Universidade Gama Filho

Taeko Suda Mazakina

MINERAÇÃO DE DADOS NO JUDICIÁRIO FEDERAL

São Paulo

2009

Taeko Suda Mazakina

MINERAÇÃO DE DADOS NO JUDICIÁRIO FEDERAL

Taeko Suda Mazakina

MINERAÇÃO DE DADOS NO JUDICIÁRIO FEDERAL

Monografia
Apresentada

à Universidade Gama Filho como requisito parcial para obtenção do título de especialista em Engenharia de Software.

_____
Orientador: Prof. Júlio Celso Noguchi

Taeko Suda Mazakina

MINERAÇÃO DE DADOS NO JUDICIÁRIO FEDERAL

Monografia julgada e aprovada:

_____

Prof. Orientador: Júlio Celso Noguchi

Resumo

Este trabalho tem o objetivo de elucidar a importância da utilização da mineração de dados no poder judiciário. Expondo os benefícios que a metodologia traria através da extração de informações existentes no grande volume da base de dados, auxiliando na agilidade em extrair as informações fundamentais para disponibilizar ao usuário final, e também auxiliar nas decisões processuais.

Tem como meta apresentar os principais conceitos da tecnologia de KDD, muito difundida atualmente, com ênfase na análise das funcionalidades, principais técnicas e metodologias utilizadas na mineração de dados para a obtenção de conhecimentos, que poderiam ser utilizadas como ferramenta de apoio à tomada de decisão.

Palavras-chave: Mineração de Dados, data mining, KDD (descoberta de conhecimento na base de dados), sistema de apoio à tomada de decisão, Judiciário Federal, Tribunal Regional Federal.

Lista de abreviaturas e siglas

AG	- Algoritmos genéticos
BI	- Business Inteligence
CJF	- Conselho de Justiça Federal
DSS	- Decision Support System
KDD	- Knowledge Discovery in Database – Descoberta de Conhecimento em Banco de dados
MINSUP	- Suporte Mínimo
MINCONF	- Medida Confiança
RNA	- Redes Neurais Artificiais
SAD	- Sistema de apoio à Decisão
TFR	- Tribunal Federal de Recursos
TRF	- Tribunal Regional Federal

Lista de ilustrações e tabelas

Figura 1: Pirâmide do Conhecimento.....23
Figura 2 – Visão Geral de um processo de Data Mining.....25
Figura 3 – Exemplo Fase I – Árvore Decisão.....31
Figura 4 - Exemplo de uma árvore de classificação.....33
Figura 5 – Configuração básica de um SAD.....49
Figura 6 – Exemplo Sistema de Apoio à Decisão.....52

Tabela 1 - Estratégias de Data Mining.....28
Tabela 2 – Amostra de Leitores.....53

Sumário

1. Introdução 09
    1.1- Tema 09
    1.2- Delimitação do tema 09
    1.3- Apresentação/Formulação do problema 09
    1.4 – Referencial Teórico 09
    1.5 – Justificativa 10
    1.6 – Objetivos 10
        1.6.1 – Objetivo Geral 10
        1.6.2 – Objetivos Específicos 10
    1.7 - Metodologia do trabalho 11
    1.8 - Descrição dos capítulos 11
2 – Referencial Teórico 11
Capítulo 1 – Introdução 14
Capítulo 2 – Mineração de Dados 15
    2.1 – O processo de KDD: Conceitos básicos 15
    2.2 – Principais atividades na área KDD 16
    2.3 - Caracterização do processo de KDD 16
    2.4 – Etapas de pré-processamento 19
    2.5 – Etapa de Mineração de Dados 20
        2.5.1 – Principais objetivos de um Data Mining 20
        2.5.2 – Objetivos do Data Mining 25
        2.5.3 – A origem do Data Mining 26
        2.5.4 – Estratégias ou tarefas Data Mining 26
        2.5.5 - Algoritmo (técnicas) de Data Mining 28
            2.5.5.1 – Árvore de decisão 30
            2.5.5.2 - Técnica APRIORI 32
            2.5.5.3 – Clusterização 34
            2.5.5.4 - Técnica de regressão 36
            2.5.5.5 – Redes Neurais 37

2.5.5.6 – Lógica Nebulosa 38
2.5.5.7– Algoritmos genéticos 40
2.5.5.8 – Mining de Texto 41
2.5.5.9 – Indução de Regras 42
2.5.5.10 - Análise Estatística de séries temporais 43
2.5.5.11 – Visualização 44
2.6 – As etapas do Data Mining 44
2.6.1 – Entendimento do problema 44
2.6.2 – Entendimento dos dados 44
2.6.3 – Preparação dos dados 44
2.6.4 – Modelagem do problema 45
2.6.5 – Avaliação do modelo 45
2.6.6 – Divulgação ou publicação do modelo 45
2.7 - Algumas aplicações práticas de técnicas de data mining 45
2.8 – Etapas de Pós-Processamento 47
Capítulo 3 – Sistema de apoio à Decisão (SAD) 48
3.1 - Introdução ao Sistema de Apoio à Decisão (SAD) 48
3.2 – Modelo de Aplicação através da Mineração de dados 53
Capítulo 4 – Tribunal Regional Federal 3ª Região 55
4.1 - Histórico do Tribunal Regional Federal (TRF) 55
5 – Conclusão 57
6 - Referências bibliográficas 61

1 – Introdução

1.1 - Tema:

Mineração de dados

1.2 – Delimitação do Tema

Mineração de dados no judiciário federal

1.3 – Apresentação/Formulação do problema

O problema existente na área judiciária federal é o grande volume de dados, a dificuldade de agilizar o processo de análise dos mesmos e a elaboração da decisão, acarretando com isso uma grande morosidade.

A mineração de dados, através da utilização de suas ferramentas, poderá contribuir na agilização do apoio à tomada de decisões na esfera da justiça federal.

1.4 – Referencial Teórico

- Alessandro Marco Rosini Ângelo Palmisano, Alessandro Marco Rosini – 2003
  Administração de Sistemas de informação e a Gestão do conhecimento;

- Efraim Turban, James C.Wetherbe, Ephraim Mclean
  Tecnologia da Informação para Gestão
  Transformando os negócios na economia digital.

- Ronaldo Goldschmidt, Emmanuel Passos
  Data Mining um guia prático (2005)
  Conceitos, técnicas, ferramentas, orientações de utilização e aplicações
  Descreve um tratamento completo do tópico, cobrindo os aspectos práticos e fornecendo a base teórica.

- Fábio Vinicius Primak
  Decisões com BI (Business Intelligence)

- Roberto Ferreira Lima Silva
  E-RH em um ambiente global e multicultural

1.5 – Justificativa

A cada dia novos investimentos estão sendo realizados em pesquisa para o desenvolvimento de técnicas computacional e algoritmos para explorar estas bases. Isto permite um suporte cada vez maior aos tomadores de decisões.

Dessa forma, cabe ressaltar que o processo de descobrimento de conhecimento em base de dados (KDD) é dependente de uma nova geração de técnicas e ferramentas de análise de dados envolvendo diversas etapas. A principal etapa desse processo chama-se Data Mining ou Mineração de Dados.

A mineração de dados vai muito além da simples consulta a um banco de dados, no sentido de que permite aos usuários explorar e inferir informação útil a partir dos dados, descobrindo relacionamentos escondidos no banco de dados.

1.6 – Objetivos

1.6.1 – Objetivo Geral

Analisar ou explorar na base de dados informações, previsões sobre fatos futuros que auxiliem em processos decisórios da organização, não de competitividade por ser um órgão público, mas de tentar atingir uma agilidade dos meios de disponibilizar os serviços e atingir a satisfação do jurisdicionado.

1.6.2 – Objetivos Específicos

- Apresentar a atual situação da justiça federal sob o ponto de vista da tomada de decisão;

- Apresentar um modelo de mineração de dados que agilize o processo de apoio à tomada de decisão na justiça federal;

- Apresentar um modelo de sistema de apoio à decisão na esfera da justiça federal;

- Identificar os aspectos da mineração de dados, analisando técnica a ser utilizada para agilizar o processo de apoio à tomada de decisão na justiça federal.

1.7 - Metodologia do trabalho

O método a ser utilizado para o desenvolvimento da pesquisa, será baseado em estudos efetuados na leitura, análise e interpretação de livros, pesquisas em materiais disponíveis na Internet, para auxiliar na tarefa de analisar, interpretar e relacionar os dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto, no que tange o aproveitamento das informações existente na base de dados, visando a agilização dos resultados dentro do judiciário federal.

1.8 - Descrição dos capítulos

Capítulo 1 – Introdução
Capítulo 2 – Mineração de Dados
Capítulo 3 – Sistema de Apoio à Decisão
Capítulo 4 – Tribunal Regional Federal da 3ª Região

2 – Referencial Teórico

De acordo com Goldschmidt e Passos (2005), um sonho de todo especialista em sistemas foi o de ver as pessoas se tornarem mais racionais na tomada de decisões, pelo menos a partir do dia em que pudessem contar com um acesso rápido e fácil a esses vastos repositórios de dados. Descobrir conhecimento é extrair dos dados o que eles implicam em termos de riscos – a evitar – e oportunidades – a serem aproveitadas. Finalmente a máquina pode ajudar nas questões em que as regras do negócio devem ficar explícitas e ser bem compreendidas. E não apenas na hora de tomar cada decisão, mas também para o planejamento das atividades a médio e longo prazos.

Torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada

contexto de aplicação. Para atender a este contexto, surge uma nova área denominada Descoberta de Conhecimento em Base de Dados. A expressão “mineração de dados”, mais popular é na realidade uma das etapas da descoberta de conhecimento.

Para Rosini (Administração de Sistemas de informação e a Gestão do conhecimento, 2003, pág.18): o sistema de apoio à decisão auxilia a direção a tomar decisões semi-estruturadas ou com rápidas mudanças...Quando se observa a sua estrutura nota-se que é o sistema que mais trabalha com a análise: incorpora a construção explícita de uma variedade maior ou menor de modelos de análise de dados...São interativos , onde seus usuários podem modificar as condições assumidas pelo sistema e modifica sua base de dados secundária. Em resumo suas principais características são:

- focaliza a decisão, ajuda a alta decisão das empresas no processo de tomada de decisão;
- enfatiza a flexibilidade, adaptabilidade e respostas rápidas;
- permite que os usuários inicializem e controlem os inputs (entradas) e outputs (saídas);
- oferece suporte e ajuda para a solução de problemas cujas soluções podem não estar especificadas em seu desenvolvimento;
- dá suporte a estilos individuais de tomada de decisão dos gerentes que com ele trabalhem;
- usam sofisticados modelos de análise e modelagem de dados.

De acordo com Efraim Turban, James C.Wetherbe, Ephraim Mclean (Tecnologia da Informação para Gestão, pág. 370): A expressão data mining, ou exploração de dados, identifica um conjunto de técnicas sofisticadas de análise que consegue resultados mesmo a um volume imenso de informação. O data mining proporciona a identificação de novos padrões e relacionamentos por meio da utilização de softwares que faz a maior parte da exploração de dados. Os agentes inteligentes são uma ferramenta-chave no descobrimento de relações anteriormente ignoradas, principalmente em estruturas complexas de dados.

Em uma definição bem genérica, o sistema de apoio à decisão (SAD) é um sistema de informação baseado em computador que combina modelos e dados, em uma tentativa de solucionar problemas semi-estruturados com grande envolvimento por parte do usuário. O SAD pode ser entendido muito mais como abordagem ou filosofia do que como uma metodologia.

De acordo com Fábio V.Primak (Decisões com BI – pág.33) um dos grandes problemas dos especialistas em análise de informação é a transformação dos dados em informação. Uma das respostas é a combinação de estatística convencional e técnicas de inteligência artificial, que resulta em uma técnica muito comentada nos dias de hoje, o Data Mining.

Para Roberto Ferreira Lima Silva (2009, pág.93) ressalta que para a implementação de um BI, é essencial que todos os sistemas de informação (Inteligência Artificial, banco de dados e suas inovações) sejam incluídos e conectados. As diversas possibilidades proporcionados pelos conglomerados de informações abrem novos potenciais de uso e aproveitamento do capital intelectual da empresa. Expandindo a definição de SAD para incluir sistemas que podem apoiar tomada de decisão, analisando vastas quantidades de dados, incluindo dados de toda a empresa, obtidos de sistemas integrados , além de dados e transações pela web.

Capítulo 1 – Introdução

Os constantes avanços na área da Tecnologia da Informação têm viabilizado o armazenamento de grandes bases de dados. Tecnologias como a Internet, sistemas gerenciadores de banco de dados, leitores de códigos de barras, dispositivos de memória secundária de maior capacidade de armazenamento e de menor custo e sistemas de informação em geral são alguns exemplos de recursos que têm viabilizado a proliferação de inúmeras bases de dados de natureza comercial, administrativa, governamental e científica.

A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Torna-se imprescindível o desenvolvimento de ferramentas que o auxiliem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação.

Na mineração de dados o objetivo é encontrar relações entre os dados que ainda não sejam conhecidas. É importante mencionar que a mineração de dados não precisa ser aplicada apenas a grandes bases de dados, sendo também possível obter conhecimento valioso a partir de bases de dados modestas.

Para muitos autores, a mineração de dados é considerada uma etapa de um processo maior, denominada KDD – Descoberta de conhecimento em base de dados, o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados.

O objetivo deste trabalho é apresentar os principais conceitos da tecnologia de KDD, dando ênfase a mineração de dados com suas funcionalidades e principais técnicas utilizadas para obtenção de conhecimento; apresentar as principais metodologias para a preparação dos dados para mineração; processos de limpeza, integração, seleção e transformação dos dados; etapas fundamentais para o sucesso da mineração. Para serem utilizados como ferramentas de apoio à tomada de decisão (SAD), no âmbito da Justiça Federal, contribuindo na agilização do atendimento ao jurisdicionado.

Capítulo 2 – Mineração de Dados

2.1 – O processo de KDD: Conceitos básicos

Algumas definições:

Segundo Fayyad (FAYYAD et al., 1996a):

> KDD é um processo , de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados.
> A etapa de pós-processamento abrange o tratamento do conhecimento obtido na mineração de dados. Tal tratamento, nem sempre necessário, tem como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto.
> A expressão modelo de conhecimento indica qualquer abstração de conhecimento, expresso em alguma linguagem, que descreve algum conjunto de dados.

Segundo Amarral (2001):

> Historicamente, a noção de encontrar padrões úteis em dados em seu estado bruto tem recebido diversos nomes, inclusive extração de conhecimento de informação, coleta de informação, arqueologia de dados ou padronização de dados. Esse processo surgiu em 1989 para encontrar o conhecimento existente na base de dados e enfatizar o alto nível das aplicações dos métodos de prospecção de dados .

"KDD é um processo não trivial de identificação válida do padrão dos dados, potencialmente útil e fundamentalmente compreensível." (FRAWLEY, 1991)

Os sistemas de software KDD incorporam teorias, algoritmos, e métodos de todos estes campos. KDD enfoca o processo global de descoberta de conhecimento de dados, incluindo como os dados são armazenados (*data warehouse*)[1] e acessados como algoritmos podem ser escalados para conjuntos de dados volumosos e ainda

[1] Um repositório grande de dados históricos, que pode ser integrado para apoio à decisão. (Glossário Projeto e Modelagem de Banco de Dados)

poder rodar eficazmente, como resultados podem ser interpretados e visualizados, e como a interação global homem-máquina pode ser modelada e suportada. KDD coloca uma ênfase especial em achar padrões compreensíveis que podem ser interpretados como conhecimento útil ou interessante.

"O desenvolvimento de sistemas que utilizem conhecimentos extraídos de bases de dados tem propiciado valiosas ferramentas de apoio à decisão." (WEISS & INDURKHYA, 1998)

2.2 – Principais atividades na área KDD

Podem ser organizadas em três grandes grupos:

- Desenvolvimento tecnológico – Esse item abrange todas as iniciativas de concepção, aprimoramento e desenvolvimento de algoritmos, ferramentas e tecnologias de apoio que possam ser utilizados na busca por novos conhecimentos em grandes bases de dados;
- Execução de KDD – Esse item refere-se às atividades voltadas à busca efetiva de conhecimento em bases de dados. As ferramentas produzidas pelas atividades de desenvolvimento tecnológico são utilizadas na execução de processo de KDD;
- Aplicação de resultados – Uma vez obtidos modelos de conhecimento úteis a partir de grandes bases de dados, as atividades se voltam à aplicação dos resultados no contexto em que foi realizado o processo de KDD. Exemplos comuns de aplicação de resultados são as alterações em estratégias de negócios que tenham como objetivo procurar tirar proveito do conhecimento obtido. Tais alterações podem variar desde o posicionamento de produtos nas gôndolas de um mercado até políticas estratégicas corporativas.

2.3 - Caracterização do processo de KDD

Basicamente uma aplicação de KDD é composta por três tipos de componentes: o problema em que será aplicado o processo de KDD; os recursos disponíveis para a

solução do problema; os resultados obtidos a partir da aplicação dos recursos disponíveis em busca da solução do problema.

a)- O problema a ser submetido ao processo de KDD. Este componente pode ser caracterizado por três elementos:

- "Conjunto de dados pode ser observado sob os aspectos intensional e extensional" (DATE, 2000; ELMASRI & NAVATHE, 1989). O aspecto intensional se refere à estrutura ou esquema do conjunto de dados, portanto, as características ou atributos do conjunto de dados. Os casos ou registros compõem o aspecto extensional do conjunto de dados. É importante destacar que o processo de KDD não requer que os dados a serem analisados pertençam a Data Warehouse. No entanto, o tratamento e a consolidação dos dados necessários à estruturação e carga neste tipo de ambiente é extremamente útil e desejável ao processo de KDD;
- O especialista no domínio da aplicação representa a pessoa ou o grupo de pessoas que conhece o assunto e o ambiente em que deverá ser realizada a aplicação de KDD. Em geral, pertencem a esta classe analistas de negócios interessados em identificar novos conhecimentos que possam ser utilizados em áreas de atuação. Costumam deter o conhecimento prévio do problema ("*background knowledge*")[2]. As informações prestadas são de fundamental importância, pois influenciam desde a definição dos objetivos do processo até a avaliação dos resultados;
- Os objetivos da aplicação compreendem as características esperadas do modelo de conhecimento a ser produzido ao final do processo. Tais objetivos retratam, restrições e expectativas dos especialistas no domínio da aplicação acerca do modelo de conhecimento a ser gerado. Os objetivos devem ser refinados ao longo do processo de KDD em função dos resultados intermediários obtidos.

[2] tradução: conhecimento de fundo (google)

b)- os recursos disponíveis para a solução do problema. Entre eles podem ser destacados: o especialista em KDD, as ferramentas de KDD e plataforma computacional disponível:

- O especialista em KDD representa a pessoa ou o grupo de pessoas que possui experiência na execução de processos de KDD. Suas atribuições variam desde a identificação e a utilização do conhecimento prévio existente sobre o problema até o direcionamento das ações do processo, que englobam a seleção e a aplicação das ferramentas disponíveis, além da avaliação dos resultados obtidos;
- Ferramentas de KDD refere-se a qualquer recurso computacional que possa ser utilizado no processo de análise de dados. Pode ser desde um ambiente de software que integre diversas funcionalidades de tratamento e análise de dados até algoritmos isolados que possam ser adaptados ao processo de KDD;
- A plataforma computacional, indica os recursos computacionais de hardware (processadores e memória) disponíveis para a execução da aplicação de KDD. São os equipamentos disponibilizados para o processo.

c)- Os resultados obtidos a partir da aplicação dos recursos no problema. Compreende, os modelos de conhecimento descobertos ao longo da aplicação de KDD e o histórico das ações realizadas:

- “A expressão modelo de conhecimento indica qualquer abstração de conhecimento, expresso em alguma linguagem, que descreva algum conjunto de dados.” (FAYYAD et al., 1996a) É muito comum que, durante o processo de KDD, sejam realizadas comparações entre os modelos de conhecimento obtidos. O modelo de conhecimento com maior precisão na classificação dos clientes possui maiores chances de ser eleito como principal resultado gerado pelo processo de KDD;
- Os históricos sobre como os modelos de conhecimento foram gerados também enquadram-se como resultados do processo de KDD. São de

fundamental importância no controle do processo, pois permitem uma análise crítica e uma revisão das ações realizadas.

2.4 – Etapas de pré-processamento

A descoberta de conhecimento em base de dados é caracterizada como um processo composto por três etapas operacionais básicas: Pré-processamento, Mineração de Dados e Pós-processamento.

A etapa de pré-processamento compreende todas as funções relacionadas à captação, à organização e ao tratamento dos dados. Essa etapa tem como objetivo a preparação dos dados para os algoritmos da etapa da Mineração de Dados. As principais funções de pré-processamento dos dados são:

- Seleção de Dados – essa função, também denominada Redução de Dados, compreende, a identificação de quais informações, dentre as bases de dados existentes, devem ser consideradas durante o processo de KDD. A seleção dos dados pode ter dois enfoques distintos: a escolha de atributos ou a escolha de registros que devem ser considerados no processo de KDD;
- Limpeza dos Dados – abrange qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados. Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD;
- Codificação dos Dados – nessa função, os dados devem ser codificados para ficarem numa forma que possam ser usados como entrada dos algoritmos de Mineração de Dados. A codificação pode ser numérica – categórica, que transforma valores reais em categorias ou intervalos; ou categórica – numérica, que representa numericamente valores de atributos categóricos;
- Enriquecimento dos dados – a função de enriquecimento consistem em conseguir de alguma forma mais informações que possa ser agregada aos

registros existentes, enriquecendo os dados, para que estes forneçam mais informações para o processo de descoberta de conhecimento. Podem ser realizadas pesquisas para complementação dos dados, consultas a bases de dados externas, entre outras técnicas.

2.5 – Etapa de Mineração de Dados

2.5.1 – Principais objetivos de um Data Mining

São muitas as definições e conceitos de Data Mining encontradas na literatura:

> A proposta de mineração de dados é descobrir padrões em dados de forma que esse conhecimento seja aplicado para a solução de problemas. O termo mineração de dados pode causar um certo desconforto, dada a ampla gama de sentidos em que o mesmo pode ser usado. Soluções típicas de problemas incluem: detecção de fraudes, baterias de análises, segmentação de mercado (classificação de tipos), melhoramento de procedimentos operacionais, melhoramento de serviços, análise de mercado (AMARRAL, 2001);
>
> A mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados (JIAWEI HAN, 2001);
>
> Mineração de dados em poucas palavras, é a análise de dados indutiva (JESUS MENA , 1999);
>
> Mineração de dados, é a exploração e análise de dados, por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes (MICHAEL J.A.BERRY, 1997);
>
> Data Mining é uma técnica para determinar padrões de comportamento, em grandes bases de dados, auxiliando na tomada de decisão (SILVA, 2000);
>
> Data Mining é um processo que encontra relações e modelos dentro de um grande volume de dados armazenados em um banco de dados (RODDRIGUES, 2000);
>
> Para uma das maiores autoridades em Data Mining do mundo, o pesquisador Gregory Piatetsky – Shapiro, conforme relato a “Negócios Exame”: Data Mining é a extração de informações potencialmente úteis e previamente desconhecidas de grandes bancos de dados.

O data mining pode ser feito por pessoas que não são programadoras. O garimpeiro muitas vezes é o usuário final com pouco ou nenhum conhecimento de programação, mas que habilitados por meio de por meio de “detalhamento sucessivo de dados” e outras ferramentas potentes de consulta pode fazer perguntas *ad hoc*[3] e obter respostas rápidas. As ferramentas de data mining podem ser combinadas com planilhas e outras ferramentas de desenvolvimento de software de usuário final, tornando relativamente fácil analisar e processar os dados garimpados. O data mining aparece sob diferentes nomes, como extração de conhecimento, imersão em dados, arqueologia de dados, exploração de dados, processamento de padrões de dados, dragagem de dados e colheita de informação. No data mining, “encontrar o filão”, muitas vezes significa obter resultados valiosos inesperadamente.

Na mineração, são definidos as técnicas e os algoritmos a serem utilizados no problema em questão.

O processo de aplicação de mineração de dados envolve vários estágios, o principal estágio antes de se iniciar a busca do conhecimento, é definir claramente a que resultado se quer chegar. Uma vez definido o resultado, é preciso definir que técnicas utilizar e como aplicar essas técnicas para obtenção do resultado (conhecimento) desejado.

A partir da seleção dos dados, estes deverão ser organizados e armazenados em uma nova base de dados para análise. Durante a carga dos dados na nova base de dados, estes podem sofrer algum tratamento prévio para evitar resultados inesperados na mineração de dados. É necessário tratar os dados quando existem distorções, como valores discrepantes gerados devido a erro na entrada de dados, ou falta de valores para alguns campos importantes para a mineração e que não eram tão importantes na entrada de dados.

A realização de uma análise prévia dos dados, através de alguns métodos estatísticos, é essencial para tentar identificar atributos mais relevantes ou dependências que possam facilitar ou dificultar a etapa de mineração de dados. Após a

[3] Ad hoc é uma expressão latina que quer dizer "com este objetivo". Geralmente significa uma solução designada para um problema ou tarefa ... (google)

análise, pode ser necessário realizar a transformação dos valores de alguns atributos para melhorar os resultados obtidos.

A mineração de dados utiliza um conjunto de técnicas, estatísticas e também inteligência artificial. Em geral um processo de descoberta de conhecimento consiste em uma iteração das seguintes etapas:

- Preparação: é o passo onde os dados são preparados para serem apresentados às técnicas de data mining. Os dados são selecionados (quais dados que são importantes), purificados (retirar inconsistências e dados incompletos) e pré-processados (reapresentá-los de uma maneira adequada para o data mining). Este passo é realizado sob a supervisão e conhecimento de um especialista, pois o mesmo é capaz de definir quais dados são importantes, assim como o que fazer com os dados antes de utilizá-los no data mining;
- Mineração: os dados preparados são processados, ou seja, onde será feita a mineração de dados propriamente dita. O principal objetivo deste passo é transformar os dados de uma maneira que permita a identificação mais fácil de informações importantes;
- Análise de Dados: o resultado da mineração é avaliado, visando determinar se algum conhecimento adicional foi descoberto, assim como definir a importância dos fatos gerados. Para esse passo, várias maneiras de análise podem ser utilizadas, por exemplo: o resultado da mineração pode ser expresso em um gráfico, em que análise dos dados passa a ser uma análise do comportamento do gráfico.

A figura 1 demonstra alguns passos mais importantes em uma representação gráfica do processo de mineração:

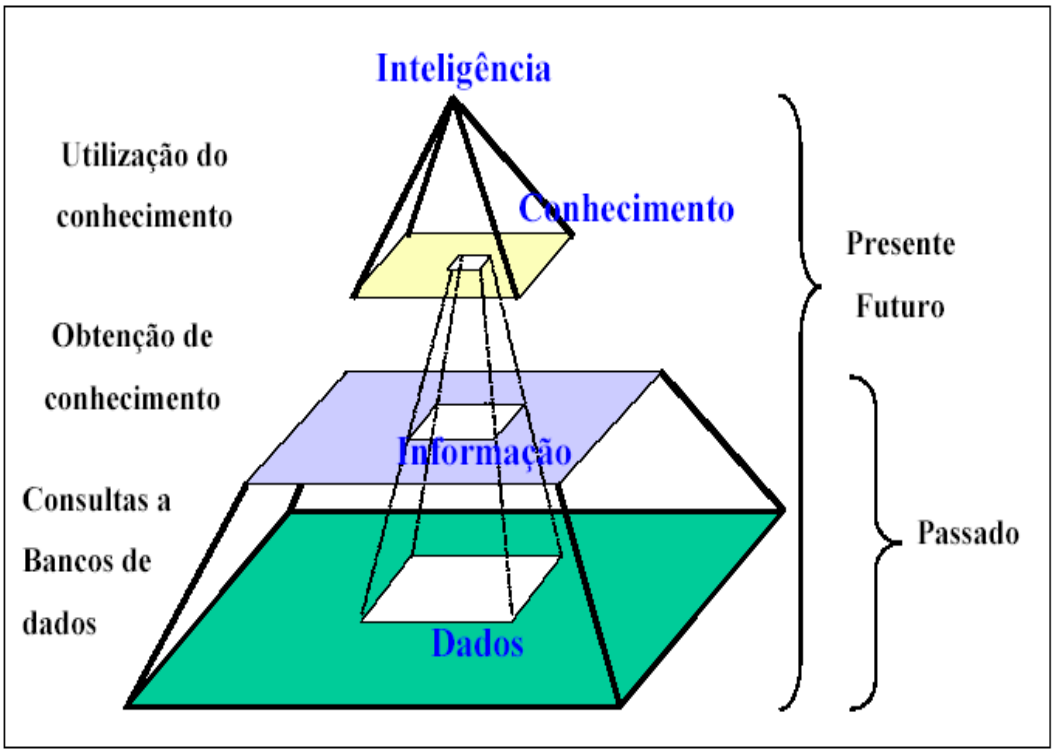


Figura 1: Pirâmide do Conhecimento

A base deste projeto é a busca por conhecimento oculto. O que torna possível essa busca é principalmente a capacidade de processamento, aliado as descobertas científicas na área de mineração de dados.

Descobrir padrões e tendências escondidos em grandes massas de dados não é um processo trivial. Na mineração este processo envolve o uso de diversas tarefas e técnicas. As tarefas são classes de problemas, que foram definidas através de estudos na área. As técnicas são grupos de soluções (algoritmos) para os problemas propostos nas tarefas. Cada tarefa apresenta várias técnicas, e algumas técnicas podem ser utilizadas para solucionar tarefas diferentes.

A mineração de modelos de classificação em bases de dados é um processo composto por duas fases: aprendizado e teste. Na fase de aprendizado, um algoritmo classificador é aplicado sobre um conjunto de dados de treinamento. Como resultado, obtêm-se a construção do classificador propriamente dito. Tipicamente, o conjunto de treinamento corresponde a um subconjunto de observações selecionadas de maneira aleatória a partir da base de dados que se deseja analisar. Cada observação do

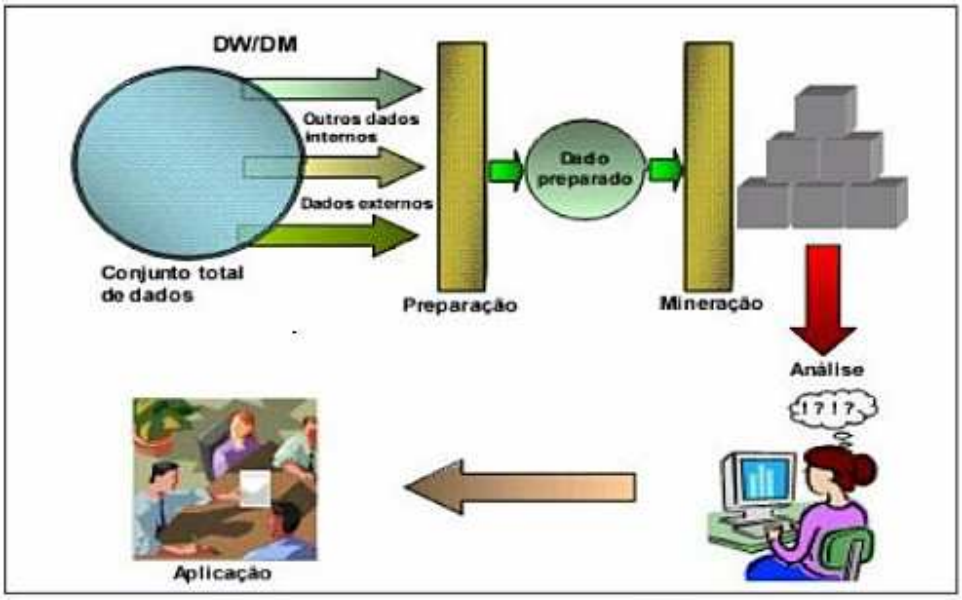
conjunto de treinamento é caracterizada por dois tipos de atributo: o atributo classe, que indica a classe a qual a observação pertence; e os atributos *preditivos*[4], cujos valores serão analisados para que seja descoberto o modo como eles se relacionam com o atributo classe.

Segundo Vessoni (2005):

> A idéia por trás do data mining pode causar um certo desconforto devido a ampla gama de objetivos em que o mesmo pode ser usado: uma empresa de varejo interessada em oferecer a melhor oferta para seus consumidores regulares; a receita federal pesquisando transações fraudulentas em remessas de moeda estrangeira; a análise de crédito de um banco, decidindo quais clientes devem receber a próxima mala direta de um novo financiamento; a classificação de clientes de uma operadora de telefonia, sugerindo qual plano se adapta melhor a cada um deles; e outros. Estes são apenas alguns exemplos das inúmeras aplicações do data mining.

Na Figura 2, numa visão geral os passos principais de um projeto de Data Mining. No esquema estão os grandes blocos do projeto, com as fases de preparação, mineração, análise e aplicação. A fase de preparação consiste de atividades que vai desde a construção de um banco de dados separado, para os dados sujeitos ao Mining até a atividade de carregar o banco de dados para o processo de Mining. Este processo de preparação dos dados é determinante no sucesso do Data Mining e costuma consumir muito tempo e recurso. A fase de mineração é responsável por criar o Data Mining definir amostras ou população e selecionar dados para treinar o modelo. Além disso, aqui deverá ser definida a formatação requerida pela ferramenta. Por exemplo: redes neurais exigem dados na forma dicotômica (sim/não) e árvore de decisão demanda agrupamentos como bom, médio e ruim. Por fim cria os previsores ou atributos-chaves para a análise do negócio.

[4] ação que se toma antes de um acontecimento para evitar que ele aconteça ou para reduzir seus efeitos. (google)



Visão geral de um processo de Data Mining

Figura 2 – Visão Geral de um processo de Data Mining

Existem várias técnicas que podem ajudar as empresas a encontrar informações para fomentar a sua tomada de decisão.

2.5.2 – Objetivos do Data Mining

O principal objetivo é a extração de valiosas informações dos dados, para a descoberta do “ouro escondido”. Esse “ouro” são as valiosas informações que os dados contém. Pequenas alterações nas estratégias oriundas das descobertas das ferramentas de Data Mining, podem transformar-se em significativas diferenças no caixa da empresa. Com a ampliação do uso dos Data Warehouses, as ferramentas de Data Mining tornaram-se primordiais. No entanto é importante lembrar que o uso de um Data Warehouse não é necessário para a aplicação de uma ferramenta de Data Mining. Basta que se tenham dados.

Várias ferramentas de análise de dados, tipo geradores de relatórios ou análises estatísticas, usam o termo Data Mining nos seus softwares computacionais. Produtos com bases em inteligência artificial também se intitulam ferramentas de Data Mining. Porém, o que denomina um verdadeiro Data Mining? O principal objetivo é a

descoberta do conhecimento, que com sua metodologia extrai informações preditivas das bases de dados.

2.5.3 – A origem do Data Mining

Para Freitas (2000), “Data Mining é um campo interdisciplinar, que emergiu da interseção entre várias áreas, principalmente aprendizado de máquina” (uma subárea da inteligência artificial, estatística e banco de dados).

- Inteligência Artificial ou IA: é uma disciplina com base nos fundamentos da heurística, diferentemente da estatística, sua tentativa é a de imitar a maneira como o homem pensa na resolução dos problemas estatísticos;
- Estatística: envolve conceitos como variância, distribuição normal, desvio simples, análise de regressão, análise de conjuntos, intervalos de confiança e análise de discriminante, todos usados para o estudo dos dados e seus relacionamentos. Essas são as bases fundamentais onde se apóiam as mais avançadas análises estatísticas. E a análise estatística clássica desempenha um papel fundamental na essência das atuais ferramentas e técnicas de Data Mining;
- Banco de Dados: uma das técnicas mais utilizadas para melhorar a base de dados é o Data Warehouse, que pode ser definido como um conjunto de tecnologias que propiciam a conversão de uma grande quantidade de dados em informação útil, transformando um banco de dados operacional em um ambiente que permite o uso dos dados estrategicamente. É um ambiente e não um produto.

2.5.4 – Estratégias ou tarefas Data Mining

Dando continuidade ao processo, que se inicia com a escolha das ferramentas (algoritmos) que serão utilizados. Essa escolha depende basicamente do objetivo do processo de KDD: classificação, agregação, regras associativas, ou outra.

- Classificação: corresponde à descoberta de um conjunto de regras de decisão que permitem classificar novas instâncias a partir de modelos obtidos dos dados existentes. Para a classificação é necessário um prévio

conhecimento das classes das instâncias disponíveis para que possa ser obtido um modelo que seja capaz de classificar novas instâncias. Por exemplo, uma população pode ser dividida em categorias para avaliação de concessão de crédito com base em um histórico de transações de créditos anteriores. Em seguida, uma nova pessoa pode ser enquadrada, automaticamente, em uma categoria de crédito específica, de acordo com suas características. Esta tarefa é considerada preditiva, pois uma vez que as classes são definidas, ela pode prever automaticamente a classe de um novo dado;

- Agregação: também conhecida como “clusterização”, refere-se ao procedimento de agrupar as instâncias de acordo com suas características, ou atributos. Assim, deseja-se que instâncias com valores similares para os atributos fiquem em um mesmo grupo e instâncias com atributos muito diferentes sejam colocadas em grupos diferentes. Por exemplo: uma população inteira de dados sobre tratamento de uma doença pode ser dividida em grupos baseados na semelhança de efeitos colaterais produzidos; acessos a web realizados por um conjunto de usuários em relação a um conjunto de documentos podem ser analisados para revelar *clusters*[5] ou categorias de usuários. Essa tarefa é considerada descritiva, ou seja, ela é usada para identificar padrões em dados históricos;
- Associação: procura-se com esta tarefa, identificar associações entre valores de atributos de instâncias na base de dados. A aplicação mais conhecida para a tarefa de associação é a obtenção de regras de associação a partir de uma base de dados de vendas para tratar o problema da “análise da cesta de compras”. As regras de associação obtidas identificam os produtos que são comprados juntos com uma certa freqüência. Por exemplo, uma análise das transações de compra em um supermercado pode encontrar itens que tendem a ocorrerem juntos com

[5] clusters é usada para particionar os registros de uma base de dados em subconjuntos. (Data mining um guia prático)

uma mesma compra (como café e leite). Os resultados desta análise podem ser úteis na elaboração de catálogos e layout de prateleiras de modo que produtos a serem adquiridos na mesma compra fiquem próximos um do outro. Essa tarefa é considerada descritiva;

- Estimativa: também conhecida como regressão, objetiva definir um valor (numérico) de alguma variável desconhecida a partir dos valores de variáveis conhecidas. Exemplos de aplicação são: estimar a probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnósticos de exames; predizer quantos carros passam por um determinado pedágio, tendo alguns exemplos contendo informações como: cidades mais próximas, preço do pedágio, dia da semana, rodovia em que o pedágio está localizado, entre outros. Essa tarefa é preditiva.

Tarefas (estratégias)	**Algoritmos**
Classificação	Árvores de classificação, redes neurais
Agregação (Clusterização) Vide Nota[6]	Métodos estatísticos, redes neurais
Associação	Métodos estatísticos, teoria de conjuntos
Estimativa (Regressão)	Métodos de regressão, redes neurais

Tabela 1 - Estratégias de Data Mining

2.5.5 - Algoritmo (técnicas) de Data Mining

Depois de escolher a tarefa a ser utilizada, deve-se também escolher o algoritmo (técnica) que será utilizado na mineração de dados. Para cada tarefa existem diferentes algoritmos que trabalham com diferentes tipos de dados e que fornecem resultados

[6] Classificação versus Agregação (clusterização): Na tarefa de classificação, os registros são subdivididos e colocadas em classes pré-difinicas. Já na clusterização, não há necessidade que se definam essas classes, pois estas são identificadas durante o processo, de forma automática. Neste caso, os registros são agrupados com fase em similaridades. Na clusterização, não há atributo especial. A importância de cada atributo em geral é considerada equivalente à dos demais.

diferentes. Alguns algoritmos são mais simples e outros mais sofisticados, como os algoritmos que utilizam lógicas difusas e redes neurais. Assim, é necessário um conhecimento do resultado que cada algoritmo pode fornecer, e como funciona internamente, para que a escolha possa ser acertada.

Dependendo do algoritmo escolhido, pode ser necessário também fornecer alguns parâmetros para a sua execução. Esses parâmetros normalmente dependem do conhecimento do usuário em relação ao negócio que está sendo considerado. É importante ressaltar que não existe um método de mineração de dados universal, portanto a escolha de um algoritmo particular para uma aplicação é, de certa forma, uma arte.

Podem ser destacados alguns critérios utilizados para a avaliação e a escolha da técnica mais adequada para atingir um determinado objetivo: robustez, grau de automação, velocidade, poder explanatório, acurácia, quantidade de pré-processamento necessário, facilidade de integração, habilidade para lidar com muitos atributos, facilidade de compreensão do modelo, facilidade de treinamento, facilidade de aplicação, capacidade de generalização, utilidade e disponibilidade.

Conclui-se, portanto, que não há critérios universais aplicáveis para a escolha e utilização de técnicas de mineração de dados. Isso porque cada técnica possui critérios específicos que devem ser levados em consideração.

Assim, de acordo com Passari (2003):

“é também extremamente difícil comparar as técnicas entre si, já que operam de maneira distinta. Avaliá-las é por meio da medição de sua habilidade em desempenhar as tarefas para a quais foram construídas.”

> Apesar de cada técnica de mineração de dados ter sua própria abordagem, elas compartilham algumas características em comum: conforme “aprendem” a partir dos dados de treinamento coletados, ela melhora, gradativamente, a sua performance; e existe sempre uma fase de treinamento, onde o modelo “aprende” os padrões e os relacionamentos (essa fase de treinamento é seguida pela fase implementação, quando o modelo é posto à prova).

De maneira geral, na fase de Data Mining, as ferramentas especializadas buscam padrões nos dados. Essa pesquisa pode ser efetuada pelo sistema automaticamente, de forma livre (roams – percorrer/vasculhar o banco de dados) ou interativamente com um analista responsável pela geração de hipóteses, chamada análise direcionada (*directed analyses*) ou também chamado aprendizado supervisionado (*supervised learning*), onde temos como que um 'professor' que 'ensina' o sistema indicando, por exemplo, quando uma premissa foi ou não correta.

Várias ferramentas distintas, como árvore de decisão, redes neurais, sistemas baseados em regras e programas estatísticos, tanto de forma isolada quanto em combinação, podem ser aplicadas ao problema. Geralmente, o processo de busca é interativo, de maneira que os analistas revêem o resultado, formam um novo conjunto de questões para aprimorar a busca em um determinado aspecto das descobertas, e realimentam o sistema com novos parâmetros.

2.5.5.1 – Árvore de decisão:

O objetivo da tarefa de classificação é construir um modelo que seja capaz de gerar classificações para novos dados (tarefa preditiva). Devem ser considerados dois tipos de atributos que caracterizam o objeto: atributos preditivos, cujos valores irão influenciar no processo de determinação da classe; e atributos objetivos, que indicam a classe a qual o objeto pertence. Desta forma a classificação visa descobrir algum tipo de relacionamento entre os atributos preditivos e objetivos.

A principal técnica utilizada é a árvore de classificação (classification tree). Uma árvore de decisão é um fluxograma (flow-chat) semelhante a uma estrutura de árvore, onde cada nó interno denota um teste em atributo, cada ramo representa o resultado do teste e cada folha representa a distribuição dos registros.

É importante observar que uma árvore de decisão pode ser utilizada com duas finalidades: previsão (exemplo: descobrir se um cliente será um bom pagador em função de suas características) e descrição (fornecer informações interessantes a respeito das relações entre os atributos preditivos e o atributo classe numa base de dados).

Tomando como exemplo uma aplicação que analisa dados de clientes, visando a aprovação ou não (atributo objetivo) de crédito para empréstimo pessoal. Neste banco de dados existem pessoas adimplentes e inadimplentes sendo cada classe caracterizada por algum tipo de padrão.

Neste processo, os clientes cujo campo resultado venha a ser o valor não, representarão os inadimplentes. Para poder preencher este campo, serão consideradas as características dos clientes (atributo preditivo) existentes no banco. Neste exemplo os atributos relevantes preditivos são tempo e profissão. O processo pode ser dividido em duas fases:

Fase I: um modelo é construído, descrevendo um conjunto pré-determinado de classes (neste exemplo, SIM ou NÃO). Em seguida, um conjunto de treinamento é analisado por um algoritmo de classificação, que gera uma saída de um modelo baseado numa árvore de classificação.

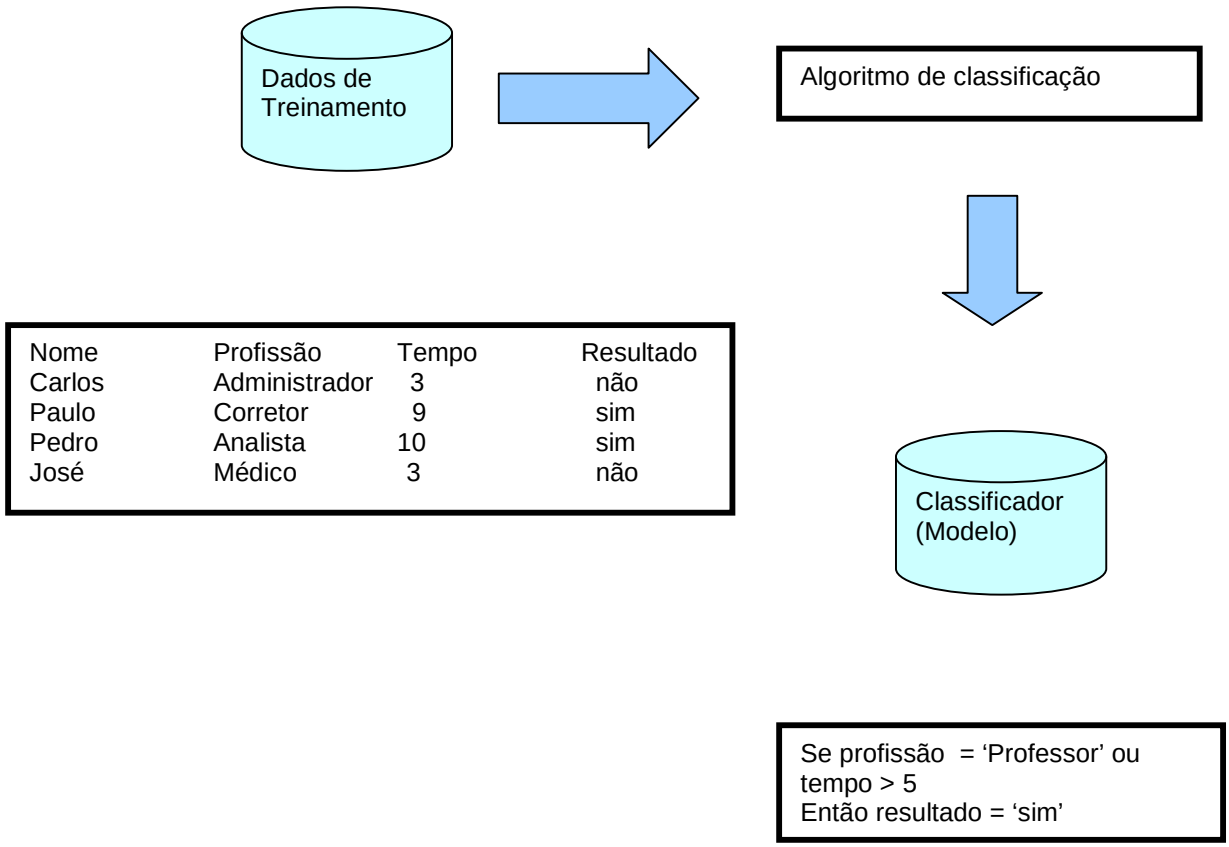


Figura 3 – Exemplo Fase I – Árvore Decisão

Fase II: o modelo gerado pela fase I é utilizado para classificação. Depois disso, é realizado um teste de exatidão e se esta for aceitável, as regras poderão ser utilizadas para a classificação de novos casos.

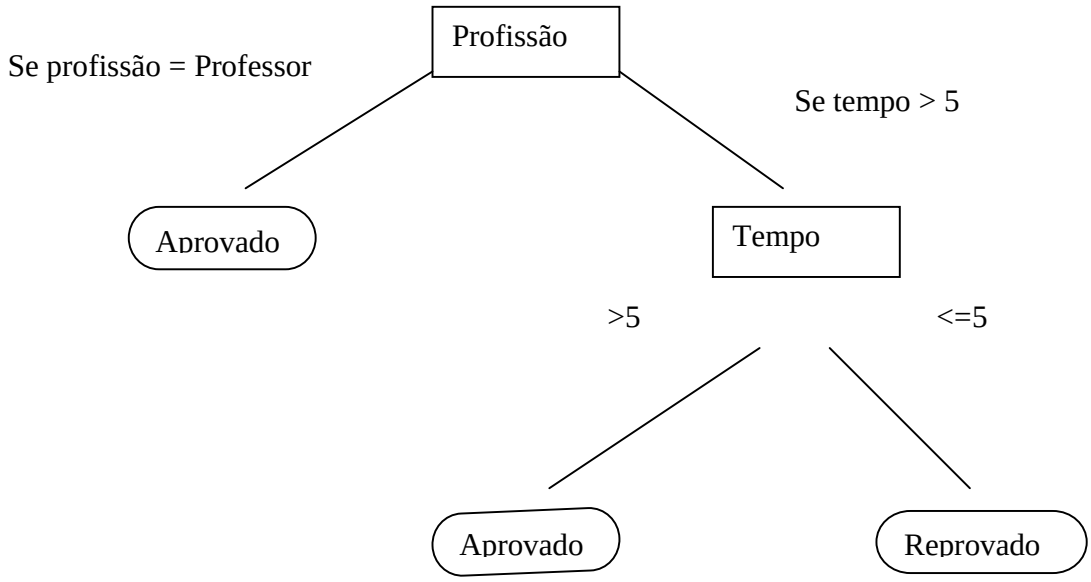


Figura 4 - Exemplo de uma visualização de uma árvore de classificação

2.5.5.2 - Técnica APRIORI

Na área de marketing é conhecida como grupos de afinidade ou análise de cestas de venda esta funcionalidade objetiva determinar que “coisas” estão relacionadas, estão juntas. Descobrir quais são as regras de associação condicionadas a valores de atributos que ocorrem juntos em um conjunto de dados.

Genericamente, uma regra de associação é representada pela notação X => Y (X implica em Y), onde X e Y são conjuntos de itens distintos. Neste caso, um item é representado por um dos conceitos existentes no domínio da aplicação. O objetivo desta técnica é representar, com determinado grau de certeza, uma relação existente entre o antecedente e o conseqüente de uma regra de associação. A associação é uma tarefa descritiva.

O objetivo é saber se determinado produto X implica na compra do produto Y. Esta implicação é baseada em dois fatores: suporte e confiança.

O suporte de uma regra representa o percentual das transações em que tal regra aparece.

O fator confiança, ao invés de considerar todas as transações, trabalha apenas com as que possuem o antecedente da regra. Assim, a confiança é calculada dividindo-se o número de vezes em que o conseqüente da regra aparece pela quantidade de transações.

Um algoritmo de extração de regras de associação deve gerar regras que possuam suporte e confiança especificados pelo usuário. Observe que as regras podem ser compostas por um ou mais itens. Dependendo do tamanho da base de dados e dos fatores de suporte e confiança, inúmeras regras são geradas. Essas regras devem ser avaliadas pelo usuário especialista, para que somente as mais relevantes possam ser utilizadas na tomada de decisão.

Na estatística trabalha-se desta forma, ou seja, a partir de uma hipótese busca-se verificar se ela é válida ou não. Na mineração de dados, apesar de ser necessária a definição da tarefa a ser realizadas e dos dados a serem analisados, normalmente não existe uma hipótese prévia a ser verificada.

O algoritmo mais conhecido utilizado para este fim é o algoritmo chamado APRIORI, que foi desenvolvido com o objetivo de tratar o problema de encontrar padrões referentes a produtos que são comprados juntos com uma certa freqüência.

Os algoritmos podem ser decompostos basicamente em duas etapas:

a) encontrar todos os conjuntos de itens freqüentes (que satisfazem à condição de suporte mínimo).
b) A partir do conjunto de itens freqüentes, gerar as regras de associação (que satisfazem à condição de confiança mínima).

Como a tarefa do item (a) demanda maior custo computacional e, uma vez gerados todos os conjuntos de itens freqüentes, a tarefa (b) se torna mais imediata, esforços de otimização têm sido concentrados na etapa (a).

Conclui-se que o número de regras possíveis pode se tornar muito grande, principalmente quando se tem muitos conjuntos freqüentes de tamanho superior a dois, inviabilizando qualquer análise por parte dos usuários de mineração de dados. Assim, são necessários critérios, ou medidas, que possam identificar, das possíveis regras de associação, quais são as mais interessantes..

A medida confiança (do inglês confidence) - MinConf, é a mais utilizadas na aplicação do algoritmo Apriori, através desta medida verifica-se a ocorrência de transações em que todos os itens da regra aparecem, sobre as transações em que os itens do antecedente estão presentes. Em termos práticos, a confiança expressa o percentual de transações em que, ocorrendo o antecedente, ocorre o conseqüente. Os valores possíveis para a confiança variam entre 0 (0%) e 1 (100%). Como exemplo, 100% de confiança significa que em todas as transações em que os itens do antecedente estão presentes, estão presentes também os itens do conseqüente.

Com os padrões de regras de associação obtidas, alguma ação deve ser realizada pelo usuário(o tomador de decisão) para que tais regras sejam validadas no ambiente da empresa e possam trazer a vantagem competitiva. Estas decisões estratégicas normalmente estão relacionadas às vendas ou à exposição dos produtos nas gôndolas.

2.5.5.3 - Clusterização

A tarefa de *clusterização*[7] é descritiva, ou seja ela visa identificar padrões em uma massa de dados. Consiste na busca de uma similaridade entre os dados tal que permita definir um conjunto finito de classes ou categorias que os contenha e os descreva. É também denominada indução não supervisionada. A clusterização pode ser definida como uma das tarefas básicas da Mineração de Dados que auxilia o usuário a realizar agrupamentos naturais de registros em um conjunto de dados.

[7] A principal diferença entre a classificação e a clusterização (agregação) é que nesta última as classes não são previamente definidas. A idéia é que o algoritmo de clusterização identifique automaticamente comportamentos similares em uma base de dados, dividindo a massa de informação em clusters. Após o processo de clusterização, o analista deve estudar os padrões identificados a fim de determinar se eles podem ser transformados em conhecimento estratégico.

A análise de clusters envolve, portanto, a organização de um conjunto de padrões (usualmente representados na forma de vetores de atributos ou pontos em um espaço multidimensional – espaço de atributos) em clusters, de acordo com alguma medida de similaridade. Intuitivamente, padrões pertencentes a um dado cluster devem ser mais similares entre si (compartilham um conjunto de propriedades comuns) do que em relação a padrões pertencentes a outros clusters.

Em geral, o processo de clusterização requer que o usuário determine qual o número de grupos a ser considerado. Com base neste número, os registros de dados são então separados nos grupos de forma que registros similares fiquem nos mesmos grupos e registros diferentes em grupos distintos. Uma vez, tendo esses grupos, é possível fazer uma análise dos elementos que compõem cada um deles, identificando as características comuns aos seus elementos e, desta forma, podendo criar um rótulo que represente cada grupo.

As técnicas de clusterização mais populares são:

- Particionamento: produzem agrupamentos simples. É importante destacar que esses algoritmos são efetivos se o número de clusters k puder ser razoavelmente estimado, se os clusters forem de forma convexa e possuírem tamanho e densidade similares. Os métodos tentam fazer os k clusters tão compactos e separados quanto possível. Quando os clusters são compactos, densos e bastante separados uns dos outros, os métodos trabalham melhor; mas quando existem grandes diferenças nos tamanhos e geometrias dos diferentes clusters, o método por particionamento pode dividir desnecessariamente grandes clusters para minimizar a distância total calculada;
- Hierarquia: criam uma decomposição hierárquica da base de dados. A decomposição é representada por um dendograma, uma árvore que iterativamente divide a base de dados em subconjuntos menores até que cada subconjunto consista de somente um objeto. Cada nó da árvore representa um cluster da base de dados.

2.5.5.4 - Técnica de regressão

A estimativa também conhecida como regressão é considerada uma tarefa preditiva, seu objetivo é prever um valor numérico desconhecido a partir de alguns atributos conhecidos, utilizando uma massa de dados histórica como modelo.

As técnicas mais comuns de estimativa são baseadas nos mesmos métodos da classificação, ou seja, utilizam árvores de decisão. A idéia básica é a geração de modelos que possam estimar o valor (numérico) de determinado atributo.

A técnica regressão linear tem como objetivo fornecer uma previsão de certos dados de acordo com uma série histórica, que deve seguir um modelo linear, ou seja, deve se 'encaixar' melhor por uma reta que representa os dados. geralmente, os problemas que a regressão linear auxilia estão relacionados à previsão da quantidade de itens em um determinado momento ou à previsão populacional.

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas (ou qualitativas) de modo que uma variável pode ser prevista a partir da outra ou outras. Pode ser utilizada em várias áreas como computação, estatística e medicina conforme alguns exemplos abaixo:

- Concentrações de soluções de proteína de arroz integral e absorbâncias médias corrigidas;
- Relação entre textura e aparência;
- Número de acessos ao disco (disk I/O) e o tempo de processamento para vários programas.

A regressão é usada basicamente com duas finalidades: de previsão (prever o valor de y a partir do valor de x) e estimar o quanto x influencia ou modifica y.

O caso mais simples de regressão é quando temos duas variáveis e a relação entre elas pode ser representada por uma linha reta (chamamos de Regressão linear simples). Esta metodologia é utilizada como técnica de mineração de dados e tem por objetivo prever um valor numérico desconhecido a partir de alguns atributos conhecidos, utilizando uma massa de dados histórica como modelo.

As técnicas mais comuns são baseadas nos métodos do modelo de classificação, mas que os atributos devem ser numéricos para desenvolver uma fórmula matemática que ajuste estes dados. Quando um novo dado é inserido no banco, será

aplicada esta fórmula calculada nos valores da tupla e assim serão definidos os valores dos atributos objetivos.

A análise de regressão possui essencialmente quatro passos: seleção das variáveis regressoras ou preditoras, diagnóstico para verificar se o modelo ajustado é adequado, aplicação de medidas remediadoras quando as condições do modelo não são satisfeitas e validação do mesmo.

Apesar de cada técnica de mineração de dados ter sua própria abordagem, elas compartilham algumas características em comum: conforme “aprendem” a partir dos dados de treinamento coletados, ela melhora, gradativamente, a sua performance; e existe sempre uma fase de treinamento, onde o modelo “aprende” os padrões e os relacionamentos (essa fase de treinamento é seguida pela fase implementação, quando o modelo é posto à prova) (PASSARI, 2003).

2.5.5.5 – Redes Neurais

As redes neurais são técnicas derivadas de pesquisas, na área da inteligência artificial, que utilizam a regressão generalizada. Essas técnicas fornecem “métodos de aprendizagem”, pois são conduzidas a partir de amostragens de testes, utilizadas para inferências e aprendizagem iniciais. Com esse método de aprendizagem, respostas a novas entradas podem ser passíveis de serem interpoladas a partir das amostras conhecidas. Essa interpolação, no entanto, depende do modelo mundial desenvolvido através do método de aprendizagem.

As RNAs (redes neurais artificiais) são sistemas de processamento de informações, compostos por muitos elementos computacionais simples, que interagem por meio de conexões, que recebem pesos distintos. “ Inspiradas na arquitetura do cérebro humano, as RNAs exibem algumas características, como a habilidade de aprender padrões complexos de informação e generalizar a informação aprendida” (BAETS e VENUGOPAL, citados por PASSARI, 2003).

Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão que conectam as camadas.

Segundo Din (1998), as RNAs utilizam um conjunto de elementos de processamento (ou nós) análogos aos neurônios no cérebro. Estes elementos de processamento são interconectados em uma rede que pode identificar padrões nos dados uma vez expostos aos mesmos, ou seja, a rede aprende através da experiênica, tais como as pessoas.

De maneira geral, as RNAs são modelos que relacionam dados de entrada com suas respectivas saídas. A partir da observação de exemplos e seu constante treinamento, obtêm-se uma matriz de pesos, os quais representam as ligações entre os neurônios de entrada e saída, imitando o que ocorre nas interconexões entre as células nervosas do cérebro humano. A grande vantagem do uso das RNAs relaciona-se com a adaptabilidade, que permite seu refinamento e minimização dos erros de previsão.

Segundo Freiman e Pamplona (2005):

> Com a característica de poder realizar previsões, além de outras possibilidades, as RNAs representam, portanto, uma importante alternativa aos tradicionais procedimentos estatísticos. Isso porque possui características próprias que, entre outras coisas, facilitam o seu uso em situações onde são exigidas inferências de relações não lineares complexas, entre as variáveis de entrada e de saída, de um modelo previsor.

Essa ferramenta de previsão, em relação às tradicionalmente usadas, exige do usuário uma maior atenção referente a alguns aspectos: seleção das variáveis de entrada, tipos de função de ativação, definição do número de camadas necessárias para um melhor processamento dos dados e cuidados especiais na interpretação dos resultados gerados, pois as RNAs são, normalmente, de difícil interpretação.

2.5.5.6 – Lógica Nebulosa

Técnicas de lógica nebulosa são utilizadas para capturar informações vagas, que, em geral, são descritas na sua forma natural, e convertê-las em um formato numérico, para facilitar as suas análises. Em termos operacionais, essas técnicas utilizam a teoria dos conjuntos nebulosos, que tem mostrado ser muito apropriada para se trabalhar com vários tipos de dados e informações, superando, muitas vezes,

os resultados obtidos com o emprego das tradicionais técnicas estatísticas e probabilísticas.

Segundo Azevedo e Demasi (2000):

> O uso da lógica nebulosa no processo de Data Mining pode ser útil quando se deseja descobrir informações em uma base de dados onde predominam dados numéricos que precisam ser classificados em conjuntos. E isso devido ao fato de as operações clássicas de união, interseção, complemento e negação da álgebra booleana para decidir a respeito da pertinência de um elemento a determinado conjunto só permitirem duas possibilidades: pertence ou não pertence. Essa classificação está sujeita à perda de informação, que pode vir a comprometer a confiabilidade do resultado final, visto que, no mundo real, na maioria das vezes, elementos podem pertencer parcialmente a um conjunto.

Segundo Takemura (2004):

> De forma objetiva e preliminar, pode-se definir Lógica Fuzzy, Difusa ou Nebulosa, como sendo uma técnica capaz de capturar informações vagas, descritas, em geral, em uma linguagem natural, e convertê-las para um formato numérico, de fácil manipulação pelos computadores de hoje.
>
> Lógica Fuzzy pode ser definida como a lógica que suporta os modos de raciocínio aproximados, em vez de exatos, como estamos naturalmente acostumados a trabalhar. Ela está baseada na teoria dos conjuntos nebulosos, e difere dos sistemas lógicos tradicionais em suas características e detalhes.

Segundo Srikant e Agrawal (1996):

> Em grandes bancos de dados que armazenam centenas de informações, o resultado de uma mineração pode trazer associações interessantes entre os itemsets da base. Por exemplo, em uma rede de supermercados, deseja-se descobrir os itens que são comercializados juntos. Para tratar essa questão, pode-se utilizar o algoritmo Apriori e encontrar as relações existentes entre os produtos. Porém, quando os dados contidos no datawarehouse forem atributos quantitativos (salário, idade, peso, etc.), é preciso transformar esses dados numéricos em binários, para realizar a mineração.

A solução proposta para encontrar associações entre dados numéricos é o algoritmo Fuzzy, que utiliza regras de associação da forma: “se X = A, então Y = B”. X e Y são atributos, A e B são variáveis nebulosas (alto, médio, baixo) que qualificam X e Y, respectivamente. As regras de associação geradas devem atender a um “grau de confiança” e a um “grau de suporte” para ser considerada interessante.

O “grau de confiança” corresponde ao cálculo que determina o quanto significante é uma associação. Caso esse “grau” atinja um valor estabelecido pelo especialista de negócios, a regra será denominada significante. Já o “grau de suporte” define se uma associação tem alguma representatividade na base de dados, ou se trata de um caso isolado. Se esse “grau” atingir um valor estabelecido pelo especialista de negócios, a associação será denominada representativa.

A utilização do algoritmo nebuloso no processo de data mining, como foi mostrado, além de descobrir uma quantidade maior de associações, tende a trazer resultados mais confiáveis, pois na manipulação de atributos quantitativos, a divisão em conjuntos é feita de uma forma a não desprezar dados possivelmente interessantes, enquanto que o algoritmo Apriori mostrou-se ineficiente nos testes aplicados, em relação ao Fuzzy, uma vez que as associações descobertas podem variar, com uma pequena modificação no limite da classificação dos itemsets.

Outra observação relevante está relacionada à flexibilidade de o s mineradores baseados em Fuzzy trabalharem com bases onde os atributos são numéricos, já que estes atributos podem ser definidos dentro de grupos que permitem a sua classificação.

2.5.5.7– Algoritmos genéticos

Os algoritmos genéticos estão relacionados com técnicas de otimização, onde se utilizam combinações de processos (exemplo: combinação genética, mutação e seleção natural). Essas técnicas estão associadas, sobretudo, com o conceito de seleção natural.

Enquanto os métodos de otimização e busca convencionais trabalham geralmente de forma seqüencial, avaliando a cada instante uma possível solução, os AGs trabalham com um conjunto possíveis de soluções simultaneamente. São classes

de procedimentos de pesquisa aleatórios capazes de realizar pesquisas adaptativas e robustas sobre uma ampla gama de topologias de espaço de pesquisa.

As soluções produzidas por algoritmos genéticos (AG) são diferenciadas das maiorias das outras técnicas de pesquisa através das seguintes características:

- Utiliza um conjunto de soluções durante cada geração ao invés de uma única solução;
- a pesquisa no espaço de strings representa uma pesquisa paralela maior no espaço de soluções codificadas;
- a memória da pesquisa realizada é representada unicamente através de conjunto de soluções disponíveis;
- um algoritmo genético é um algoritmo aleatório, uma vez que mecanismos de pesquisa utilizam operadores de probabilidade;
- ao prosseguir de uma geração para a seguinte, um AG encontra o equilíbrio próximo ao ótimo entre aquisição e exploração de conhecimento, manipulando soluções codificadas.

AGs são utilizados para resolver problemas e agrupar problemas. Sua capacidade para resolver problemas em paralelo fornece uma ferramenta poderosa para Mineração de Dados. As deficiências de AGs incluem a grande superprodução de soluções individuais, o caráter aleatório do processo de pesquisa e a elevada demanda no processamento computacional. Em geral, uma substancial demanda computacional é exigida para alcançar qualquer coisa significativas com algoritmos genéticos.

2.5.5.8 – Mining de Texto

Nas últimas décadas, a maior parte do trabalho em processamento de dados tem usado dados estruturados. A grande maioria dos sistemas lê e armazena dados em banco de dados relacionais. Os esquemas são organizados em linhas e colunas.

Entretanto, existem grandes quantidades de dados que residem em texto de forma livre. Apenas recentemente o trabalho na análise em texto tomou rumo significativo. As empresas agora estão comercializando produtos que focalizam a análise do texto.

Os dados digitados no computador normalmente são inseridos com pressa. A linguagem inclui abreviações, jargão, palavras com erros de digitação e gramática incorreta. Uma das opções seria a contratação de pessoas para ler os textos e determinar como cada uma dever ser categorizada, o trabalho seria de forma manual e muito tedioso.

Uma opção viável, desenvolvida nos últimos anos, é aplicar uma solução de software. Este processa o texto e determina os conceitos provavelmente representados no texto. Essa não é uma simples busca de texto. Sinônimos são mapeados para o mesmo conceito. Algumas palavras são mapeadas para conceitos diferentes, dependendo do contexto. O software utiliza uma ontologia que relaciona palavras e conceitos entre si. Após cada garantia ser categorizada de várias maneiras, é possível obter informações agregadas e úteis.

2.5.5.9 – Indução de Regras

A Indução de Regras (*Rules Induction*) se refere à detecção de tendências dentro de grupos de dados ou de “regras” sobre os dados. As regras são, então, apresentadas aos usuários como uma lista “não encomendada”, ou seja, sem que obedeça algum critério previamente estabelecido.

É o processo de analisar uma série de dados e, a partir dela, gerar padrões. O processo é, em sua essência, semelhante àquilo que um analista faria em uma análise exploratória.

Segundo Freitas, 2000):

> Consiste na descoberta de regras de previsão, do tipo SE...ENTÃO, onde a parte SE (a “condição”) da regra específica alguns valores de atributos previsores e a parte ENTÃO da regra prevê um valor para um determinado atributo cuja previsão é desejada. Por exemplo: suponha que se tenha um banco de dados de vendas de produtos,com dados sobre produtos vendidos e os clientes que compraram aqueles produtos. Assuma que os dados incluem atributos tais como a idade, sexo do cliente e o tipo do produto comprado. Analisando esses dados, um sistema de Data Mining poderia descobrir uma regra de previsão do tipo SE...ENTÃO, tal como: SE (idade_cliente < 18) E (sexo_cliente = “M”) ENTÃO (produto_comprado_videogame).

Idealmente as regras descobertas deveriam satisfazer três propriedades, a saber:

- fazerem previsões corretas, ou seja, na maioria das vezes que a parte “SE” da regra é verdadeira, a parte “ENTÃO” da regra também é verdadeira;
- serem compreensíveis para o usuário, ou seja, as regras representam conhecimento em um alto nível de abstração, tal como a regra acima, ao invés de equações matemáticas complexas e não compreensíveis pelo usuário;
- serem úteis para a tomada de decisão, o que está relacionado ao fato da regra expressar conhecimento novo ou surpreendente para o usuário. No exemplo acima, o usuário poderia usar a regra descoberta para, por exemplo, fazer uma mala direta direcionada, enviando uma propaganda de um novo videogame apenas para clientes que têm menos de 18 anos e são do sexo masculino.

Vários algoritmos e índices são usados para executar esse processo. Na indução de regras, a grande maioria do processo é feito pela máquina e uma pequena parte é feita pelo usuário.

2.5.5.10 - Análise Estatística de séries temporais

A estatística é a mais antiga tecnologia em Data Mining, e é parte da fundamentação básica de todas as outras tecnologias. Ela incorpora um envolvimento muito forte do usuário, exigindo engenheiros experientes, para construir modelos que descrevam o comportamento dos dados através dos métodos clássicos de matemática. Interpretar os resultados dos modelos requer especialistas (expertises). O uso de técnicas estatísticas requer um trabalho muito forte de máquinas/engenheiros.

A análise de séries temporais é um exemplo disso, apesar de freqüentemente ser confundida como um gênero mais simples de Data Mining chamado previsão (forecasting).

Enquanto que a análise de séries temporais é um ramo altamente especializado da estatística, o *forecasting* é, de fato, uma disciplina menos rigorosa, que pode ser satisfeita, embora com menos segurança, através da maioria das outras técnicas de Data Mining.

2.5.5.11 – Visualização

As técnicas de visualização são um pouco mais difíceis de definir, sendo que muitas pessoas a definem como “ferramentas complexas de visualização”, enquanto outras como simplesmente a capacidade de geração de gráficos.

Nos dois casos, a visualização mapeia o dado que está sendo minerado de acordo com dimensões específicas. Nenhuma análise é executada pelo programa de Data Mining além da manipulação da estatística básica. O usuário, então, interpreta o dado através do monitor de vídeo.

2.6 – As etapas do Data Mining

A implementação de um sistema de Data Mining pode ser dividida em seis fases interdependentes para que o mesmo atinja seus objetivos finais:

2.6.1 – Entendimento do problema

A fase inicial do projeto deve ter como objetivo identificar as metas e necessidades partindo de uma perspectiva do problema, e então convertê-las para uma aplicação de data mining e um plano inicial de “ataque” ao problema.

2.6.2 – Entendimento dos dados

Esta fase tem como principal atividade a extração de uma amostra dos dados a serem usados e avaliar o ambiente em que os mesmos se encontram.

2.6.3 – Preparação dos dados

Criação de programas de extração, limpeza e transformação dos dados para utilização pelos algoritmos de Data Mining, É nessa etapa que os dados são adaptados para serem inseridos no algoritmo escolhido para o processamento.

2.6.4 – Modelagem do problema

Seleção dos algoritmos dentre os apresentados a serem utilizados e processamento efetivo do modelo. Alguns algoritmos precisam dos dados em formatos específicos, o que acaba causando diversos retornos à fase de preparação dos dados.

2.6.5 – Avaliação do modelo

Ao final da fase de modelagem, diversos modelos devem ter sido avaliados sob a perspectiva do analista responsável. Então, o objetivo passa a ser avaliar os modelos com a visão do problema, certificando-se que não existem falhas ou contradições com relação às regras do problema.

2.6.6 – Divulgação ou publicação do modelo

A criação e a validação do modelo permitem o avanço de mais um passo, no sentido de tornar a informação gerada acessível. Isto pode ser feito de várias formas, desde a criação de um software específico para tal, até a publicação de um relatório para uso interno.

2.7 - Algumas aplicações práticas de técnicas de data mining

Para mostrar a relevância do uso das técnicas de data mining nos mais diversos setores da sociedade, neste item são apresentados alguns exemplos de sucesso onde foram aplicadas as referidas técnicas:

a) a Wal-Mart constitui uma das maiores cadeias varejistas dos Estados Unidos. “É conhecida tanto por sua política de baixos níveis de estoque e ressuprimento constante de produtos (baixos lotes e alta freqüência), como por sua política agressiva com os concorrentes regionais. Utilizando ferramentas de data mining, que auxiliam na previsão de cada item transacionado nas lojas da empresa, essa empresa modificou seus sistemas de ressuprimento automático de produtos. Além disso, identificou padrões de consumo, em cada loja, para a escolha do mix de produtos a ser ofertado” (RODRIGUES, 2005).

Identificou também um hábito curioso dos consumidores. Há cinco anos, ao procurar eventuais relações entre o volume de vendas e os dias das semana, o

software de data mining apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas. Uma investigação mais detalhada, revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana;

b) “o Banco Itaú costumava enviar mais de um milhão de malas diretas aos correntistas, com uma taxa de resposta de apenas 2%. Com um banco de dados contendo as movimentações financeiras de seus três milhões de clientes, durante 18 meses, e utilizando ferramentas de data mining, conseguiu reduzir em um quinto a conta com despesas postais e, ainda, aumentou sua taxa de resposta para 30%” (RODRIGUES, 2005);

c) “o Bank of America usou técnicas de data mining para selecionar, entre seus 36 milhões de clientes, aqueles com menor risco de dar calote num empréstimo. A partir dos resultados obtidos, enviou cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e, portanto, precisassem de empréstimos financeiros para ajudar os filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. Como resultado final, em três anos o banco lucrou 30 milhões de dólares “(“Data mining overview”, 2005);

d) “a empresa American Express, a partir da definição de estratégias de marketing com o auxílio de técnicas de KDD, fez aumentar as vendas, com utilização de cartão de crédito, em cerca de 20%” (FAYYAD et al., citados por Romão, 2005);

e) “empresas de telecomunicações dos Estados Unidos, a partir da utilização de malas diretas personalizadas com data mining, obtiveram reduções da ordem de 45% nas taxas de serviço com novos consumidores (RODRIGUES, 2005). Relacionado a esse mesmo setor”, segundo o “Data mining overview” (2005), atualmente, existe uma explosão nos crimes contra a telefonia celular, dentre os quais, a clonagem. Assim, técnicas de data mining poderiam ser utilizadas para detectar hábitos dos usuários de celulares. Quando um telefonema fosse feito e considerado pelo sistema como uma exceção, o programa poderia fazer uma chamada para confirmar se foi ou não uma tentativa de fraude;

f) “no vestibular PUC-RJ, utilizando as técnicas de data mining, um programa de obtenção de conhecimento, depois de examinar milhares de alunos, forneceu a

seguinte regra: se o candidato é do sexo feminino, trabalha e teve aprovação com boas notas, então não efetiva matrícula. Uma reflexão justifica essa regra: de acordo com os costumes do Rio de Janeiro, uma mulher com idade para realizar o vestibular, se trabalha é porque precisa, e nesse caso deve ter feito, também, inscrição para ingressar na universidade pública gratuita. Se teve boas notas, provavelmente, foi aprovada na universidade pública, onde efetivará a matrícula. Claro que há exceções: pessoas que moram em frente à PUC, pessoas mais velhas, pessoas de alto poder aquisitivo e pessoas que voltaram a estudar por outras razões. Mas a grande maioria obedece à regra anunciada". ("Data mining overview", 2005);

g) "algumas aplicações desenvolvidas pelo Data Mining Center (Universidade do Alabama) estão voltadas à utilização de técnicas de mineração de dados para efetuar previsão de fenômenos naturais. Dentre os projetos, está o desenvolvimento do AMSU (Advanced Microwave Sounding Unit), que é um radiômetro de microondas utilizado para detectar temperaturas em diferentes níveis da atmosfera. Com base nesse tipo de informação, é possível estimar velocidades de ventos radiais, que, combinadas com outros fatores, podem ser utilizadas para detectar ciclones tropicais" (SILVA, 2003).

2.8 – Etapas de Pós-Processamento

Abrange o tratamento do conhecimento obtido na mineração de dados. Tem como objetivo facilitar a interpretação e a avaliação da utilidade do conhecimento descoberto. Entre as principais funções da etapa estão: elaboração e organização, podendo incluir a simplificação, de gráficos, diagramas, ou relatórios demonstrativos; além da conversão da forma de representação do conhecimento obtido.

Em geral, é nesta etapa que o especialista em KDD e o especialista no domínio da aplicação avaliam os resultados obtidos e definem novas alternativas de investigação dos dados.

Capítulo 3 – Sistema de apoio à Decisão (SAD)

3.1 - Introdução ao Sistema de Apoio à Decisão (SAD)

Sistema de Apoio à Decisão (SAD) ou Decision Support System (DSS) é um termo que descreve sistemas que apóiam, não substituem, gerentes em suas atividades de tomar decisões. O sistema envolve atividades as quais tentam proporcionar aos profissionais a 'melhor' decisão.

SAD é um sistema interativo, sob controle parcial do usuário, que oferece dados e modelos para o suporte à discussão e à solução de problemas semi-estruturadas ou não estruturadas.

Devem ter seus dados e modelos organizados em função da decisão, flexibilidade e capacidade de adaptação às mudanças no ambiente e no estilo do responsável pela tomada de decisão.

Características essenciais de um SAD segundo Sprague e Carlson:

- Sistemas baseados em computador;
- Ajudam profissionais que tomam decisões;
- Possuem uma interação direta com o usuário;
- São baseados em modelos de aplicação e modelo de dados;
- Tentam modelar e dar soluções a problemas semi-estruturados.

A figura abaixo ilustra a configuração básica de um SAD:

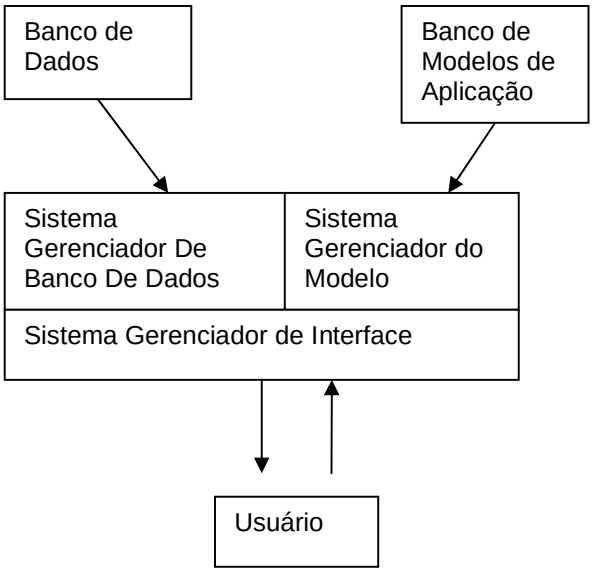


Figura 5 - Configuração básica de um SAD

A arquitetura de um SAD engloba um planejamento de hardware, software e interface com o usuário que venha de encontro com as possibilidades da organização e com a sua cultura. Para isso existem diversas preocupações relacionadas à análise, extração e armazenamento da base de dados, bem como o formato como essas informações serão disponibilizadas de forma que o usuário possa aproveitar ao máximo as informações ali contidas.

“Os sistemas de apoio a decisão e as ferramentas que apóiam esta categoria de informação surgiram e evoluíram a níveis totalmente acessíveis. Os SADs são baseados em computadores que auxiliam o processo de tomada de decisão utilizando modelos para resolver problemas não estruturados.” (FREITAS, 2001)

“Um dos fatores básicos do nascimento do SAD foi a mudança do modo de abordagem de um problema: passou-se de uma análise orientada a partir dos dados e direcionada ao problema e à decisão, para uma análise onde o ponto de partida é o tomador de decisão e o problema.” (LEVINE, 1989)

Analisa-se antes o problema em seu contexto e seu ambiente, a fim de melhor compreendê-lo para, então, caminhar em direção à solução. Este conceito permitiu o

desenvolvimento dos SAD. A questão passa a ser a representação dos processos pelos quais um sistema desenvolve comportamentos e a compreensão da ação inteligente.

A forma como os dados serão analisados é de grande importância quando consideramos a construção de um SAD. Assim, entende-se que uma representação estatística tem maior aproveitamento quando utilizada para demonstrar informações em sistemas estratégicos. Isso porque revelam resultados frutos de comparações.

Para Strehlo (1996):

> As metodologias baseadas em modelagem de dados partem de uma estruturação de bases de dados operacionais, isto é, bases de dados geradas a partir das operações das empresas, em direção à estruturação de uma base de dados gerenciais. É uma abordagem que se tem utilizado na concepção de depósitos de dados (data warehouses – DW).

Atualmente há dois tipos básicos de sistemas de apoio à decisão: orientados por modelo e orientados por dados.

- Os Sads orientados por modelo constituem-se em sistemas autônomos isolados dos principais sistemas organizacionais de informação e usam algum tipo de modelo para executar análises “se/não” e outros tipos. Sua capacidade de análise se baseia em teoria ou em modelo bem fundamentado, combinado com uma boa interface de usuário, que o torna fácil de usar;
- O segundo tipo de SAD é o orientado por dados. Esses sistemas analisam grandes repositórios, encontrados em grandes sistemas organizacionais. Eles dão apoio à tomada de decisão pela permissão ao usuário de extrair e de analisar informações úteis anteriormente ocultas em grandes bancos.

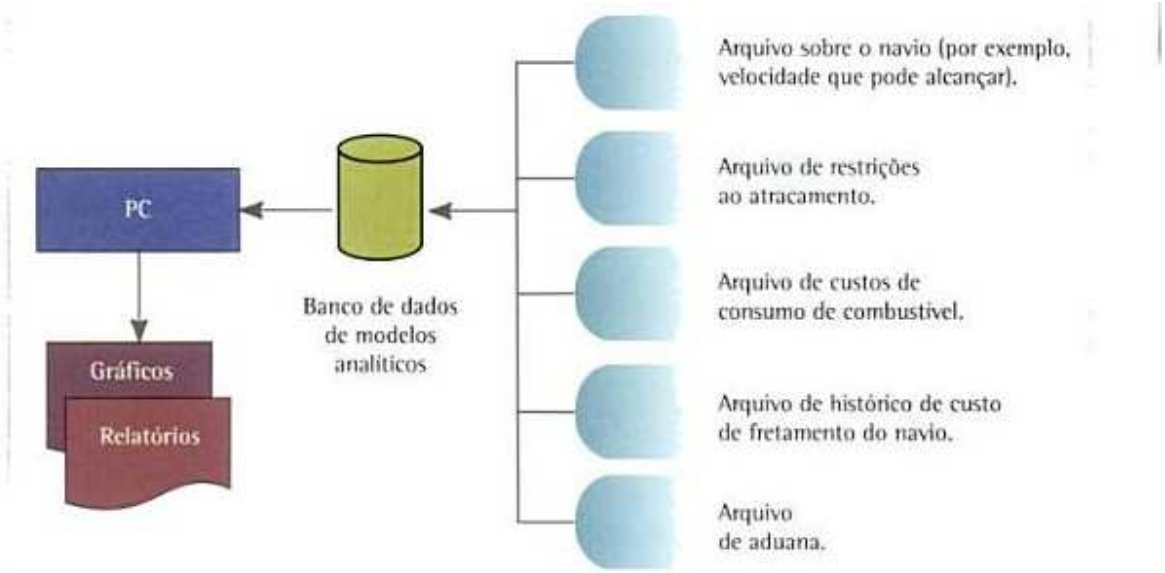
Elas pressupõem que a partir da disponibilidade de uma grande massa de dados estarão fornecendo toda a informação gerencial necessária. No entanto, estarão fornecendo dados em excesso e pouca informação. Muitos gerentes de empresas

acham-se perdidos e acabam por ignorar grandes quantidades de informação por entender que elas são excessivas.

As tecnologias para armazenamento de informação são tão comuns quanto numerosas. Junte-se a isso, a vontade dos empreendedores de extrair o máximo de vantagem de suas informações. Esses elementos tornam a mineração de dados e a busca de conhecimento a partir de banco de dados uma área de conhecimento em crescente expansão nos dias de hoje. Será raro em um futuro próximo, uma organização que não invista nas tecnologias do conhecimento.

A busca pelo conhecimento nunca foi fácil. A utilização de equipamentos de computadores e de técnicas avançadas de inteligência artificial não substituem as habilidades abstratas humanas na interpretação de qualquer tipo de informação. Os softwares de mineração de dados auxiliam em muito e minimizam o trabalho exaustivo do homem na análise de imensas quantidades de dados, tornando a informação mais clara e a busca pelo conhecimento mais fácil.

A figura abaixo apresenta o sistema de apoio a decisão (SAD). Esse sistema captura informações de diversos bancos de dados que apontam restrições previstas por normas e legislações a fim de compilar informações úteis para a formulação de relatórios. Esses relatórios gerados fornecem apoio para a tomada de decisão.



fonte Laudon e Laudon 2004[8]

Figura 6 – Exemplo Sistema de Apoio à Decisão

No campo de apoio à decisão, o data mining utilizado de forma consciente em conjunto por gerentes e engenheiros da informação resulta em vasta gama de novos e totalmente inesperados conhecimentos que não o seriam de forma alguma localizados por qualquer uma das partes isoladamente. Gerentes têm o conhecimento sobre o dia-a-dia e o universo de suas atividades,mas não é viável que analisem toda a informação colhida em suas atividades. O data mining realiza o trabalho pesado e exaustivo que seria impossível a qualquer agente humano ou ao menos não poderia ser realizado em tempo hábil.

Dados históricos são importantes porque grande quantidade de informação fica armazenada. Trabalhar somente com informações atuais pode impedir que se detecte tendências e padrões de comportamento ao longo do tempo. Informações históricas são cruciais para se entender o condicionamento dos negócios.

[8] Kenneth Laudon e Jane Laudon – Sistemas de Informações Gerencias (publicado 2004) são autores citados por Roberto F.L.Silva no livro E-Rh em um ambiente global.

3.2 – Modelo de Aplicação através da Mineração de dados

Exemplo de modelo de aplicação através de mineração de dados utilizando classificação:

Considerando a seguinte situação de uma editora.

A editora X deseja tomar uma decisão de como aumentar as vendas de uma de suas revistas. Quatro categorias de interesse foram escolhidas para categorizar os leitores: Culinária, Esportes, Saúde e Beleza. Para descobrir o perfil de seus leitores, a revista usará um questionário com dez perguntas.

As respostas desse questionário devem proporcionar a descoberta do conhecimento sobre os leitores da revista.

O espaço amostral é de 2000 leitores.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
L1	A	b	C	d	a	E	a	B	d	E
L2	D	c	E	d	c	A	a	C	d	D
...										
L2000	A	b	C	d	e	A	b	C	e	B

Tabela 2 – Amostra de leitores

Modelo de aplicação:

Se Q1=”a” e Q3=”b” e Q4=”e” e Q7=”c” então classificar leitor como leitor de culinária;

Se Q2=”b” e Q3=”d” e Q5=”a” então classificar leitor como leitor de esportes;

Se Q3=”b” e Q8=”d” e Q9=”a” então classificar leitor como leitor de saúde;

Se Q4=”b” e Q6=”d” e Q10=”a” então classificar leitor como leitor de beleza.

Quais foram as informações utilizadas para resolver o problema:

- Número de perfis previamente escolhidos = 4;
- Amostra de 2000 leitores;
- Tipos de perfis: Beleza, Saúde, Esportes e Culinária;
- Regras que determinavam a classificação dos leitores.

A Mineração de dados ou descoberta de dados está na quantidade de pessoas em cada classe.

Decisões que podem ser tomadas:

- Criar novos assuntos dentro da revista;
- Aumentar reportagens de determinado assunto.

Capítulo 4 – Tribunal Regional Federal 3ª Região

4.1 - Histórico do Tribunal Regional Federal (TRF)

Com a Constituição de 1988 foram criados cinco Tribunais Regionais Federais (TRF's) que absorveriam, de forma descentralizada, os processos a serem julgados pelo extinto Tribunal Federal de Recursos (TFR).

Os Tribunais iniciaram suas atividades em 1989, recebendo as soluções de informática anteriormente adotadas pelo TFR.

Em 1992, a Lei nº 8.472 de 14/10/92, dispôs sobre a composição e a competência do Conselho da Justiça Federal, e definiu em seu art. 2º:

> As atividades de recursos humanos, orçamento, administração financeira, controle interno e informática, além de outras atividades auxiliares comuns que necessitem de coordenação central, na Justiça Federal de Primeiro e Segundo Graus, serão organizadas em forma de sistema, cujo órgão central será o Conselho da Justiça Federal (CJF).

A partir de então, embora as atividades diárias sejam desenvolvidas por cada regional de forma bastante autônoma, os projetos maiores são, em geral, definidos de forma conjunta. Não há subordinação dos órgãos de informática ao CJF, mas procura-se o consenso na forma de autuação dos TRF's mediada pelo CJF, de forma a permitir intercâmbio entre esses órgãos.

Os primeiros projetos foram feitos sem nenhuma formalização: apenas reuniões de levantamento de dados com os usuários e desenvolvimento propriamente dito. Havia à época uma atuação bastante próxima da Assessoria de Organização e Métodos, principalmente porque os procedimentos internos do TRF, por ser um órgão novo, estavam em definição.

O TRF é órgão de 2ª Instância, ou seja, recebe os processos judiciais em grau de recurso (reavalia decisões de 1ª Instância) e os processos judiciais que envolvem a União. A decisão judicial não é monocrática: um mínimo de 4 juízes participam do julgamento de cada processo, e chegam ao Acórdão. O TRF, como instituição pública deve permitir à população acesso aos seus indicadores de desempenho.

O maior problema existente é o grande volume de processos judiciais a serem julgados e a dificuldade de agilizar o processo de análise dos mesmos e a elaboração da decisão.

O procedimento de análise dos processos judiciais e a elaboração da decisão envolve: identificação dos problemas apresentados no teor do processo (causa), conhecimento da legislação e conhecimento de jurisprudência (decisões já existentes sobre assuntos similares). É fator crítico de sucesso, a retenção e disseminação do conhecimento, fornecer soluções de classificação de processos, mecanismos de busca para facilitar a atualização dos servidores/desembargadores.

Em 2004, o Conselho da Justiça Federal – CJF, através da resolução nº 398/2004, instituiu o Sistema Nacional de Estatística da Justiça Federal – *Sinejus*[9].

Além disso, determinou que os TRF's responsáveis por cada uma das cinco regiões judiciais federais do Brasil disponibilizassem estas informações.

Esta determinação requer que cada Tribunal desenvolva formas de padronização e recuperação das informações estatísticas, bem como forma de divulgação à população, permitindo o acompanhamento da movimentação processual, da produtividade e da prestação jurisdicional. Portanto torna-se fator crítico de sucesso, o registro de indicadores e cruzamento de informações estatísticas e acesso a dados históricos.

[9] Este sistema estabeleceu a terminologia e os critérios para a divulgação de dados e indicadores que integram as informações estatísticas da Justiça Federal.

5 – Conclusão

O objetivo deste trabalho foi descrever a importância da metodologia do processo de KDD e sua etapa de mineração de dados, que são cada vez mais utilizadas e difundidas. Podem ser aplicadas em bases de dados imensas e repletas de informações úteis para futuras aplicações, as quais cada vez mais necessitam de agilidade e confiança para tomadas de decisões a curto e longo prazo.

Conclui-se que o interesse por mineração de dados, em particular por suas funcionalidades, tem aumentado gradativamente, principalmente pela alta demanda de transformar grandes quantidades de dados em informações úteis. A mineração de dados vem sendo uma ferramenta muito importante aprimorando e objetivando ações futuras, conduzindo agilidade, confiança, prevenção e comparação.

O uso de Data Mining para construção de um modelo pode trazer as seguintes vantagens para a organização:

- Modelos são de fácil compreensão: pessoas sem conhecimento estatístico (por exemplo, magistrados) podem interpretar o modelo e compará-lo com a produtividade de seu gabinete;
- Grandes bases de dados podem ser analisadas: grandes conjunto de dados, de até vários gigabytes de informação podem ser analisados com Data Mining;
- Data Mining descobre informações não esperadas: como muitos modelos diferentes são validados, alguns resultados inesperados podem surgir. Em diversos estudos, descobriu-se que combinações de fatores particulares apresentaram resultados inesperados;
- Variáveis não necessitam de recodificação: Data Mining lida tanto com variáveis numéricas (quantitativas) quanto categóricas (qualitativas). Estas variáveis aparecem no modelo exatamente da mesma forma em que aparecem na base de dados;
- Modelos são precisos: os modelos obtidos por Data Mining são validados por técnicas de estatística. Dessa forma, as predições feitas por modelos são precisas.

No poder judiciário, especificamente no TRF da 3ª região, um novo posicionamento a ser buscado, não de competitividade por ser um órgão público, mas sim de satisfação dos clientes internos e os jurisdicionados é a de garantir a agilização dos meios de disponibilizar informações processuais claras e atuais, utilizando-se de meios via Internet ou intranet, por haver pólos de atuação em localizações geográficas bem distantes da unidade Central. Tornar eficiente a prestação jurisdicional e o acesso da população ao Judiciário, é garantia constitucional das mais relevantes.

Por fim, pode-se dizer que o projeto é um estudo para o desenvolvimento de um grande trabalho de mudança na gestão do conhecimento.

5.1 - Apresentar a atual situação da justiça federal sob o ponto de vista da tomada de decisão;

Atualmente observa-se uma quantidade expressiva de processos e um número reduzido de servidores. Aliado a isso, existe uma grande formalidade que deve ser obedecida no trâmite processual. O resultado é que para a população verifica-se grande morosidade no julgamento e para os servidores, sobrecarga de trabalho.

Um dos objetivos desejado seria que os TRF's trabalhassem de forma equânime, para facilitar a tramitação de processos entre as instâncias. Utilizando-se das informações básicas cadastradas inicialmente pela 1ª Instância, evitando com isso a redundância e a propagação de novas bases de dados contendo informações similares. Soluções de Data warehouse e cruzamento de informações seriam uma boa estratégia.

5.2 - Apresentar um modelo de mineração de dados que agilize o processo de apoio à tomada de decisão na justiça federal;

A maior parte do trabalho em processamento de dados nas últimas décadas tem usado dados estruturados. A grande maioria dos sistemas em uso hoje lê e armazena dados em bancos de dados relacionais. Entretanto, existem grandes quantidades de dados históricos que residem em texto de forma livre. Neste contexto foi desenvolvida uma solução de software intitulada "mining de texto" (descrito no item cap.2.5.4.8).

O software processa o texto e determina os conceitos provavelmente representados no texto. Não se trata da de uma busca simples de texto. Sinônimos são

mapeados para o mesmo conceito. Algumas palavras são mapeadas para conceitos diferentes, dependendo do contexto. Cabe aqui a intervenção do usuário para que seja tomado uma decisão, a soma de alguns dados pode resultar em fatos diferentes. Após cada garantia ser categorizada de várias maneiras, é possível obter informações agregadas e úteis. Seria uma forma de agilizar o andamento processual, dando o mesmo tratamento a ações processuais jurisdicionais que tratam do mesmo assunto (Por exemplo: revisão da aposentadoria).

Foram criados algumas metas que tem por objetivo o julgamento de pelos menos 'X' processos por mês. Foi feito um levantamento estatístico por gabinete (desembargador) e quantidade de processos aguardando julgamento, para se chegar a uma estimativa do que seria o 'ideal' no atendimento ao jurisdicionado, e qual o tempo gasto para a concretização do mesmo. Através da técnica de associações é possível a criação de modelos padrões obedecendo o trâmite processual.

5.3 - Apresentar um modelo de sistema de apoio à decisão na esfera na justiça federal;

Muitos conhecimentos encontram-se escondido na vasta quantidade de dados (incluindo-se dados históricos), que ficam disponíveis em banco de dados. Este estudo mostrou como o Data Mining pode transformar esses dados brutos em informações valiosas a fim de auxiliar a tomada de decisão. Quando a organização emprega o Data Mining é capaz de: criar parâmetros para entender o comportamento dos dados referentes a pessoas envolvidas com o órgão; identificar afinidades entre dados, entre pessoas e produtos, ou serviços e analisar hábitos para detectar comportamentos fora do padrão.

Uma boa utilização seria para avaliar dados processuais de agrupados por determinada matéria ou assunto, juntamente com o objeto do processo, analisando os procedimentos coerentes no trâmite processual, possibilitando com isso uma igualdade no resultado do andamento processual.

5.4 - Identificar os aspectos da mineração de dados, analisando técnica a ser utilizada para agilizar o processo de apoio à tomada de decisão na justiça federal;

Os sistemas em si não requerem nenhuma direção gerencial e não propõem a perfeição; necessitam ser alimentados com dados para fornecerem informações para a tomada de decisão. Por esse motivo, às vezes, não os levam a sério, e os sistemas acabam caindo no descrédito, simplesmente por incompreensão tecnológica, pois falta exatamente um gerenciador (ou mais) que detenha, além do conhecimento técnico avançado e adequado às novas regras da administração pública.

O sucesso da administração depende do corpo funcional, o qual por sua vez depende de um eficaz controle interno, onde são monitorados os excessos de gastos, desperdícios, retrabalhos, descontroles, entre outros.

Várias técnicas apresentadas neste estudo poderiam dar subsídios para a concretização destes objetivos, tais como: árvore de decisão (cap. 2.5.4.1); clusterização (cap. 2.5.4.3); técnica de regressão (cap. 2.5.4.4); mining de texto (cap. 2.5.4.8); indução de regras (cap. 2.5.4.9).

A qualidade impõe a definição de um conceito de padronizações e demais ferramentas necessárias para que se evite que os administradores públicos percam a credibilidade perante a sociedade. Neste sentido, deve-se ter uma filosofia transparente orientada para atingir metas, prevenção de riscos, técnicas e inovações tecnológicas modernas, treinamento – principalmente – e, conseqüentemente, qualificação do corpo técnico.

Assim, sincronizando-se o uso eficaz de métodos científicos, criatividade e capacidade, os órgãos podem utilizar-se com eficiência a tecnologia, gerenciando, com seus servidores, programas de desenvolvimento indispensáveis ao atendimento dos interesses da sociedade e da instituição.

6 - Referências bibliográficas

AMARRAL, Fernando C. **Data Mining Técnicas e Aplicações para o Marketing** Ed. Berkeley Brasil. 2001;

AZEVEDO, Denise; DEMASI, Pedro. **Consulta a Banco de Dados Utilizando Conceitos Nebulosos**. Out. 2000.
Disponível em <http://www.nce.ufrj.br/labic/downloads/ricim_out_2000.pdf> ;

BERRY Michael J.A.; Gordon Linoff; **Data Mining – Techniques for Marketing, Sales and Customer Support**; John Wiley & Sons, Inc, 1997;

DIN – Departamento de Informática – UEM – Universidade Estadual de Maringá. GSI – Grupo de Sistemas Inteligentes – Mineração de Dados, 1998.
Disponível em: http://www.din.uem.br/ia/mineracao/tecnologia/ferramentas.html;

FAYYAD, U.M.; Piatetsky-Shapiro, G.; Smyth, P. **From Data Mining to Knowledge Discovery: An Overview.** Knowledge Discovery and Data Mining, Menlo Park: AAAI Press, 1996a;

FREIMAN, J. P.; PAMPLONA, E. de O. **Redes neurais artificiais na previsão do valor de commodity do agronegócio**. In: Encuentro Internacional de Finanzas, 5. Anais…, Santiago, Chile, 2005. 14p;

FREITAS O. G. e Rodrigues, A. M. **Sistema de Apoio a Decisão Usando a Tecnologia Data Mining**, CBComp 2001;

GOLDSCHMIDT Ronaldo; Emmanuel Passos. **Data Mining um guia prático**, 2005;

http://www.feg.unesp.br/ceie descoberta de conhecimento em banco de dados para apoio a tomada de decisão;

JIAWEI Han, Micheline Kamber. **Data Mining – Concepts and Techniques**; Morgan Kauffmann Publishers, Inc, 2001;

LEVINE, A. L. and Pomerol, J. C. **Sistemas Interativos de Apoio a Decisão e Sistemas Especialistas**, Ed. Hermºs, 1989;

MENA Jesus; **Data Mining Your Website**; Digital Press, 1999;

PASSARI, A. F. L. **Exploração de dados atomizados para previsões de vendas no varejo utilizando redes neurais**. São Paulo: USP, 2003. (Dissertação de Mestrado);

PRIMARK Fábio Vinícius - **Decisões em BI (Business Intelligence)**;

RODRIGUES, A. M. **Escavando Dados no varejo. Centro de Estudos em Logística – COPPEAD** – Universidade Federal do Rio de Janeiro, 2000;

SILVA, E.M. **Avaliação do Estado da Arte e Produtos Data Mining**. UCB – Universidade Católica de Brasília, 2000);

SILVA Roberto Ferreira Lima. - **E-RH em um ambiente global e multicultural**;

SPRAGUE, R.H. e Watson, H.J. **Sistema de Apoio à Decisão: colocando a teoria em prática** (segunda edição);

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh. **Mining Quantitative Association Rules in Large Relational Tables**, 1996.
Disponível em <http://www.acm.org>;

SQL Magazine, Data Mining, Edição 10;

SQL Magazine, Data Mining, Edição 26;

TAKEMURA, Roberto. **Lógica Difusa.**
Disponível em: <http://www.din.uem.br/ia/controle/fuz_prin.htm>;

TEOREY Toby, Sam Lightstone, Tom Nadeau – **Projeto e Modelagem de Banco de Dados**, 2006.

FOLHA DE APROVAÇÃO DA MONOGRAFIA

UNIVERSIDADE GAMA FILHO
CURSO DE PÓS-GRADUAÇÃO LATO SENSU
TECNOLOGIA DA INFORMAÇÃO

Utilidade da mineração de dados no Judiciário Federal

Esta monografia será examinada e aprovada para a obtenção do título de Especialização em Engenharia de Software no Programa de Pós-Graduação da Universidade Gama Filho

_____

Prof. Julio Celso Noguchi – Orientador

_____

Examinador

_____

Examinador

SÃO PAULO/SP
2009