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

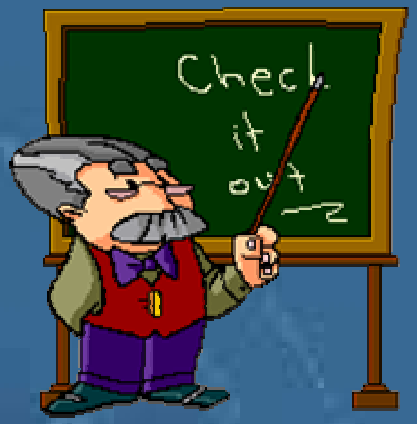
TEXT MINING

Aluno
José Lino Uber

Orientador
Paulo Roberto Dias

Dezembro/2004

Roteiro

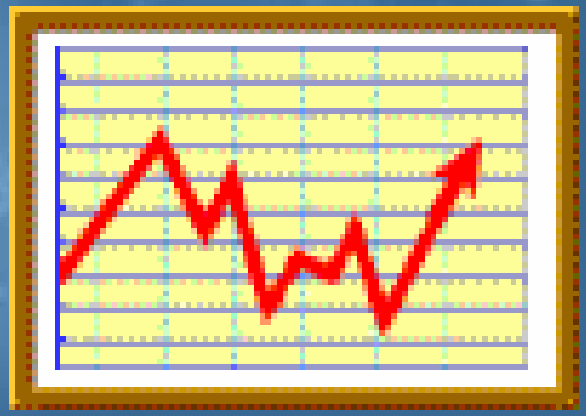


- Introdução
- Objetivo
- Conceitos
- Motivação / Tipos de informações que podem ser filtradas
- Metodologia
- Técnicas
- Desenvolvimento
- Especificação
- Implementação
- Resultados e discussão
- Conclusão
- Extensões

Introdução

Organizações e pessoas acumulam grandes volumes de informações textuais e não sabem como gerenciá-las de forma eficiente, perdendo tempo e conhecimento

As ferramentas de Text Mining podem ajudar a melhorar o negócio através da análise de informações textuais, oferecendo conhecimento novo e útil



Introdução

Vários fatores têm contribuído para o grande volume de informações armazenadas em banco dados. A queda nos custos de armazenamento pode ser vista como a principal causa deste crescimento. Outro fator é a disponibilidade de computadores de alto desempenho a baixo custo.

Para se obter conhecimento nesta bases de dados, existem algumas formas de realizar a mineração dos dados. Neste trabalho será estudada a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

Introdução

A metodologia CRISP-DM 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.

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,** esta foi a variável analisada.

Objetivo

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.

Conceitos

- O que é Text Mining?

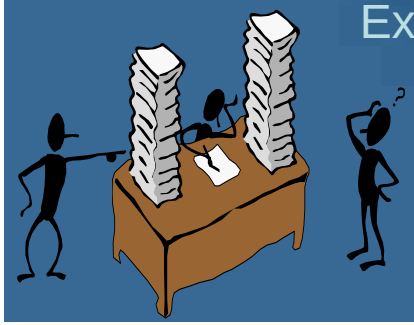
  É uma tecnologia para a análise de textos que permite diminuir a "sobrecarga de informações", descobrir padrões, associações e regras, e realizar análises qualitativas ou quantitativas.

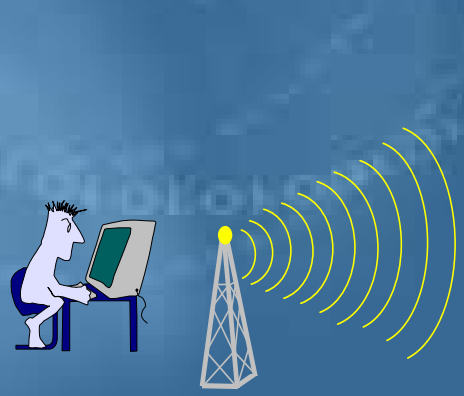
- Qual a sua importância?

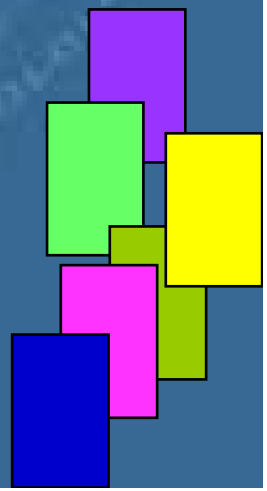
  Auxiliar na busca de informações específicas, agilizando processos com uso de inteligência.

- Permite

  Recuperação de informações
  Extração de dados
  Classificação
  Extração de resumos de textos







Conceitos

- *Stopwords: São* palavras que não demonstram a mínima relevância, não possuem representatividade alguma. Exemplo as vogais.

- *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.

Conceitos

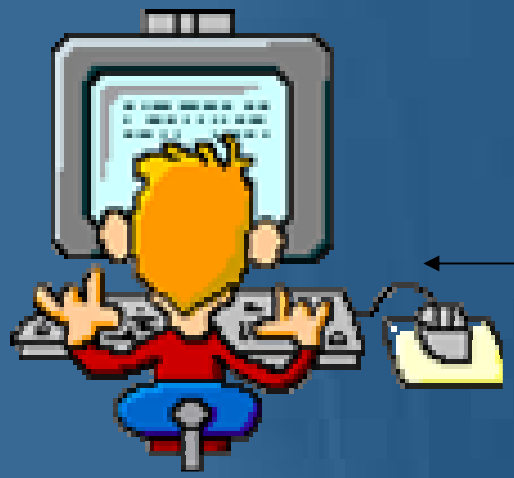
- *Collocations: São* agrupamentos de palavras onde o significado é composto pela soma dos significados das partes mais algum componente semântico adicional. Exemplo: guarda-volume, onde as duas palavras juntas tem um significado. Separadas representam duas outras coisas.

- *Stemming:* consiste em reduzir todas as palavras ao mesmo *stem*, por meio da retirada dos afixos da palavra, permanecendo apenas a raiz dela. Por exemplo, quando a palavra “referência” é transformada no *stem* “referênc”, ao invés do *stem* considerado correto “refer”.

Motivações

Sobrecarga de Informações

Motivações



Desenvolvimento

- Malharia Circular
- Malharia Retilínea
- Tinturaria

Usuário

Consulta / Análise

Universo de registros

Que tipos de informações podem ser filtradas ?

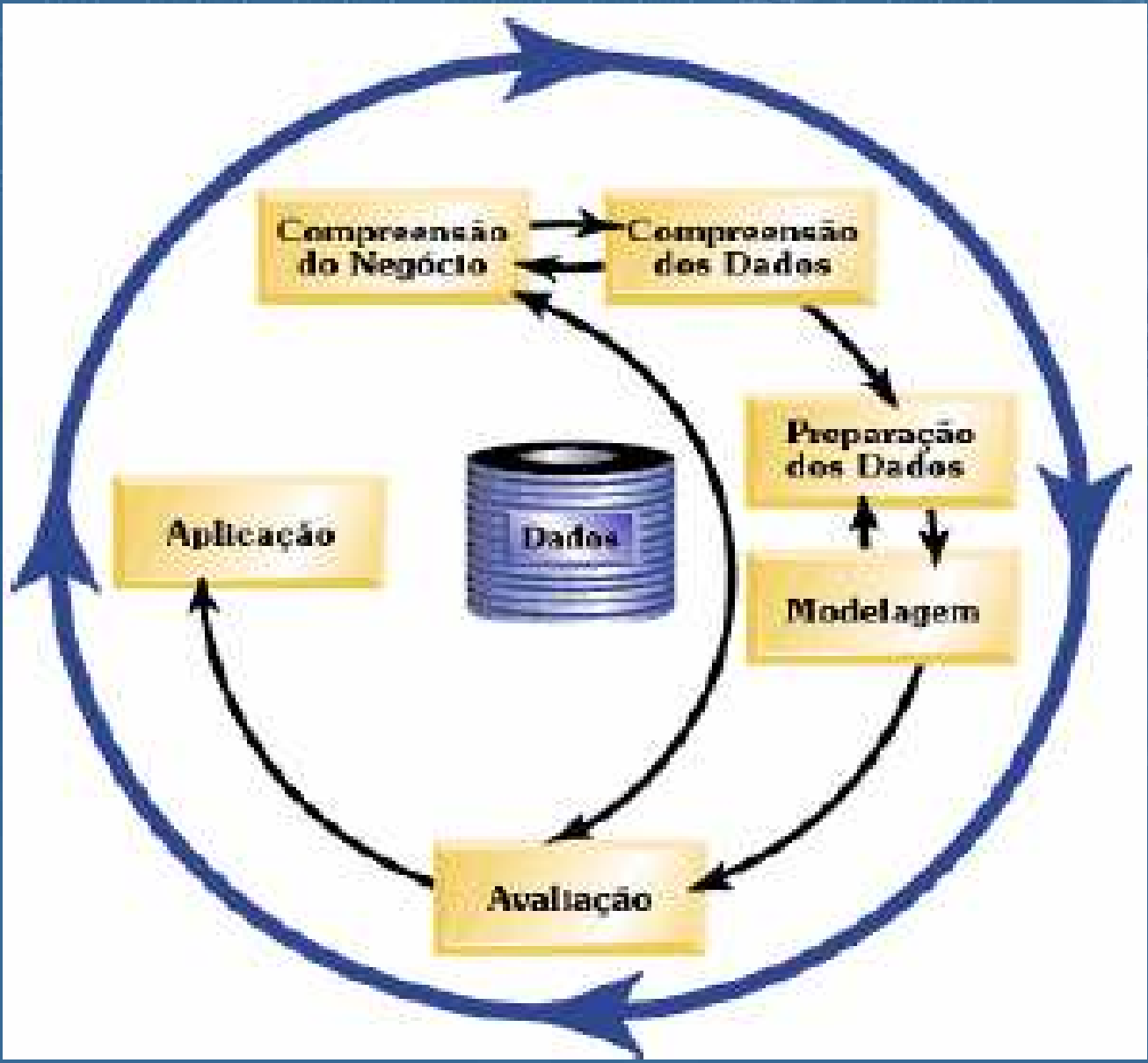
- E-mails
- Textos livres resultantes de pesquisas
- Arquivos eletrônicos (txt, doc, pdf)
- Páginas Web
- **Campos textuais (memos) em Bancos de Dados**
- Documentos eletrônicos, digitalizados a partir de papéis
- Outros ...

Metodologia CRISP-DM para DCBD

- 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.

- DCBD – Descoberta de conhecimento em bando de dados

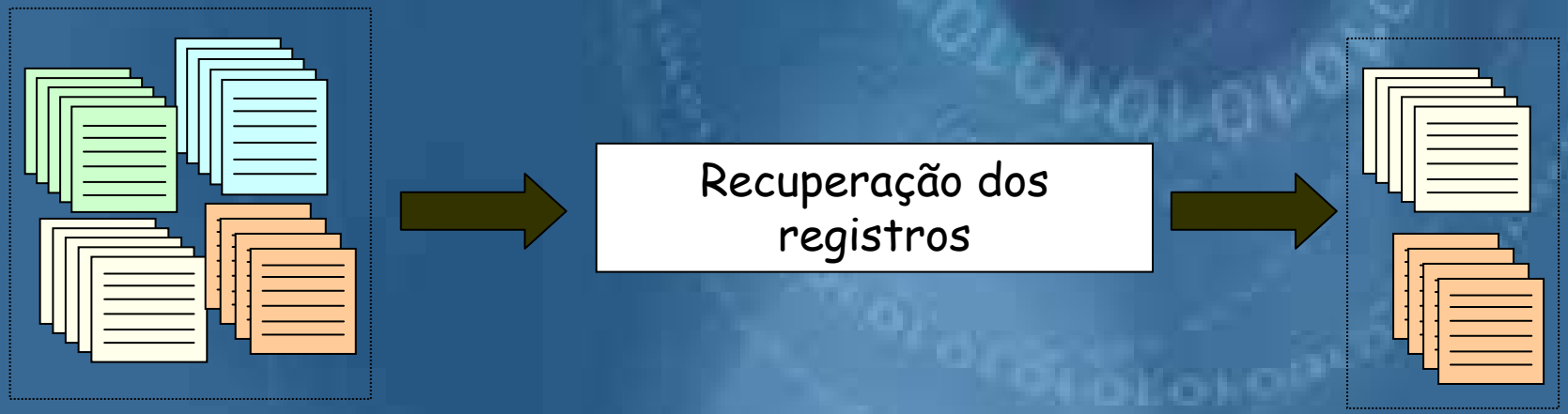
Processo para DCBD segundo CRISP-MD



Metodologia - Etapas

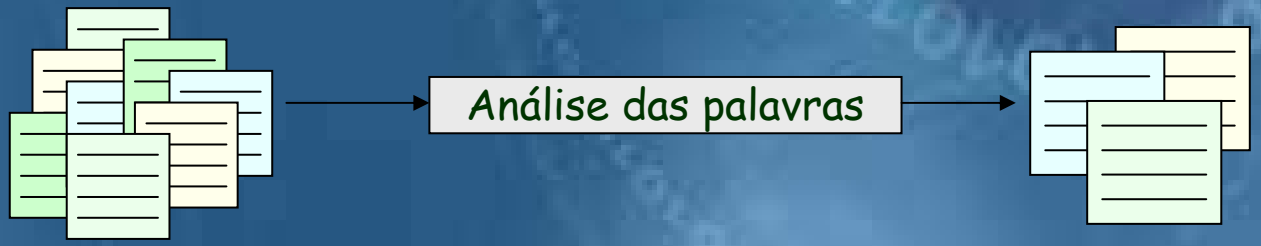
1º Passo: Compreensão do negócio

(Ficha de Atendimento)



Metodologia – Etapas

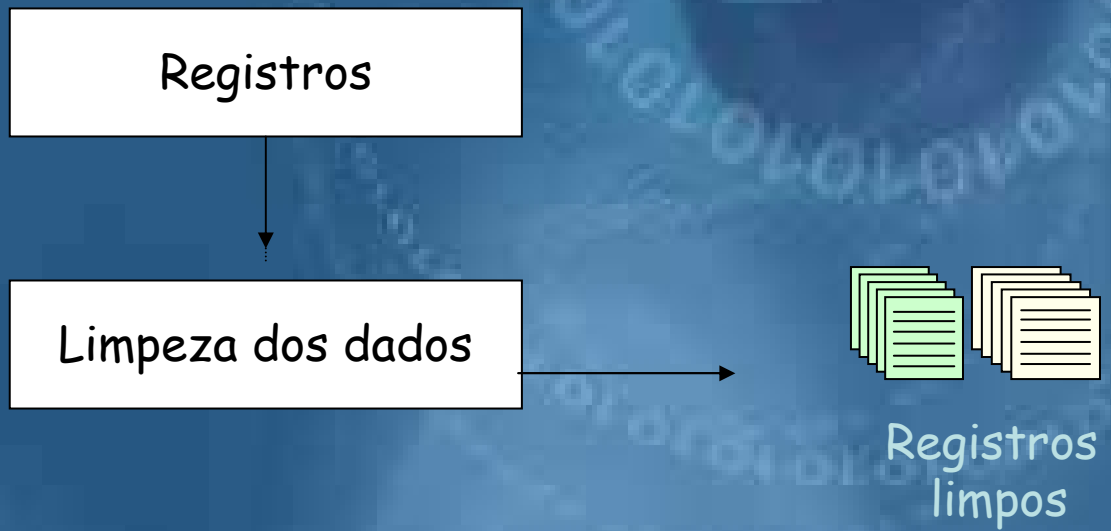
2º Passo: Compreensão dos dados



Metodologia – Etapas

3º Passo: Preparação dos dados

Limpeza dos dados



Metodologia – Etapas

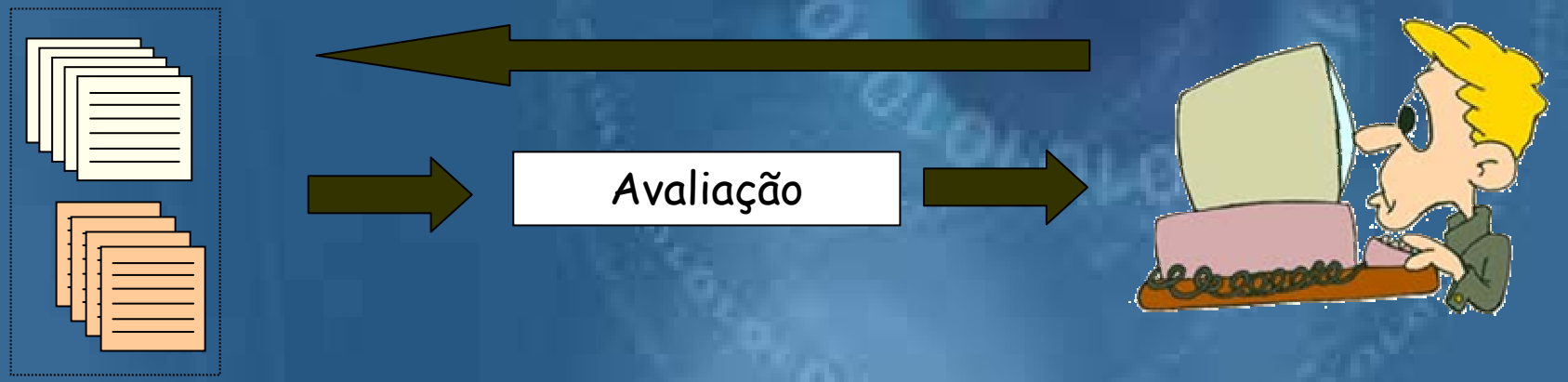
4º Passo: Modelagem

Algoritmo de mineração de textos

Metodologia – Etapas

5º Passo: Avaliação

Revisão dos passos seguidos



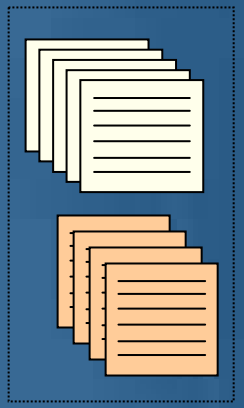
Registros classificados

Analista de Negócio ou Especialista

Metodologia – Etapas

6º Passo: Aplicação

Resultado do conhecimento obtido



Registros avaliados

Registros classificados corretamente

Técnicas

- Recuperação da informação
- Indexação automática
- Extração de informações
  - Sumarização
  - Clustering
  - **Classificação de Dados (FOCO)**

Técnicas

Recuperação da informação

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.

Técnicas

Extração de informações

Encontrar valores implícitos nos textos.

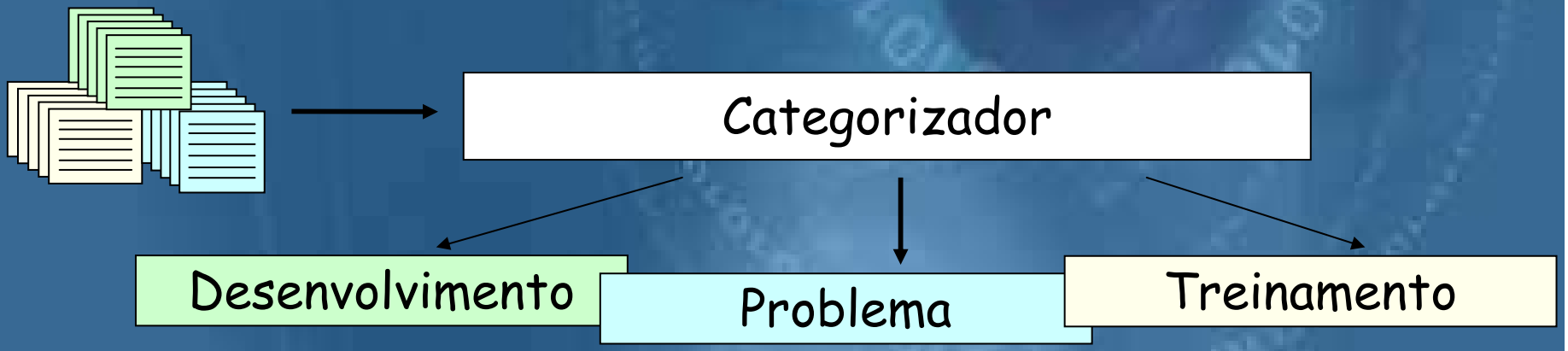
José da Silva é funcionário da Empresa ABC, reside na Rua X, número 32, na cidade de Porto Alegre e ....

Nome: José da Silva
Empresa: ABC
Endereço: Rua X, 32
Cidade: Porto Alegre

Técnicas

Classificação de Dados (FOCO)

Encontrar o assunto de um texto



Desenvolvimento

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.

- Requisitos não funcionais
  - Desempenho
  - Banco de dados
  - Visualização

Desenvolvimento

- Requisitos funcionais
  - Lista de palavras.
  - Lista de palavras excluídas (*stopwords*)
  - Criação e remoção de categorias;
  - Lista de palavras chaves (*keywords*).
  - Lista de freqüência.
  - Criação de gráfico.
  - Busca de registros por palavras.
  - Busca de frases.

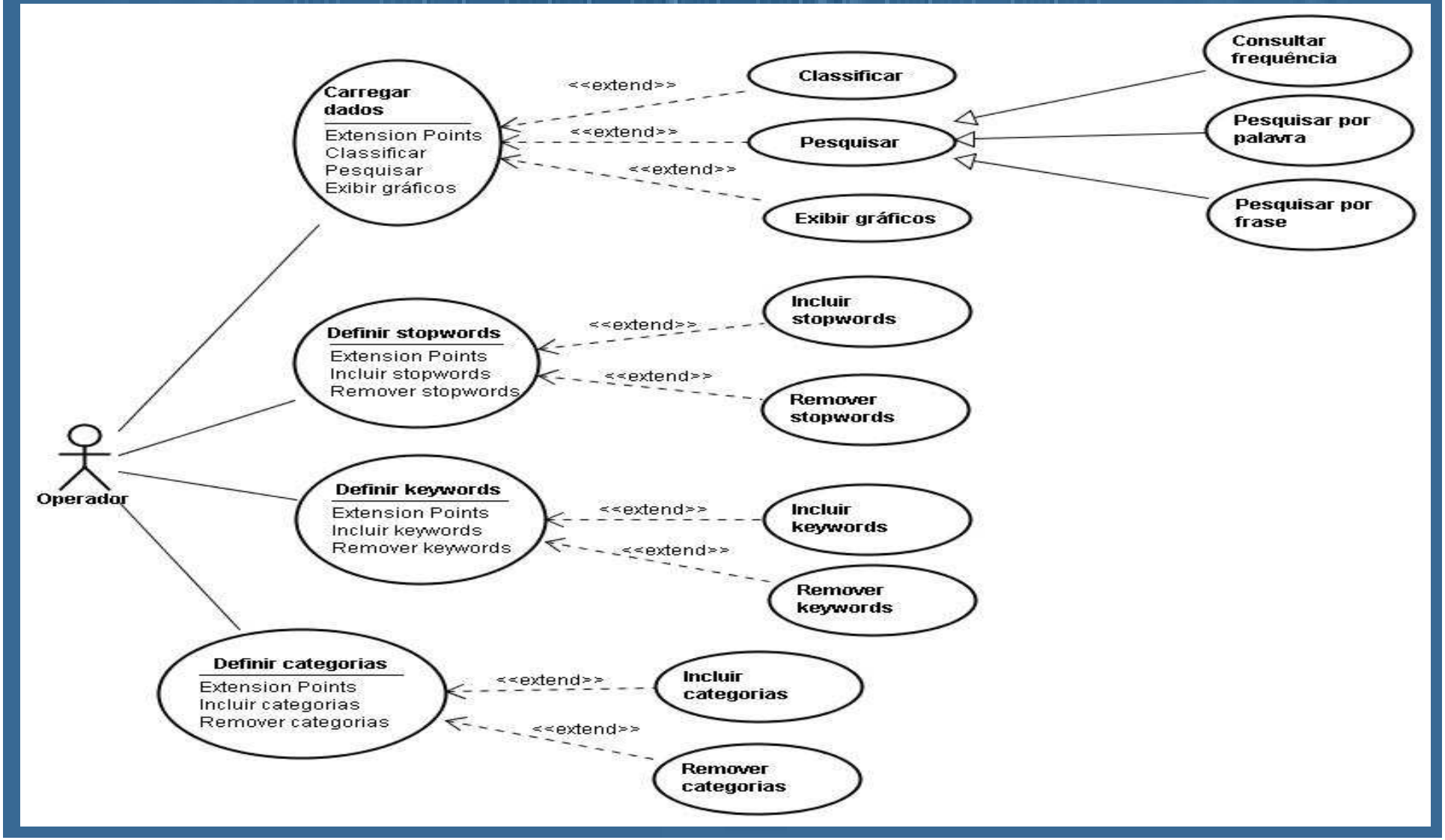
Especificação

- Seguindo a metodologia CRISP-DM, inicializou-se compreendendo o funcionamento das Fichas de Atendimento (FA). Para isto verificou-se todas as tabelas envolvidas na (FA).

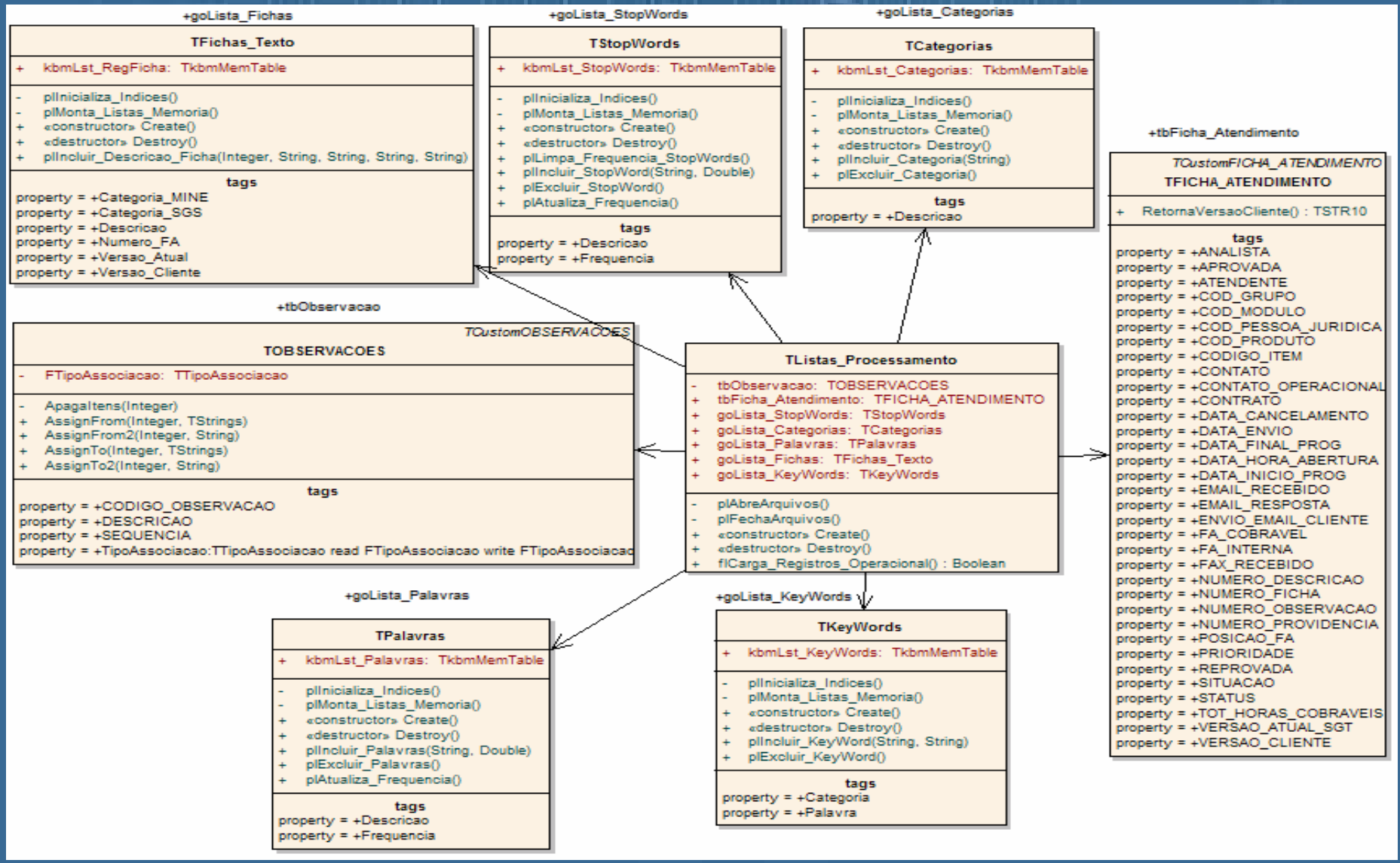
- Os dados encontravam-se limpos e organizados, sem erros de grafia, portanto a etapa de preparação dos dados não foi realizada. A modelagem foi realizada com a ajuda da ferramenta Jude Community para a especificação do Diagrama de Caso de Uso, e Enterprise Architect para especificar o Diagrama de Classes.

- Na etapa de avaliação constatou-se que o modelo mostrou-se adequado e eficiente apontando as Fichas de Atendimento que estavam com a Situação incorreta em relação ao texto contido na descrição.

Especificação – Caso de uso



Especificação - Classe



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.

O software foi batizado de ***MINING OF INFORMATION***.

Permite que o usuário obtenha conhecimento dos textos de forma interativa.

Implementado na linguagem de programação DELPHI. O ambiente de programação adotado foi o Borland Delphi 6.0, devido as facilidades de construção de interfaces.

Resultados e discussão

- Falta de referências bibliográficas

Conclusões

- Text Mining pode ser muito útil para apoiar processos de tomada de decisão.
- As pesquisas em Text Mining são recentes, e o interesse em sua realização tem sido cada vez maior.
- Com a construção desse software, minimizou-se os esforços dos gerentes e diretores na determinação de tarefas e prioridades.

Extensões

- 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);
- Permitir que dados não-estruturados possam ser utilizados, possibilitando ao usuário utilizar textos que contenham delimitadores;
- Implementar outras técnicas de mineração, permitindo ao usuário uma comparação entre os métodos, identificando o melhor método;
- Automatizar a forma de seleção de palavras chaves (*keywords*);
- Implementar técnica referente a árvore de decisão.

SOFTWARE IMPLEMENTADO

MINING OF INFORMATION