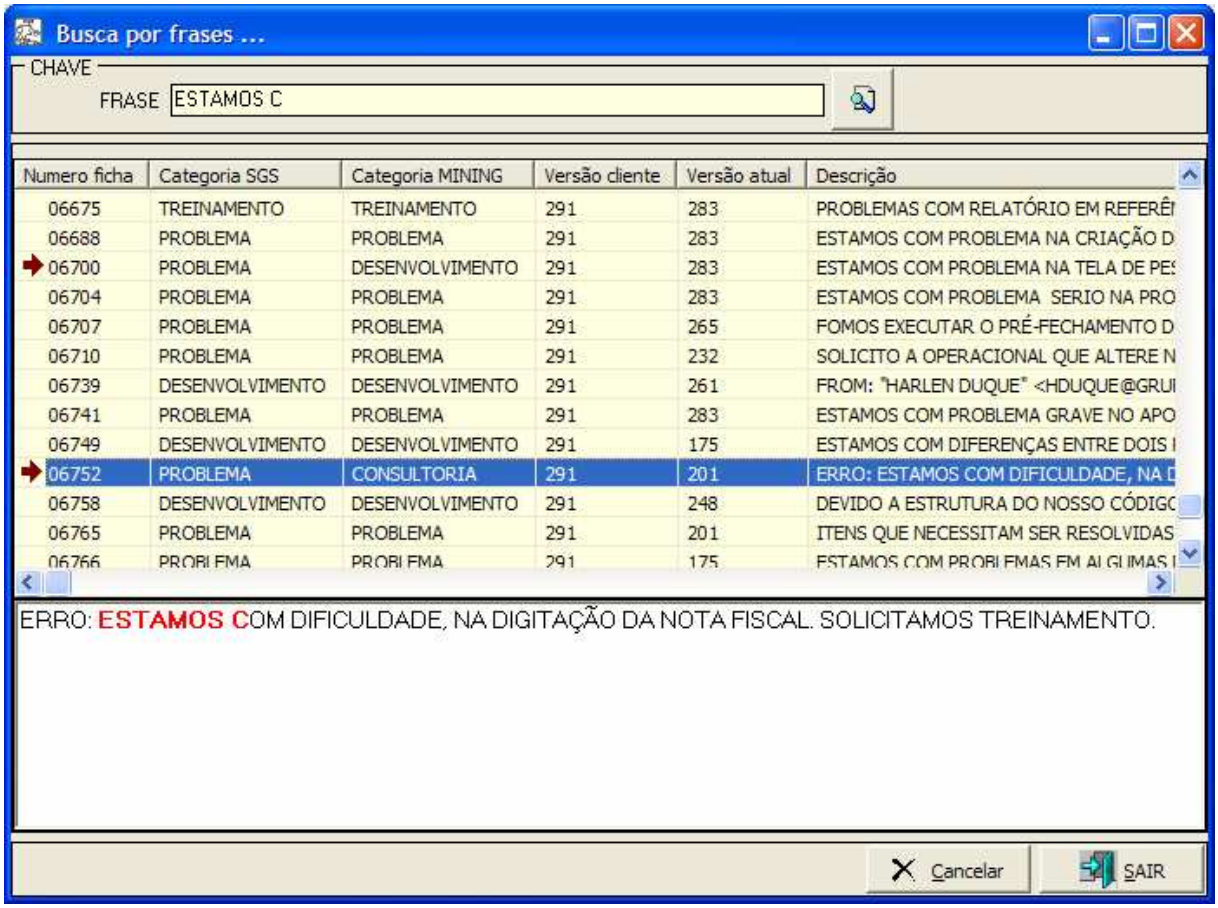


Figura 21 – Resolução do problema proposto

Outro exemplo seria o item 4.3.2.1.7 descrito neste trabalho, que apresenta graficamente a quantidade de registros por categoria. Conforme figura 18 p. 44, pode-se ver que de 6.103 registros lidos, 4496 são de problemas com as versões enviadas para os clientes. E 40% destes registros são de problemas com o software (telas, erros de manipulação dos dados, etc). Uma sugestão para melhor atender os clientes, seria criar um setor de qualidade dentro da empresa. Este setor seria responsável por testar as principais funcionalidades do software e detectar os erros antes de enviar a versão para o cliente.

Na dissertação de Uber (2004), também ficou claro que é a técnica utilizada mostrou-se eficiente detectando ocorrências policiais cadastradas de forma equivocada.

5 CONCLUSÕES

As empresas possuem uma grande quantidade de informação disponível para análise. Essa análise torna-se inviável caso não seja realizada com o auxílio de técnicas e ferramentas computacionais. O local onde as informações são coletadas também merece atenção especial, pois ela deve ser confiável. As fontes mais confiáveis são aquelas provenientes de bases de dados cujas informações são compiladas.

Na etapa de avaliação do software desenvolvido, o mesmo mostrou-se adequado e eficiente apontando as Fichas de Atendimento que estavam com a Situação incorreta em relação ao texto contido na descrição. Chegando ao término da construção desse software, pode-se salientar que os objetivos do trabalho foram atingidos, pois minimizou-se os esforços dos gerentes e diretores na determinação de tarefas e prioridades.

O objetivo desse trabalho foi mostrar como as técnicas de descoberta de conhecimento em texto, descoberta de conhecimento em base de dados, recuperação de informações e extração das informações existentes. Para atingir esse objetivo foram identificadas as técnicas e métodos de descoberta de conhecimento.

Text Mining pode ser muito útil para apoiar processos de tomada de decisão. As pesquisas são recentes, e o interesse em sua realização tem sido cada vez maior.

É importante salientar que as metodologias apresentadas nesta pesquisa são colocadas por seus idealizadores como propostas e merecem aperfeiçoamento.

5.1 EXTENSÕES

Durante a elaboração desta monografia, e implementação do software alguns pontos deixaram de ser cobertos. Isso ocorre não só neste, mas em todo trabalho científico, pois nem sempre todos os problemas de desenvolvimento podem ser resolvidos de uma única vez. Optou-se em resolver os pontos mais importantes para a solução do problema.

Devida a complexidade da maioria dos métodos optou-se pelo método mais rápido de implementação, não focando a otimização do algoritmo.

Como possíveis extensões desse trabalho enumera-se:

a) automatizar a forma de seleção de palavras chaves (*keywords*);

b) permitir que outros formatos de textos (MS-WORD, Acrobat, HTML, XML, e outros) sejam utilizados, bem como outros bancos de dados (SQL Server, MySQL, e outros);
c) permitir que dados não-estruturados possam ser utilizados, possibilitando ao usuário utilizar textos que contenham delimitadores;
d) implementar outras técnicas de mineração, permitindo ao usuário uma comparação entre os métodos, identificando o melhor método;
e) implementar técnica referente a árvore de decisão.

REFERÊNCIAS

CARVALHO, L. A. V. – **Data mining:** a mineração de dados no marketing, medicina, engenharia e administração. São Paulo: Érica, 2001.

CHAVES. Marcirio Silveira. **Um estudo e apreciação sobre algoritmos de *stemming.*** In: JORNADAS IBEROAMERICANAS DE INFORMÁTICA, 9., 2003, Cartagena de Indias, Colômbia. Disponível em: <http://xldb.di.fc.ul.pt/~mchaves/pg_portugues/public/stemming.pdf> . Acesso em: 16 out. 2004.

CRISP-DM 1.0. Winter Park, Florida, [2000]**.** Disponível em: <http://www.crisp-dm.org>. Acesso em: 20 mar. 2004.

DIXON, Mark. **An overview of document mining technology.** Australia., 1997. Disponível em: <http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps>. Acesso em: 02 jun. 2004.

FAYYAD, U. **Advances in knowledge discovery and data mining**. Cambridge: MIT Press, 1996.

FAUSETT, L. **Fundamentals of neural networks.** Florida: Prentice Hall, 1994.

GONÇALVES, Márcio. **Extração de dados**. Rio de Janeiro: Axcel, 2003.

LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses**.** In: OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA), 4., 2000. **Proceedings**... Pelotas: EDUCAT, 2000. Disponível em: <http://www.inf.ufrgs.br/~wives/portugues/publicacoes.html>. Acesso em: 20 set, 2003.

NUGGEST® **Kdnuggets.com** (KD stands for Knowledge Discovery) is the leading source of information on Data Mining, Web Mining, Knowledge Discovery and Decision Support Topics, New York, NY**,** [2001]. Disponível em: <http://www.kdnuggets.com/polls/2003/data_mining_techniques.htm>. Acesso em: 16 jun. 2004.

ROMÃO, Wesley. **Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia**. 2002. 253 f. Tese (Doutorado em Engenharia da Produção) – Universidade Federal de Santa Catarina, Florianópolis.

SANCHEZ , A. **Definicion e historia de los corpus**. In: A. SANCHEZ et al (Org.). CUMBRE - Corpus Linguistico de Espanol Contemporâneo. Madrid: SGEL, 1995.

SANTOS, Maria Angela Moscalewski Roveredo dos. **Extraindo regras de associação a partir de textos**. 2002. 51 f. Dissertação (Mestrado em Informática Aplicada) - Universidade Católica do Paraná, Curitiba.

SILVA, Edilberto M. **Descoberta de conhecimento com o uso de *text mining*:** cruzando o abismo de Moore. 2002. 175 f. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação) - Universidade Católica de Brasília, Brasília.

TAN, A H. ***Text mining***: the state of the art and the challenges. Singapore: Kent Ridge Digital Labs, 1999. Disponível em: <http://citeseer.ist.psu.edu/cache/papers/cs/12174/http:zSzzSztextmining.krdl.org.sgzSzdocszSztext_mining_KDAD99.pdf/tan99text.pdf>. Acesso em: 30 mar. 2004.

UBER, Jacqueline. **Validação das ocorrências policiais com base na descrição textual do boletim de ocorrência utilizando *text mining***. 2004. 70 f. Projeto de Dissertação (Mestrado em Ciências da Computação) – Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis.

WIVES, Leandro K. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva.** 2000. 116 f. Exame de qualificação (Mestrado em Ciências da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre.

WIVES, Leandro K. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de "*Clustering*".** 1999. 84 f. Dissertação (Mestrado em Informática) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.