UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

A Utilização de Raciocínio Baseado em Casos para a Análise de Crédito e Cobrança

por

MARCELO COSTA ISOLANI

Dissertação submetida à avaliação, como requisito parcial, para a obtenção do grau de mestre em Ciência da Computação.

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Porto Alegre, setembro de 2002.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Isolani, Marcelo Costa

A Utilização de Raciocínio Baseado em Casos para a Análise de Crédito e Cobrança/ por Marcelo Costa Isolani. – Porto Alegre: PPGC da UFRGS, 2002.
83 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2002. Orientador: Oliveira, José Palazzo Moreira.

1. Inteligência Artificial. 2. Raciocínio Baseado em Casos. 3. Data Warehouse. 4. Business Intelligence. I. Oliveira, José Palazzo Moreira. II. Título.

Agradecimentos

À minha Fernanda, que durante todo este tempo suportou todos os “efeitos colaterais” destes anos de estudo e trabalho mais do que eu mesmo.

À minha família, principalmente a minha mãe e meu pai, que se foi quando eu tinha apenas 9 anos, mas que sempre nos mostraram que o caminho da vida é o estudo e a Vó Gloria, que sempre nos foi uma mãe mais do que uma tia.

Aos meus companheiros de turma, principalmente ao Fernando, Tatiana e Ailton com os quais partilhei a maior parte destes momentos, onde brincamos, brigamos, estudamos, compartilhamos problemas acadêmicos, profissionais e pessoais, ou seja, nos tornamos amigos.

A GVT, principalmente a Paulo Pontes e Ivo Murbach, que me ajudaram no desenvolvimento deste trabalho.

Sumário

Lista de Abreviaturas **5**

Lista de Figuras **6**

Lista de Tabelas **7**

Resumo **8**

Resumo **8**

Abstract **9**

1 Introdução **10**

2 Definições **14**

2.1 Business Intelligence 14
2.1.1 Desenvolvimento de BI *15*
2.1.2 Sistemas de Front-End *16*
2.1.3 Sistemas de Informações Gerenciais *19*
2.2 Sistemas de Back-End 21
2.2.1 Data Warehouse *21*
2.2.2 Data Mart *33*
2.2.3 Fatores Críticos de Sucesso em Projetos de DW e DM. *33*
2.2.4 Data Mining *33*
2.2.5 Redes Neurais *41*
2.3 Raciocínio Baseado em Casos 42
2.3.1 O Funcionamento de um CBR *44*
2.3.2 O Desenvolvimento de um CBR *47*

3 Estudo de Caso - Análise de Crédito e Cobrança a partir da Similaridade dos Casos **49**

3.1 Estudo da aplicação de Análise de Risco e Inadimplência 49
3.1.1 Análise do Processo de Negócio *50*
3.2 Utilizando um CBR 53
3.2.1 Aquisição do Conhecimento *55*
3.2.2 Representação de Conhecimento *56*

4 Implementação **68**

4.1 Aspectos da Implementação 68
4.1.1 Dados 68
4.1.2 Funcionamento *69*
4.1.3 Resultados Obtidos com o Sistema *75*

5 Conclusão **78**

Bibliografia **80**

Lista de Abreviaturas

AI	Artificial Intelligence (Inteligência Artificial)
ANATEL	Agência Nacional de Telecomunicações
BD	Banco de Dados
BI	Business Intelligence (Inteligência de Negócios)
CBR	Case-Based Reasoning (Raciocínio Baseado em Casos)
CI	Competitive Intelligence (Inteligência Competitiva)
DDD	Discagem Direta à Distância
DDI	Discagem Direta Internacional
DM	Data Mart
DSS	Decision Support System (Sistema de Suporte à Decisão)
DW	Data Warehouse
EDP	Eletronic Data Processing (Processamento Eletrônico de Dados)
EDP	Eletronic Data Processing (Processamento Eletrônico de Dados)
ETL	Extraction, Transformation and Load (Extração Transformação e Carga)
GED	Gestão Eletrônica de Documentos
KDD	Knowledge Discovery Database (Descoberta de Conhecimento em Banco de Dados)
KM	Knowledge Management (Gerenciamento de Conhecimento)
KMS	Knowledge Management System (Sistema de Gerenciamento do Conhecimento)
ODS	Operational Data Store (Armazenamento de Dados Operacionais)
OLAP	On-line Analytic Processing
OLTP	On-line Transaction Processing
SGBD	Sistemas Gerenciadores de Banco de Dados
SMS	Short Message Service (Serviço de Mensagens Curtas)
SQL	Structured Query Language (Linguagem Estruturada de Consultas)
STFC	Sistema Telefônico Fixo Comutado

Lista de Figuras

FIGURA 1.1 - Mapa do Sistema Telebrás (adaptado ANATEL) 10
FIGURA 1.2 - Mapa STFC após privatização do setor (adaptado ANATEL) 11
FIGURA 2.1 - Visão das Relações entre BI, CI e KMS (adaptado de [BAR2001]) 15
FIGURA 2.2 - Componentes de um ambiente de BI [BAR2001] 16
FIGURA 2.3 - Modelo "Star Schema" [KIM98] 23
FIGURA 2.4 - Estratégia para implementação gradativa de *DM* [BAR00] 23
FIGURA 2.5 - Base de um *Data Mining* [DAT2000A] 35
FIGURA 2.6 - Classificação linear de dados de empréstimo [AMA2001] 37
FIGURA 2.7 - Regressão para o conjunto de dados de empréstimo [AMA2001] 38
FIGURA 2.8 - Divisão do conjunto de dados de empréstimos em três grupos 39
FIGURA 2.9 - Visão Esquemática de uma Rede Neural [BAR2001] 42
FIGURA 2.10 - O Ciclo do CBR adaptado [KOL93] 45
FIGURA 3.1 - Macro Processo de Negócio 50
FIGURA 3.2 - Visão do *timeline* de ações - Cenário de Cobrança 52
FIGURA 3.3 - Funcionamento de um índice (memória plana) 59
FIGURA 3.4 - Indexando Atributos 60
FIGURA 3.5 - Pesando Atributos 61
FIGURA 4.1 - Modelo Esquemático de Leitura 69
FIGURA 4.2 - Funcionalidade de Carga 70
FIGURA 4.3 - Consulta a Base de Casos Cliente 71
FIGURA 4.4 - Tela de Pesos dos Atributos. 71
FIGURA 4.5 - Formulário de Consulta Crédito (cliente) 72
FIGURA 4.6 - Formulário de Consulta Cobrança (fatura) 73
FIGURA 4-7 - Esquema de Consulta a Dados de Cobrança 73
FIGURA 4.8 - Tela de Geração de Campanha 74
FIGURA 4.9 - Clientes Inseridos no processo de Campanha 74
FIGURA 4.10 - Resultado Campanha Inadimplentes com clientes adimplentes 75
FIGURA 4.11 - Resultado Campanha Inadimplentes com clientes inadimplentes 76
FIGURA 4.12 - Gráfico comparativo das Campanhas 77

Lista de Tabelas

TABELA 2.1 – Comparativo entre DSS e EDP 17
TABELA 3.1 - Diferenças entre as técnicas [LOR98] 54
TABELA 3.2 - Exemplo de um caso de Análise para Crédito 57
TABELA 3.3 - Exemplo de um caso de Análise para Cobrança 58
TABEL 3.4 - Fatores para Verificação de Índices e Pesos 60
TABELA 3.5 - Pesos dos Atributos para Crédito 61
TABELA 3.6 - Pesos para Análise de Cobrança 62
TABELA 3.7 - Comparativo de Similaridade entre dois casos hipotéticos 64
TABELA 3.8 - Distribuição dos Casos atributo IDADE 65
TABELA 3.9 – Relacionamento para o atributo Bairro 67
TABELA 3.10 - Relacionamento para o atributo Profissão 67

Resumo

Data Warehouse (*DW*) é um processo que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas e *ad-hoc*, relatórios analíticos e de suporte de decisão. Já um *Case-Based Reasoning* (CBR) é uma técnica de Inteligência Artificial (AI – *Artificial Intelligence*) para a representação de conhecimento e inferência, que propõe a solução de novos problemas adaptando soluções que foram usadas para resolver problemas anteriores. A descrição de um problema existente, ou um caso é utilizado para sugerir um meio de resolver um novo problema, avisar o usuário de possíveis falhas que ocorreram anteriormente e interpretar a situação atual.

Esta dissertação tem por objetivo apresentar um estudo do uso de um *DW* combinado com um CBR para a verificação de "risco" de inadimplência no setor de telecomunicações. Setor este que devido as grandes mudanças que ocorreram no mercado, que passam desde a privatização do setor e a entrada de novas operadoras fixas e celulares, criando um ambiente de concorrência, anteriormente inexistente, possibilitando assim ao cliente trocar de operadora ou até mesmo deixar a telefonia fixa e ficar somente com a celular, e vai até ao fato da estabilização econômica e as novas práticas de mercado, que determinou a baixa das multas, tornando assim compensador aos clientes deixar as faturas vencidas a perder juros de aplicações ou pagar juros bancários para quitar a sua dívida, visto que a empresa telefônica só pode aplicar as sanções com o prazo de 30 dias.

Este trabalho mostra o desenvolvimento de um *CBR* para aplicação na área de Crédito e Cobrança, onde são detalhados os vários passos, a utilização do mesmo junto ao um *DW*, o que proporciona a comparação com desenvolvimento de outros sistemas similares e as diferenças (vantagens e desvantagens) que isso traz ao mesmo.

Palavras-chave: Inteligência Artificial, Data Warehouse, Raciocínio Baseado em Casos, Crédito e Cobrança.

TITLE: "USING CASE BASED REASONING TO CREDIT & COLLECTION ANALYSIS"

Abstract

Data Warehouse (*DW*) date is a process that agglutinates data from heterogeneous sources, including historical and external data to take care of to the necessity of structuralized and ad-hoc consultations, analytical reports and of decision support. A Case-Based Reasoning (CBR) is one technique of Artificial Intelligence (AI) for the representation of knowledge and inference, which aims the solution of new problems, considers adapted solutions that had been used to decide previous problems. The description of an existing problem, or a case, is used to suggest a way to fix a new problem, to inform the user of possible imperfections that had occurred in the past and to interpret the current situation.

This work has for objective to present a study of the use of a *DW* combined with a CBR for the verification of risk of insolvency in the sector of telecommunications. Sector this that which had the great changes that had occurred in the market, that they pass since the privatization of the sector and the entrance of new fixed and cellular operators, creating a competition environment, previously inexistent, thus making possible the customer to change of operator or even though to leave the fixed telephony and to be only with the cellular one, and goes until the o new fact of the economic stabilization and the practical ones of market, that determined low of the fines, thus becoming the compensator the customers to keep the invoices not paid, losing interests of applications or to pay to interests bank clerks for quitting its debt, since the telephonic company only can apply the sanctions with the stated period of 30 days.

This work shows a CBR development for application in the area of Credit & Collection, where the some steps are detailed, the it uses with an *DW*, what it provides to the comparison with development of other similar systems and the differences (advantages and disadvantages) that it brings.

Keywords: Artificial Intelligence, Data Warehouse, Case-Based Reasoning, Credit and Collection

1 Introdução

Nos últimos anos o governo brasileiro adotou diversas medidas voltadas à privatização do setor de telecomunicações visando à redução da atuação do Estado, em atividades que podem ser desenvolvidas pela iniciativa privada, gerando com isso, perspectivas de melhoria no acesso e no atendimento da população, no que diz respeito aos serviços de telecomunicações.

As variáveis que passaram a existir no setor, que antes era monopolizado pela União, sistema Telebrás e concessionárias estaduais conforme FIGURA 1.1, deram início a uma série de mudanças no mercado de telecomunicações.

A promulgação da Lei Mínima (telefonia móvel), da Lei Geral das Telecomunicações e a abertura de Licitação para exploração da "Banda B" da telefonia celular representaram os primeiros passos para a implantação dessa nova fase.

Entretanto, sem desconsiderar cada passo dado rumo à abertura do mercado nesse setor, ressalta-se que o marco principal foi o leilão das empresas que compunham o Sistema Telebrás, realizado no mês 07/98, abrangendo toda a operação de serviço fixo comutado e de longa distância.

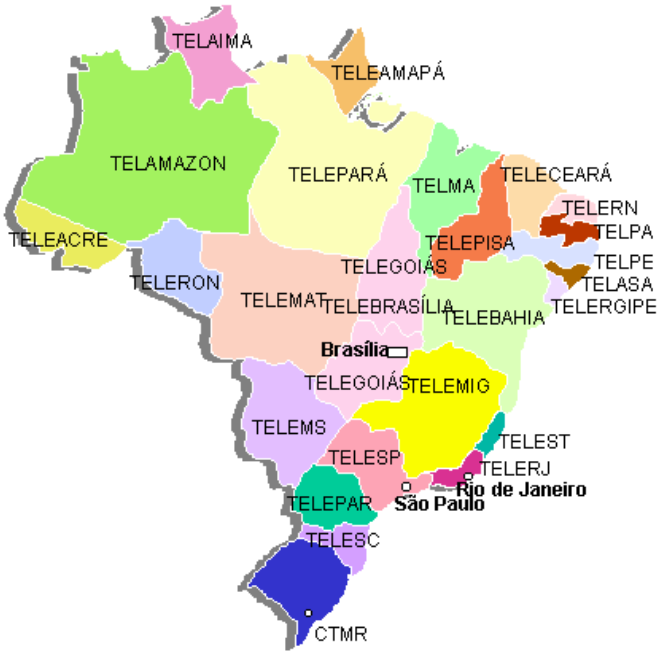


FIGURA 1.1 - Mapa do Sistema Telebrás (adaptado ANATEL)

Vários grupos formados por grandes empresas internacionais, nacionais, fundos de pensão e investidores de diferentes setores da economia participaram de maneira frenética desse momento histórico para o futuro das telecomunicações no país.

O Governo Federal criou a ANATEL – Agência Nacional de Telecomunicações, órgão incumbido, doravante, de outorgar as concessões, regulamentar a atuação das empresas concessionárias e fiscalizar sua atuação pós-privatização.

O mapa das concessões do STFC – Sistema Telefônico Fixo Comutado foi dividido em três regiões conforme FIGURA 1.2.

FIGURA 1.2 - Mapa STFC após privatização do setor (adaptado ANATEL)

Neste modelo passou a existir duas operadoras STFC por região, chamadas de *incumbent*, que são aquelas já existentes e privatizadas, e as empresas "*espelhos*", que são as novas empresas compradoras de concessão junto a ANATEL. Ambas possuem normas e objetivos a serem cumpridos para que mantenham a concessão, o chamado Plano de Universalização.

Com isso surge um novo cenário também mercadológico, onde: a) os clientes passaram a ter opção de operadoras; b) disputa pelo mercado destes clientes e, c) a maior modificação consubstancia-se no fato de que o telefone deixa de ser um patrimônio pessoal para se tornar somente um serviço público.

Isto altera a forma com que o então usuário e o agora cliente, passa a ter com este mercado, pois deixa a estar a mercê de um serviço monopolizado, para poder escolher a operadora, avaliando vantagens (sejam estas tecnológicas ou financeiras), tendo o apoio do Código de Defesa do Consumidor e a própria ANATEL para defender os seus direitos.

Além da modificação do cenário de telefonia fixa, também cresce a guerra no mercado de telefonia celular e no mercado das *carriers*, empresas encarregadas em prover serviços de ligações de Longa Distancia Nacional (DDD) e Internacional (DDI).

Com isso, além de muitas oportunidades, as empresas deste novo mercado passam a conviver com alguns problemas. Por exemplo, o mercado depara-se com a inadimplência, principalmente nas empresas espelhos, *carriers* (empresas encarregadas de telefonia de longa distância) e celulares, que passam a disputar, após a saturação em alguns produtos, mercados de classes econômicas antes não atendidas e que não possuíam a "cultura da telefonia".

Não somente estes clientes, mas também aqueles que mesmo tendo esta "cultura da telefonia", deixaram de honrar seus compromissos por despreocupação, agora com opção de passar para uma outra operadora ou a fazer ligações utilizando diferentes códigos de acesso.

Outro ponto agravante neste contexto é a estabilização econômica e os altos juros bancários praticados no mercado. Isso provocou a prática de multas muito baixas, comparando-se aos juros bancários, fazendo com que os clientes, quando necessitassem, optar por não pagar a conta telefônica em detrimento a pagar juros bancários ou a outros serviços (água, luz ou telefones celulares).

Toda esta modificação, apesar de ter sido prevista, aconteceu em um prazo muito curto, cerca de cinco anos, fazendo com que todo o mercado se agitasse.

Desta forma, meio a todas as variáveis, surge à necessidade de ferramentas para estar analisando o perfil destes clientes e os fatores que os levam à inadimplência, assim como, de alguma forma, dimensionar os riscos destes.

Por conseguinte, este trabalho tem como objetivos estudar essas técnicas em conjunto para a resolução de problemas não convencionais de administração de organizações e desenvolver uma aplicação de exemplo do uso de duas técnicas:

- ***Business Intelligence*** (BI): que de uma forma ampla pode ser entendido como a utilização de variadas fontes de informação para se definir as estratégias de competitividade nos negócios da empresa. Neste caso os fatores do problema em questão. O universo empresarial hoje padece de um mal clássico, possui uma montanha de dados, mas enfrenta grande dificuldade na extração de informações a partir dela. Assim esta ferramenta tem um papel fundamental, no estudo de casos feito, para que o especialista possa aferir comportamentos e também tecnicamente falando, para que seja facilitada a carga de informações para o CBR.

- ***Case-Based Reasoning*** (CBR): que pelo problema em questão, onde há uma grande quantidade de variáveis a serem julgadas, esta aplicação tem uma série de vantagens como a fácil aquisição de conhecimento, pelos próprios bancos de dados e ocorrências já registradas pela organização.

A primeira parte do trabalho traz a introdução, que visa nivelar o conhecimento e o entendimento sobre os aspectos relevantes do mercado e a problemática que será estudada e algumas variáveis que motivaram o desenvolvimento do estudo.

A segunda parte faz uma revisão teórica de conceitos como *DW* e CBR. O terceiro capítulo faz uma análise estudo de caso, com todas as suas variáveis e aspectos relevantes levantados durante o estudo do processo de Crédito e Cobrança. O quarto capítulo explica as técnicas utilizadas, seu funcionamento e seus resultados. O quinto, e último capítulo traz as conclusões gerais do trabalho.

2 Definições

2.1 Business Intelligence

Conceitualmente, ***Business Intelligence*** – BI (Inteligência de Negócios), de uma forma abrangente, pode ser entendido com a utilização de variadas fontes de informação para se definir estratégias de competitividade nos negócios da empresa. O universo empresarial, nos dias atuais, sofre um mal, que também atinge as pessoas, que é o sobrecarga de informação. Ou seja, possui uma quantidade grande de dados, mas isso dificulta a tomada de decisão, na medida em que a alta e média gerência se sentem impotentes no processo de busca, recuperação e análise destas informações.

O foco de um sistema de BI é o processo de tomada de decisão. Alguns sistemas de BI se especializam em informações sobre clientes, aproximando dos sistemas de CRM (*Customer Relationship Management*). Segundo [NON97], o uso criativo de redes de comunicação e bancos de dados facilita o processo de combinação. Sistemas de BI apresentam recursos para ordenar, categorizar e estruturar informação. A principal diferença entre o BI e o GED (Gerenciamento Eletrônico de Documentos) é que o primeiro baseia-se em registros bem formatados de bancos de dados, enquanto que o segundo lida com documentos em sua maioria não-estruturados e nos mais diversos formatos.

Esta comparação também pode ser feita quando se compara BI, CI (*Competitive Intelligence*) e KMS (*Knowledge Management System)*. [BAR2001] entende BI, conforme demonstra a Figura 3, como sendo um "guarda chuva conceitual" que envolve tanto CI como KMS. Segundo o autor KMS objetiva estabelecer uma aproximação integrada e colaborativa para capturar, criar, organizar e usar todos os ativos de informação de uma empresa. Enquanto a técnica BI é mais compartimentada, objetiva e focada em estruturas definidas, a de KMS trabalha o ativo de informações independentemente de forma, estrutura e domínio. Já comparando BI e CI, [BAR2001] define CI como o objetivo de explora o "outro lado da trincheira" obtendo informações detalhadas sobre os competidores e o mercado. E a estratégia está em focar exatamente as informações que se deseja, buscar estas informações nos mais variados tipos de publicações e transformar estas informações em algo estruturado para o armazenamento em *DW* específicos para CI. O autor ainda faz a seguinte afirmação:

> "*Podemos entender CI como um BI aplicado ao mundo fora de nossas fronteiras empresarias, focado primariamente em informações textuais e factuais que dizem respeito aos movimentos do mercado de nossas concorrentes. A estratégia de CI passa pelas vizinhanças dos conceitos emergentes de KMS e se acopla aos de Mining ...*".

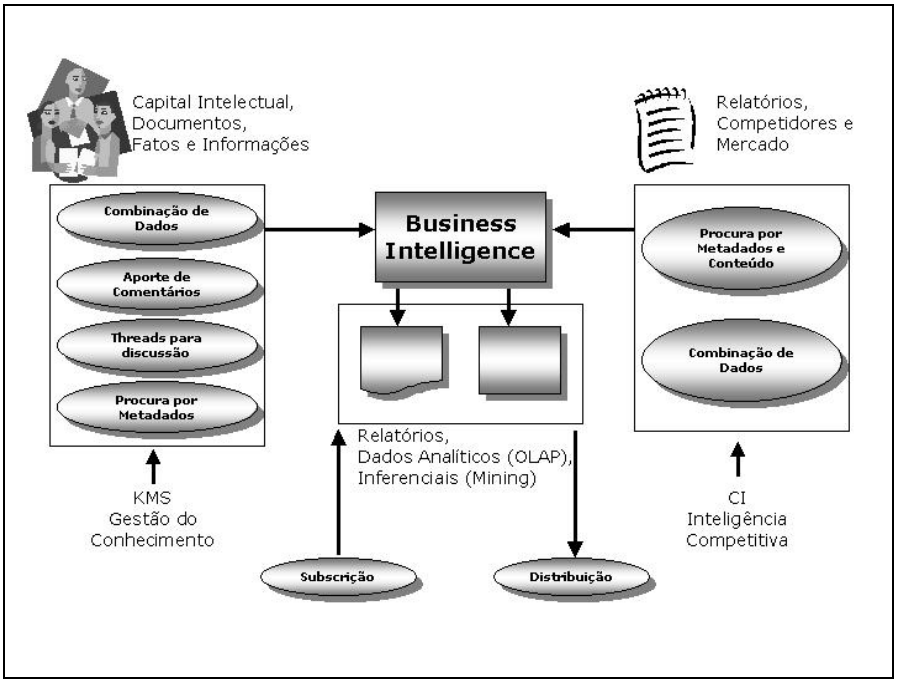


FIGURA 2.1 - Visão das Relações entre BI, CI e KMS (adaptado de [BAR2001])

2.1.1 Desenvolvimento de BI

Os conceitos de BI podem ser entendidos como o apoio e subsidio aos processos de tomada de decisão baseados em dados trabalhados especificamente para a busca de vantagens competitivas. Os dados armazenados nos sistemas tradicionais legados estão formatados e estruturados de uma maneira que dificulta o seu tratamento informacional. Desta forma, um BI deve ser entendido como o processo de desenvolvimento de:

- **Sistemas de *Front-End***:
  - SAD (Sistemas de Apoio à Decisão);
  - EIS (Executive Information Systems); e
  - OLAP (*On-Line Analytical Processing*): ferramentas de consulta analítica;
  - OLTP (*On-Line Transaction Processing*): ferramentas de consulta a dados operacionais.

- **Sistemas de Back-end:**
  - Data Warehouse (*DW*);
  - Data Mart (*DM*);
  - Data Mining: ferramentas de mineração de dados.

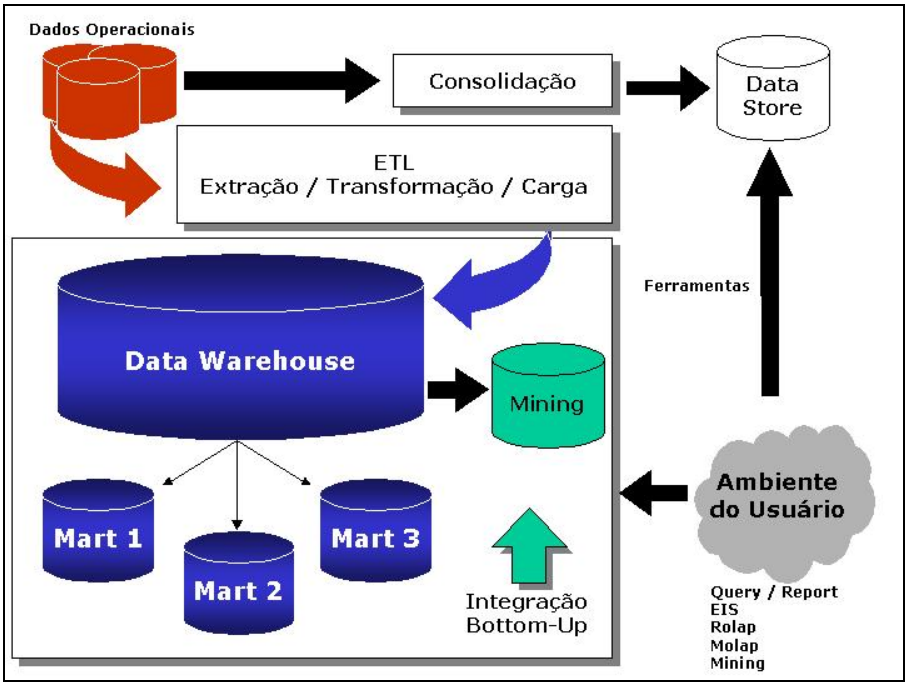


FIGURA 2.2 - Componentes de um ambiente de BI [BAR2001]

2.1.2 Sistemas de Front-End

2.1.2.1 Sistemas de Apoio à Decisão

A principio um Sistema de Suporte à Decisão – DSS (*Decision Support System*) identifica qualquer sistema que pretenda suporte as pessoas que tem a tarefa de tomar decisões em situações não estruturadas. DSS existe para ser um recurso, para aumentar a capacidade dos tomadores de decisão e não para substituir o seu julgamento. Objetivam as decisões onde julgamentos são requeridos, mas em que os algoritmos não suportam completamente. Mesmo sem determinação exata, somente com definições iniciais, há a noção que o sistema deva ser baseado em recursos computacionais, que deva operar de forma interativa e *on-line* e que, preferencialmente, tenha um recurso gráfico para as saídas de dados.

2.1.2.2 Definições

LITTLE (apud [TUB99]) define um DSS como "*um modelo baseado em um conjunto de procedimentos para o processamento de dados e julgamentos para auxiliar um administrador em sua tomada de decisão*". Argumentando ainda que para o sucesso deste sistema este deve ser, simples, robusto, fácil de controlar, adaptativo, completo nas situações importantes e de fácil interação. Implícita nesta definição está à concepção de que o sistema é baseado em recursos computacionais e serve como extensão da capacidade do usuário em solucionar problemas.

ALTER (apud [TUB99]) define um DSS pela comparação do mesmo com um tradicional sistema de EDP (*Eletronic Data Processing*) em cinco dimensões, conforme demonstrado na TABELA 2.1.

TABELA 2.1 – Comparativo entre DSS e EDP

Dimensão	DSS	EDP
Uso	Ativo	Passivo
Usuário	Administração	Técnico
Meta	Eficiência	Eficiência Mecânica
Horizonte de Tempo	Presente e Futuro	Passado
Objetivo	Flexibilidade	Consistência

BONCZEK (apud [TUB99]) define um DSS como um sistema baseado em recursos computacionais consistindo de três componentes:

- **um sistema de linguagem**: um mecanismo para prover comunicação entre o usuário e os outros componentes do DSS;

- **um sistema de conhecimento:** o repositório de conhecimento do domínio do problema incluído em um DSS, assim como os dados e os procedimentos;

- **um sistema de processamento dos problemas:** é a conexão entre os dois componentes anteriores, contendo uma ou mais capacidade de manipulação geral de problemas, requerida para a tomada de decisão.

[TUB99] explica que a definição formal de um DSS não provém um foco consistente, devido ao fato de que cada um tenta limitar o escopo. Além disso, ignoram a questão central de um DSS que seria: *suportar e melhorar a tomada de decisão*.

2.1.2.3 Características de um DSS

Devido ao fato de não haver um consenso no que, exatamente, seria um DSS, obviamente não há um acordo sobre quais são as características padrão de um DSS, mas pode-se dizer que o ideal seria:

- um DSS prover suporte para tomadores de decisão, principalmente em situações não estruturadas ou semi-estruturadas, trazendo informações computadorizadas para comb inar com o julgamento humano;

- suportar os vários níveis de administração (estratégica, tática e operacional);

- suportar tanto para indivíduos quanto para grupos. Uma quantidade menor de problemas estruturados, muitas vezes envolve diversos indivíduos de diferentes departamentos e de diferentes níveis organizacionais;

- prover suporte para decisões seqüenciais e interdependentes;

- suportar todas as fases do processo de tomada de decisão: inteligência, planejamento, escolha e implementação;

- suportar uma variedade de estilos e processo de tomada de decisão;

- serem adaptativos através do tempo. Sendo que o tomador de decisão deve ser reativo, ser capaz de enfrentar as mudanças de condições de forma rápida e adaptar o DSS para encontrar estas modificações. Assim o DSS deve ser adaptável para que o usuário possa adicionar, excluir, combinar, modificar ou re-arranjar elementos básicos;

- permitir aos usuários "conforto" ao trabalharem com um DSS. Interatividade, forte capacidade gráfica e uma interface lingüística que pode melhorar consideravelmente a efetividade de um DSS;

- melhorar a efetividade do processo de tomada de decisão (exatidão, conveniência e qualidade);

- permitir ao tomador de decisão controle completo sobre todos os processos de decisão. Um DSS especialmente objetiva suportar e não tomar o lugar do tomador de decisão;

- tornar os usuários aptos a construir e modificar sistemas simples por si mesmos. Já os sistemas mais abrangentes podem ser construídos com o auxilio de especialistas em sistemas de informação;

- utilizar modelos para analisar situações de tomada de decisão. A função de modelagem deve disponibilizar a "experimentação" com diferentes estratégias e sob diferentes configurações;

- prover o acesso a uma variedade de fonte de dados, formatos e tipos podendo ir de sistemas de informações geográficas a sistemas orientados a objetos.

Estas características disponibilizam ao tomador de decisão a fazer melhor, com mais consistência e são providas por um conjunto de componentes principais.

2.1.2.4 Componentes de um DSS

Um DSS é composto pelos seguintes subsistemas:

- Administração *de Dados*: inclui o banco de dados, onde estão contidos os dados das situações relevantes e são administrados por um Sistema de Administração de Banco de Dados (DBMS);

- Administração *de Modelo:* um pacote que inclui modelos financeiros, estatísticos, científicos ou outros modelos quantitativos;

- Administração *do Conhecimento:* este sistema pode suportar qualquer dos outros subsistemas ou age como um componente independente. É este subsistema que tem como objetivo prover a inteligência para aumentar a capacidade na tomada de decisão;

- *Interface com o usuário:* modulo de interação entre o usuário e os comando para o DSS;

- *Usuário*: o próprio usuário é considerado como um subsistema. Pesquisadores declaram que algumas das contribuições do DSS derivam da interação entre o recurso computacional e o tomador de decisão;

2.1.3 Sistemas de Informações Gerenciais

Sistemas de Informações Gerenciais – EIS (*Executive Information System*) dá aos executivos, ferramentas para analisar a grande quantidade de informações na tomada de decisões em "tempo real" sem o auxílio de outras pessoas, como especialistas, por exemplo. Podendo, por exemplo, os executivos de marketing responder as suas próprias questões e incluir suas próprias variáveis. O sistema é altamente intuitivo e é mantido pela própria área onde é usado, que continuamente o atualiza e o melhora.

2.1.3.1 Definições

Os termos "sistemas de informações gerenciais" e sistemas de suporte gerenciais significam coisas diferentes a pessoas diferentes. Muitas vezes os termos podem ser usados de forma intercambiada, mas as definições baseadas em DELONG (apud [TUB99]) distinguem estes dois tipos de sistemas:

- *Sistemas de Informações Gerenciais (EIS)*: é um sistema baseado em recursos computacionais que esta a serviço dos executivos de uma empresa, provendo as informações que os mesmos necessitam. Este sistema tem como objetivo prover acesso rápido às informações e acesso direto a relatórios gerenciais. O EIS é muito interativo, suportado por gráficos, relatando as exceções e a função de *drill-down* (esta função será detalhada a seguir).

- *Sistemas de Suporte Gerencial (ESS)*: é um sistema de suporte global, que dá suporte além de um EIS, incluindo comunicação, automatização de escritório, suporte de análise e inteligência.

2.1.3.2 Características de um EIS

As características desejadas de um EIS e alguns de seus benefícios são apresentados a seguir, entretanto nem todas as implementações dos sistemas incluem todas essas características:

Qualidade da Informação

- flexível;
- correta;
- instantânea;
- relevante;
- completa;
- validada;

Interface com o Usuário

- sofisticada interface gráfica com o usuário;
- interatividade;
- acesso seguro à informação;
- tempo de resposta curto;
- acessível de vários lugares;
- procedimento de acesso confiável;
- minimiza o uso de comandos, usando outros sistemas, como mouse e tela interativa;
- desenhado para administrar o estilo individual dos usuários;
- contém menu de alto ajuda;

Capacidade Técnica

- acesso para agregar informação “global”;
- acesso a correio eletrônico;
- uso “pesado” de dados externos;
- interpretações escritas;
- indicador de problemas;
- hipertexto e hipermídia;
- análise *ad- hoc*;
- apresentação e análise multidimensional;
- apresentação de informação de forma hierárquica;
- incorporar gráfico e texto na mesma exibição;
- mostrar tendências, relações e desvios;
- acesso a dados históricos;
- organizado em torno dos fatores críticos de sucesso;

Benefícios

- facilita a realização dos objetivos organizacionais;
- facilita o acesso a informação;
- permite ao usuário ser mais produtivo;
- aumenta a qualidade no processo de tomada de decisão;
- provém uma vantagem competitiva;
- “economiza” tempo do usuário;

- melhora a capacidade de comunicação;
- melhora a qualidade da comunicação;
- provém melhor controle na organização;
- permite a antecipação de problemas e oportunidades;
- permite o planejamento;
- permite encontrar a causa para alguns tipos de problemas;
- encontra a necessidade dos executivos.

2.2 Sistemas de Back-End

2.2.1 Data Warehouse

Para [INM92] um *DW* é uma coleção de dados orientada por assunto, integrada, variante no tempo e não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão. Da mesma forma, [HAC95] afirma que o objetivo de um *DW* é fornecer uma "imagem única da realidade do negócio". De uma forma geral, sistemas de *DW* compreendem um conjunto de programas que extraem dados do ambiente de dados operacionais da empresa, um banco de dados que os mantém e sistemas que fornecem estes dados aos usuários.

[INM94] coloca o *DW* como epicentro da infra-estrutura estratégica para a empresa, suportando informacional promovendo uma sólida plataforma de dados históricos integrados para serem analisados com visão corporativa. Sendo que estes revitalizam sistemas da empresa, pois:

- permitem que sistemas mais antigos continuem em operação;

- consolidam dados inconsistentes dos sistemas mais antigos em conjuntos coerentes;

- extraem benefícios de novas informações oriundas das operações coerentes;

- provém um ambiente para planejamento de novos sistemas de cunho operacional.

É importante considerar, no entanto, que um *DW* não contem apenas dados resumidos, podendo conter também dados primitivos. É desejável prover ao usuário a capacidade de aprofundar-se num determinado tópico, investigando níveis de agregação menores ou mesmo o dado primitivo, permitindo também a geração de novas agregações ou correlações com outras variáveis. Além do mais, é extremamente difícil prever todos os possíveis dados resumidos que serão necessários: limitar o conteúdo de um *DW* apenas a dados resumidos significa limitar os usuários apenas às consultas e análises que eles puderem antecipar frente a seus requisitos atuais, não deixando qualquer flexibilidade para novas necessidades.

2.2.1.1 Abordagens Principais

Os primeiros projetos de *DW* muito provavelmente caminharam pelas linhas metodológicas de duas principais fontes: [INM92] e [KIM98] já que estes dois autores se constituíram nas referencias mais fortes em termos de projetos de *DW* e *DM*.

As metodologias apresentam têm algumas diferenças, conforme detalhado a seguir.

2.2.1.1.1 INMON

A abordagem deste autor se concentrou inicialmente no estilo mais tradicional de construção de Bando de Dados, muito próximo daquele surgido nos primeiros projetos de BD, onde se buscava uma forte integração entre todos os dados da empresa, que habitavam em áreas funcionais diferentes. Isso seria representado em um modelo único, integrado e conciso, mas que se mostrou rígido e de difícil consecução.

O ciclo metodológico básico se iniciava pela análise de requerimentos de negócio, seguido pela modelagem de dados e pela análise das fontes desses dados, fator preponderante no sucesso de qualquer *DW*. Abordagem muito parecida com qualquer projeto de banco de dados.

O diferencial da abordagem deste autor está baseado no oferecimento de um conjunto de ferramentas fundamentais para se estabelecer um projeto de *DW*, principalmente com ampla abrangência.

2.2.1.1.2 KIMBALL

Com um estilo mais simples e incremental. A metodologia chamada "*Star Schema*" (modelo estrela) aponta para projetos de *DM* separados, que deverão ser integrados na medida da sua evolução.

O autor já "rebatizou" essa abordagem de "*Super Marts"*. Os projetos serão menores, independentes, focando áreas ou assuntos específicos e terão a sua conexão com o passar do tempo, desde que mantidas a compatibilidade dimensional entre chaves das tabelas. Sendo muito similar aos projetos de banco de dados relacionais recentemente utilizados nas empresas, com enfoque departamental ou por assunto.

Tem como desvantagem a possibilidade de se produzir diversos *DM*, sem uma perfeita conexão entre os mesmos, além de uma provável duplicação de esforços na fase de extração, preparação e carga de dados (ETL).

Esta abordagem tem como essência a etapa de projeto de *DM*. Centrada na modelagem dimensional, com o conceito de **"modelo estrela"** (*Star Schema*), essa abordagem transforma os dados em tabelas *Fatos* (onde se concentram os dados de interesse, passiveis de manipulação numérica e estatística), e também em tabelas *Dimensão* (tabelas satélites que possuem as chaves de entrada do modelo, além de informações descritivas de cada dimensão), conforme mostra a FIGURA 2.3.

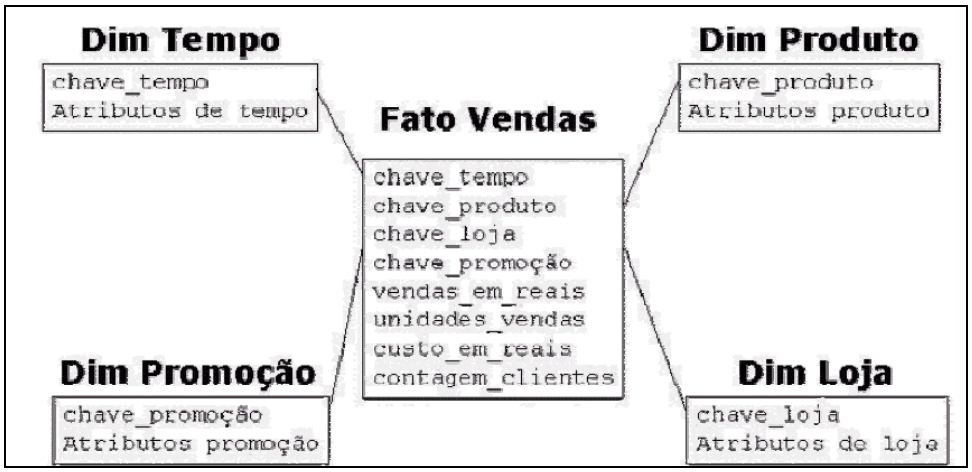


FIGURA 2.3 - Modelo "Star Schema" [KIM98]

2.2.1.1.3 Convergência das Abordagens

O que se pode notar atualmente é que os *DW* começam a ser construídos pelos primeiros *DM* demandados. Mas a compatibilidade de dimensões dos *"Super Marts"* planejados e alguns dos padrões de dados das tabelas "fatos" são fundamentais para minimizar possíveis rupturas.

Os *DM* seguintes deverão ser integrados mantendo a conformidade entre as dimensões ao longo do tempo, conforme FIGURA 2.4.

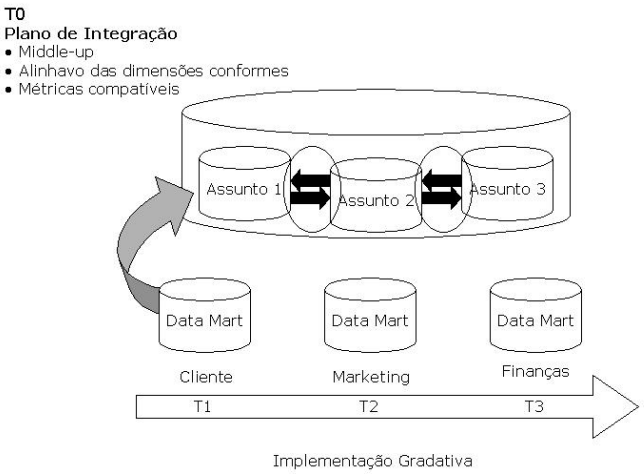


FIGURA 2.4 - Estratégia para implementação gradativa de *DM* [BAR00]

2.2.1.2 Aplicações típicas

[INM94] coloca que as aplicações típicas podem ser classificadas em dois grandes conjuntos:

a) **aplicações do negócio**: constituem as aplicações que dão suporte ao dia a dia do negócio da empresa, que garantem a operação da empresa, também chamadas de sistemas de produção;

b) **aplicações sobre o negócio**: são as aplicações que analisam o negócio, ajudando a interpretar o que ocorreu e a decidir sobre estratégias futuras para a empresa - compreendem os sistemas de suporte à decisão e sistemas de informações executivas.

Uma arquitetura de dados adequada para dar suporte aos dois tipos de aplicações baseia-se em dois ambientes de bancos de dados: os bancos de dados operacionais - para dar suporte às **aplicações do negócio** - e os bancos de dados para suporte à decisão - para dar suporte às **aplicações sobre o negócio**, conforme citado anteriormente.

2.2.1.3 Características do *Data Warehouse*

[INM92] descreve as seguintes características do *DW*:

- **Orientado por temas**: refere-se ao fato do *DW* armazenar informações sobre temas específicos importantes para o negócio da empresa. Exemplos típicos de temas são: produtos, atividades, contas, clientes, etc. Em contrapartida, o ambiente operacional é organizado por aplicações funcionais;

- **Integrado:** refere-se à consistência de nomes, das unidades das variáveis, etc., no sentido de que os dados foram transformados até um estado uniforme;

- **Variante no tempo:** refere-se ao fato do dado em um *DW* referir-se a algum momento específico, significando que ele não é atualizável, enquanto que o dado de produção é atualizado de acordo com mudanças de estado do objeto em questão, refletindo, em geral, o estado do objeto no momento do acesso. Em um *DW*, a cada ocorrência de uma mudança, uma nova entrada é criada, para marcar esta mudança;

- **Não volátil:** significa que o *DW* permite apenas a carga inicial dos dados e consultas a estes dados. Após serem integrados e transformados, os dados são carregados em bloco para o *DW*, para

que estejam disponíveis aos usuários para acesso. No ambiente operacional, ao contrário, os dados são, em geral, atualizados registro a registro, em múltiplas transações. Esta volatilidade requer um trabalho considerável para assegurar integridade e consistência através de atividades de recuperação de falhas e bloqueios. Um *DW* não requer este grau de controle típico dos sistemas orientados a transações.

2.2.1.4 Arquitetura do Data Warehouse

Nos últimos anos, o conceito de *DW* evoluiu rapidamente de um considerável conjunto de idéias relacionadas para uma arquitetura voltada para a extração de informação especializada e derivada a partir dos dados operacionais da empresa. O estudo de uma arquitetura descrevendo o ambiente de *DW* permite compreender melhor a estrutura geral de armazenamento, integração, comunicação, processamento e apresentação dos dados que servirão para subsidiar o processo de tomada de decisão nas empresas.

2.2.1.5 Camadas do *Data Warehouse*

Segundo [INM94] o *DW* possui as seguintes camadas:

- **Camada de Bancos de Dados Operacionais e Fontes Externas:** Corresponde aos dados das bases de dados operacionais da organização junto com dados provenientes de outras fontes externas que serão tratados e integrados para compor o *DW*;

- **Camada de Acesso à Informação:** É a camada com a qual os usuários finais interagem. Representa as ferramentas que o usuário utiliza no dia a dia, tal como Excell, SAS e outras. Também envolvem o *hardware* e *software* utilizado para obtenção de relatórios, planilhas, gráficos e outros. A cada dia surgem sistemas mais sofisticados para manipulação, análise e apresentação dos dados, incluindo-se ferramentas *de Data Mining* e visualização;

- **Camada de Acesso aos Dados:** Esta camada é responsável pela ligação entre as ferramentas de acesso à informação e os bancos de dados operacionais. Esta camada se comunica não só com diferentes SGBD e sistemas de arquivos de um mesmo ambiente como também, idealmente, com outras fontes sob diferentes protocolos de comunicação, no que se chama acesso universal de dados;

- **Camada de Metadados (Dicionário de Dados):** Metadados são as informações sobre os dados mantidos pela empresa (descrições de registro em um programa COBOL, comandos CREATE do SQL, informação em um diagrama E-R (Entidade - Relacionamento), dados em um dicionário de dados - são exemplos de metadados). Para poder manter a funcionalidade de uma ambiente de *DW* é

necessário ter disponível uma grande variedade de metadados, desde dados sobre as visões dos usuários até dados sobre os bancos de dados operacionais. Idealmente o usuário deve poder ter acesso aos dados de um *DW* sem que tenha que saber onde residem estes dados ou a forma como estão armazenados;

- **Camada de Gerenciamento de Processos:** A camada de gerenciamento de processos está envolvida com o controle das diversas tarefas a serem realizadas para construir e manter as informações do dicionário de dados e do *DW*. Esta camada é responsável pelo gerenciamento dos processos que contribuem para manter o *DW* atualizado e consistente;

- **Camada de Transporte ou *Middleware*:** Esta camada gerencia o transporte de informações pelo ambiente de redes. É usada para isolar aplicações, operacionais ou informacionais, do formato real dos dados nas duas extremidades. Também inclui a coleta de mensagens e transações e se encarrega de entregá-las em locais e tempos determinados;

- **Camada do *DW*:** O *DW* corresponde aos dados usados para fins "informacionais". Em alguns casos, *DW* é simplesmente uma visão lógica ou virtual dos dados, podendo de fato não envolver o armazenamento destes dados. Em um *DW* que exista fisicamente, cópias dos dados operacionais e externos são de fato armazenadas, de modo a prover fácil acesso e alta flexibilidade de manipulação;

- **Camada de Gerenciamento de Replicação:** Esta camada inclui todos os processos necessários para selecionar, editar, resumir, combinar e carregar o *DW* e as correspondentes informações de acesso a partir das bases operacionais e fontes externas. Normalmente isto envolve programação complexa, mas cada vez mais são disponibilizadas ferramentas para facilitar estes processos. Esta camada pode também envolver programas de análise da qualidade dos dados e filtros que identificam padrões nos dados operacionais.

2.2.1.6 Desenvolvimento do *Data Warehouse*

Os passos seguintes demonstram algumas abordagens e estratégias utilizadas e que devem ser consideradas no desenvolvimento de um *DW*.

2.2.1.6.1 *Abordagens para o Desenvolvimento de um DW*

O sucesso do desenvolvimento de um *DW* depende fundamentalmente de uma escolha correta da estratégia a ser adotada, de forma que seja adequada às características e necessidades específicas do ambiente onde será implementado. Existe uma variedade de abordagens para o desenvolvimento de *DW*, devendo-se fazer uma

escolha fundamentada em pelo menos três dimensões: escopo do *DW* (departamental, empresarial, etc), grau de redundância de dados, tipo de usuário alvo.

O escopo de um *DW* pode ser tão amplo quanto aquele que inclui todo o conjunto de informações de uma empresa ou tão restrito quanto um *DW* pessoal de um único gerente. Quanto maior o escopo, mais valor o *DW* tem para a empresa e mais cara e trabalhosa é sua criação e manutenção. Por isso, muitas empresas tendem a começar com um ambiente departamental e só após obter um retorno de seus usuários expandir seu escopo.

Quanto à redundância de dados, há essencialmente três níveis: o *DW* virtual, o *DW* centralizado e o *DW* distribuído. O *DW* virtual consiste em simplesmente prover os usuários finais com facilidades adequadas para extração das informações diretamente dos bancos de produção, não havendo assim redundância, mas podendo sobrecarregar o ambiente operacional. O *DW* central constitui-se em um único banco de dados físico contendo todos os dados para uma área funcional específica, um departamento ou uma

empresa, sendo usados onde existe uma necessidade comum de informações. Um *DW* central normalmente contém dados oriundos de diversos bancos operacionais, devendo ser carregado e mantido em intervalos regulares. O *DW* distribuído, como o nome sugere, possui seus componentes distribuídos por diferentes bancos de dados físicos, normalmente possuindo um grau de redundância alto e por conseqüência, procedimentos mais complexos de carga e manutenção.

2.2.1.6.2 Estratégia Evolucionária

[INM92] relata que os *DW*, em geral, são projetados e carregados passo a passo, seguindo uma abordagem evolucionária. Os custos completos de uma implementação, em termos de recursos consumidos e impactos no ambiente operacional da empresa justificam esta estratégia.

Muitas empresas iniciam o processo a partir de uma área específica da empresa, que normalmente é uma área carente de informação e cujo trabalho seja relevante para os negócios da empresa, criando os *DM*, para depois ir crescendo aos poucos, seguindo uma estratégia *botton-up* ou assunto-por-assunto.

Outra alternativa é selecionar um grupo de usuários, prover ferramentas adequadas, construir um protótipo do *DW*, deixando que os usuários experimentem com pequenas amostras de dados. Somente após a concordância do grupo quanto aos requisitos e funcionamento, é que o *DW* será de fato carregado com dados dos sistemas operacionais na empresa e dados externos.

2.2.1.6.3 Aspectos de Modelagem

[INM94] afirma que os principais aspectos de modelagem são:

- A especificação de requisitos do ambiente de suporte à decisão associado a um *DW* é fundamentalmente diferente da especificação

de requisitos dos sistemas que sustentam os processos usuais do ambiente operacional de uma empresa;

- Os requisitos dos sistemas do ambiente operacional são claramente identificáveis a partir das funções a serem executadas pelo sistema. Requisitos de sistemas de suporte à decisão são, por sua vez, indeterminados. O objetivo por trás de um *DW* é prover dados com qualidade; os requisitos dependem das necessidades de informação individuais de seus usuários. Ao mesmo tempo, os requisitos dos sistemas do ambiente operacional são relativamente estáveis ao longo do tempo, enquanto que os dos sistemas de suporte à decisão são instáveis: dependem das variações das necessidades de informações daqueles responsáveis pelas tomadas de decisões dentro da empresa.

No entanto, embora as necessidades por informações específicas mudem com freqüência, os dados associados não mudam. Imaginando-se que os processos de negócio de uma empresa permaneçam relativamente constantes, existe apenas um número finito de objetos e eventos com as quais uma organização está envolvida. Por esta razão, um modelo de dados é uma base sólida para identificar requisitos para um *DW*.

De qualquer forma, [INM92] salienta que é um erro pensar que técnicas de projeto que servem para sistemas convencionais serão adequadas para a construção de um *DW*. Os requisitos para um *DW* não podem ser conhecidos até que ele esteja parcialmente carregado e já em uso.

Outra questão interessante, discutida por [KIM98], é a adequação do modelo ER (Entidade-Relacionamento) ao tipo de transação de sistemas OLTP. O principal objetivo da modelagem, neste caso, é eliminar ao máximo, a redundância, de tal forma que uma transação que promova mudanças no estado do banco de dados, atue o mais pontualmente possível. Com isso, nas metodologias de projeto usuais, os dados são "fragmentados" por diversas tabelas, o que traz uma considerável complexidade à formulação de uma consulta por um usuário final. Por isso, salienta [KIM98], esta abordagem não parece ser a mais adequada para o projeto de um *DW*, onde estruturas mais simples, com menor grau de normalização devem ser buscadas.

2.2.1.6.4 *Etapas do Desenvolvimento de um DW*

Na verdade, é difícil apontar no momento, uma metodologia consolidada e amplamente aceita para o desenvolvimento de *DW*. Na literatura e nos casos de sucesso relatados sobre implementações em empresas, são propostas no sentido de construir um modelo dimensional a partir do modelo de dados corporativo ou departamental (base dos bancos de dados operacionais da empresa), de forma incremental. [INM92], salienta que, de fato, um *DW* é construído de uma maneira "heurística", confirmando a estratégia evolucionária discutida no item anterior.

De qualquer forma, a metodologia a ser adotada é ainda bastante dependente da abordagem escolhida, em termos de ambiente, distribuição, etc. A seguir, serão apresentadas as etapas sugeridas para um desenvolvimento do tipo estrela.

Segundo [KIM98], desenvolver um *DW* é uma questão de casar as necessidades dos seus usuários com a realidade dos dados disponíveis. [KIM98] ainda aponta um conjunto de pontos fundamentais no projeto da estrutura de um *DW*. São os seguintes os chamados pontos de decisão, que constituem definições a serem feitas e correspondem, de fato, a etapas do projeto:

- os processos, e por conseqüência, a identidade das tabelas de fatos;
- a granularidade de cada tabela de fatos;
- as dimensões de cada tabela de fatos;
- aos fatos, incluindo fatos pré-calculados;
- os atributos das dimensões;
- como acompanhar mudanças graduais em dimensões;
- as agregações, dimensões heterogêneas, minidimensões e outras decisões de projeto físico;
- duração histórica do banco de dados;
- a urgência com que se dá a extração e carga para o *DW*.

[KIM98] recomenda que estas definições sejam realizadas na ordem descrita. Esta metodologia segue a linha *top-down*, pois começa identificando os grandes processos da empresa.

2.2.1.7 Relacional x Multidimensional

Bancos de dados relacionais encontram em sua flexibilidade o potencial para consultas pré-definidas, um de seus pontos fortes. Bancos de dados relacionais são sabidamente mais flexíveis quando são usados com uma estrutura de dados normalizada. Uma típica consulta OLAP, no entanto, "atravessa diversas e requer diversas operações de junção para reunir estes dados. O desempenho dos sistemas de banco de dados relacionais tradicionais é melhor para consultas baseadas em chaves do que consultas baseadas em conteúdo".

Para atender os requisitos deste tipo de transações, fornecedores de SGBDs relacionais têm adicionado funcionalidades a seus produtos. Estas funcionalidades incluem extensões às estruturas de armazenamento e aos operadores relacionais e esquemas de indexação especializados. Estas técnicas podem melhorar o desempenho para recuperações por conteúdo através da pré-junção de tabelas usando índices ou pelo uso de listas de índices totalmente invertidas.

A maioria das ferramentas de acesso aos *DW* explora a natureza multidimensional dos dados. Por isso, estruturar os dados em bancos de dados relacionais tradicionais em esquemas do tipo estrela ou floco de neve tornou-se uma abordagem bastante comum. Estes esquemas podem usar múltiplas tabelas e ponteiros para simular uma estrutura multidimensional. Também é possível usar algum outro

mecanismo não relacional para armazenar algumas das agregações pré-calculadas enquanto outras são obtidas dinamicamente. Esta abordagem goza dos benefícios de um mecanismo relacional, tirando vantagem do cálculo prévio de algumas agregações. Normalmente a tabela central de fatos é bem grande enquanto as das demais dimensões são bem menores, conforme **"modelo estrela"** descrito anteriormente.

2.2.1.8 Granularidade

Granularidade, segundo [INM94] se refere ao nível de detalhe em que as unidades de dados são mantidas no *DW*. Quanto maior o nível de detalhes, menor o nível de ***granularidade***. Esta é uma questão fundamental no projeto de um *DW* porque afeta diretamente o volume de dados armazenados no *DW* e ao mesmo tempo, o tipo de consulta que pode ser atendida. O volume de dados contidos no *DW* é balanceado de acordo com o nível de detalhe de uma consulta.

2.2.1.9 Povoando um *Data Warehouse*

A extração, limpeza, transformação e carga de dados dos sistemas existentes na empresa para o *DW* (processo este conhecido como ETL). Segundo [INM94] constituem tarefas críticas para o seu funcionamento efetivo e eficiente. Diversas técnicas e abordagens têm sido propostas, algumas bastante genéricas e outras especialmente voltadas para a manutenção de integridade dos dados num ambiente caracterizado pela derivação e replicação de informações.

[INM94] afirma, ainda, que os produtos oferecidos no mercado procuram automatizar processos que teriam de ser feitos manualmente ou utilizando ambientes de programação de mais baixo nível. De fato, não existe uma ferramenta única capaz de oferecer suporte aos processos de ETL, ferramentas especializam-se em questões específicas.

2.2.1.10 Extração de Dados

[INM94] acredita que as várias alternativas para extração permitem balancear desempenho, restrições de tempo e de armazenamento. Por exemplo, se a fonte for um banco de dados *on-line*, pode-se submeter uma consulta diretamente ao banco para criar os arquivos de extração. O desempenho, das aplicações ligadas às fontes tende, a prejudicar as transações "*on-line*" e consultas para extração competirem entre si. Uma solução alternativa é criar uma cópia corrente dos dados das fontes a partir da qual se faria então a extração. Como desvantagem desta solução, é o espaço adicional de disco necessário para armazenar a cópia.

Outra alternativa é examinar o ciclo de processamento de algumas transações *"off-line"* que atuem nas fontes. Os programas que criam os arquivos de extração para a carga do *DW* podem ser incorporados a um ponto apropriado deste esquema de processamento.

2.2.1.11 Transformação e Filtros

Os processos de transformação invocam as regras que convertem valores de dados das fontes para valores do ambiente global e integrado do *DW*. Algumas ferramentas permitem ao usuário controlar a maioria das atividades de exportação e transformação através de parâmetros e *scripts*, constituindo uma filtragem avançada. Outras ferramentas atuam como *shells* onde programas específicos de extração e filtragem escritos em linguagens de programação, como C ou COBOL, são inseridos.

A maioria das ferramentas comercial oferece alguma maneira de filtrar dados para garantia de qualidade, durante os processos de extração e transformação. Entretanto, ferramentas específicas para limpeza de dados oferecem mecanismos bem mais sofisticados. As melhores ferramentas para garantir qualidade são ainda aquelas desenvolvidas para áreas específicas como engenharia civil ou farmácia.

2.2.1.12 Incorporando Modificações

Sempre que modificações relevantes são feitas nas fontes de informações, estes novos dados são extraídos e traduzidos para o modelo de dados do *DW*, onde são integrados com os dados já existentes.

A detecção e extração das modificações dependem das facilidades disponíveis pela fonte. Se esta fonte for sofisticada, como um SGBD relacional que possui *triggers*[1], este processo é relativamente fácil. No entanto, em muitos casos a fonte não dispõe de recursos avançados para detecção e captura das modificações. Nestes casos, existem basicamente três alternativas, para detectar e extrair modificações:

- A aplicação que utiliza a fonte de informação é alterada de modo a enviar notificações de alteração para o *DW*. Esta alternativa requer que o código existente seja modificado. No entanto, na maioria dos casos esta opção é impraticável devido à complexidade do código e ao grande tempo necessário para sua alteração;

- O arquivo de *log* do sistema é analisado de modo a obter as modificações relevantes. O problema com esta alternativa é que normalmente são necessários privilégios do administrado do Banco de Dados para acessar o *log* e muitos administradores relutam em prover este acesso pois coloca em risco a segurança do sistema;

- As modificações são determinadas através da comparação do *dump* corrente da fonte com um *dump* anterior. A questão desta alternativa é a *escalabilidade*, ou seja, à medida que os dados fontes aumentam é necessário um número muito maior de comparações. Deste modo, torna-se necessário implementar este algoritmo do modo mais eficiente possível.

[1] Processos disparados automaticamente pela execução de um determinado evento no banco de dados.

Nas abordagens onde são mantidos o *DW* empresarial e *DM*. A necessidade de se estabelecer uma estratégia que coordene a entrega de novos dados a todos os bancos de dados. É preciso considerar a incorporação de um servidor de replicação na arquitetura de distribuição dos dados. Um servidor de replicação é uma aplicação sofisticada que seleciona e separa dados para distribuição para cada *DM*, aplicando restrições de segurança, transmitindo uma cópia dos dados para os locais adequados e criando um *log* de todas as transmissões.

2.2.1.13 Derivação e Sumarização

Diferentes alternativas também existem para prover suporte a dados. Uma abordagem segundo [INM92].é derivar os dados durante o processo de carga e armazená-los no ambiente relacional corporativo. Uma alternativa é fazer a derivação quando o servidor de replicação distribui os dados para os *DM*. Ou então, derivar os dados quando o usuário submeter uma consulta ou lançar uma simulação.

As ferramentas mais simples são os produtos para consultas e geradores de relatórios básicos. Em geral, oferecem uma interface gráfica para geração de consultas, permitindo o uso de menus e botões para a especificação de elementos de dados, condições, critérios de agrupamento, sem que seja necessário aprender uma linguagem especializada para acesso ao banco. O processamento estatístico, neste caso, é limitado a médias, totais, desvios padrão e algumas outras funções básicas de análise. Estes geradores de relatório não atendem usuários que precisem mais do que uma visão estática dos dados e que não pode mais ser manipulada. Ferramentas OLAP podem oferecer a este tipo de usuário maior capacidade de manipulação, permitindo analisar o porque dos resultados obtidos. Estas ferramentas, muitas vezes, são baseadas em bancos de dados multidimensionais, o que significa que os dados precisam ser extraídos e carregados para as estruturas proprietárias do sistema, já que não há padrões abertos para o acesso de dados multidimensionais.

Outra solução descrita é o OLAP relacional, que vai diretamente ao *DW* usando chamadas de consultas SQL padrão. As ferramentas *front-end* permitem efetuar requisições multidimensionais, mas o programa de OLAP relacional transforma consultas em rotinas SQL. O usuário recebe resultados cruzados de tabelas em forma de planilha multidimensional ou de outra forma que suporte a rotação e manipulação. Os defensores do OLAP relacional argumentam que ele utiliza padrões abertos de SQL e que traz os dados em um nível no nível mais detalhado, mais prontamente acessíveis. Por outro lado, os bancos multidimensionais argumentam que em uma estrutura multidimensional nativa alcança-se melhor desempenho e flexibilidade.

O OLAP não é uma solução imediata, configurar o programa de OLAP e ter acesso aos dados requer uma clara compreensão dos modelos de dados da empresa e das funções analíticas necessárias aos executivos e outros analistas de dados.

Comparativamente ao OLAP, [INM94] afirma que Sistemas de Informações Executivas apresentam uma visualização de dados mais simplificada, altamente consolidada e, na maior parte das vezes, estática. Até porque, em geral, os executivos não dispõem do tempo e da experiência para executar uma análise OLAP.

2.2.2 Data Mart

Muitas empresas iniciam o processo a partir de uma área específica da empresa, que normalmente é uma área carente de informação e cujo trabalho seja relevante para os negócios da empresa, criando os chamados *DM*, para incrementa-los gradativamente, seguindo uma estratégia "*botton-up"* ou assunto a assunto.

Também pode ser feita a seleção um grupo de usuários, prover ferramentas adequadas, construir um protótipo do *DW*, deixando que os usuários experimentem com pequenas amostras de dados. Somente após a concordância do grupo quanto aos requisitos e funcionamento, é que o *DW* será de fato carregado com dados dos sistemas operacionais na empresa e dados externos.

DM também podem ser criados como subconjunto de *DW* maior, em busca de autonomia, melhor desempenho e simplicidade de compreensão.

2.2.3 Fatores Críticos de Sucesso em Projetos de *DW* e *DM*.

- Foco bem definido: é fundamental num projeto desses que se saiba exatamente o que se deseja obter, mesmo que isso leve tempo para ser garimpado;

- Patrocinador forte: fundamental que num projeto que possa atingir áreas correlatas, haja uma presença forte do usuário;

- Dados necessários: antes de definir quaisquer compromissos com os patrocinadores do projeto, verificar com cuidado a fonte de dados;

- Alto envolvimento dos usuários;

- Bom time de projeto;

- Definição de uma boa arquitetura tecnológica;

- Estimulo e acompanhamento da utilização da aplicação.

2.2.4 Data Mining

Data Mining (mineração de dados) é o processo de analisar dados por diferentes perspectivas e resumi-los em informação aproveitável, que poderá ser utilizada melhorar os ingressos, reduzir custos ou outras ações úteis para a empresa. Uma ferramenta de ***Data Mining*** é ferramenta analítica que permite analisar dados por diferentes dimensões ou ângulos, categorizá-los e resumir os relacionamentos encontrados. Tecnicamente, ***Data Mining*** é o processo de encontrar correlações ou padrões entre milhares de campos em grandes bases de dados. Informações, que aparentemente estão camufladas ou escondidas, permitindo agilidade na tomada de decisões.

Uma empresa que emprega a técnica de ***Data Mining*** é capaz de:

- Criar parâmetros para entender o comportamento do consumidor;
- Identificar afinidades entre as escolhas de produtos e serviços;
- Prever hábitos de compras;
- Analisar comportamentos habituais para se detectar fraudes.

A análise automatizada e antecipada oferecida pelo ***Data Mining***, vai muito além da simples análise de eventos passados, que é fornecida pelas ferramentas de retrospectiva típicas de sistemas de apoio à decisão como consultas no padrão SQL do universo transacional (OLTP). [SAN2001]

Por esta característica o ***Data Mining*** é extremamente adequado para analisar estes grupos de dados, que seriam difíceis de serem analisados usando apenas OLAP, devido à grandeza, alta densidade ou não serem intuitivos, para serem compreendidos facilmente.

O ***Data Mining*** pode ser considerado como sendo uma forma de KDD (*Knowledge Discovery in Databases*), área de pesquisa em evidência no momento, que envolve AI. A construção de modelos estatísticos costuma ser o método tradicionalmente utilizado para derivar tendências [DAT2000A]. Entretanto, estatísticas tradicionais são limitadas se se considerando:

- a análise se torna trabalhosa quando o número de variáveis a serem investigadas cresce;

- métodos estatísticos possuem condições que limitam o número de casos

- a utilizar, fazendo com que apenas uma pequena parte do universo esteja disponível para a análise;

- quando os relacionamentos dos dados são não lineares, torna-se difícil aplicar métodos estatísticos tradicionais.

Além disso, poucas empresas dispõem de pessoal preparado para esta tarefa tão especializada, mesmo com a ajuda de estatísticos, pode-se levar semanas para projetar e construir os modelos.

A estrutura de um ***Data Mining*** está baseada em Estatística, AI e em *Machine Learning*, conforme apresentado na FIGURA 2.5.

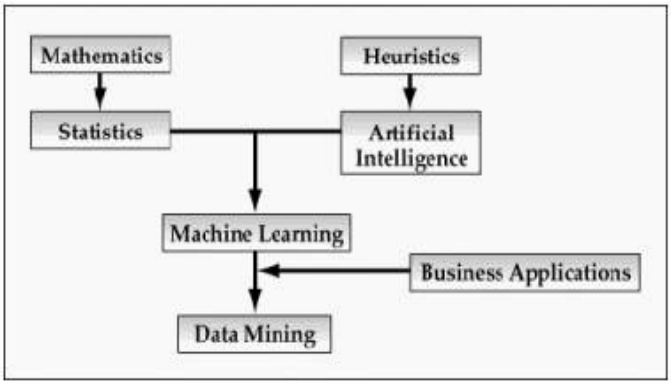


FIGURA 2.5 - Base de um *Data Mining* [DAT2000A]

2.2.4.1 O processo de Data Mining

Como todos os processos de descoberta de conhecimento em banco de dados (KDD – *Knowledge Discovery in Database*) existem duas grandes fases [AMA2001], estas duas fases possuem inúmeros passos, os quais envolvem um número elevado de decisões a serem tomadas pelo usuário, ou seja, é um processo interativo, pois ao longo do processo de KDD, um passo será repetido tantas vezes quantas se fizerem necessárias para que se chegue a um resultado satisfatório, sendo que de forma geral os passos principais são:

- **preparação de dados**: que envolve os seguintes passos:

  - definição do objetivo do problema, que é o conhecimento desejado pelo usuário final, ou seja, é definido o tipo de conhecimento que se deseja extrair do banco de dados. Nessa fase, são analisados o conhecimento da aplicação e a verificação do conhecimento anterior;

  - criação de um conjunto de dados-alvo. Nesse passo, seleciona-se um conjunto de dados ou focaliza-se em um subconjunto de atributos ou de instancias de dados, em que a descoberta deverá ser efetuada. Muitas vezes, o sucesso depende da correta escolha dos dados que formam o conjunto de dados-alvo. Para isso, são utilizados técnicas, linguagens, ferramentas e comandos convencionais de bancos de dados como o padrão SQL, por exemplo;

  - limpeza e pré-processamento dos dados. Nesse caso, deve-se fazer a limpeza dos dados de maneira que os incorretos ou incompletos sejam desprezados. Com isso, é feita uma purificação dos dados usando operações básicas, como as de definição de ruídos. Nela são coletadas as informações necessárias para a modelagem e correção do ruído e para estratégias de manipulação de campos de dados perdidos, considerando as seqüências de informações de tempo e as mudanças de conhecimento;

- redução e projeção de dados, que constitui em encontrar as características úteis que representam as dependências dos dados no objetivo do processo. Muitas vezes, pode não ser necessário representar todas as faixas de valores de um determinado problema. Assim, pode-se reagrupar esses valor em faixas mais abrangentes, desta forma diminuindo a complexidade do problema;

- **mineração de dados**: que também tem alguns passos a serem seguidos:

  - escolha das tarefas de mineração de dados, sendo o passo ao qual decide-se qual o objetivo do processo de mineração de dados. Os objetivos são diversificados, tais como:

    - classificação;
    - regressão;
    - clusterizacao.

  - escolha dos algoritmos de mineração de dados, neste passo são selecionados os métodos para serem usados na busca de padrões dos dados. Este passo inclui a decisão de quais modelos e parâmetros são mais apropriados para aquisição do tipo de conhecimento que é desejado.

  - Mineração de dados, propriamente dita, que é caracterizada pela busca de padrões de interesse em uma forma particularmente representativa ou em um conjunto dessas representações. Como exemplo, pode-se citar:

    - Regras de classificação;
    - Arvores de decisão;
    - Regressão;
    - Clusterização.

  - Interpretação de padrões da exploração, onde os dados de saída definidos nos passos anteriores são analisados e interpretados pelos especialistas do domínio. Caso seja necessário, pode-se repetir qualquer um dos passos anteriores para se obter a correta interpretação dos padrões;

  - Consolidação do conhecimento descoberto, onde se incorpora o conhecimento no desempenho do sistema, na documentação do conhecimento do relatório para as partes interessadas. Neste passo, faz-se também a verificação e a resolução de conflitos potenciais com o prévio conhecimento extraído.

O processo de KDD pode envolver interações significativas e retornar a qualquer dos passos, independente da fase a que ele pertença.

Um problema importante no processo de mineração de dados é que as informações nos objetos de dados são geralmente corrompidas ou esquecidas. Portanto, técnicas estatísticas devem ser aplicadas para estimas a confiança dos relacionamentos descobertos.

2.2.4.2 Tarefas em mineração de dados

Como dito anteriormente a etapa de Mineração de Dados é caracterizada pela busca de padrões de interesse em uma forma particularmente representativa ou em um conjunto dessas representações, entretanto, independentemente dos objetivos as tarefas primárias são as mesmas.

2.2.4.2.1 Classificação

Esta é uma função de aprendizado, em que o dado é mapeado em uma das diversas classes pré-definidas.

Como exemplo dos métodos de classificação usados no processo de KDD, pode-se citar a classificação de tendências do mercado financeiro e a identificação automática de objetos de interesse em grandes bases de dados de imagem [FAY96].

A FIGURA 2.6 mostra uma partição simples dos dados em duas regiões distintas de classes, então se, por exemplo, um banco quisesse usar a região de classificação para uma definição linear, não poderá ser considerada uma perfeita separação das classes (**X** os clientes sem empréstimo e **O** os clientes com empréstimo).

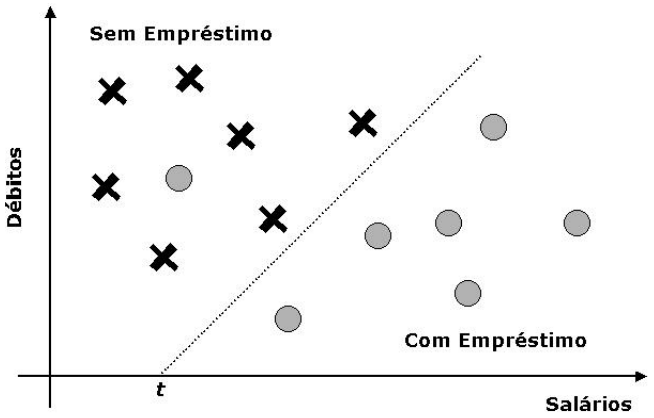


FIGURA 2.6 - Classificação linear de dados de empréstimo [AMA2001]

2.2.4.2.2 Regressão

Esta função de aprendizado mapeia os dados com pré-elaboração variável de valores reais. Como exemplos de aplicação de regressão, pode-se citar:

- predizer a soma de biomassa presente em uma floresta, fornecida por medidas de sensores remotos de microondas;

- estimar a probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames;

- predizer a demanda de consumo de um novo produto em função da despesa feita;

- predizer séries temporais, em que as variáveis de entrada podem ser versões da variável de pré-elaboração.

Partindo do mesmo exemplo anteriormente utilizado, a FIGURA 2.7 mostra o resultado da regressão linear simples, em que o "débito total" é visto como uma função linear de renda e a pré-elaboração é pobre, devido a uma correlação fraca entre as duas variáveis.

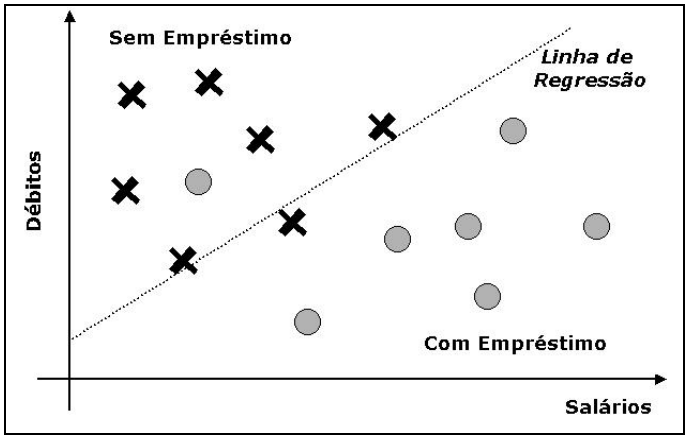


FIGURA 2.7 - Regressão para o conjunto de dados de empréstimo [AMA2001]

2.2.4.2.3 Clusterização

É uma tarefa comum na descrição onde exista um numero finito de categorias ou agrupamentos ou "*cluster*" para descrever os dados. As categorias podem ser mutuamente exclusivas e exaustivas ou consistir em uma representação como categorias hierárquicas ou sobrepostas.

Esta aplicação pode ser utilizada para a descoberta de grupos homogêneos para consumidores de mercado e identificação de subcategorias do espectro de medidas [FAY96].

A FIGURA 2.8, mostra a *clusterização* do conjunto de dado do exemplo utilizado em três grupos[2]. Outro ponto importante na mesma figura é que os grupos estão sobrepostos, permitindo, desta forma, que pontos pertençam a mais de um grupo.

O relacionamento do grupo é a tarefa de estimativa da probabilidade que consiste em técnicas de estimativas de dados da função de probabilidade multivariada de todas as variáveis e campos do banco de dados [SIL86].

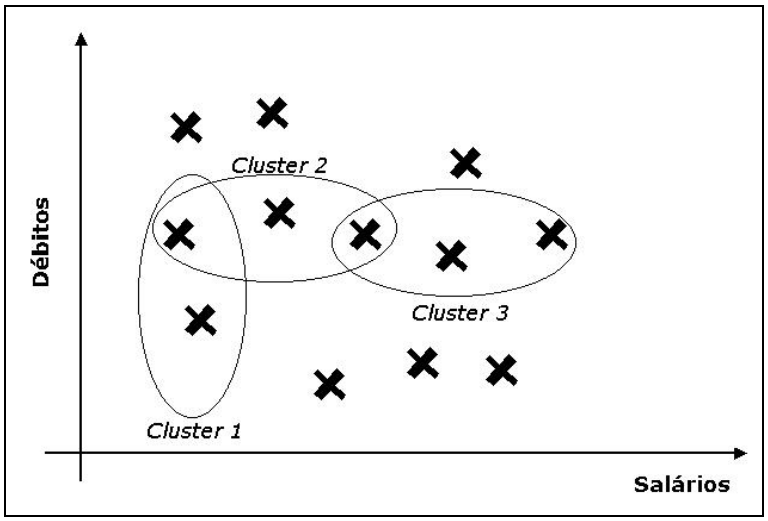


FIGURA 2.8 - Divisão do conjunto de dados de empréstimos em três grupos **[AMA2001]**

Outras tarefas que ainda podem ser citadas são:

- **Agregação**: tarefa que envolve métodos para encontrar uma descrição compacta de um subconjunto de dados;

- **Modelagem de Dependência**: encontra um modelo que descreve as dependências significativas entre variáveis. Os modelos de dependência existem em dois níveis:

    - o nível estrutural do modelo especifico: geralmente na forma gráfica, cujas variáveis são localmente dependentes entre si;
    - o nível quantitativo do modelo, o qual especifica a solidez usando a escala numérica.

[2] Nesse caso, todos os pontos passam a ser representado pó X, para indicar que os membros das classes não mais são conhecidos.

- **Detecção de mudanças e desvio**: focaliza a descoberta das modificações mais significativos no banco de dados, partindo de medidas anteriores ou valores normativos [FAY96].

O ***Data Mining*** faz uso da simulação, que representa a ação de construir um modelo, ou cenário, de uma situação onde se sabe a resposta e aplicá-lo a outra situação onde a resposta é desconhecida. Aplicando-se técnicas para executar a simulação é possível achar padrões relevantes e de interesse do usuário. As técnicas e os algoritmos mais utilizados pelo **Data Mining** para análise da informações são representadas por: [DAT2000A] [AMA2001] e [THE2000]:

- **métodos estatísticos**: usados em problemas de descoberta de conhecimento, em que o interesse está centrado em uma simples variável de saída *y* e uma coleção pré-editada. Todos os modelos assumem a viabilidade dos treinados e têm como objetivo encontrar um modelo para prognosticar o valor *y*, partindo de *x*;

- **regressões lineares**: os modelos clássicos de elaboração e classificação são regressões lineares e a análise linear de discriminante, respectivamente. Nesses modelos, o termo 'linear' é derivado do fato da superfície de regressão ou classificação de um plano;

- **árvores de decisão e regras**: têm forma de representação relativamente simples, fazendo com que o modelo inferido seja relativamente fácil de ser compreendido pelo usuário. Apesar do fácil entendimento, quando se restringe as informações a uma árvore particular ou às regras de representação pode-se reduzir significativamente a forma funcional do modelo;

- **aprendizagem relacional:** utiliza um padrão mais flexível de linguagem de lógica de primeira ordem. Esta técnica pode facilmente encontrar fórmulas como $x = y$. Muitas pesquisas em modelos de métodos de avaliação para aprendizado relacional constituem pura lógica;

- **algoritmos genéticos**: simulam o processo de seleção natural (assim como na biologia), onde os 'organismos' que melhor se adaptam no meio ambiente. Assim estes algoritmos usam a mesma propriedade para desenvolver seus modelos. Vários modelos são estudados, mas apenas aqueles que se mostram mais hábeis para o encontro da solução desejada são desenvolvidos. Daí o fato de existirem mais de um conjunto de considerações inteiramente diferentes que podem ser usadas em uma mesma solução do problema. Sendo difícil encontrar uma solução matematicamente perfeita para um problema, porém podem existir soluções próximas a perfeição.Usados bastante na área de otimização;

- **Redes Neurais:** também são usadas redes neurais, analisada no tópico a seguir.

2.2.5 Redes Neurais

É um tipo de tecnologia cada vez mais utilizada no processo de *Data Mining*. A grande vantagem desta técnica reside em sua habilidade de aprendizagem partindo das suas experiências, não ficando restritas a uma ordem seqüencial pré-fixada.

Consistem em algoritmos e procedimentos computacionais que "imitam" a capacidade de aprendizagem do cérebro humano. Esta técnica formada por **nós**[3] cujo processamento se assemelha a estrutura dos neurônios, daí a comparação. Não pode ser considerada uma técnica estatística por não apresentar a robustez de uma e também não oferece estimativas definidas e o comportamento de uma Rede Neural, com certa massa de dados, nem sempre se repetirá com outra. No entanto, é um recurso que permite o emprego estatístico em modelos não lineares.

Em uma Rede Neural, os **nós** são conectados como uma rede e funcionam paralelamente. A primeira fase de nós é composta por nós de entrada, que recebem os *inputs* das variáveis fornecidas pelo banco de dados, transformam-no de acordo com uma função (chamada de função de ativação), produzindo, assim, uma informação de saída que será enviada a próxima fase de nós. Esta, por sua vez, receberá diversas informações dos nós de entrada como seu *input*. Essa fase é formada por nós ocultos, que, em redes neurais mais complexas, podem formar diversas camadas. Por fim, tem-se os **nós** de saída, que processam as informações recebidas e produzem uma resposta, mas não enviam para outro **nó**, por ser o final da rede. Se esta é uma rede de classificação, o **nó** de saída já é a classe final. Para o caso de modelos de previsão, o nó de saída já representa um valor *preditivo*.

Assim como os neurônios cerebrais, as redes precisam ser treinadas. Este treinamento consiste em estimar valores dos parâmetros das funções de ativação $y=f(x,\Theta)$ de tal forma que os valores preditos sejam o mais próximo possível dos valores observados. Isso é obtido por meio de uma amostra aleatória do conjunto de dados, contendo entradas (x) e saídas (y). primeiramente escolhe-se valores iniciais de Θ. Posteriormente, uma nova estimativa é obtida quando as observações (xi, yi) da amostra aleatória de treinamento são consideradas. Esse procedimento vai sendo repetido até que se acerte a convergência, isto é, um ponto ótimo global.

Partindo das estimativas dos parâmetros, obtidas da amostra de treinamento, ajusta-se o modelo com os dados completos. Tal modelo pode ser avaliado com uma amostra de validação, possibilitando a verificação do ajuste de tal forma independente. Entretanto, deve-se tomar muito cuidado para que a rede não sofra uma sobrecarga de informação ou treinamento, ou seja, para que a rede não comece a explicar todas as relações entre os **nós** e acabe representando ate o erro aleatório no modelo. Já que sempre irá existir um erro, um ruído nos dados, sendo importante tentar controla-lo, e não esperar que ele desapareça. Este treinamento superestimado pode ser evitado

[3] Na literatura também podem ser encontrada a nomenclatura de *nodos* ou até mesmo como *neurônios*.

simplesmente determinando-se um limite máximo de erro tolerado. Quando a rede atingir este limite, o treinamento é dado como finalizado.

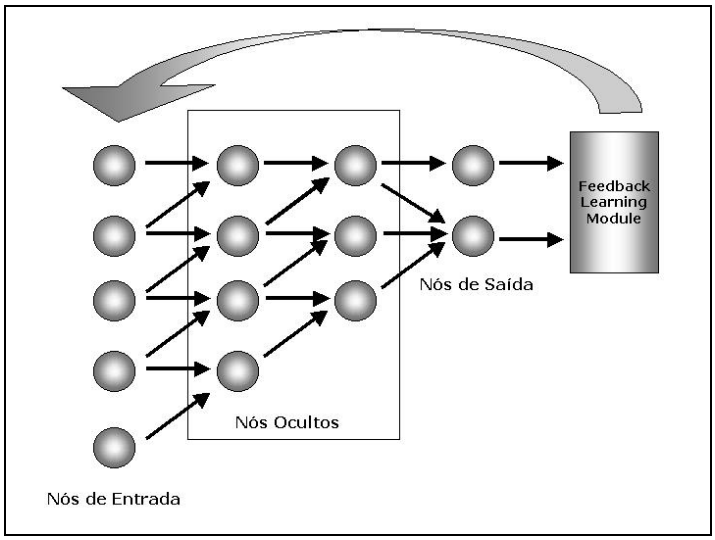


FIGURA 2.9 - Visão Esquemática de uma Rede Neural [BAR2001]

2.3 Raciocínio Baseado em Casos

Um sistema de CBR – Raciocínio Baseado em Casos (*Case Based Reasoning*), constitui-se de quatro processos considerados fundamentais:

- **Recuperação**: buscar todos os casos anteriores;

- **Reutilização**: dado um novo caso o sistema busca similaridade entre os casos recuperados;

- **Adaptação**: no caso do caso problema a ser encontrado não seja similar é necessário adaptá-lo antes de demonstrá-lo;

- **Retenção de casos**: esta solução adaptada torna-se um novo é considerado um novo caso é incorporado à base de casos já existente.

Desta forma, um CBR pode parecer ser simples, entretanto tudo depende do domínio da aplicação a qual o sistema será desenvolvido. Diversas pesquisas [KOL93] [WAT95] mostram e definem varias técnicas para desenvolver um CBR.

A idéia de que um CBR está no mesmo processo de solução de problemas humanos [VAN89], onde o uso de experiências passada tem sido apoiada pela ciência cognitiva e psicológica. Nesta abordagem, o aprendizado é um produto do ato de vivenciar novos fatos tentando compreendê-los e integrá-los ao conhecimento já existente.

Cada experiência ou caso corresponde ao reconhecimento de quais aspectos do problema são relevantes e como influenciam na solução de um problema.

Um CBR é uma técnica de AI que se propõe exatamente a representar o conhecimento e a inferência, propondo a solução de novos problemas adaptando soluções passadas na resolução de novos problemas e teve o seu inicio a partir exatamente dos estudos cognitivos da memória feitos, inicialmente, propondo uma variedade de estruturas de memória dinâmicas para organizar e viabilizar o processo de aprendizado.

[KOL93] apresenta sistemas especialistas que utilizavam medidas de similaridade e técnicas de adaptação para reaplicar antigas soluções em novos problemas do mesmo domínio.

Quando se compara a outras técnicas de representação e inferência em sistemas especialistas baseados em modelos, um sistema de CBR possui as seguintes vantagens:

- permite ao "argumentador" propor soluções para os problemas rapidamente, evitando o tempo necessário para descobrir a origem das questões a partir do nada;

- proposição de soluções em domínios que não são completamente entendidos;

- aplicando técnicas de banco de dados para representar os casos e indexá-los podem ser gerenciados grandes volumes de informação [LOR98];

- sistemas de CBR podem aprender automaticamente através da seleção e aquisição de novos casos, tornando a manutenção da base de conhecimento muito fácil. Sendo esta característica algo muito importante em um CBR, já que o aprendizado ocorre naturalmente pelo produto da solução de um novo problema. Quando um problema é resolvido com sucesso, a experiência é mantida para resolver problemas semelhantes no futuro.

É claro que também há desvantagens na utilização de CBR:

- um CBR pode ser incitado a utilização de casos anteriores de forma irrestrita, sem levar em conta uma validação destes com uma nova situação;

- um CBR induz à solução de um novo problema;

- muitas vezes as pessoas, especialmente novatos, não se lembram de um conjunto apropriado de casos quando eles estão argumentando;

Apesar destas desvantagens, o sucesso deste tipo de sistema está repercutindo também na administração de empresas, onde na visão dos administradores, a utilização de CBR facilita o processo de aquisição de conhecimento em diversos pontos:

- os especialistas sentem-se mais a vontade em explicar ao engenheiro de conhecimento as situações já vividas em suas rotinas a descrever regras;

- os especialistas que fornecem os casos, de certa maneira, participem da construção do sistema;

- os administradores apreciam a capacidade que um CBR tem em armazenar conhecimento e poder utilizá-lo a qualquer momento, podendo prevenir erros e não perder tempo solucionando problemas que já foram resolvidos, não sendo necessário fazer qualquer tipo de re-análise.

Este tipo de sistema tem sido muito utilizado para o desenvolvimento de um grande número de aplicações nos mais diversos domínios, tais como:

- **diagnósticos**: [SIM92] sistemas de diagnóstico baseado em casos tentam recuperar casos antigos que tenham sintomas parecidos com o novo problema e sugere o diagnóstico para este problema baseado nesta recuperação de casos;

- **help desk** [DEA93]: sistemas utilizados na área de serviços ao consumidor, tratando de problemas com os produtos;

- **suporte à decisão** [TSA97]: na tomada de decisões quando as pessoas se deparam com problema complexo, elas geralmente procuram por problemas análogos a procura de soluções possíveis ao seus problemas. Assim, sistemas de CBR têm sido desenvolvidos para auxiliar no processo de recuperação de problemas, encontrando problemas parecidos ao problema do usuário;

- **projetos** [COS93]: no domínio de projetos, sistemas de CBR têm sido desenvolvidos para projetos arquitetônicos e industriais. Estes sistemas auxiliam o usuário em apenas uma fase do projeto (somente na recuperação de casos). Para o suporte a todo projeto seria necessário combinar CBR com outras técnicas de raciocínio.

2.3.1 O Funcionamento de um CBR

Como já dito anteriormente, um sistema de CBR resolve problemas através de uma seqüência de etapas, conforme demonstra na FIGURA 2.10.

- **Recuperação**: recupera, na base de casos, o mais parecido com o novo problema. Identifica e pesquisa índices, calcula a similaridade entre o caso recuperado e o novo problema:

  - Indexação: um bom índice é um fator muito importante para que o CBR possa recuperar casos relevantes. O fundamental para se

atingir este objetivo é ter conhecimento de como o caso deve ser indexado de uma forma onde o processo de recuperação seja preciso e eficiente;

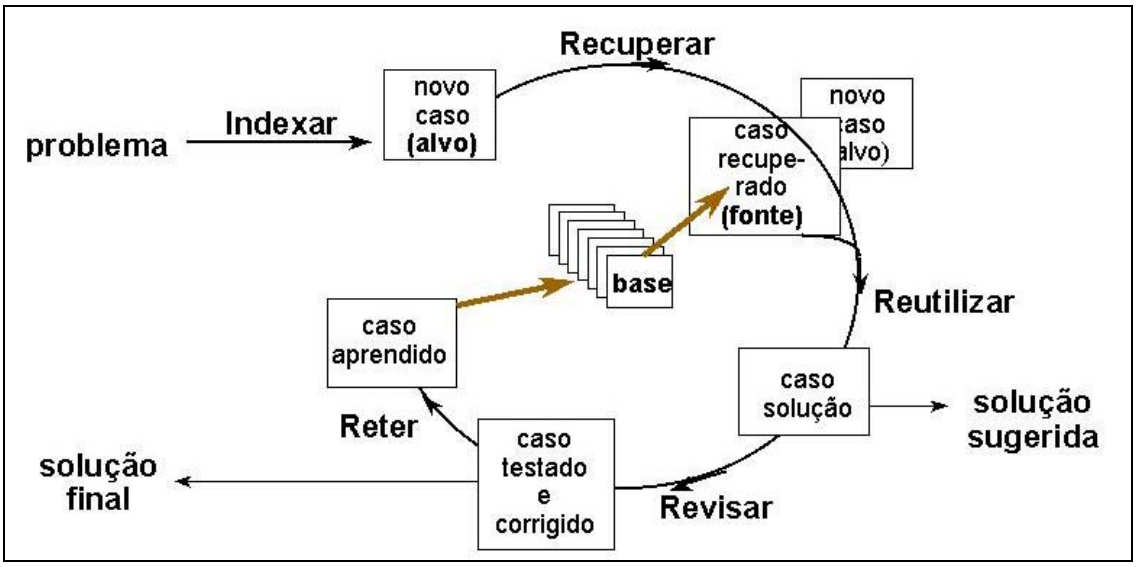


FIGURA 2.10 - O Ciclo do CBR adaptado [KOL93]

- Identificação das características: pode ser considerado somente como sendo as informações / características de entrada, sendo que as características não conhecidas podem ser desprezadas ou solicitadas para serem explicitadas pelo usuário do sistema;

- Comparação: é uma sub-tarefa que pode ser resumida da busca de um conjunto de possíveis casos na base de dados;

- Seleção: após achar o conjunto de possíveis casos na base, é feita uma busca usando as características problema (índices) de uma maneira direta ou indireta. Podendo ser feita de três maneiras:

    a) seguindo os índices diretos apontados para as características do problema;
    b) procurando uma estrutura de índices; e
    c) procurando um modelo de domínio de conhecimento geral.

Os tipos de aplicação mais usados na tarefa de recuperação têm sido [LOR98]:

- *Método de Recuperação Indutivo*: muito usado quando o objetivo da recuperação pode ser bem definido. Os casos podem ser indexados pelas principais características e a árvore decisória maximiza a tarefa de recuperação;

- *Baseada em conhecimento*: aplica o conhecimento para localizar casos relevantes. Muito semelhante a sistemas especialistas esta solução híbrida tem sido usada para recuperações mais precisas;

- *Vizinhos mais próximos* (KNN): combina casos recuperados com base no somatório de pesos das características do novo problema.

  Os casos com o total de comparações com alguma similaridade métrica são retornados do processo de comparação.

- **Reutilização**: também considerada como adaptação, visto que após os casos ser recuperado na base, o sistema CBR adapta a solução do caso para solucionar o problema novo. A adaptação de um caso recuperado no contexto pode ser vistas sob o aspecto das diferenças entr e o caso passado e o atual e também visto sob qual parte de um caso recuperado pode ser transferido a um caso novo. Para esta tarefa de **reutilização**, é necessário:

  - *Copiar*: as similaridades consideradas relevantes e a solução dos casos recuperados é transferida para o novo caso com a sua solução;

  - *Adaptar*: segundo [ALL94] há duas maneiras de fazer reuso dos casos do passado: a) reutilizando a solução dos casos passados (esta solução é chamada **estrutural**); e b) reutiliza o método passado que construiu a solução (solução **derivacional**), ou seja, "o como" o problema foi resolvido.

- **Revisão:** é necessário revisar (adaptar) a solução do caso recuperado gerada pelo processo de reutilização quando a solução não pode ser aplicada diretamente no novo problema. Isso mostra uma oportunidade de aprender com a falha;

- **Retenção:** é o processo de incorporar tudo que for útil no novo problema na biblioteca de casos. Isto envolve decidir que informação armazenar e de que forma armazenar, como indexar o caso de futuras recuperações e integrar o novo caso à biblioteca de casos. Para isso os passos devem ser seguidos:

  - *Extração*: a base de casos é atualizada não importando de que maneira o problema foi resolvido. Caso o problema tenha sido solucionado através de um caso anterior, um novo caso pode ser construído. Caso o problema tenha sido resolvido através de outros métodos, com a intervenção de um especialista, por

    exemplo, o novo caso tem que ser construído para que o sistema aprenda;

- *Índice*: verificar qual o índice utilizar para futuras recuperações e como estruturar a procura destes índices;

- *Integrar*: esta tarefa final da atualização da base de conhecimento com o novo caso.

2.3.2 O Desenvolvimento de um CBR

Para construir um CBR deve-se definir a forma de representação dos casos na base de casos, a forma de indexação e comparação dos casos aos novos problemas, a forma de julgamento do caso e caso haja a geração de uma solução como a mesma deve ser incluído na base juntamente com a descrição de um novo caso.

Assim, o principal objetivo de um CBR é atingido através da:

- experiência do sistema, que pode ser conseguida com o aumento na base de casos relevantes;

- habilidade do sistema de compreender as novas situações, tendo como base as experiências antigas;

- qualidade de sua avaliação e correção;

- capacidade de adaptação;

- habilidade de integrar apropriadamente novas experiências na "memória".

Assim, para desenvolver um CBR as seguintes etapas devem ser cumpridas:

- selecionar uma base de informações, que contenha implicitamente o conhecimento da instituição na solução do problema. A confiabilidade do CBR está intimamente ligada a esta base, quanto mais completa e corretamente definida estiver esta base;

- definir quais são os atributos relevantes e podem ser utilizados para a solução do problema. Os atributos que definem a forma de solução do problema pelo especialista, ou evidenciados pela aplicação de métodos de descoberta do conhecimento;

- definição de índices para os casos, para que seja possível quando necessários. A indexação é o maior problema no desenvolvimento de um sistema CBR, onde se deve decidir o que armazenar em um novo caso, encontrando uma estrutura apropriada para a descrição dos

  conteúdos dos casos e decidir com uma memória de caso de ser organizada e indexada para a recuperação e reutilização dos casos;

- definição de métodos de recuperação dos casos para a verificação da similaridade entre os casos contidos na base e os novos problemas que irá resolver. É necessário definir quando um problema é suficientemente semelhante a um caso armazenado na base a ponto de se saber se a solução pode ser também aplicada a este novo;

-

- definir como a solução associada a um caso na base de casos deve ser adaptada para contemplar um novo problema. A adaptação quando necessária é baseada em conhecimento, tudo isso partindo de “feeling” e experiências do analista;

- definir qual será o processo de aprendizado. Se a solução adaptada para o problema teve ou não êxito, o sistema deve avaliar o armazenamento ou não da solução.

O estudo de caso será detalhado posteriormente para a descrição das técnicas utilizadas no desenvolvimento de um CBR.

3 Estudo de Caso - Análise de Crédito e Cobrança a partir da Similaridade dos Casos

Este trabalho foi totalmente voltado à aplicação de um CBR na análise de risco dos clientes e também no relacionamento com o cliente, onde um sistema de CBR além de ter uma aplicação bastante útil, apresenta um diferencial muito grande em relação às demais técnicas apresentadas.

3.1 Estudo da aplicação de Análise de Risco e Inadimplência

Em uma empresa de telecomunicações, de acordo com as normas da agência reguladora – ANATEL (Agência Nacional das Telecomunicações) – é vetado a qualquer empresa negar o fornecimento do serviço de telecomunicações aos clientes que por ventura possam vir a solicitar pelo serviço.

Desta forma, o risco de inadimplência destas empresas é muito alto e o investimento feito em equipamentos e infra-estrutura pode tornar o negócio inviável. Nestas circunstâncias, é de extrema importância tornar o processo de cobrança uma atividade mais "inteligente", utilizando técnicas para reduzir o risco e com o menor esforço ter o maior retorno, assim como é feito, por exemplo, com as técnicas de *Data Mining* para a identificação de potencial de compras dos clientes (emissão de malas diretas, *up-selling* e *cross-selling*).

Mesmo que a empresa pudesse utilizar a análise de crédito para negar ou não o serviço, esta não pode ser considerada a única fonte de única informação devido a distorções. Por exemplo, digamos que um médico, cuja renda mensal é considerável tenha acabado de contrair uma dívida alta junto a um banco ou instituição financeira, o *bureau* de crédito pode indicar este como um caso de alto risco. Fato este que para a situação em questão, pode não ser determinante, pois o valor de qualquer tipo de serviço de telecomunicações não pode ser considerado como um fator que onere o orçamento familiar deste profissional.

Também se tem que levar em conta o fator regulador da ANATEL, órgão este que determina as sanções que podem ser aplicadas, em quais tempos, fazendo com que a empresa tenha que ser o mais eficiente a fim de:

- nos casos de alto risco ter mais tempo para recuperar o montante devido;

- identificar grandes distorções como sendo fraudes e tomar ações imediatas;

- ter previsibilidade no fluxo de caixa da empresa.

Torna-se inexorável que diversos fatores, mesmo que algumas vezes sejam conhecidos, sejam analisados para identificar onde estão os maiores riscos, quais são os clientes, regiões, cidades e fatores socioeconômicos.

3.1.1 Análise do Processo de Negócio

Pode-se dividir o "momento" de se analisar um cliente em dois: a entrada do cliente na empresa (crédito) e durante a sua utilização mensal (cobrança). Mas todo o processo, mostrado, seqüencialmente, na FIGURA 3.1, deve ser integrado e o sucesso do segundo depende muito da eficácia do primeiro e de todos os outros processos intermediários.

FIGURA 3.1 - Macro Processo de Negócio

3.1.1.1 Atendimento e Venda

Todo o processo se inicia quanto o cliente entra em contato com a empresa, ou vice-versa e inicia-se o atendimento, onde todas as dúvidas do cliente são respondidas, informações sobre produtos e serviços são dadas.

Caso haja um interesse deste cliente, inicia-se então o processo de venda, este é um processo, que apesar de aparentemente simples, é extremamente delicado e importante, visto que é exatamente neste processo onde as informações dos clientes serão coletadas.

Informações muito importantes deste cliente são solicitadas e devem ser analisadas de imediato para evitar uma assinatura fraudada (*fraud subscription*). É exatamente neste ponto onde as informações dos clientes devem ser validadas também para **Crédito**. Desta forma, as informações sobre a veracidade das informações são enviadas, *on-line*, para um *bureau* de crédito com o intuito de fazer tal verificação.

São usadas informações combinadas com o CPF para melhorar a qualidade e a segurança da informação (como por exemplo, data de nascimento, nome dos pais, etc). Esta validação é muito importante no processo de **Fraude e Garantia de Receita [4].**

[4] Processo, dentro de uma empresa de telecomunicações, que tem como função a análise de todo o processo de negócio e análise dos sistemas, com o intuito de determinar pontos falhos onde possam haver fraudes e que podem gerar uma evasão de receita, podendo ser por dolo ou não de empregados da própria empresa ou parte dos clientes desta.

Outro fator importante, é que apesar desta informação realmente poder ter um grau de eficácia não pode ser determinante. Mesmo porque as maiores distorções acontecem em nas extremidades. Por exemplo, imaginando que a classe de crédito seja feita por uma variação de letras de **A** a **F**, onde A seria o melhor cliente, aquele que teoricamente não teria dívidas e teria uma grande capacidade de endividamento e F seria aquele onde o cliente teria dívidas comprovadamente não pagas ou então estaria endividado o bastante ao ponto de não ser recomendável dar o crédito ao mesmo.

Entretanto, ocorrem casos onde um cliente classificado como **A**, ou seja, não teria nenhum impedimento jurídico ou endividamento, é um cliente de cujo poder aquisitivo teria um impacto caso uma conta telefônica mensal tivesse um valor de R$ 50,00 (cinqüenta reais), por exemplo. Em contrapartida, e um caso mais comum, que já foi citado acima, ocorre quando o cliente tem um poder aquisitivo alto, mas é dado como **E** ou **F**, ou seja, tem um grau de endividamento relativamente alto e teoricamente este cliente seria um cliente com um alto risco para a empresa. Mas este cliente é um profissional liberal, tem uma renda familiar alta e a representatividade de uma conta média de R$ 100,00 (cem reais), não representaria, percentualmente, um impacto no seu orçamento familiar.

3.1.1.2 Instalação, Uso e Faturamento.

Efetuada a venda, inicia-se o processo de instalação, que apesar de onde a sua qualidade impacta diretamente no processo de cobrança, pois neste processo o *workflow* deixa de estar dentro dos sistemas e as tarefas devem ser efetuadas fisicamente, dependendo assim de tarefas e controles manuais para as mesmas.

Isso pode gerar diferenças entre o que está nos sistemas e o que fisicamente está instalado. Tendo o serviço disponibilizado o cliente inicia o uso e a análise deste uso é determinante para que as fraudes sejam identificadas antes que a evasão de receita seja efetuada, novamente entra o processo de **Fraude e Garantia de Receita.**

Dias antes da data de vencimento, então, inicia-se o processo de faturamento. Por determinação do órgão regulador, todas as empresas prestadoras de serviço de telecomunicações devem disponibilizar aos seus assinantes pelo menos seis datas, a escolha do assinante, para o vencimento de suas faturas, cada uma destas diferentes datas são conhecidas como **Ciclos de Faturamento**[5].

Após todo o processo de faturamento, o cliente recebe a sua fatura em casa o pagamento, como outra determinação ANATEL é que esta fatura deve ser entregue cinco dias antes da data de vencimento.

3.1.1.3 Cobrança

Aguarda-se então a data de vencimento e a confirmação de pagamento daquela fatura por parte do cliente. Caso isso não venha ocorrer, inicia-se o processo de

[5] **Data de Vencimento** da fatura, escolhida pelo cliente.

cobrança, propriamente dito. São estipuladas regras, onde as ações de cobrança são distribuídas dentro de uma linha de tempo, como mostra a FIGURA 3.2.

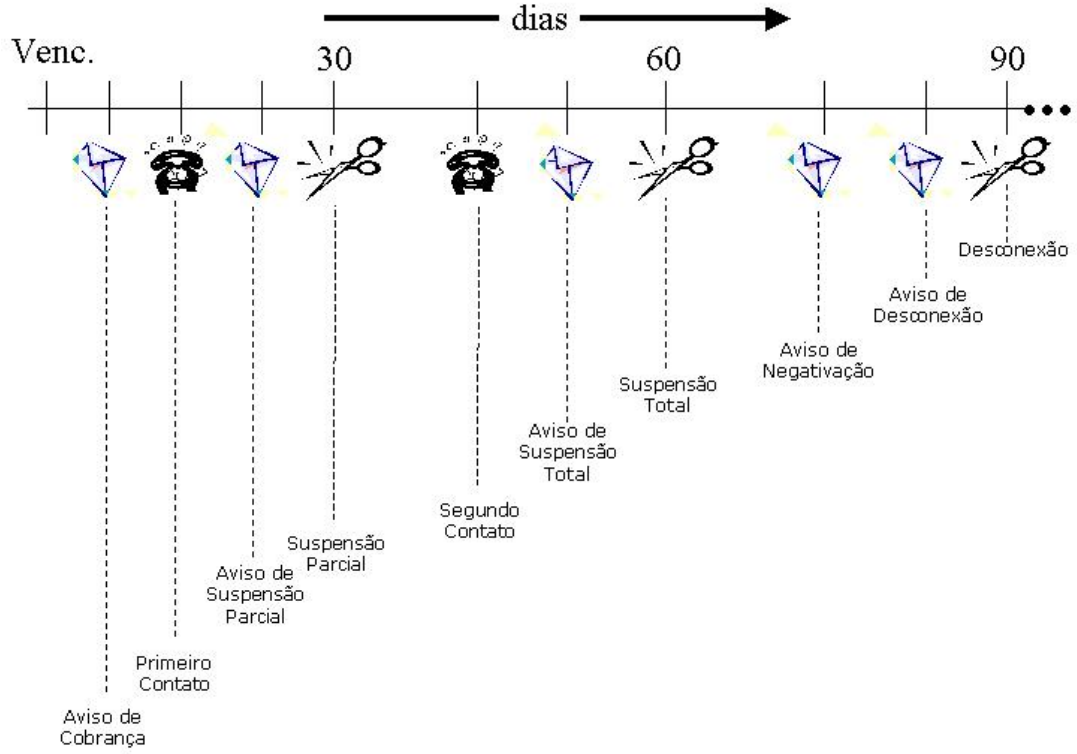


FIGURA 3.2 - Visão do *timeline* de ações - Cenário de Cobrança

Nota-se no esquema, as ações de cobrança iniciam-se após a data de vencimento e estendem-se durante cerca de 90 (noventa) dias com três pontos bem marcantes dentro deste "*timeline*" que seriam onde seriam efetivadas as sanções, suspensão parcial, suspensão total e desconexão que ocorrem respectivamente em 30 (trinta), 60 (sessenta) e 90 (dias). Deve-se notar também que estes são os prazos mínimos estipulados pelo órgão regulador, mas nada impede da empresa majorar este prazo ou até mesmo, de acordo com a sua conveniência, deixar de executar algumas sanções.

Após este período de 90 (noventa) dias fica a critério da empresa desconectar o cliente fisicamente de sua infra-estrutura e tomar as ações necessárias contra o devedor/cliente para conseguir reaver o valor da dívida, como enviá-los para "negativação"[6]. Estas ações de negativação nos órgãos de proteção de crédito, agências de cobrança, têm abordagens bem mais agressivas e em último caso a companhia pode tentar reaver este montante através de meios jurídicos.

Entretanto, durante o prazo inicial de 90 (noventa) dias, podem ser notadas ações como campanhas ativas de contato telefônico, onde são feitos contatos diretamente com o cliente, não só com o objetivo de cobrar, mas também de verificar se

[6] Envio dos dados do cliente aos órgãos de proteção ao crédito, que tem o intuito de dificultar movimentações financeiras.

houve algum problema que levou a inadimplência e oferecer condições para que esta situação seja regularizada.

Também há as ações de carta, que podem ser divididas em dois tipos:

- **cartas de cobrança**: onde é exposta a situação do cliente, valor total da dívida e outras informações relevantes. O que muda dentro destas cartas é a abordagem, indo desde uma carta apenas explicativa até uma carta mais enfática;

- **cartas de aviso**: são ações previstas pelas normas, onde deve constar a sanção e a data que o mesmo sofrerá.

Apesar de neste *timeline* mostrar somente estes tipos de ação, não há nenhuma regra que estipule as mesmas, ficando a cargo da empresa qualquer outro tipo de contato, que por ventura, queira fazer com os clientes. Como, por exemplo, SMS (*short message service*), ou seja, enviar lembretes ao celular do cliente, caso este possua e tenha informado, ou o *hot line*[7] no próprio telefone do cliente.

Estas ações não necessariamente devem ser "ações padrões" para todos os clientes que entrarem em cobrança. Desde que respeite as regras de sanção, a empresa pode o quão rígida será a ação de cobrança, quantidade de passos, abordagem, etc., dependendo de variáveis que ela possa estabelecer. Como, por exemplo, não colocar grandes corporações ou repartições públicas dentro destes cenários e sim solicitar que vendedores ou gerentes de conta entre em contato, com as mesmas, como o objetivo de aprimorar o relacionamento com as mesmas[8].

Assim, diversas variáveis devem ser analisadas, para que o processo de cobrança seja mais eficiente e se possível, deveria iniciar antes mesmo do cliente se tornar inadimplentes, fazendo uma campanha antes da primeira conta, explicando detalhadamente a conta telefônica, propondo uma negociação antecipada ou até mesmo, para casos de suspeita de fraude, o corte do serviço deste cliente, mas para isso seria necessária a determinação do grau de risco de cada um deles.

3.2 Utilizando um CBR

Uma aplicação de cobrança, e suas variáveis podem ser tratadas como uma regra bem conhecida e que podem ser aplicadas a todos os mercados. Entretanto, quanto mais focado ao setor, mais eficiente será o processo.

---

[7] Serviço que disponibiliza ao cliente mensagem quanto o mesmo tenta utilizar o telefone pela primeira vez ou então o transfere para um atendimento automático, de onde o cliente pode solicitar falar com o serviço de atendimento da empresa.

[8] Há somente uma exceção no caso da sanção dos clientes, a agência permite que um cliente tenha seu serviço suspenso ou até mesmo cortado imediatamente após o vencimento, no caso do cliente ter serviços não ligados à prestação de serviços de telecomunicações (uso e mensalidade) sendo cobrados na fatura em questão (e.g., a taxa de instalação do serviço ou qualquer outro tipo de cobrança de serviço de terceiros como jornais, revistas, anúncios, etc).

O setor de telecomunicações, como visto acima, tem suas características de cobrança, principalmente o fato de não poder minimizar o risco, suspendendo os serviços assim que nota a inadimplência e se nega a vender a clientes que demonstrarem qualquer anotação que os desabone em *bureaus* de credito.

Uma aplicação de Raciocínio Baseado em Casos torna-se interessante devido ao fato de que estas empresas estão, neste momento, concorrendo por clientes e tentando maximizar os seus lucros, gerando uma grande quantidade de ações para atingir os clientes certos e mantê-los, assim fazendo com que a aplicação de conceitos como CRM – *Customer Relantionship Management* (Gerenciamento de Relacionamento com o Cliente) sejam bastante utilizadas. Com isso, as informações de cada cliente estão ficando mais e mais detalhadas o que propicia a implantação de um CBR, sendo que as principais vantagens seriam:

- apesar do grande volume de dados e da complexidade de uma rotineira avaliação de risco, um sistema como o CBR, é muito vantajoso na aquisição do conhecimento, sem que para isso seja necessário a modelagem de diversas regras comparativas, o que seria comum em outros sistemas de AI;

- o conhecimento dos próprios casos já constantes na base podem ser reutilizados;

- com a atualização sistemática da base de dados, a base de conhecimento aprende automaticamente, aprendendo rapidamente as nuances de modificação, a alteração do mercado, por exemplo.

[LOR98] demonstra uma comparação entre sistemas de CBR e outros sistemas, conforme mostra a TABELA 3.1, além disso, faz-se importante salientar que a construção de um sistema de CBR, no caso da aplicação em questão torna-se interessante também por motivos como:

TABELA 3.1 - Diferenças entre as técnicas [LOR98]

	CBR	Redes Neurais	Sistema de Regras
Aplicabilidade	Melhor aplicado em estruturas de simbólicos.	Melhor aplicado quando os dados não podem ser representados simbolicamente.	Melhor aplicado em sistemas fáceis de extrair de regras do especialista.
Justificativa da Decisão	Retorna o motivo da decisão	Não retorna uma explicação ou justificativa para a solução proposta.	Retorna o motivo da decisão percorrendo as regras que foram disparadas para chegar a solução.
Aprendizado	Permite o aprendizado através da aquisição de novos casos.	Ocorre através de cálculos	O aprendizado ocorre somente através da inclusão de novas regras ou da modificação das regras.

- rapidez de desenvolvimento;
- fácil utilização e manutenção;
- rápida aquisição do conhecimento e manutenção deste, sem a necessidade de adequações.

De acordo com o que já foi visto anteriormente, a construção de um CRB é feita através dos seguintes passos:

- aquisição da base de conhecimento;
- definição dos atributos da informação que são relevantes e podem ser utilizados para a solução do problema;
- definição de índices para os casos, para que seja possível a sua recuperação quando necessário;
- definição dos métodos de recuperação de casos para a verificação da similaridade entre os casos contidos na base e os novos problemas que o sistemas irá resolver;
- definição da forma como a solução associada ao caso recuperado deve ser adaptada para ser reutilizada pelo novo problema;
- definição do processo de aprendizado.

Cada um dos passos é explicado nas seções seguintes.

3.2.1 Aquisição do Conhecimento

Como um método de construir sistemas “inteligentes”, um CBR tem um apelo por parecer relativamente simples e natural. Nesta etapa, onde o objetivo é exatamente extrair o conhecimento que o especialista utiliza na resolução dos problemas.

Este especialista seria um indivíduo que tem é muito treinado na solução de um determinado problema e pode identificar com rapidez os todos os elementos fundamentais de sua especialidade, conhecendo todos os controles e manipulações e referentes a mesma. Entretanto, deve-se ressaltar que é muito difícil encontrar especialistas para relatar todo conhecimento que eles usam para a solução dos problemas.

A aquisição do conhecimento é desenvolvimento em seis etapas [LOR98], que serão detalhados a seguir.

3.2.1.1 Identificação e estudo do problema

Não houve um especialista, em particular, envolvido no trabalho, mas uma série de pessoas especializadas em Crédito e também em Cobrança (que são processos totalmente distintos).

Além de uma série de reuniões, com os mais diversos enfoques foram feitos com o intuito de identificar a causa e resolver o problema foram feitas. Onde se teve a oportunidade de discutir todos os pontos do processo de negócio.

3.2.1.2 Análise do Conhecimento

Para que se possa, após a análise do conhecimento feita com o especialista deve ser feita a extração das informações para que efetivamente possa ser gerada a base de casos. Entretanto, nem sempre esta geração é simples, dependendo da maneira que os dados se apresentam:

- não disponíveis em uma fonte externa: ou seja, todos os casos se apresentam apenas na memória do especialista;
- semi-disponíveis em uma fonte externa: estes dados podem ter um controle inicial, mas se apresentarem incompletos;
- disponíveis e contém erros: os dados estão completos, mas muitos casos se apresentam conflitantes ao que deveriam se apresentar;
- totalmente disponíveis;

Neste caso, a base de informações estava ***totalmente disponível*** uma vez que de acordo com a arquitetura de sistemas da empresa não ter um só sistema que cruze todo o processo, desde o *atendimento/venda* até o *faturamento* um *DW* é utilizado para armazenar os dados analíticos dos clientes e de seus *ciclos de faturamento* e que tem uma rotina de sumarização e armazenamento para estes dados.

O interessante é que o dado apesar de estarem em um *DW* o seu nível de *granularidade* no que tange ao *DM* de marketing está em um nível de detalhe analítico, o que permitiu a extração de todos os dados dos clientes e para a parte do primeiro faturamento foi utilizado um *DM* de faturamento que também contém toda a informação das faturas. E este tipo de *DW* com uma granularidade alta, também pode ser chamado de ODS.

3.2.2 Representação de Conhecimento

O modelo de um CBR não necessita de um modelo completo do domínio, uma vez que se destinam a modelar os próprios episódios em que o problema foi conhecido

e solucionado. Permitindo, assim, o desenvolvimento de sistemas em domínios onde o conhecimento é pouco dominado ou muito complexo.

Como, já dito anteriormente, como neste estudo de caso, não há um conhecimento bem estruturado e em profunda modificação pelos fatores externos o CBR vêm de encontro com as necessidades apresentadas.

3.2.2.1 Representando Casos

Como a base de informações, *DW* existente, possui todos os dados se apresentavam bem estruturados e razoavelmente completos, além de grande parte das variáveis apontadas pelos especialistas, como relevantes eram informações obrigatórias dos sistemas, utilizou-se esta informação diretamente como casos. Também há outro ponto importante que são das informações descritivas, as informações da solução (que no caso não necessariamente é uma solução, mas uma situação) também constavam já estruturadas no sistema.

Não houve de um indicativo, por parte do especialista, de quais os campos que levam, ou não, á **inadimplência**, mas sim uma combinação de variáveis, como o processo sintomático de uma doença, por exemplo. As características utilizadas para a solução do problema, vão depender do momento em que está sendo feita a análise, conforme o processo de negócio demonstrado anteriormente.

Um caso exemplo é mostrado na TABELA 3.2.

TABELA 3.2 - Exemplo de um caso de Análise para Crédito

Atributo	Valor (Exemplo)
Telefone Celular	Sim
Endereço e-mail	Não
Ciclo de Faturamento	10
Plano de Preço	Plano Básico
Faixa Etária	25 a 30 anos
Quantidade de Pessoas na Casa	3 a 5 pessoas
Tecnologia	Cabo
Mercado	Residencial
Sexo	Masculino
Forma de Pagamento	Pagamento em Fatura
Classe Econômica	C
Profissão	Veterinário
Bairro	Centro

Já para o caso de determinação de "risco"[9] da fatura, algumas informações a mais serão adicionadas para fins de análise. Devendo-se notar que são, na verdade, dois CBRs totalmente independentes, para a determinação de pesos e valores. Obviamente os valores serão similares, mas esta separação fez-se necessário para que não houvesse

[9] Na verdade o risco não pode ser determinado, mas o fator apresentado, na verdade, é o grau de similaridade da mesma em relação a uma outra também inadimplente.

a distorção na análise, principalmente no primeiro caso que estaria incompleto ao comparar com o segundo, mostrado na TABELA 3.3.

3.2.2.2 Geração da Base de Dados

A geração de casos neste trabalho foi feita a partir dos aspectos relevantes de cada cliente e a situação de cobrança apresentada por este no momento da criação da base. Deve-se notar que com isso, "congelou-se" a situação destes clientes.

Assim, alguns problemas devem ser sanados:

- os casos deverão ser constantemente atualizados, como uma rotina de ETL de um *DW*, para que seja feita a atualização da inteligência do sistema;

- aproveitando-se esta rotina, os casos que se apresentarem idênticos, não deverão ser carregados, com o intuito de melhorar a performance de ambos CBR (Clientes e Faturas).

Como o estudo de caso foi feito com uma base temporal para análise, tal função não foi contemplada no protótipo.

TABELA 3.3 - Exemplo de um caso de Análise para Cobrança

Característica	Valor (Exemplo)
Telefone Celular	Sim
Endereço e-mail	Não
Ciclo de Faturamento	10
Plano de Preço	Plano Básico
Faixa Etária	25 a 30 anos
Quantidade de Pessoas na Casa	3 a 5 pessoas
Tecnologia	Cabo
Mercado	Residencial
Sexo	Masculino
Forma de Pagamento	Pagamento em Fatura
Classe Econômica	C
Profissão	Veterinário
Bairro	Centro
Dias para Pagamento	31 a 45 dias
Quantidade de Reclamações	2
Valor da Fatura	R$ 150 e R$ 500,00

3.2.2.3 Indexação dos Casos

Os índices são criados para que se possa ter um acesso mais rápido e direto os casos mais relevantes, minimizando assim o gasto computacional com a comparação e conseqüentemente melhorando o tempo de resposta do sistema, conforme esquema, mostrado na FIGURA 3.3.

Estes índices podem ser selecionados automaticamente ou não, mas os processos automáticos, por sua vez, quantificam as diferenças entre os casos e o relacionamento entre as características do problema [LOR98]. Como o especialista, neste caso, não conseguiu determinar com exatidão os índices para a aplicação, foi necessária a utilização de análise matemática para tal função, atributo a atributo feita conforme detalhado a seguir.

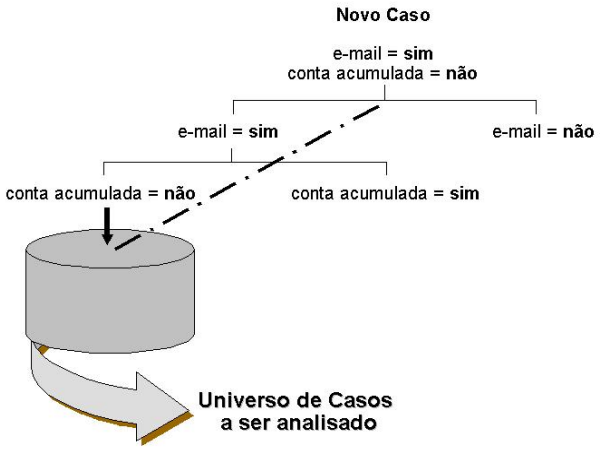


FIGURA 3.3 - Funcionamento de um índice (memória plana)

Para a criação dos índices foram adotados métodos automáticos, que tem em sua lógica a seguinte lógica:

- todos casos, com seus respectivos valores foram inseridos em uma tabela cruzada, para que fosse possível a análise de como os casos estavam distribuídos dentro dos valores possíveis;

- dentro desta tabela verificou-se para quantas atributos estes valores eram relevantes. Ou seja, a concentração dos casos dentro dos valores possíveis, conforme TABEL 3.4. ;

- desta forma, todos os casos foram contados e calculados para todos os casos e chegou-se a uma determinação de quais os atributos que tinham maior influência na determinação de cada situação. Entretanto, como o índice tem a função de agrupar os casos foi necessário também verificar o comportamento deste atributo em relação aos valores possíveis, verificando o total geral de casos em relação aos valores possíveis dentro do atributo, elegendo os atributos onde os casos se dividiam, nesta análise geral, de forma mais homogênea;

- após esta análise foi estabelecido que somente os dois maiores atributos analisados para índice serão selecionados pelo sistema;

TABEL 3.4 - Fatores para Verificação de Índices e Pesos

	% dos valores em um atributo	Fator Peso
Discriminante	De 90 a 100 %	10
Altamente Determinante	De 81 a 90 %	7,5
Determinante	De 71 a 80 %	5
Pouco Determinante	De 61 a 70 %	3,5
Muito Relevante	De 51 a 60 %	3
Relevante	De 31 a 50 %	2,5
Pouco Relevante	De 16 a 30 %	1
Sem Relevância	De 0 a 15 %	0,5

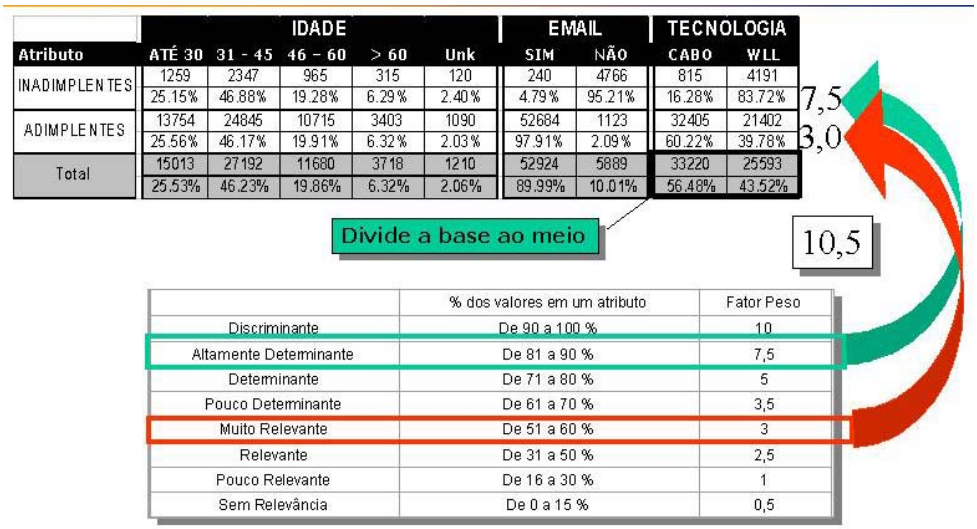


FIGURA 3.4 - Indexando Atributos

3.2.2.4 “Pesando” as características

Inicialmente, a mesma análise utilizada para a determinação dos índices foi usada, entretanto sem a preocupação de verificar se os valores de um atributo dividiam ou não a base na significativamente. Para a determinação deste fator, foi analisado também se os atributos não estavam concentrados em um único atributo-valor, pois caso isso ocorresse não poderia ser determinado nada relevante. Mas deve-se notar que, se em apenas uma das situações (adimplente ou inadimplente) o conjunto atributo-valor estivesse concentrado em um valor totalmente diferenciado das demais (e que não fosse desconhecido) este peso poderia ser mantido, pois apesar de não poder determinar em um primeiro momento qual a possível situação provável.

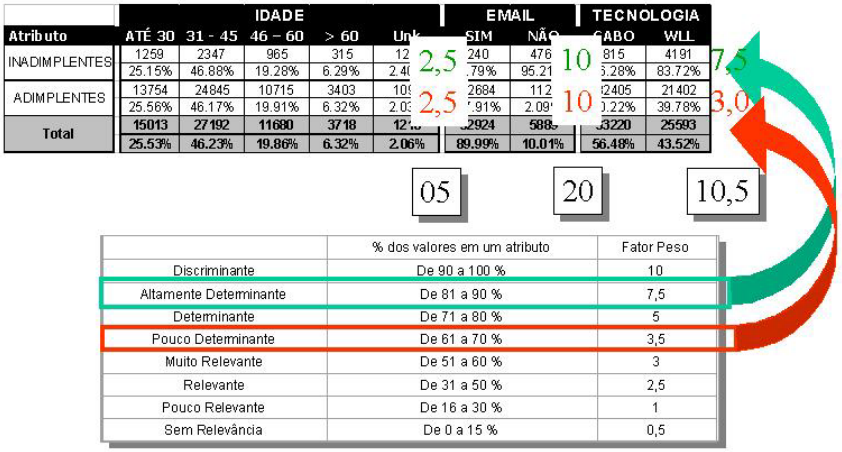


FIGURA 3.5 - Pesando Atributos

Após esta etapa foram atribuídos aos atributos para o maior atributos peso iguais a 10, a partir daí, aplicando se uma fórmula matemática (regra de três simples) determinou-se os demais pesos[10], conforme demonstra as TABELA 3.5 e TABELA 3.6.

TABELA 3.5 - Pesos dos Atributos para Crédito

Atributo	Peso
Celular	3,25
Ciclo de Faturamento	2,5
Classe de Crédito	2,75
Bairro	4,25
E-mail	10
Estado Civil	4,25
Forma de Pagamento	10
Idade	2,5
Mercado	8,75
Pessoas na Casa	4,25
Plano de Preço	3,25
Profissão	3
Sexo	3
Tecnologia	6,25

3.2.2.5 Recuperando Casos

Uma das etapas mais importantes no CBR é a característica que o mesmo tem em recuperar casos relevantes. Para isso os passos anteriores (indexar e pesar os casos) são extremamente importantes. A recuperação é baseada na similaridade entre caso alvo e casos fontes, sendo que o mesmo pode ser medido de forma **explícita** ou **indireta**.

[10] Estes pesos mostram a situação inicial do sistema, devendo variar conforme o sistema "aprende com o tempo".

- **Medida explícita**: este é o tipo que normalmente é utilizado pela facilidade de implementação:
    - independente da estratégia de recuperação ou da organização da memória;
    - k vizinhos mais próximos (*knn – k nearst-neigborhood*).

- **Medida indireta**:
    - dependente da estratégia de recuperação e/ou da organização da memória;
    - memória dinâmica (hierárquica)

TABELA 3.6 - Pesos para Análise de Cobrança

Atributo	Peso
Celular	3
Ciclo de Faturamento	1,75
Classe de Crédito	2,75
Conta Acumulada	10
Dias para Pagamento	2,5
Bairro	4,25
E-mail	10
Estado Civil	4,25
Forma de Pagamento	10
Idade	2,5
Valor da Fatura	6,25
Mercado	8,75
Pessoas na Casa	4,25
Plano de Preço	3,25
Primeira Fatura	7,5
Profissão	3
Quantidade de Reclamações	3
Sexo	3
Tecnologia	7,5

Desta maneira foram calculados os percentuais relativos de cada conjunto classe–atributo. Tomando-se o exemplo do atributo **IDADE**, onde os valores variam de 0 (até 30 anos) a 3 (maior que 60 anos) e levando-se em consideração que o valor mais similar seria 0 e o mais distante seria 1 (100%), encontrou-se qual seria faixa percentual de variação entre os valores dos atributos linearmente. Este raciocínio pode ser feito tanto nos atributos que tem valores discretos,como nos casos onde os atributos apresentam valores *booleanos*, como será demonstrado a seguir:

- **Atributos Discretos**: Para o caso do atributo **IDADE**, que é um caso de atributo contínuo, a faixa de variação entre os atributos seria de aproximadamente 25% (100% / 4). Hipoteticamente se em um caso problema o valor informado fosse até 30 anos (valor 0) e o valor do

atributo em um caso problema a ser analisado fosse de 46 a 60 anos (valor 3), a discrepância entre estes fatores seria 3 ou 75%;

- **Atributos Booleanos:** tomando-se como exemplo o atributo **TECNOLOGIA**, que tem como valores possíveis, no caso CABO (0) e RÁDIO (1), o fator de variação ficaria sempre em 100%.

Para calcular a similaridade entre os casos, o ideal seria poder determinar o quanto cada valor está percentualmente próximo *knn*, seguindo a seguinte fórmula:

$$Simil(X,Y) = \frac{\sum_{i=1}^{n} w_i \times sim_i(valor(a_{xi}), valor(a_{yi}))}{\sum_{i=1}^{n} w_i}$$

wi – peso do atributo
a_{xi} e a_{yi} – atributo-valor nos casos a serem comparados
sim_i – diferença entre os atributos-valores

Desta forma, quando se compara dois casos, atributo a atributo o seu % de similaridade se reduz a cada atributo analisado, conforme demonstrado na . Por exemplo, o atributo Débito Automático, cujo peso é 10 e os valores possíveis são:

1. Fatura para pagamento;
2. Débito Automático em Conta Corrente;
3. Débito Automático em Cartão de Crédito.

O percentual de similaridade neste caso, para este atributo seria de 3,42 em 10,25 possíveis, conforme mostrado na seqüência, ou seja, uma redução de 6,83 pontos percentuais.

$$Simil(X,Y) = \frac{\sum_{i=1}^{n} w_i \times sim_i(valor(a_{xi}), valor(a_{yi}))}{\sum_{i=1}^{n} w_i}$$

$$Simil(X,Y) = \frac{10 \times 33,33}{97,5}$$

$$Simil(X,Y) = 3,42$$

Este procedimento seria efetuado para todos os atributos um percentual de similaridade entre os casos seria de **64,16%** (100% - 35,84%).

TABELA 3.7 - Comparativo de Similaridade entre dois casos hipotéticos

Atributo	Peso	Caso 1	Caso 2	% de Discrepância no atributo	Redução Discrepância x Peso
Celular	3	SIM	NÃO	100	3,08
Ciclo de Faturamento	1,75	0	6	83,3	1,50
Classe de Crédito	2,75	1	4	50	1,41
Conta Acumulada	10	NÃO	NÃO	0	0,00
Dias para Pagamento	2,5	entre 15 e 30	entre 15 e 30	0	0,00
Bairro	4,25	QUEBEC	CALIFORNIA	40	1,74
E-mail	10	SIM	NÃO	100	10,26
Estado Civil	4,25	SOLTEIRO	SOLTEIRO	0	0,00
Forma de Pagamento	**10**	**DÉB. AUT. CONTA**	**FATURA**	**33,33**	**3,42**
Idade	2,5	21 a 30	21 a 30	0	0,00
Valor da Fatura	6,25	> 100	100 e 500	60	3,85
Mercado	8,75	RESIDENCIAL	RESIDENCIAL	0	0,00
Pessoas na Casa	4,25	2	3 a 5	20	0,87
Plano de Preço	3,25	FRANQUIA 1	FRANQUIA 1	0	0,00
Primeira Fatura	7,5	SIM	SIM	0	0,00
Profissão	3	ADVOGADO	SERVENTE	66	2,03
Reclamações	3	0	0	0	0,00
Sexo	3	MASC.	MASC.	0	0,00
Tecnologia	7,5	CABO	WLL	100	7,69
SOMA DOS PESOS	**97,5**				**35,84**

3.2.2.6 Tratamento de exceções

Há alguns fatores, denominados desconhecidos (*Unknowm*), que devem ser considerados de maneira especial quanto à comparação dos valores. A forma que estas exceções foram tratadas é descrita a seguir.

3.2.2.6.1 Valor desconhecido no caso problema

Neste caso, quando o usuário deixar de informar um determinado valor de atributo, seja porque este não pode verificar a existência de nenhum dos fatores ou não tenha conhecimento para tentar identificá-lo, o sistema simplesmente não poderá verificar este atributo, visto que se isso ocorrer não se estaria distanciando, de nenhuma forma, este caso dos demais casos onde os valores existem. Para que não se tomasse uma atitude “otimista”, onde o caso seria considerado similar, ou “pessimista”, que onde se teria uma atitude totalmente contrária classificando o mesmo como distante, optou-se pela política de se atribuir a este fator um valor que seria a metade do peso para aquele atributo.

Desta forma, quando um ou mais atributos sejam informados como desconhecidos o sistema apontará o melhor caso, mas em contrapartida diminuirá o grau de certeza apresentado de acordo com o percentual daquele atributo que foi informado como desconhecido.

3.2.2.6.2 Valor desconhecido na base de casos

Ocorre quando o usuário informar um valor qualquer para um atributo, o sistema deverá compará-lo com um valor desconhecido em um caso existente na base de casos. Para sanar esta situação, criou-se dentro banco de dados uma tabela, que

armazena todos os valores médios relativos ao atributo-valor de cada atributo. Isto é calculado pela média dos valores válidos para este conjunto.

Tomando, por exemplo, um caso hipotético no caso do atributo **IDADE**, para a classe **INADIMPLENTE** existe uma distribuição dos valores, conforme TABELA 3.8.

TABELA 3.8 - Distribuição dos Casos atributo IDADE

Atributo	ATÉ 30	31 - 45	46 – 60	> 60	Unk[11]
INADIMPLENTES	1259	2347	965	315	120

A quantidade de casos de cada faixa é multiplicada pelo seu valor, que como dito anteriormente, varia de **0** (até 30 anos) a **3** (maior que 60 anos). Todos estes valores foram somados e divididos pelo número de casos válidos. Desta forma para este conjunto classe-atributo o cálculo ficou:

$$\textbf{Valor médio} = [\Sigma (Val_{atrib} * Qtde_{casos})] / Casos_{válidos}$$

$$\textbf{Valor médio} = [\ (0*1259)_{até30} + (1*2347)_{31a45} + (2*965)_{46a60} + (3*315)_{mais60}] \ / \ 4886$$

$$\textbf{Valor médio} = [\ 0_{\ até30} + 2347_{31a45} + 1930_{46a60} + 945_{\ mais60}] \ / \ 4886$$

$$\textbf{Valor médio} = [\ 5222\ ] \ / \ 4886$$

$$\textbf{Valor médio} = 1{,}069$$

Assim sendo, caso o usuário informasse um valor, por exemplo, 3 (mais de 60 anos), e um dos casos este atributo estivesse desconhecido (para este conjunto nenhum caso se apresenta desconhecido) o valor utilizado para o cálculo de similaridade seria **1,069**.

Deve-se salientar que estes valores podem ficar desatualizados, então atualiza-los sempre que novos casos sejam inseridos na base, para manter a confiabilidade dos cálculos é uma rotina importante.

3.2.2.6.3 Casos Especiais

Neste ponto do trabalho, alguns problemas tiveram que ser resolvidos, como sendo que os dois principais eram como calcular a similaridade dos atributos ***bairro*** e ***profissão***;

[11] Os casos desconhecidos não entram no cálculo de valor médio.

Em ambos os estudos (Cliente ou Fatura) os dois atributos apresentavam considerável distribuição, os valores de pesos e índices tornavam-se irrelevantes. Entretanto, na análise percentual de clientes por bairro/profissão, podia-se notar claramente a concentração de inadimplência em alguns valores.

Assim, encontrou-se a necessidade estabelecer uma similaridade entre estes bairros/profissões para que os cálculos automáticos de índice e pesos pudessem ser feitos, assim como facilitar o cálculo de similaridade e foram solucionados da seguinte forma:

BAIRRO

Para este atributo foram analisadas diferentes formas de calcular a similaridade:

- **proximidade nos valores de CEP**: como todos os clientes, obrigatoriamente tem que ter o endereço de instalação do cliente completo na base, pensou-se em utilizar o mesmo para estar, pelos 3 algarismos intermediários do CEP (86-**041**-410, e.g.) estar fazendo o cálculo da similaridade dentro da cidade. Entretanto, além de ser extremamente complexo, um logradouro, por exemplo, que percorra uma distância considerável dentro da cidade, apesar de ter a mesma numeração estaria percorrendo, conseqüentemente, mais de um bairro e distorcendo a análise;

- **distância do bairro ao centro da cidade**: o segundo levantamento a ser feito foi tentar atribuir um fator de distanciamento do bairro ao centro da cidade. Entretanto, esta hipótese foi abandonada logo no início, pois pelos relatórios, pode-se verificar que este fator não era relevante, visto que existem bairros, que apesar de estarem afastados do centro da cidade analisada tinham um índice de inadimplência baixo;

- **valor venal médio (IPTU)**: foi feito um levantamento da classe econômica dos bairros em questão e notou-se uma forte similaridade. Este fator refere-se a diversos itens que podem ser relacionados, como, por exemplo, o valor do metro quadrado construído para aquela região, que constam no IPTU (Imposto Predial Territorial Urbano), mas que se mostrou ineficiente pela falta de atualização deste tipo de dados na grande maioria das cidades;

- **classe econômica estimada**: que é um levantamento “visual”, feito por especialistas, das construções e de seus tipos (comerciais, residenciais ou industriais) que predominam nestes bairros, qual o padrão destas, quantidade de carros na garagem, etc.

Assim sendo, optou-se pelo fator “classe econômica estimada” para a determinação de similaridade, pois se demonstrou muito eficiente na análise dos casos.

A TABELA 3.9 é um exemplo de qual o valor atribuído a alguns bairros, sendo que na interface com o usuário do sistema, o mesmo continua a selecionar o nome do bairro.

TABELA 3.9 – Relacionamento para o atributo Bairro.

BAIRRO	CLASSE
AGARI	A
ALPES	B
ALPES III	C
ALPHAVILLE	C
AMARO	C

PROFISSÃO

Precisou-se também encontrar uma similaridade nos mesmos. Para isso utilizou-se o fator de "*grau de escolaridade*" necessário para exercer determinada profissão, conforme TABELA 3.10.

TABELA 3.10 - Relacionamento para o atributo Profissão

PROFISSÃO	GRAU DE INSTRUÇÃO
MÉDICO	3
PEDREIRO	1
AUXILIAR DE ENFERMAGEM	2
CONTADOR	3
SERVIÇOS GERAIS	1

Esse objeto de análise demonstrou-se satisfatório. Entretanto, sabe-se que também está sujeito a uma margem de erro maior do que a do ***bairro***, pois existem profissões que não tem qualquer relação com o grau de escolaridade, como por exemplo, agricultor, pecuarista, juiz classista, etc., o que fatalmente irá prejudicar a análise. Para estes casos, a escolaridade foi colocada na média (segundo grau) com o intuito de diminuir a distância do calculo de similaridade no momento da análise.

Com isso, além de aumentar a performance do sistema, o peso de cada atributo ficou mais aderente ao analisado pelo especialista e melhorando assim e o cálculo de similaridade, assim como o entendimento por parte do usuário do sistema.

Foi feita, também, uma validação posterior via *DW*, que comprovou a similaridade de distribuição dos atributos pela solução dada.

4 Implementação

Neste capitulo, será apresentado o protótipo que foi implementado, detalhando todo o processo de carga, recuperação de casos e geração de arquivos para serem utilizados dentro do processo de cobrança, que foi o principal objetivo deste estudo.

4.1 Aspectos da Implementação

Todos os sistemas da empresa em questão estão desenvolvidos sobre uma arquitetura de banco de dados relacionais (ORACLE 8.1.3), sobre este sistema está montado um *DW* com todas as informações necessárias para que todos os dados fossem carregados. Entretanto, como a base de dados, demandaria uma grande capacidade de armazenamento, foi definida uma amostra para que este estudo de caso fosse estruturado.

Assim, toda a estrutura do *DW*, foi extraída para um BD relacional MS ACCESS com o intuito de facilitar o manuseio desta. Também foi escolhido um universo de clientes relacionados somente a uma cidade e que estivessem em um determinado período de tempo.

4.1.1 Dados

Como dito a base de dados estava toda disponível em um *DW*, onde são armazenados os dados, em um alto grau de granularidade, de todos os clientes. No momento em que o cliente é cadastrado nos sistemas operacionais da companhia, as informações do mesmo são enviadas ao *DM* Mercadológico.

Após isso, quando são geradas as faturas, os dados de faturamento do cliente, também são enviados ao *DM* Financeiro, além dos dados relativos à atividade de CRM com relação às vezes em que o cliente entrou em contato com a companhia para tirar dúvidas ou reclamar de problemas diversos (problemas técnicos ou relativos ao faturamento).

Deve-se notar também, que para evitar uma grande quantidade de informações semelhantes com relação a clientes idênticos e também para fazer com que o CBR tenha uma maior eficiência, há uma rotina de sumarização dos dados, com o intuito de se melhorar a qualidade e agilidade do CBR.

Também deve se considerar a periodicidade com que estes dados serão atualizados, pois como há uma grande de quantidade de casos similares e o aumento da área de cobertura da empresa não se apresentar tão dinâmica, os dados podem ser

carregados semanalmente ou até mesmo mensalmente, quando a rotina de sumarização faria sentido e os dados teriam um grau maior de relevância.

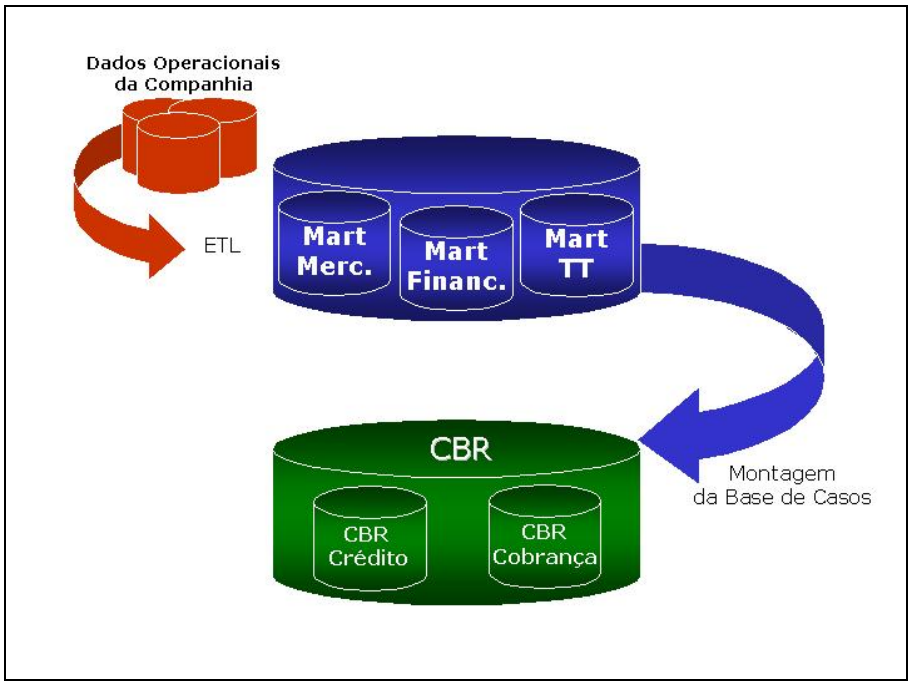


FIGURA 4.1 - Modelo Esquemático de Leitura

Fica óbvio que os dados poderiam ser carregados para a base de casos diretamente dos dados operacionais. Assim, caso opte-se pela carga direta dos dados operacionais da companhia a rotina de ETL deveria ser "duplicada", ou seja, duplicação de carga, de validação, etc. E como juntamente com o CBR o especialista utiliza o *DW* para agregar conhecimento, desta forma se garante a consistência entre os dados dos *DM*'s (*DW*) e do CBR.

4.1.2 Funcionamento

Como explicado anteriormente, o sistema, apesar de ser um só, na verdade visa atender a dois momentos do processo:

- **Crédito**: momento de análise do risco do cliente na hora da venda;
- **Cobrança**: momento de estabelecimento dos cenários de cobrança ao qual o cliente será aplicado.

4.1.2.1 Carregando os casos

Para iniciar a utilização do sistema é necessário carregar os dados. Isso pode ser feito separadamente para os dados relativos a análise de **Cliente** e de **Faturas.**

O sistema disponibiliza guias específicas para que o usuário/administrador do sistema faça esta carga somente no momento em que queira inserir todos os novos casos, constantes no *DW*, no CBR. Fazendo assim com que o sistema tenha mais casos para comparação.

Esta operação é feita na função conforme demonstrado na FIGURA 4.2.

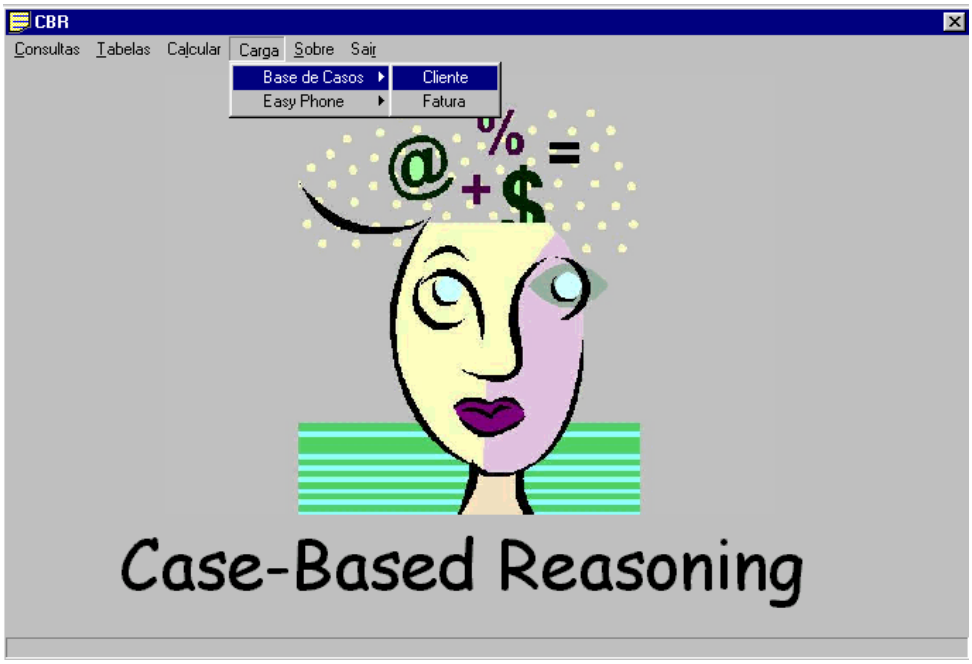
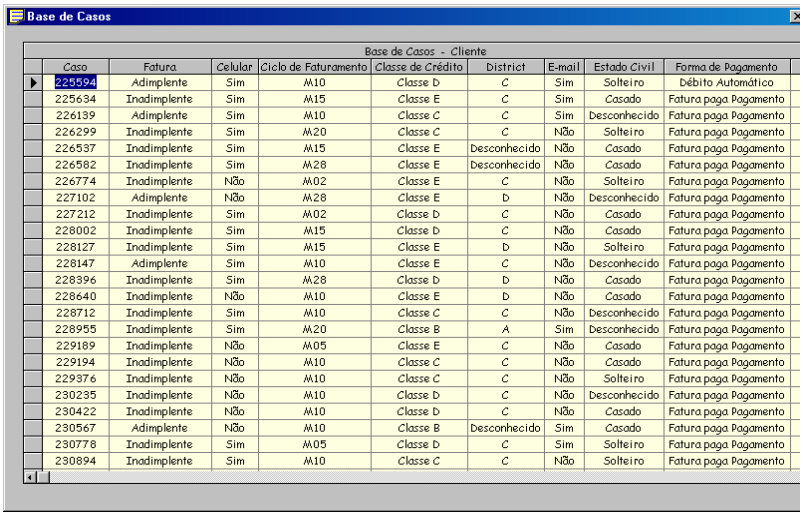


FIGURA 4.2 - Funcionalidade de Carga

Após a carga efetuada, pelo menu **CONSULTA**, **BASE DE CASOS CLIENTE,** o usuário pode analisar todos os casos que constam na base, conforme a FIGURA 4.3, se foram inseridas com êxito, etc. Esta consulta também pode ser utilizada, para que após a consulta do melhor caso, o especialista ou o usuário comparem o caso analisado com o caso retornado pelo sistema, visto que o ID do caso mais semelhante consta no formulário de consulta.



Caso	Fatura	Celular	Ciclo de Faturamento	Classe de Crédito	District	E-mail	Estado Civil	Forma de Pagamento
225594	Adimplente	Sim	M10	Classe D	C	Sim	Solteiro	Débito Automático
225634	Inadimplente	Sim	M15	Classe E	C	Sim	Casado	Fatura paga Pagamento
226139	Adimplente	Sim	M10	Classe C	C	Sim	Desconhecido	Fatura paga Pagamento
226299	Inadimplente	Sim	M20	Classe C	C	Não	Solteiro	Fatura paga Pagamento
226537	Inadimplente	Sim	M15	Classe E	Desconhecido	Não	Casado	Fatura paga Pagamento
226582	Inadimplente	Sim	M28	Classe E	Desconhecido	Não	Casado	Fatura paga Pagamento
226774	Inadimplente	Não	M02	Classe E	C	Não	Solteiro	Fatura paga Pagamento
227102	Adimplente	Não	M28	Classe E	D	Não	Desconhecido	Fatura paga Pagamento
227212	Inadimplente	Sim	M02	Classe D	C	Não	Casado	Fatura paga Pagamento
228002	Inadimplente	Sim	M15	Classe D	C	Não	Casado	Fatura paga Pagamento
228127	Inadimplente	Sim	M15	Classe E	D	Não	Solteiro	Fatura paga Pagamento
228147	Adimplente	Sim	M10	Classe E	C	Não	Desconhecido	Fatura paga Pagamento
228396	Inadimplente	Sim	M28	Classe D	D	Não	Casado	Fatura paga Pagamento
228640	Inadimplente	Não	M10	Classe E	D	Não	Casado	Fatura paga Pagamento
228712	Inadimplente	Sim	M10	Classe C	C	Não	Desconhecido	Fatura paga Pagamento
228955	Inadimplente	Sim	M20	Classe B	A	Sim	Desconhecido	Fatura paga Pagamento
229189	Inadimplente	Não	M05	Classe E	C	Não	Casado	Fatura paga Pagamento
229194	Inadimplente	Não	M10	Classe C	C	Não	Casado	Fatura paga Pagamento
229376	Inadimplente	Não	M10	Classe C	C	Não	Solteiro	Fatura paga Pagamento
230235	Inadimplente	Não	M10	Classe D	C	Não	Desconhecido	Fatura paga Pagamento
230422	Inadimplente	Não	M10	Classe D	C	Não	Casado	Fatura paga Pagamento
230567	Adimplente	Não	M10	Classe B	Desconhecido	Sim	Casado	Fatura paga Pagamento
230778	Inadimplente	Sim	M05	Classe D	C	Sim	Solteiro	Fatura paga Pagamento
230894	Inadimplente	Sim	M10	Classe C	C	Não	Solteiro	Fatura paga Pagamento

FIGURA 4.3 - Consulta a Base de Casos Cliente

4.1.2.2 Calculando Pesos, Índices e Valor-Médio.

Antes de se iniciar a utilização do CBR, após a primeira carga, é necessário que sejam calculados os índices, pesos e valor-médio, caso contrário o sistema estará impossibilitado de efetuar os cálculos. Depois disso, a periodicidade destes cálculos ficam a critério do administrador do sistema. Entretanto, é interessante que seja feito com certa freqüência para que a comparação seja mais eficiente.

Há um menu especifico **CALCULAR** onde estes cálculos podem ser facilmente iniciados pelo usuário. Foi disponibilizada uma tela, para que todos os atributos e seus respectivos pesos sejam analisados pelo usuário/especialista, conforme mostrado na FIGURA 4.4. Para visualizar os outros atributos basta utilizar a barra de rolagem que consta na tela.

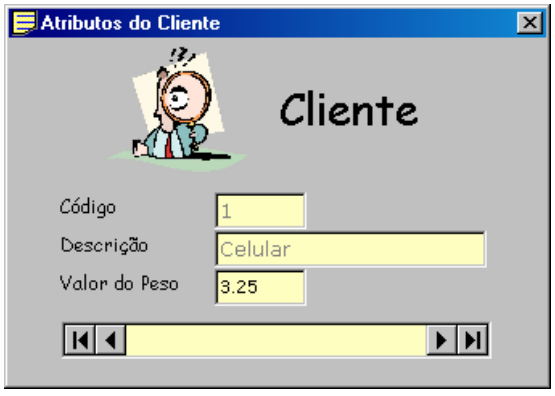


FIGURA 4.4 - Tela de Pesos dos Atributos.

4.1.2.3 Comparando os Casos

A entrada de dados é necessária para alimentar o sistema e pode ser feita de duas maneiras:

- para os casos de crédito, onde é necessário comparar, somente um novo caso à base de casos já existente utiliza-se um formulário apresentado no próprio sistema e que pode ser acessado através do menu **CONSULTAS**, depois a opção **MELHOR CASO** e depois **CLIENTE**, acessando assim a tela mostrada na FIGURA 4.5;

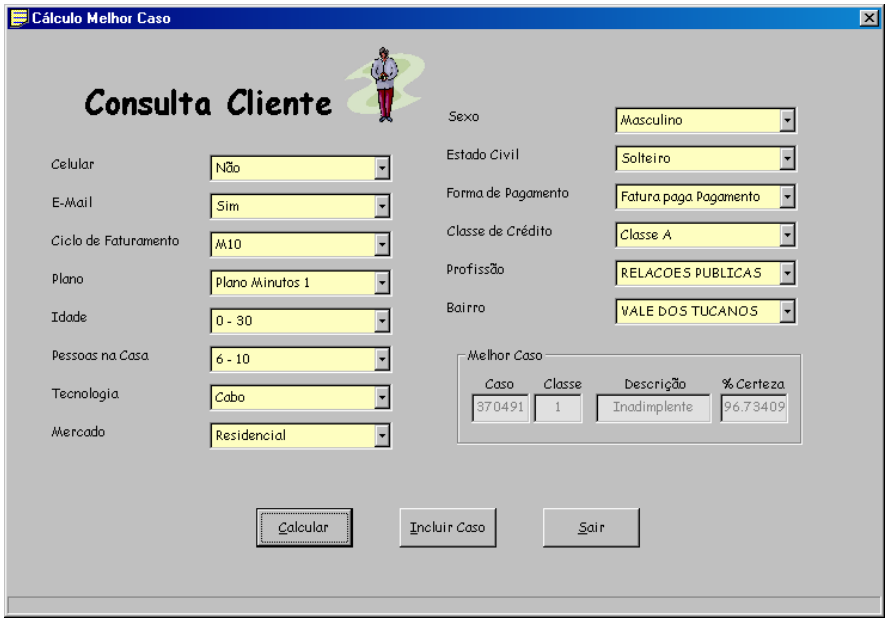


FIGURA 4.5 - Formulário de Consulta Crédito (cliente)

- para os casos de consulta a casos de cobrança, também pode ser utilizado um formulário muito similar a consulta de casos de cliente, para que casos isolados sejam analisados, conforme mostra a FIGURA 4.6.

Como o objetivo desta ferramenta é iniciar “a cobrança” dos clientes antes mesmo da data de vencimento, o processo acontece da seguinte maneira:

- quando houver uma nova carga no DM Financeiro e as novas faturas dos clientes sejam disponibilizadas, todos os dados necessários para a recuperação dos casos daqueles clientes são carregados no CBR (em uma tabela temporária);

- os dados, destas novas faturas, são comparados com a base de casos existentes;

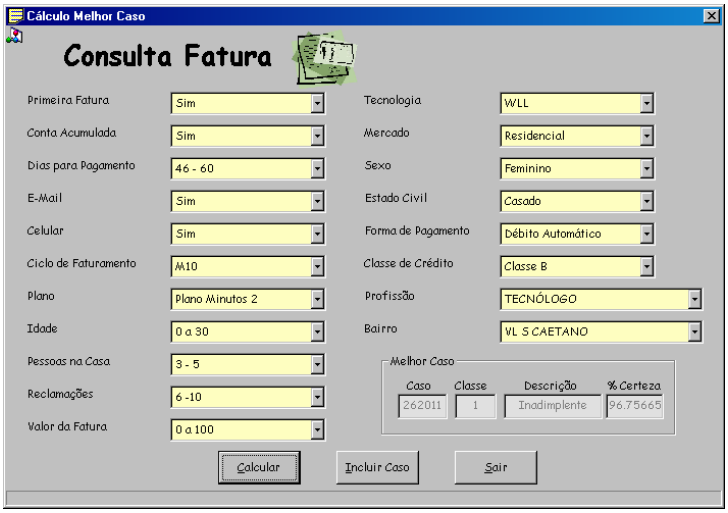


FIGURA 4.6 - Formulário de Consulta Cobrança (fatura)

- após esta comparação, o usuário do sistema pode gerar, já a partir do sistema uma campanha de contato com estes clientes, obedecendo o grau de similaridade estipulado pelo usuário;

- desta forma o sistema exporta o arquivo no formato definido a ser carregado no *predictive dialer* e gerar ligações aos clientes.

O modelo esquemático deste processo é mostrado na FIGURA 4-7.

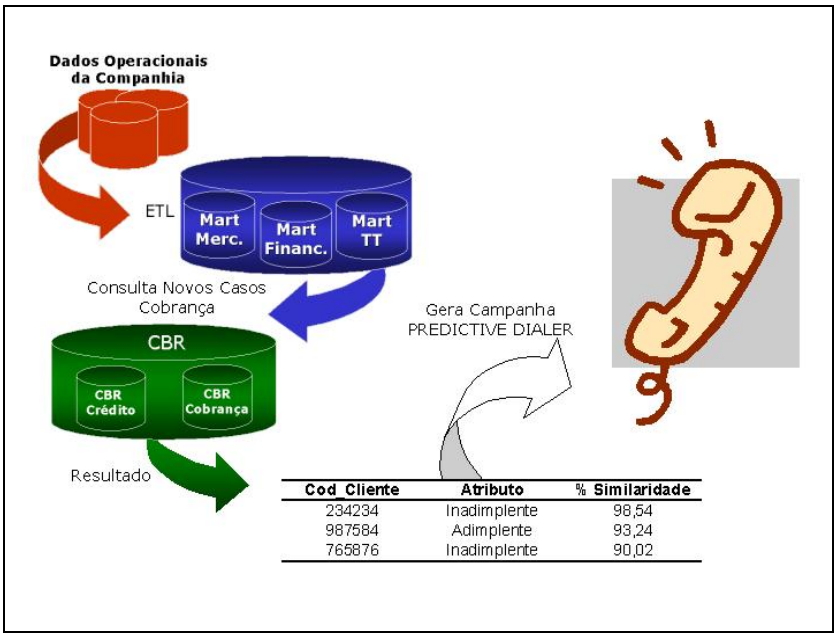


FIGURA 4-7 - Esquema de Consulta a Dados de Cobrança

Na prática, o usuário do sistema acessar o gerador a partir do menu **CARGA**, opção **GERAR INVOICE** e informar o percentual de similaridade desejado para a campanha na tela, conforme FIGURA 4.8.

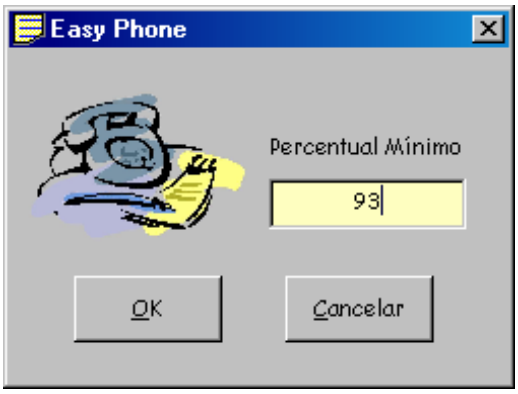


FIGURA 4.8 - Tela de Geração de Campanha

Assim como na carga dos clientes, os dados a serem utilizados para a campanha também são visualizados após a carga, conforme FIGURA 4.9.

O mais interessante é gerar este tipo de campanha para clientes **FPD**[12] (*First Payment Default*), pois além destes clientes apresentarem um risco maior, iniciar as atividades de cobrança antecipadamente pode identificar possíveis problemas que afetaram os clientes e dar mais opções aos mesmos (e.g. parcelando a fatura ou prorrogando o prazo para pagamento, verificando problemas de faturamento, ajustando a fatura, emitindo boleto para pagamento, etc.)

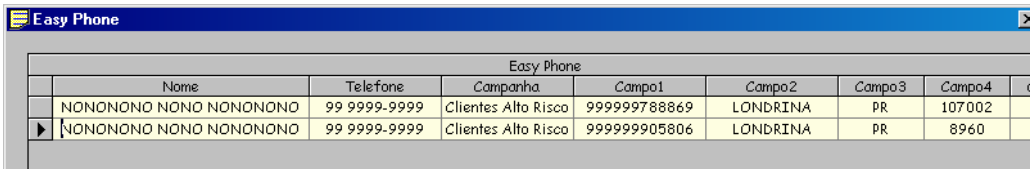


FIGURA 4.9 - Clientes Inseridos no processo de Campanha

Uma ação como esta, pode ser vista não somente como um serviço a seus clientes, pelo menos para os não fraudadores, mas também para a própria companhia, pois não somente reduz o nível de inadimplência, a previsão de fluxo de caixa, como também reduz os níveis de reclamação, diminuindo o fluxo de chamadas nos CALL CENTERS e demais problemas.

Vale salientar que todos estes fatores são medidos mensalmente pela agência, devendo a empresa tomar muito cuidado com os mesmos, pois a ocorrência de, por exemplo, grandes níveis de reclamação de conta, podem determinar a aplicação de sanções, geralmente pesadas multas (Resolução 217 – ANATEL).

[12] Clientes que estão sendo faturados pela primeira vez, não tendo nenhuma outra conta paga ou não.

4.1.3 Resultados Obtidos com o Sistema

A verificação do resultado do sistema não pôde ser feita durante o processo em produção. Entretanto, fez-se uma comparação de alguns casos novos e o acompanhamento destes para ver se a o fator de similaridade dos clientes iriam se confirmar no decorrer do tempo.

Foram inseridos dois grupos de faturas (casos) reais, sendo um grupo INADIMPLENTE e outro ADIMPLENTE, para que o sistema gerasse a campanha. Assim, foram usados cinco fatores de similaridade distintos para gerar campanhas com a mesma fatura e o resultado destas massas de testes.

4.1.3.1 Casos Adimplentes

Inicialmente, foram inseridas no sistema, dados de faturas sabidamente ADIMPLENTES e geradas as campanhas para inadimplentes. O resultado esperado seria que ***nenhuma*** destas faturas estivesse no arquivo gerado pela campanha. Para isso foram feitos testes com vários fatores de similaridade e o resultado obtido está mostrado na FIGURA 4.10

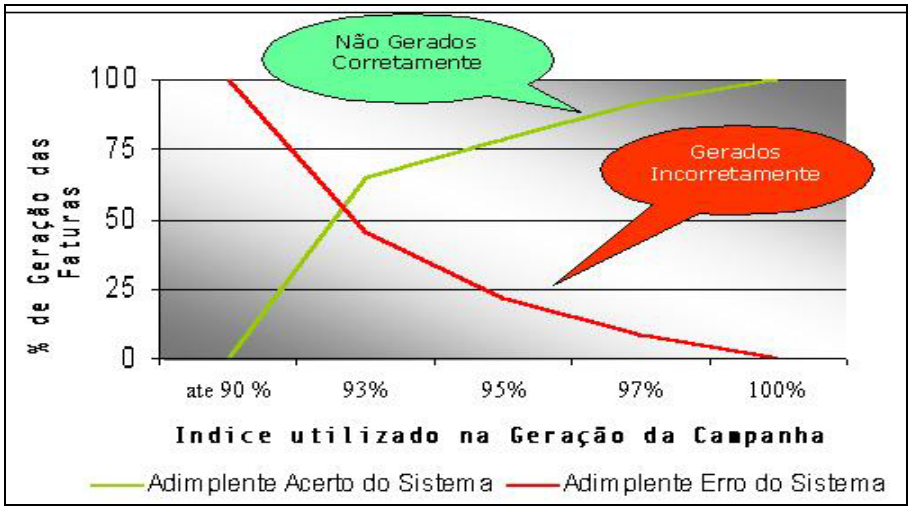


FIGURA 4.10 - Resultado Campanha Inadimplentes com clientes adimplentes

Nota-se claramente que com o fator de similaridade de 90% o sistema gera todas as faturas erradamente, pois o fator de similaridade é satisfatório para todas as faturas. A medida que o fator é majorado o sistema passa a ter um maior grau de acerto.

4.1.3.2 Casos Indimplentes

Após esta etapa, contrariamente a etapa inicial, foram inseridas no sistema, dados de faturas sabidamente INADIMPLENTES e geradas as campanhas para inadimplentes. O resultado esperado seria que ***todas*** estas faturas estivessem no arquivo

gerado pela campanha. Para isso também foram feitos testes com vários fatores de similaridade e o resultado obtido está mostrado na FIGURA 4.11.

Nota-se, que com o fator de similaridade de 90% o sistema gera todas as faturas erradamente, pois o fator de similaridade é satisfatório para todas as faturas. A medida que o fator é majorado o sistema passa a não gerar todas as faturas erradamente, devido a pequenas distorções de similaridade.

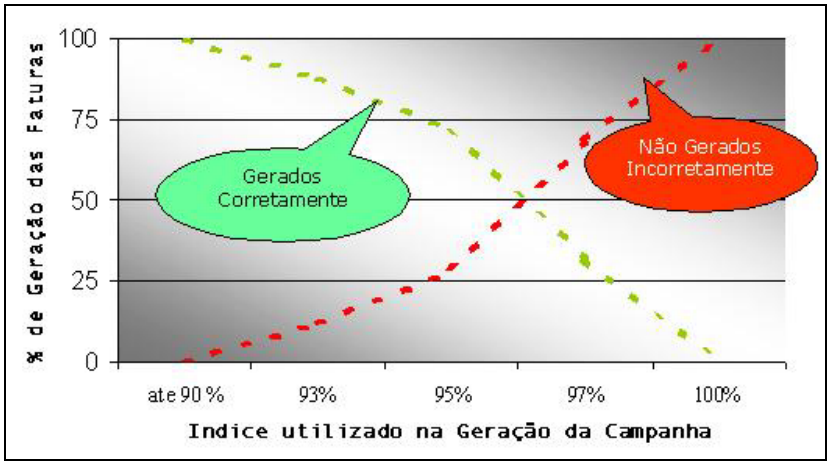


FIGURA 4.11 - Resultado Campanha Inadimplentes com clientes inadimplentes

4.1.3.3 Resultado Comparativo

Sobrepondo os dois gráficos, pode-se notar que o fator de similaridade encontrado onde haverá um maior número de contatos feitos com “eventuais inadimplentes” do que com inadimplentes, neste momento do sistema, é de 95%, conforme mostra a FIGURA 4.12.

Nota-se que este fator deverá ser avaliado constantemente, já que com o tempo, a conseqüente “aprendizagem!” do sistema e a maturação do mesmo, este fator de similaridade tende a sofrer alterações, podendo chegar a um fator próximo de 100%, além também do percentual de erro do sistema (a não inserção de clientes inadimplentes e inclusão de adimplentes) tende a diminuir.

Outro ponto que deve ser salientado, é que a análise do fator de similaridade também deve levar em conta a quantidade de contatos que poderão ser feitos pela força de trabalho ou capacidade do sistema de geração de ligações consiga suportar naquele determinado momento. Por exemplo, em um hipotético ciclo de faturamento, são processadas cem mil novas faturas, mas a força de trabalho disponível para que se faça

campanha ativa no período que antecede a data de vencimento das faturas teria capacidade de efetuar somente dois mil contatos. Entretanto, caso se utilize o fator de similaridade de 95%, conforme visto, estar-se-ia gerando cerca de 15 mil contatos.

Assim, o fator de similaridade também pode ser utilizado para controlar a quantidade de contatos efetuados, mesmo porque haverá uma eficiência maior.

Ou seja, pela amostragem verificou-se que a utilização do CBR para estar sugerindo uma campanha inicial de contato com os clientes com maior risco, baseando-se nos casos já existentes na base, com o intuito de antecipar o processo de cobrança pode trazer benefícios consideráveis tanto para a área de cobrança, que poderá estar sendo mais eficiente em suas ações e até mesmo suspendendo imediatamente clientes suspeitos de fraude, quanto para a área de relacionamento com os clientes, que poderá estar maximizando a utilização de seus recursos.

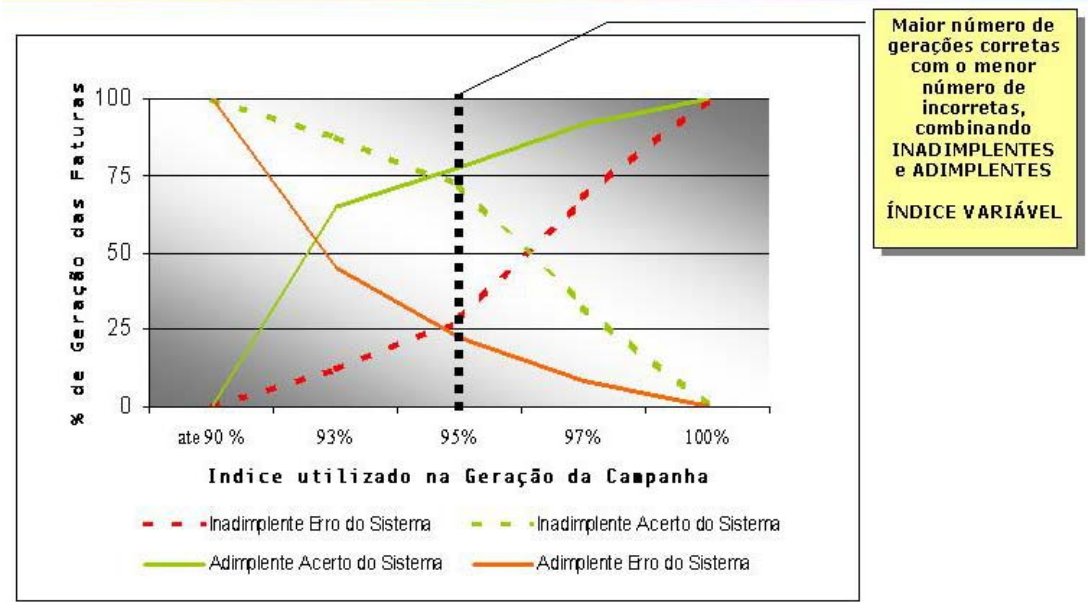


FIGURA 4.12 - Gráfico comparativo das Campanhas

5 Conclusão

Como dito no capítulo introdutório deste trabalho, o mercado de telecomunicações tem mudado muito nos últimos anos, o que obriga a todas as empresas, cada vez mais, a conhecer os seus clientes, mantê-los e principalmente saber quais destes clientes devem, ou não, ter vantagens dadas pela empresa.

Certamente as propostas de BI (*Business Intelligence*) sendo utilizadas a outras propostas como CRM (*Customer Relantionship Manager*) não podem mais ser entendidas como soluções miraculosas, mais sim como soluções imprescindíveis, sem as quais as organizações poderão perder o "bonde da história" e desaparecer. Assim, cada dia mais as empresas estão preocupadas em antecipar tendências, analisar o mercado, fatores econômicos que podem afetar diretamente o negócio e principalmente a movimentação da concorrência.

Com o presente estudo de caso ficou claro como as ferramentas se completam (BI, CRM e CBR) e como a utilização racional das mesmas pode gerar resultados satisfatórios em um tempo muito reduzido quando necessário.

As ferramentas de CRM, apesar de não abordadas neste trabalho, têm grande importância no momento em que fornecem uma preocupação muito grande em manter todas as informações do histórico do cliente dentro da empresa. A ferramenta de BI foi fundamental para suprir os especialistas com informações e possibilitar a análise dos dados, em várias dimensões, permitindo aos mesmos comprovar, com dados, algumas inferências que tinham sobre o comportamento dos clientes no problema dado.

Agora, como uma surpresa, pode-se colocar o fato do CBR, que teve um funcionamento muito bom no que se propôs inicialmente, mas como um "gerador de conhecimento", visto que em determinados momentos, ao gerar pesos e índices, por exemplo, na verificação da veracidade das informações geradas, descobriu-se que o sistema estava correto e "instigou" ao especialista a achar uma resposta para aquela "questão".

Além disso, um CBR mostrou-se muito adequado para a aplicação de Crédito e Cobrança, trazendo muitas vantagens, como por exemplo, a facilidade de aquisição do conhecimento (diferentemente se tivesse optado por uma técnica de *Data Mining*), a reutilização do conhecimento e principalmente a rápida e simples modelagem/desenvolvimento do sistema. O CBR também é muito fácil de ser mantido, como demonstrado no desenvolvimento do trabalho, bastando para isso gerar uma nova carga de dados, calcular índices e pesos, tudo isso sem haver qualquer nova interferência de um especialista ou de um engenheiro do conhecimento.

O sistema desenvolvido é muito simples, mas isso não significa que não atenda ou que se deixou de analisar todos os dados. Pelo contrário, foi uma ferramenta onde todos os dados, até mesmo aqueles que o especialista acreditava ser de menor interferência no comportamento a ser analisado, foram inseridos e cruzados a fim de ter um maior grau de confiabilidade.

Entretanto, como o tempo de desenvolvimento solicitado era restrito, fica a ser desenvolvida a análise de outros atributos relativos aos clientes (não somente ADIMPLENTE ou INADIMPLENTE), como por exemplo:

- mês de ocorrência, pois é sabido que a inadimplência ocorrem sazonalmente;
- clientes que solicitaram negociação;
- pesquisa e complementação dos motivos pelos quais os clientes se tornaram inadimplentes;
- *churn*[13] de clientes; e
- fraudes de subscrição.

Estas informações poderiam estar auxiliando o atendente já no contato com o cliente ou até mesmo fazer com que este percentual seja levado para dentro de outros sistemas, para ser um dos fatores de análise de *score* de risco, mas para isso seria necessária uma análise mais detalhada de outros pontos do processo.

Como este trabalho foi feito com um estudo de caso baseado somente em uma cidade (Londrina), também falta a implementação da diferenciação por cidade. Mas para isso, também um estudo para a para determinar a similaridade das cidades, talvez por estado (sabendo-se, por exemplo, que o Rio Grande do Sul o índice de inadimplência é menor), mas deverá ser analisado posteriormente.

Este trabalho também fez uma ligação entre a prática e teoria, principalmente em campos que hoje não são muito explorados. o que parece ser muito importante pois a universidade, sem deixar de ser acadêmica, deve voltar os olhos ao mercado, procurar, sem ser comercial, demonstrar ganhos a sociedade em qualquer área que seja. Todo trabalho dentro de uma universidade deve resultar em algo "real" para a sociedade, caso contrário estará se fechando em si mesma e cada vez menos projetos e financiamentos virão, menos pesquisas serão feitas e este processo neste ciclo vicioso.

Trabalhos como estes, levando em consideração o cenário econômico que está sendo vislumbrado, têm um grande valor, já que as empresas podem reduzir consideravelmente seus gastos com consultas a *bureaus* de crédito, podendo estabelecer políticas mais aderentes com a sua realidade.

[13] Possibilidade do cliente trocar de operadora;

Bibliografia

[ALL94] ALLEN, B. P. N Case-Based Reasoning: business application. **Communications of the ACM,** New York, v. 37, n.3, p.40-42, Mar. 1994.

[ALT96] ALTER, S. **Information systems**: a Management Perspective. New York: Addison Wesley, 1996.

[AMA2001] AMARAL, F. C. N. do. **Data Mining**: técnicas e aplicacoes para o Marketing Direto. São Paulo: Berkeley Brasil, 2001.

[BAR2001] BARBIERI, Carlos**. Business Intelligence – Modelagem & Tecnologia**. São Paulo: Axcel Books, 2001.

[COS93] COSTAS, T.; KASHYAP, N. Case-Based Reasoning and Learning in Manufactoring with TOLTEC Planner. **IEEE Transactions on Systems, Man e Cybernetics**, New York, v 1, p. 1992-194, 1993.

[DAT2000a] DATAWARE, Knowledge Management: Linking People to Knowledge for Bottom-Line Results Online. Disponível em: <www.dataware.com> . Acesso em: jun. 2001.

[FUL98] FULD, L. M. **Forum da Fuld Co**. Resposta dada a Jan Herring em 26/10/98. Disponível em : <www. fuld.com/forum/fuld>. Acesso em: nov. 2000.

[GAL98] GALLIERS. R. D.; BAETS, W.R.J. **Information Technology and Organizacional Transformation**. Londres: John Wiley & Sons, 1998.

[INM92] INMON, W. H. **Building the DW**. USA: John Wiley & Sons Inc, 1992.

[INM94] INMON, W.H.; HACKATHON, R. D. **Using the DW**. USA: Wiley-QED Publication, 1994.

[KIM98] KIMBALL, R. **DW Toolkit**. Tradução Mônica Rosemberg. São Paulo: Makron Books, 1998.

[KOL93] KOLODNER, J. L. **Case-Based Reasoning**. San Mateo: Morgan Kaufmann, 1993.

[LOR98] LORENZI, F. **Uso da Metodologia de Raciocínio Baseado em Casos na Investigação de Irregularidades nas Internações Hospitalares**. 1998. 73f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, UFRGS, Porto Alegre.

[MER96] MERIDITH, M.; KHADER, A. Divide and Aggregate: Designing Large Warehouses. [S.l.: s.n.], 1996.

[MUR96] MURRAY, P. C. **New language for new leverage**: the terminology of knowledge management (KM). Disponível em: <http://www.ktic.com/topic6/13_term0.htm>. Acesso em: jun. 2001.

[NAI85] NAISBITT, J. **Paradoxo Global**. Rio de Janeiro: Campus, 1985.

[NON97] NONAKA, I.; TAKEUCHI, H. **Criação de Conhecimento na Empresa.** Rio de Janeiro: Campus, 1997.

[SAM94] SAMMON, W. L.; KURLAND, M. A.; SPITALNIC, R. **Business Competitor Intelligence**: methods for collecting, organizing and using information. New York: John Wiley, l994.

[SAN2001] SANTOS , J., HENRIQUES, N. A.C.; REIS, V. Data Mining / Data Warehousing. Disponível em: <http://students.fct.unl.pt/users/nach/DMDW/prologo>. Acesso em 01 jun.2001.

[SCH94] SCHEER, A. W. **Business Process Engineering – Reference Model for Industrial Enterprises**. Berlim: Springer – Verlag, 1994.

[SIL89] SILVA, L. N. **A 4ª Onda**: os novos rumos da sociedade da informação. Rio de Janeiro: Record, 1989.

[SIM85] SIMPSON, R.L. **A Computer Model of Case-Based Reasoning in Problem Solving:** an investigation in the domain of dispute mediation. Atlanta: Georgia Institute of Technology, School of Information and Computer Science, 1985. (GIT-ICS-85/18)

[SVE98] SVEIBY, K. E. **A Nova Riqueza das Organizações**: gerenciando e avaliando patrimônios de conhecimento. Rio de Janeiro : Campus, 1998.

[THE2000] THEARLING, K. **Data Mining, CRM, Decision Support, and Database Marketing**. Disponível em: <http://www3.shore.net/~kht/text/DMwhite/DMwhite.htm>. Acesso em: nov. 2000.

[TSA97] TSATSOULIS, C.; CHENG, Q. ; WEI, H. Integrating case-based reasoning and decision theory. **IEEE Expert**, Los Alamitos, p.46-55, July/Aug. 1997.

[TUB99] TURBAN, E. **Decision Support Systems and Intelligent Systems**. New Jersey: Prentice-Hall, 1999.

[VAN89] VANLEHN, K. Problem solving and congnitive skill acquisition. In: POSNER, M. (Ed.). **Foundation of Cognitive Science**. Cambridge, MA: Mit Press, 1989. p. 526-579.

[WAT94] WATSON, I.; MARIR, F. Case-Based Reasoning: a Review. **The Knowledge Engineering Review**, Saldford, v. 9, p. 327-354, 1994.

[WAT95] WATSON, I. **The Case for Case-Based Reasoning**. Saldford: University of Saldford, 1995.