UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO – BACHARELADO

GESTÃO DO CONHECIMENTO: APLICAÇÃO EM DATA MINING UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS PARA GERAÇÃO DO CAPITAL INTELECTUAL

SIDNEI SCHMITT

BLUMENAU
2007

2007/2-31

SIDNEI SCHMITT

GESTÃO DO CONHECIMENTO: APLICAÇÃO EM DATA MINING UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS PARA GERAÇÃO DO CAPITAL INTELECTUAL

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

Prof. Oscar Dalfovo, Dr. - Orientador

BLUMENAU
2007

2007/2-31

GESTÃO DO CONHECIMENTO: APLICAÇÃO EM DATA MINING UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS PARA GERAÇÃO DO CAPITAL INTELECTUAL

Por

SIDNEI SCHMITT

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

Presidente:	Prof. Oscar Dalfovo, Dr. – Orientador, FURB
Membro:	Prof. Ricardo Alencar de Azambuja, MAd. – FURB
Membro:	Prof. Cláudio Loesch, Dr. – FURB

Blumenau, 05 de dezembro de 2007.

Dedico este trabalho a meus pais e meus familiares, minha noiva e a todos os amigos, especialmente aqueles que me ajudaram diretamente na realização deste.

AGRADECIMENTOS

A Deus, pelo seu imenso amor e graça.

À minha família, que sempre me apoiou e esteve presente.

À minha noiva Roberta e aos meus amigos, pelosauxílios e cobranças.

Ao meu orientador, Dr. Oscar Dalfovo, por ter acreditado na conclusão deste trabalho.

Quanto maiores somos em humildade, tanto mais perto estamos da grandeza.

Rabindranath Tagore

RESUMO

O conhecimento tornou-se um dos fundamentais recursos para as organizações, desde o momento em que ocorreu a troca de uma economia industrial para uma economia global extremamente competitiva. Diante disto, a gestão do conhecimento surge como uma metodologia de gerenciamento que vai além do simples processo de inovação, determinando a vantagem competitiva de uma organização. O principal objetivo na gestão do conhecimento para as organizações é obter alguma vantagem competitiva sobre seus concorrentes inovando seus produtos, serviços e processos. Visando facilitar a adaptação das organizações frente às mudanças provocadas pela globalização das economias e o conseqüente acúmulo de dados armazenados por estas organizações, o presente trabalho apresenta o emprego de uma arquitetura de gerenciamento do conhecimentocom o uso de *data mining* aliado à teoria dos conjuntos aproximativos para proporcionar às organizações mais agilidade e conseqüentemente uma melhor competitividade. Este trabalho apresenta a especificação e desenvolvimento de uma ferramenta em ambiente web para o gerenciamento de capital intelectual. Dentro deste processo, o sistema possibilita o levantamento do capital intelectual existente na organização, bem como conhecer quais os seus profissionais que estão mais preparados para enfrentar o mercado.

Palavras-chave: Gestão do conhecimento. Capital intelectual. Mineração de dados. Conjuntos aproximativos.

ABSTRACT

Knowledge has become one of the key resources for the organizations, from the moment when change took place in a industrial economy to an extremely competitive global economy. Because of this, the management of knowledge emerges as a management methodology that goes beyond the simple process of innovation, determining the competitive advantage of an organization. The main objective in the management of knowledge for organizations is to obtain a competitive advantage over its competitors innovating products, services and processes. To facilitate the adjustment of organizations front to the changes brought by the globalization of economies and the consequent accumulation of data stored by these organizations, this work presents the use of an architecture of the knowledge management with the use of data mining combined with the theory of approximate sets to give organizations more agility and therefore better competitiveness. This work presents the specification and development of a tool in Web environment for the management of intellectual capital. Within this process, the system allows to check the existing intellectual capital in the organization, and know the professionals who are more prepared to face the market.

Key-words: *Management of knowledge. Intellectual capital. Data mining. Approximate sets*.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

DM – *Data Mining*

ERP – *Enterprise Resource Planning*

FURB – Fundação Universidade Regional de Blumenau

HTML – *Hyper Text Markup Language*

IBM – *International Business Machines Corporation*

IDE – *Integrated Development Environment*

J2EE – Java 2 Enterprise Edition

JSP – *Java Server Pages*

JSTL – *Java Server Pages Standard Tag Library*

KDD – *Knowledge Discovery in Database*

MBR – *Memory Based Reasoning*

RF – Requisitos Funcionais

RNF – Requisitos Não-Funcionais

SDK – *Software Development Kit*

SGBD – Sistema Gerenciador de Banco de Dados

SQL – *Structured Query Language*

TCA– Teoria dos Conjuntos Aproximativos

UML – *Unified Modeling Language*

LISTA DE SÍMBOLOS

$\varnothing$ - conjunto vazio

$\subset$ - está contido

$\Rightarrow$ - implica que

$\cap$ - intersecção

$\forall$ - para todo ou qualquer que seja

$\in$ - pertence

$\cup$ - união

SUMÁRIO

1 INTRODUÇÃO

O avanço tecnológico dos últimos anos, tornou relativamente fácil o acúmulo de informações, seja pela redução de custos ou pela evolução da capacidade e desempenho dos meios de armazenamento de dados. Prass (2004) explica que, devido a este avanço tecnológico, as organizações têm se mostrado cada vez mais eficiente em capturar, organizar e armazenar grandes quantidades de dados, obtidos a partir de suas operações cotidianas como compras, vendas, cadastro de informações e movimentações. O problema reside no fato de as organizações ainda não conseguirem usar adequadamente essa gigantesca montanha de dados para transformá-la em conhecimentos úteis, que possam ser utilizados em suas próprias atividades, sejam elas comerciais ou científicas.

Com as mudanças que estão ocorrendo atualmente no mercado mundial, a incerteza é o fator dominante dos mercados financeiros. A tecnologia proliferante e a competição múltipla tornam-se rapidamente obsoletas. Neste cenário é perceptível que o sucesso de uma instituição está na sua habilidade de criar novos conhecimentos, disseminá-los rapidamente, e embuti-los em seus novos produtos e serviços.

Gimenes (2000) cita que, a quantidade de informações comerciais ou científicas armazenadas em bancos de dados das organizações, está ultrapassando a habilidade técnica e a capacidade humana na sua interpretação. Os bancos de dados alcançaram tais proporções que não se consegue extrair as informações importantes contidas nestes bancos, utilizando-se sistemas de gerenciamento de banco de dados convencionais. Bernardes (2001) acrescenta ainda, que estas informações adquiridas e captadas devem ser analisadas para produzir novos conhecimentos, os quais poderão proporcionar a produção de novos produtos e serviços, que irão facilitar a vida do homem. As técnicas de análise existentes atualmente para avaliação das informações são manuais e não produzem o efeito desejado. Tais fatos mostram a necessidade de produzir uma ferramenta que seja capaz de analisar automaticamente as bases de dados para obter conhecimento e gerenciá-lo para que possa auxiliar os administradores e analistas nos processos de tomada de decisão e julgamento. Gimenes (2000) explica que a necessidade de transformar estes dados em informações significativas é óbvia e técnicas computacionais foram e estão sendo desenvolvidas para analisar os dados e auxiliar a encontrar o conhecimento no caos das informações.

Fayyad et al (1996) explica que o *Data Mining* (DM) ou mineração de dados, como também pode ser definido, é o processo de reconhecimento de padrões válidos ou não,

existentes nos dados armazenados em grandes bancos de dados. Gimenes (2000) acrescenta que a mineração de dados consiste basicamente na aplicação de técnicas estatísticas, muitas vezes complexas, que precisam ser analisadas por pessoas especializadas. Prass (2004) explica que a importância deste processo se dá ao fato de buscar descobrir as informações escondidas nos dados armazenados.

Pawlak (1982) explica que a teoria dos conjuntos aproximativos foi desenvolvida por Zdzislaw Pawlak no começo da década de 80 para lidar com dados incertos e vagos em aplicações de inteligência artificial. Pessoa e Simões (2003, p. 3) citam que "[...] a TCA é uma extensão da teoria dos conjuntos, que enfoca o tratamento de incerteza dos dados através de uma relação de indiscernibilidade que diz que dois elementos são ditos indiscerníveis, se possuírem as mesmas propriedades [...]".

Bernardes (2001) descreve que o maior problema das organizações na atualidade é a concorrência provocada principalmente pela globalização do mercado mundial, pois hoje as empresas competem não mais com empresas da mesma região ou mesmo do próprio país; esta competição ocorre agora em âmbito mundial e a sobrevivência de uma empresa está associada à tecnologia da informação e a seu capital intelectual.

Através da aplicação da arquitetura proposta associada aos componentes da tecnologia da informação, pode-se recuperar e armazenar o conhecimento explícito em mídia digital de uma organização, com maior eficiência proporcionando assim, o gerenciamento do capital intelectual, maior competitividade, maior adaptação e maior integração da organização. O trabalho proposto facilita a rápida adaptação da organização frente às mudanças provocadas pela globalização das economias e o conseqüente acúmulo de dados justificando assim, o emprego de uma arquitetura de gerenciamento do conhecimento com o uso de *data mining* para proporcionar maior competitividade à organização.

Diante do exposto, o presente trabalho visa desenvolver uma aplicação para efetuar a classificação e segmentação de dados, através da mineração de dados utilizando a técnica da Teoria dos Conjuntos Aproximativos (TCA).

1.1 OBJETIVOS DO TRABALHO

O objetivo geral deste trabalho é o estudo e o desenvolvimento de uma aplicação em gestão do conhecimento utilizando *data mining* baseado na teoria dos conjuntos

aproximativos, gerando o capital intelectual para as organizações.

Os objetivos específicos do trabalho são:

a) demonstrar o potencial do DM para classificação e segmentação de dados baseado na TCA;
b) selecionar os perfis mais adequados dos profissionais cadastrados como capital intelectual;
c) demonstrar graficamente o resultado da mineração dos dados.

1.2 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em quatro capítulos que estão descritos nos parágrafos abaixo.

No primeiro capítulo é feita a contextualização e justificativa do desenvolvimento do trabalho.

No segundo capítulo é disponibilizada a fundamentação teórica necessária para um razoável conhecimento das tecnologias, componentes utilizados no desenvolvimento do trabalho e pesquisa dos trabalhos correlatos.

O terceiro capítulo tem como foco o desenvolvimento do sistema baseado nos conceitos da TCA descrevendo os requisitos principais do problema como também a especificação e a implementação do sistema

O quarto capítulo apresenta as conclusões finais e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

No presente capítulo são apresentados os aspectos teóricos relacionados ao trabalho. É evidenciada a importância da informação para as organizações. São apresentadas informações teóricas sobre as etapas de implementação de DM. São apresentados alguns fundamentos teóricos sobre a TCA. Em seguida são descritos alguns trabalhos correlatos.

2.1 GESTÃO DO CONHECIMENTO

Conhecimento é a informação apropriada e interpretada pelo ser humano, possibilitando-o a ter novas idéias. Atualmente é um assunto muito procurado e pesquisado pelas organizações. Alguns dos fatores que influenciam nesses processos de busca do conhecimento são a rápida evolução da tecnologia, o acesso aos mercados globais e como lidar e extrair dados e informações da inteligência competitiva nas organizações (DALFOVO, 2007, p. 19). O autor acrescenta que o principal objetivo na gestão do conhecimento pelas organizações é obter alguma vantagem competitiva sobre seus concorrentes inovando seus produtos, serviços e processos, agregando novos valores, otimizando novos processos de produção, criando novas invenções, melhorando a qualidade de vida dos seres humanos e também, auxiliando na tarefa de preservação do meio ambiente.

"O conhecimento é mais profundo, rico e mais expansivo que os dados e a informação. O conhecimento é aplicado aos fatos ou idéias adquiridas por estudos, investigação, observação e experiência. [...]" (BERNARDES, 2001 p. 24). O conhecimento originalmente é aplicado na mente do conhecedor, mas em uma organização ele é freqüentemente encontrado não somente em documentos ou repositórios, mas também nos processos, em rotinas organizacionais, nas práticas e nas normas das organizações. É intangível e impalpável, porém sua existência é tão poderosa que transforma pessoas, organizações e países.

Bernardes (2001) acrescenta que o conhecimento sempre será um diferencial competitivo de extremo poder, sendo considerado mais precioso do que os recursos naturais, industriais ou até mesmo mais valioso do que o próprio dinheiro. Porém, é necessário atentar-se para a importância não só da aquisição, mas também a necessidade da criação e transferência, lembrando que conhecimento sem ação é nulo, ou nem pode ser considerado

como tal. Não é tarefa das mais fáceis definir o conhecimento, tendo em vista sua característica de intangibilidade. Uma das maiores barreiras para o sucesso do conhecimento é a distinção entre conhecimento e informação. Informação na verdade, consiste em dados organizados, agrupados e categorizados em modelos para criar significado, o conhecimento torna-se a informação posta para uso produtivo, capaz de habilitar as ações corretas.

Dalfovo (2007) explica que a gestão do conhecimento surge como uma metodologia de gerenciamento que vai além do simples processo de inovação, onde é necessário contemplar mercados e tendências nos processos de desenvolvimento tecnológicos, além de outros fatores que determinem a vantagem competitiva de uma organização.

O conhecimento tornou-se um dos recursos fundamentais para as organizações, desde o momento em que ocorreu a troca de uma economia industrial baseada em linha de montagem e controle hierárquico para uma economia global, descentralizada. Por isto a organização deve ser capaz de manter e reproduzir seu conhecimento essencial independente da distância geográfica, das diferenças culturais e diferenças de línguas que enfrentará com o processo da globalização. Além disto, deve ser capaz de enriquecer, abusando de sua criatividade, os seus produtos e serviços com conhecimento vindo de locais diferentes que participem da sua mão de obra global; deve também, ser capaz de manter-se acima de seus concorrentes, mesmo com o rápido ritmo de competição mundial.

Para melhor se adaptarem a este cenário, as organizações devem monitorar e armazenar as informações de seus clientes, fornecedores, empregados e concorrentes e, destas informações, extraírem conhecimentos que as tornem mais competitivas e mais moldadas às mudanças exigidas pelo mercado (BERNARDES, 2001, p. 29).

Nonaka e Takeuchi (1997) classificam o conhecimento em dois tipos:

a) tácito: é um tipo de conhecimento muito difícil de ser expresso por meio de palavras e é adquirido com a experiência, de maneira prática utilizando-se da intuição e da subjetividade. Segundo os autores, o aprendizado mais poderoso vem da experiência direta e é aquele que se articula por meio da linguagem formal, com afirmações gramaticais, expressões matemáticas, especificações, manuais, etc; envolve fatores intangíveis como, por exemplo, crenças pessoais, perspectivas, sistema de valor, intuições, emoções e habilidades individuais. Este processo pode ter inclusive sua interatividade prejudicada devido aos ruídos que na maioria das vezes existem em qualquer processo de comunicação;
b) explícito: é um tipo de conhecimento que pode ser facilmente expresso em palavras ou números e pode ser prontamente transmitido formalmente e

> sistematicamente entre pessoas. Envolve o conhecimento de fatos. É objetivo, teórico, digital e articula por meio da linguagem formal, com afirmações gramaticais, expressões matemáticas, especificações, manuais, etc. Esse foi o modo dominante de conhecimento na tradição filosófica ocidental.

Bernardes (2001) explica que, os principais objetivos do gerenciamento do conhecimento organizacional são, melhorar a produtividade e o conhecimento dos colaboradores a fim de proporcionar meios para a rápida construção e utilização da coleção do conhecimento da organização. Para isto, as instituições devem criar redes de informações que propiciam a geração de conhecimento, seja uma rede de informação formal ou informal. O autor acrescenta ainda que o conhecimento é bastante diferente da informação e o seu gerenciamento é decisivo e qualitativamente diferente do gerenciamento da informação.

O gerenciamento do conhecimento dá uma acentuada importância às pessoas, seus trabalhos práticos, culturais e também decide como e quais tecnologias deverão ser empregadas neste cenário. Por outro lado, ele freqüentemente visualiza uma primeira solução tecnológica e coloca em segundo plano as considerações pessoais dos trabalhadores e os trabalhos culturais. Este fato pode ser a origem dos baixos resultados alcançados pela tecnologia da informação nas organizações, que em sua maioria, a tem como um pequeno pedaço a solução do problema. “Diante deste cenário, a tecnologia da informação pode auxiliar o gerenciamento do conhecimento organizacional através dos seus componentes tecnológicos, cujo conjunto produz uma arquitetura de rede que consiste num tipo de tecnologia que se encaixa em uma estrutura geral para suportar o gerenciamento do conhecimento [...]” (BERNARDES, 2001, p. 33).

2.2 A IMPORTÂNCIA DA INFORMAÇÃO PARA AS ORGANIZAÇÕES

Conforme Gonçalves e Gonçalves (2001) nas décadas de 50 e 80 houve um aumento significativo da turbulência, que se acentuou na década de 90, onde restrições governamentais, insatisfação dos consumidores, invasão de concorrentes estrangeiros, aceleração do desenvolvimento tecnológico, novas relações no trabalho, pressões ecológicas e sociais foram elementos novos que a cada dia aumentavam a complexidade de gestão das organizações.

Diante do cenário atual existente, pode-se observar nos dias de hoje que estas

mudanças rápidas e alta competitividade entre as empresas levaram algumas organizações a crescerem demasiadamente e dominarem o mercado, enquanto outras, até mesmo antigas e tradicionais empresas, sendo vendidas ou simplesmente encerrando suas atividades. Gonçalves e Gonçalves (2001, p. 48) destacam que "[...] o principal elemento gerador da longevidade destas companhias estava ligado à sua sensibilidade para o ambiente, para aprender e se adaptar de forma mais rápida que os concorrentes".

Todas as empresas saudáveis geram e usam conhecimento. Ao interagirem com seu ambiente, as organizações absorvem a informação, a transformam em conhecimento e tomam decisões. "O maior valor da tecnologia na gerência do conhecimento é o de estender o alcance e aumentar a velocidade da transferência de conhecimento. [...]" (GONÇALVES e GONÇALVES, 2001, p. 56).

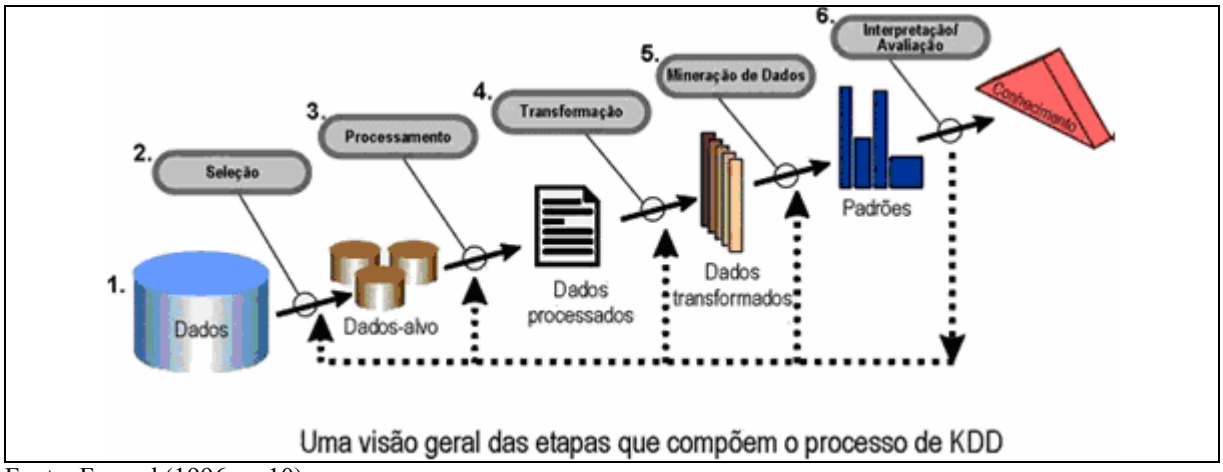
Madeira (2003) cita que para adquirir este conhecimento, as organizações investiram muito em tecnologia e armazenamento de dados a fim de conhecer melhor os seus clientes e prever comportamentos futuros com base no passado. Com isso, criaram-se imensas bases de dados para armazenar todos os dados que pareciam ser úteis. Romão, Pacheco e Niederauer (2003, p. 2) acrescentam que "a partir dos dados é possível extrair um tipo de informação mais estratégica, o conhecimento, normalmente mais resumido e em menor quantidade, mas de importância vital para se tomar decisões. [...]".

Gimenes (2000) destaca que a quantidade de informação armazenada em bancos de dados destas organizações ultrapassa a habilidade técnica e a capacidade humana na sua interpretação. Romão, Pacheco e Niederauer (2003) explicam que a aplicação de algoritmos específicos deve garantir que o tipo e forma do conhecimento obtido estejam adequados ao processo de tomada de decisões rápidas e inteligentes. O desafio está em adaptar estas técnicas tradicionais para serem viáveis diante de banco de dados, normalmente não projetados para facilitar a aplicação destas técnicas. A afirmação do autor ajuda a compreender melhor a importância e planejamento da construção da aplicação de DM.

2.3 DESCOBERTA DE CONHECIMENTOS EM BANCOS DE DADOS - DCBD

Abordar técnicas e ferramentas que buscam transformar os dados armazenados pelas organizações em conhecimento é o objetivo da área denominada *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimentos em bases de dados. O KDD é um

processo que permite que os resultados sejam alcançados e melhorados ao longo do tempo. As etapas que compõe o processo do KDD são apresentados na Figura 1.



Fonte: Fayyad (1996, p. 10).

Figura 1 – Etapas do processo KDD

Fayyad (2002 apud Almeida et al, 2004, p. 2) afirma que "[...] o processo de extração de conhecimento de dados é constituído por um conjunto de etapas cuja finalidade é obter um conhecimento específico a respeito de um determinado domínio". Fayyad (1996) explica que o processo de KDD é um conjunto de atividades contínuas que compartilham o conhecimento descoberto a partir de bases de dados. Esse conjunto é composto de cinco etapas que são resumidamente abordadas a seguir:

a) seleção dos dados: Quoniam et al. (2001) afirma que a etapa de seleção consiste em identificar e selecionar todas as fontes externas e internas de informação e selecionar o subconjunto de dados ou variáveis necessários para o processo do KDD. Jesus (2004) explica que os dados representam a fonte para a descoberta do conhecimento e que estão armazenados nos bancos de dados das organizações ainda não explorados e provenientes dos sistemas legados que através da aplicação do processo de KDD resultarão em conhecimento;

b) pré-processamento e limpeza dos dados: Almeida et al. (2004) cita que nesta etapa deverão ser realizadas tarefas que eliminem ou tratem as referências dos dados estranhos ou inconsistentes através de algoritmos específicos, a fim de balancear a base de dados, evitando que o sistema fique tendencioso. As causas que podem levar à situação de ausência de dados são, por exemplo, a não disponibilidade ou inexistência da referência do mesmo na base de dados referenciada no processo do KDD. Quoniam et al. (2001) acrescenta que esta etapa exige maior esforço, correspondendo a 60% do trabalho de DM, entre ferramentas de visualização e de reformatação dos dados;

c) transformação dos dados: Almeida et al. (2004) explica que nesta etapa é onde os dados necessitam ser armazenados e formatados adequadamente em bases e ou tabelas de dados específicas, Jesus (2004) acrescenta que os algoritmos de mineração normalmente não podem acessar os dados em seu formato nativo, seja em razão da forma como são armazenados ou pela normalização adotada na modelagem do banco. Por isto é necessária a conversão dos dados para um formato apropriado a fim de possibilitar que os algoritmos de aprendizado de DM possam ser aplicados com maior eficiência;
d) mineração de dados: Gimenes (2000) afirma que nesta etapa as ferramentas especializadas procuram, através de algoritmos especializados, os padrões existentes nos dados. Essa busca pode ser efetuada automaticamente pelo sistema ou interativamente através do auxílio do analista responsável pela geração das hipóteses. O autor acrescenta ainda que diversas ferramentas distintas, como redes neurais, árvores de decisão, sistemas baseados em regras e programas estatísticos, podem ser aplicadas isoladamente ou em combinação ao problema proposto. Ao final do processo, o sistema de DM deve gerar um relatório da análise efetuada a fim de possibilitar aos analistas verificarem os resultados obtidos;
e) interpretação/avaliação: Almeida et al. (2004) afirma que esta etapa deve ser realizada em conjunto com os analistas responsáveis. Caso o conhecimento gerado pela mineração não seja satisfatório, os analistas podem formar um novo conjunto de questões e realimentar o sistema com novos parâmetros para realizar uma nova busca pelo conhecimento desejado.

2.4 TAREFAS DESEMPENHADAS PELO DATA MINING

Jesus (2004) afirma que para possibilitar que o DM efetue a correta análise de dados na busca de conhecimento em um conjunto de dados, é extremamente importante a escolha apropriada da ferramenta ou conjunto de ferramentas de DM que serão implantadas. Assim é importante que se conheçam as tarefas desempenhadas e suas técnicas, a fim de dar suporte a sua escolha. Cada uma destas ferramentas difere quanto à classe de problemas que o algoritmo será capaz de resolver. Esse conjunto de ferramentas é resumidamente abordado a seguir:

a) classificação: Matos (2004) define que a tarefa de classificação consiste basicamente em construir um modelo que possa ser aplicado a dados não classificados visando categorizá-los em classes, onde cada registro é examinado e classificado de acordo com uma classe predefinida do processo. Jesus (2004) acrescenta que as técnicas de redes neurais artificiais, árvores de decisão, análise discriminante e regressão logística são apropriadas para a tarefa de classificação;
b) estimação: enquanto a classificação lida com resultados discretos, a estimação envolve a geração de valores ao longo das dimensões dos dados, utilizando-se dos algoritmos de estimativa. Compolt (1999) explica que fornecendo-se alguns dados, usa-se a estimativa para estipular um valor para alguma variável contínua desconhecida como receita, altura ou saldo de cartão de crédito. Ao invés de um classificador binário determinar um risco como sendo positivo ou negativo, por exemplo, a técnica gera valores aproximados dentro de uma determinada margem, obtendo-se assim a vantagem de os registros individuais poderem ser ordenados por classificação. Jesus (2004) acrescenta que as técnicas de redes neurais artificiais, algoritmos genéticos e técnicas de estatística são adequadas para estimação;
c) previsão: a previsão consiste na determinação do futuro de uma grandeza. Jesus (2004) cita que esta técnica assemelha-se a classificação e estimação, exceto pelo fato de que os registros serem classificados de acordo com alguma atividade futura prevista ou valor futuro estimado. O autor explica ainda que qualquer uma das técnicas usadas na classificação ou estimação podem ser adaptadas para o uso na previsão, onde os dados históricos são usados para construir um modelo que explica o comportamento atual observado. Quando este modelo é aplicado às entradas atuais, o resultado é a previsão de atitudes futuras. As regras de associação, raciocínio baseado em casos, redes neurais artificiais, árvores de decisão e séries temporais são adequadas. A escolha da técnica dependerá basicamente da natureza dos dados;
d) agrupamento por afinidade: também conhecida por análise de *clusters*[1] ou somente agrupamento, esta tarefa consiste em identificar possíveis agrupamentos nos dados. Um agrupamento deve ser entendido como sendo uma coleção de objetos de dados

[1] *Cluster* pode ser definido como sendo um conjunto de dados aglomerados com valores aproximados através do DM.

que são semelhantes um ao outro. Diferentes medidas de similaridade, baseadas em funções de distância podem ser especificadas para diferentes contextos de aplicação. Um bom método de *cluster* assegura que a similaridade *inter-cluster* é baixa e a similaridade *intra-cluster* seja alta. Jesus (2004) cita que a técnica mais utilizada neste caso é a análise de seleção estatística;

e) segmentação: é o processo de agrupamento de uma população heterogênea em vários subgrupos ou *clusters* mais homogêneos. Nardelli (2000) explica que o que distingue a segmentação da classificação é que na primeira existem as classes pré-definidas. A segmentação é realizada automaticamente por algoritmos que identificam características em comum e particionam o espaço definido pelos atributos. Jesus (2004) acrescenta que a segmentação é normalmente uma técnica preliminar utilizada quando nada ou pouco se sabe sobre os dados, como na metodologia da descoberta não supervisionada de relações. As técnicas utilizadas para segmentação são as redes neurais artificiais, estatísticas e algoritmos genéticos;

2.5 TÉCNICAS DE DATA-MINING

Jesus (2004) explica que existem hoje, várias técnicas conhecidas de DM, no entanto, antes de determinar qual técnica a ser usada na aplicação deve-se efetuar uma análise criteriosa dos objetivos, buscando conhecer primeiramente qual tipo de banco de dados deseja-se trabalhar ou qual tem-se para aplicar as técnicas de mineração de dados, por exemplo, se o banco de dados é do tipo relacional, que é o tipo mais conhecido e utilizado nos Sistemas de Gerenciamento de Banco de Dados (SGBD) atuais, ou se é do tipo orientado a objeto ou um banco de informações da web. Outro fator importante é saber qual o tipo de conhecimento pretende-se explorar, se é conhecimento baseado em regras de classificação, ou agrupamento ou regras de associação. Linoff (1997 apud Jesus 2004) afirma que nenhuma técnica consegue resolver todos os problemas de mineração de dados. O autor acrescenta que a familiaridade com uma variedade de técnicas é necessária para encontrar o melhor caminho para resolver um determinado problema. A seguir são abordadas algumas técnicas de DM.

2.5.1 Memory Based Reasoning (MBR)

MBR significa raciocínio baseado em casos e é uma técnica de DM dirigida, que utiliza exemplos conhecidos como modelo para efetuar previsões sobre exemplos ainda desconhecidos. O MBR é uma técnica que procura os vizinhos mais próximos dos exemplos conhecidos e combinam seus valores para atribuir valores de classificação ou de previsão. Harrison (1998 apud Jesus, 2004) explica que uma das maiores vantagens do MBR é a habilidade de ser executado em qualquer fonte de dados, mesmo sem modificações. Os dois elementos-chave no MBR são a função de distância, usada para encontrar os vizinhos mais próximos, e a função de combinação, que combina valores dos vizinhos para fazer uma previsão. O autor explica ainda que outra vantagem do MBR é a habilidade de aprender sobre novas classificações, simplesmente introduzindo novos exemplos no banco de dados durante o processo. Uma vez encontradas a função de distância e a função de combinação corretas, estas informações tendem a permanecer muito estáveis, mesmo com a incorporação de novos exemplos para novas categorias nos dados conhecidos. Esta facilidade de incorporar mudanças ao domínio à extensão faz com que o MBR destaque-se da maioria das outras técnicas de DM.

2.5.2 Redes neurais artificiais

As redes neurais ou redes neuronais como também são conhecidas, é uma classe especial de sistemas modelados seguindo a analogia com o funcionamento do cérebro humano. “Assim como no cérebro, uma rede neural é uma coleção massivamente paralela de unidades de processamento pequenas e simples, onde as interligações formam a maior parte da inteligência da rede e são formadas por neurônios artificiais ligados de maneira similar aos neurônios do cérebro humano. [...]” (MATOS, 2004, p. 6). A intensidade das interligações da rede neural pode alterar como resposta a um estímulo, o que permite a rede aprender. Na sua forma mais comum, as redes neurais conseguem aprender com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão, por isto, as redes neurais podem também ser aplicadas às situações de DM não-dirigido e às previsões em séries temporais. O autor cita ainda que uma das principais vantagens das redes neurais é a sua variedade de aplicações. “[...] As redes neurais são interessantes porque detectam padrões nos

dados de forma analógica ao pensamento humano tornando-a um fundamento interessante para uma ferramenta de *data mining*." (BARTOLOMEU, 2002, p. 84).

2.5.3 Árvores de decisão e indução de regras

Conforme Bartolomeu (2002) as árvores de decisão são normalmente usadas em aplicações de DM na classificação de dados. Este técnica consiste em dividir os registros do conjunto de dados de treinamento em subconjuntos separados, cada um descrito por uma regra simples em um ou mais campos. Primeiramente a técnica determina que deva ser escolhida a variável a ser avaliada e em seguida efetua a procura da variável mais correlacionada dentro do sistema, através da fórmula matemática denominada entropia[2], montando a árvore de busca com várias ramificações.

Gimenes (2000) acrescenta que a partir da escolha da variável a ser analisada, dado um grupo de dados com numerosas colunas e linhas, uma ferramenta de árvore de decisão pode solicitar ao usuário que escolha uma destas colunas como objeto de saída, exibindo assim, o fator mais importante correlacionado com aquele objeto de saída que será um nó da árvore de decisão. Isso significa que o usuário pode rapidamente ver qual o fator que mais direciona o seu objeto de saída, e pode entender porque o fator foi o escolhido. O autor acrescenta que uma boa ferramenta de árvore de decisão deve também permitir que o usuário explore a árvore de acordo com a sua vontade, do mesmo modo que ele poderá encontrar grupos que são de seu maior interesse. Os usuários devem também poder selecionar os dados fundamentais em qualquer nó da árvore, movendo-o para uma planilha ou outra ferramenta para análise posterior, possibilitando saber os itens que mais influenciam uma determinada variável.

2.5.4 Algoritmos genéticos

Os algoritmos genéticos aplicam mecanismos de seleção de busca, que são usados para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão. "Os

[2] Entropia remete ao fato de as equações matemáticas usarem métodos probabilísticos para serem deduzidas, sendo assim quanto maior o número de arranjos possíveis, maior será a entropia.

algoritmos genéticos são semelhantes às técnicas de estatística, pois também precisam conhecer o modelo em profundidade. [...]" (Jesus, 2004, p. 37). Compolt (1999) cita que esta técnica, além de ser apropriada para resolver os tipos de DM, pode ser usada para aprimorar MRB's e redes neurais.

2.6 A TEORIA DOS CONJUNTOS APROXIMATIVOS

Ramos, Machado e Costa (2003, p. 2) destacam que "[...] um dos grandes problemas enfrentados pelos usuários de um sistema de informação é saber o grau de coerência ou de qualidade das informações extraídas do mesmo, ou seja, a exatidão do sistema de informação". Os autores acrescentam ainda que os sistemas de avaliação e análise devem ser dotados de artifícios que permitam o tratamento da subjetividade inerente ao problema.

Pessoa e Simões (2003) afirmam que a TCA baseia-se na noção de conjunto aproximativo, que acontece quando subconjuntos de um conjunto universo têm o mesmo valor do atributo de resultado. Porém, pode acontecer que um conceito não seja definido claramente devido aos elementos serem indiscerníveis e terem valores de decisões contraditórias. Os autores acrescentam que os elementos destes subconjuntos são divididos em os que podem certamente ser classificados em pertencentes a uma desejada classe, os que não podem ser classificados e os que não pertencem a classe desejada. Se existem elementos que não podem ser classificados, o conjunto é dito aproximativo.

Pessoa e Simões (2003) explicam que a relação de indiscernibilidade pode ser mais bem compreendida em um sistema de informação, que pode estar no formato de uma tabela, por exemplo, e que pode tornar-se desnecessariamente grande, quando elementos iguais são representados muitas vezes ou quando alguns atributos são desnecessários. Por isto, a TCA trata estes problemas a partir de uma relação de equivalência de modo que apenas um objeto represente toda uma determinada classe. A relação de indiscernibilidade constitui a base matemática da TCA e pode ser entendida como binária, à medida que dois objetos possuem a mesma descrição, porém com atributos diferentes. A partir disto, pode-se afirmar que o que a TCA busca é encontrar todos os objetos que produzem um mesmo tipo de informação, ou seja, que são indiscerníveis (GOMES e GOMES, 2004, p. 91).

Nunes (2005) cita que a TCA pode hoje ser classificada como uma técnica poderosa, convergindo com áreas de grande interesse no campo das ciências cognitivas e da inteligência

artificial. O uso da TCA em DM possibilita uma extração de dados mais eficiente e precisa na mineração do conhecimento. Politi (2006, p. 236) acrescenta que "[...] Essa teoria tem se mostrado como uma base teórica para a solução de muitos problemas com mineração de dados, principalmente no que diz respeito à redução de dados".

No item 2.6.1 é apresentado o conceito de TCA, baseado e transcrito a partir de Pawlak (1982) com tradução do autor deste trabalho.

2.6.1 Conceitos básicos da teoria dos conjuntos aproximativos

Uma tabela de informação é uma tabela de dados estruturada de forma que as linhas representam os objetos *i*, e as colunas representam os atributos *j*. Nas entradas da tabela, devem ser colocados os valores correspondentes de V_{ij} conforme especificado na Tabela 1.

	Atributo 1	Atributo 2	...	Atributo j
Objeto 1	V_{11}	V_{12}	...	V_{1j}
Objeto 2	V_{21}	V_{22}	...	V_{2j}
...	...	...	...	...
Objeto i	V_{i1}	V_{i2}	...	V_{ij}

Fonte: Adaptado de Pawlak (1982).

Tabela 1 – Estrutura de uma tabela de informação

Considerando-se os conjuntos:

a) U = {objeto1,objeto2,...,objeto j} onde U é um conjunto universo finito de objetos;

b) Q = {atributo1,atributo2,...,atributo i} onde Q é um conjunto finito de atributos.

Cada atributo $q \in Q$ está associado a um conjunto de possíveis valores que qualquer objeto possa tomar, chamado de domínio do atributo e é denotado por V_q e demonstrado no Quadro 1.

$$V = \bigcup_{q \in Q} Vq$$

Fonte: Pawlak (1982).

Quadro 1 – Domínio do atributo

E ainda uma função de informação conforme descrito no Quadro 2.

$$f: UxQ \rightarrow V \text{ tal que } f(x,q) \in V_q$$

Fonte: Pawlak (1982).

Quadro 2 – Função de associação do valor ao domínio

Esta função associa cada par de objeto x e atributo q ao valor correspondente V_{xq} de seu domínio V_q. A estrutura $S=\langle U,Q,V,F\rangle$ é definida como sendo um sistema de informação.

É importante na TCA verificar se um subconjunto $P \subset Q$ de atributos de condição fornece conhecimento adequado a determinados propósitos, como diagnóstico baseado nos valores assumidos por um determinado atributo de decisão. Isto pode ser utilizado para fins de classificação, por exemplo, onde dado um sistema de informação S e $P \subset Q$, pode-se afirmar que dois objetos $x, y \in U$ são indiscerníveis para o conjunto de atributos P se, e somente se, $f(x,q) = f(y,q)$ para todo $q \in P$. Ou seja, x e y são indiscerníveis em P, se apresentam os mesmos valores para todos os atributos em P. A relação de indiscernibilidade I_P em U é definida pela condição $(x,y) \in I_P$ se x, y são indiscerníveis para o conjunto P de atributos. A fórmula para obtenção de I_P pode ser observada no Quadro 3.

$$I_P = \{(x,y) \in UxU \mid f(x,q), \forall q \in P\}$$

Fonte: Pawlak (1982).

Quadro 3 – Função de indiscernibilidade

A relação de indiscernibilidade I_P é uma relação de equivalência reflexiva, simétrica e transitiva. Portanto, efetua uma partição de U em classes de equivalência, onde cada uma das quais é um subconjunto dos elementos de U que são indiscerníveis entre si, fazendo com que cada uma destas classes seja um conjunto P-elementar em S. A família de todas estas classes é denotada por U / I_P.

$Des_P(X)$ denota a descrição do conjunto P-elementar $X \in U / I_P$ em termos dos pares (atributo, valor) conforme mostrado no Quadro 4.

$$Des_P(X) = \{(q,v) \mid f(x,q) = v, \forall x \in X, \forall q \in P\}$$

Fonte: Pawlak (1982).

Quadro 4 – Descrição do conjunto P elementar

Seja $P \subset Q$ e $Y \subset U$. Então a aproximação P-inferior de Y denotada por P_Y, a aproximação P-superior de Y denotada por P^Y e o conjunto P-fronteira de Y, denotado por $Fr_P(Y)$, são definidas conforme demonstrado no Quadro 5.

$$P_Y = \bigcup \{X \in U/I_P \mid X \subset Y\}$$

$$P^Y = \bigcup \{X \in U/I_P \mid X \cap Y \neq \varnothing\}$$

$$Fr_P(Y) = P^Y - P_Y$$

Fonte: Pawlak (1982).

Quadro 5 – Definição da aproximação de Y

O conjunto P_Y é formado por todos os elementos que certamente podem ser classificados como elementos de Y, descernindo-os mediante o conjunto de atributos P. Já P^Y é o conjunto de elementos de U que possam possivelmente, ser classificados como elementos de Y. O conjunto P-fronteira $Fr_P(Y)$ é o conjunto de elementos que podem possivelmente, mas não certamente, serem classificados como elementos de Y. Evidentemente, $P_Y \subset P^Y$ e $P_Y = P^Y$ se e somente se $Fr_P(Y) = \varnothing$.

A cada $Y \subset U$ associa-se uma precisão de aproximação do conjunto Y por P em S, definida conforme descrito no Quadro 6, onde card denota a cardinalidade do conjunto, satisfazendo $0 \leq \alpha p(Y) \leq 1$.

$$\alpha p(Y) = \frac{\text{card}\,(P_Y)}{\text{card}\,(P^Y)}$$

Fonte: Pawlak (1982).
Quadro 6 – Definição da precisão de aproximação

Seja S um sistema de informação $P \subset Q$ e seja $Y = \{Y_1, Y_2, ..., Y_n\}$ uma partição de U. O coeficiente demonstrado no Quadro 7 é chamado qualidade da aproximação da partição Y pelo conjunto de atributos P, definida como sendo a qualidade da classificação. Este coeficiente deve satisfazer a relação $0 \leq \gamma p(Y) \leq 1$.

$$\gamma_P(Y) = \frac{\sum_{i=1}^{n} \text{card}(P_{Yi})}{\text{card}(U)}$$

Fonte: Pawlak (1982).
Quadro 7 – Definição da qualidade de classificação

Sejam R e $P \subset Q$ dois conjuntos de atributos em um sistema de informação S. Diz-se que R depende de P e denota-se por $P \rightarrow R$ se $I_P \subset I_R$. Descobrir dependências entre os atributos é importância primordial em TCA para a análise do conhecimento.

Outro ponto importante é a redução de atributos, de tal modo que um conjunto reduzido de atributos forneça a mesma qualidade de classificação em relação a um conjunto original de atributos; logo, dado algum $P \subset Q$, o mínimo subconjunto $R \subset P$ tal que $\gamma_R(Y) = \gamma_P(Y)$ é chamado de Y-redução de P e é denotado por $RED\gamma(P)$. Um sistema de informação pode ter mais de uma Y-redução. A intersecção de todas as Y-reduções é chamada de Y-núcleo de P ou $CORE\gamma(P)$. Logo $CORE\gamma(P) = \cap RED\gamma(P)$ representa o núcleo da coleção dos atributos mais significativos no sistema.

Uma tabela de decisão é um sistema de informação formado por atributos de condição

em C e por atributos de decisão em D tal que Q seja determinado pela condição descrita no Quadro 8.

$$Q = C \cup D, C \cap D = \varnothing$$

Fonte: Pawlak (1982).
Quadro 8 – Definição do conjunto Q

Uma tabela de decisão é determinística se $C \rightarrow D$ caso contrário é não–determinística. Uma tabela de decisão determinística descreve unicamente as decisões a serem efetuadas quando algumas condições são satisfeitas. No caso de uma tabela de decisão não-determinística, as decisões não são univocamente determinadas pelas condições.

Pode-se derivar um conjunto de regras de decisão a partir de uma tabela de decisão. Seja $U / I_C = \{X_1, X_2,..., X_K\}$ a família de todas as classes de condição e $U / I_P = \{Y_1, Y_2, ..., Y_n\}$ a família de todas as classes de decisão. Então $Desc_C(X_i) \Rightarrow Desc_D(Y_j)$ é chamada uma regra de decisão (C, D). As regras de decisão também podem ser expressas em declarações lógicas tipo 'se ... então...', relacionando classes de condição e de decisão. O conjunto de regras de decisão para cada classe de decisão Y_j (j = 1,..., n) é denotado por $\{r_{ij}\}$, precisamente descrito conforme Quadro 9.

$$\{r_{ij}\} = \{Desc_C(X_i) \Rightarrow Desc_D(Y_j) \mid X_j \cap Y_j \neq \varnothing, i = 1,..., k\}$$

Fonte: Pawlak (1982).
Quadro 9 – Conjunto de regras de decisão para cada classe de decisão Y_j

Uma regra r_{ij} é determinística se $X_i \subset Y_j$; caso contrário é não-determinística. Regras não-determinísticas são conseqüências de uma descrição aproximada de classes de decisão ou categorias em termos de classes de condição que são na verdade os blocos de objetos indiscerníveis por atributos de condição. Isto significa que usando o conhecimento disponível, não se pode decidir se alguns objetos da região fronteira pertencem ou não a uma determinada categoria.

2.7 TRABALHOS CORRELATOS

A seguir, são apresentados três trabalhos desenvolvidos, entre os quais, de Alberto Pereira de Jesus (JESUS, 2004) e os trabalhos de Geandro Compolt (COMPOLT, 1999) e Biancca Nardelli (NARDELLI, 2000).

O objetivo principal no trabalho de Geandro Compolt (COMPOLT, 1999) foi gerar um modelo de classificação de dados utilizando técnicas de DM, mais especificamente árvores de decisão. Foi implementado um protótipo que permitia ao usuário definir um valor-prioridade para cada atributo que faria parte do modelo de classificação. Foram analisadas as características de sistemas de informação, técnicas de DM e montada uma base de dados fictícia, com informações de condições que conduziam à concessão de crédito a fornecedores.

No trabalho de Biancca Nardelli (NARDELLI, 2000) o objetivo principal foi gerar um modelo de classificação de dados utilizando técnicas de DM, mais especificamente árvores de decisão. Para esta tarefa foi implementado um protótipo que permitia ao usuário definir um valor-prioridade para cada atributo que fazia parte do modelo de classificação. Para a elaboração do protótipo, foram analisadas as características de sistemas de informação e técnicas de DM, e montada uma base de dados fornecida pela Central de Informação dos Alunos da FURB, que foi aplicado à classificação.

No trabalho de Alberto Pereira de Jesus (JESUS, 2004), o objetivo foi desenvolver um sistema de recuperação e disseminação de informações, personalizado segundo o perfil dos usuários da Biblioteca Central da FURB, por meio de técnicas de DM, através da seleção das variáveis de interesse, gerando uma nova base de dados, e utilizando técnicas de discretização e redução de dimensionalidade, a fim de agregar valor aos serviços de referência, fazendo com que os sistemas de busca sejam personalizados dinamicamente segundo o perfil do usuário.

3 DESENVOLVIMENTO

O sistema desenvolvido neste trabalho realiza a tarefa de mineração de dados, baseado na TCA. Para tanto, estudou-se o funcionamento matemático da TCA a fim de gerar o capital intelectual a partir das informações de conhecimentos informadas pelos usuários de uma organização. Este capítulo aborda a realização e análise dos requisitos que definem as características do sistema proposto.

A seguir são descritas as suas especificações.

3.1 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

Os requisitos, classificados como Requisitos Funcionais (RF) e Requisitos Não Funcionais (RNF) descrevem o que o sistema deve e o que não deve fazer. Os RF demonstram às funcionalidades e o comportamento que o sistema deve possuir em determinadas situações. Os RNF demonstram as restrições que o sistema terá sobre alguns serviços ou funções oferecidas como usabilidade, navegabilidade, portabilidade, segurança e hardware.

No Quadro 10 são apresentados os Requisitos Não Funcionais e, em seguida, no Quadro 11 são apresentados os Requisitos Funcionais que o software desenvolvido no presente trabalho deve contemplar.

Requisitos Não Funcionais
RNF01. O sistema deve apresentar os resultados obtidos na análise da TCA na mesma tela do cadastro dos valores de domínio dos atributos.
RNF02: Os eventos analisados pelo sistema devem ser visualizados em no máximo 30 (trinta) segundos em condições normais de rede.
RNF03: O sistema deve possuir um controle de acesso somente em nível de usuário.
RNF04: O sistema deve funcionar em ambiente web.
RNF05: O sistema deve ser compatível com os navegador Microsoft Internet Explorer e Mozilla Firefox.
RNF06: O sistema deve utilizar banco de dados MySQL.
RNF07: O sistema deve ser desenvolvido na linguagem JSP utilizando as bibliotecas JSTL.
RNF08: Os requisitos mínimos de hardware devem ser: processador Intel Pentium 4 2 GHz, ou similar, 512 MB de memória RAM, 100 MB de espaço em disco.
RNF09: O sistema deve disponibilizar texto de ajuda no corpo das páginas acessadas.

Quadro 10 – Requisitos não-funcionais

Requisitos Funcionais
RF01: O sistema deve permitir ao analista visualizar e selecionar o nome das tabelas do banco de dados que devem ser utilizadas para o processo de análise. UC01
RF02: O sistema deve permitir ao analista informar qual o valor de cálculo dos níveis de conhecimento que serão utilizados na seleção de usuários. UC02
RF03: O sistema deve permitir ao analista selecionar quais os atributos devem ser analisados pela TCA. UC03
RF04: O sistema deve permitir ao analista visualizar e selecionar os valores de domínio correspondentes a cada atributo selecionado. UC04
RF05: O sistema deve permitir ao analista adicionar ou remover os atributos durante ou após a entrada de valores de domínio dos atributos. UC05
RF06: O sistema deve permitir ao analista adicionar ou remover as linhas da tabela durante ou após a entrada dos valores de domínio dos atributos. UC06
RF07: O sistema deve verificar ao iniciar, se todas as tabelas de referência estão informadas corretamente. UC07
RF08: O sistema deve emitir uma mensagem de erro caso haja algum erro nas tabelas de referência informadas. UC08
RF09: O sistema deve verificar se todos valores de domínios dos atributos e se todos os valores de decisão foram informados. UC09
RF10: O sistema deve emitir uma mensagem de erro caso não sejam informados os valores de domínios dos atributos ou valores de decisão. UC10
RF11: O sistema deve emitir uma mensagem de erro sempre que houver algum erro na comunicação com o banco de dados. UC11
RF12: O sistema deve calcular os valores dos níveis de conhecimento dos usuários baseado nos atributos selecionados pelo analista. UC12
RF13: O sistema deve calcular e informar ao analista qual o atributo mais significativo obtido através da TCA. UC13
RF14: O sistema deve emitir ao analista o resultado da análise da TCA após salvar os valores informados na tabela. UC14
RF15: O sistema deve gerar um arquivo de log com os resultados obtidos no processo da análise da TCAUC 15

Quadro 11 – Requisitos funcionais

3.2 ESPECIFICAÇÃO

Neste item são apresentadas as especificações dos diagramas de casos de uso, diagramas de atividades e diagramas de classes. Em seguida é apresentado um exemplo de cálculo da TCA.

3.2.1 Diagramas de Casos de Uso

Os casos de uso têm como função, representar as principais funcionalidades que podem ser observadas em um sistema e dos elementos externos que interagem com o mesmo (BEZERRA, 2002). A seguir são listados os diagramas de casos de uso modelados na fase de especificação do sistema.

3.2.1.1 Diagramas de caso de uso do analista

O analista é o responsável pelos cadastros das configurações e informações dos valores dos atributos que fazem parte do processo de cálculo da TCA. Na Figura 2, é apresentado o diagrama de casos de uso do analista.

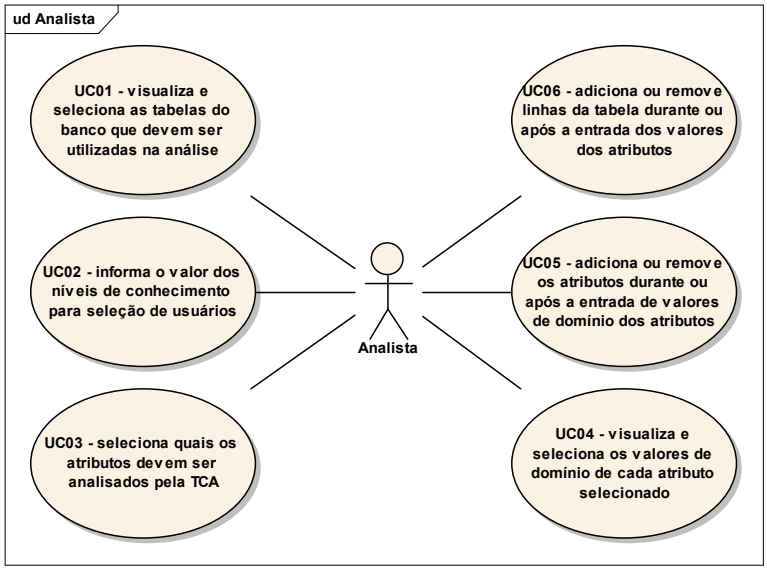


Figura 2 – Diagrama de casos de uso do analista

3.2.1.2 Diagramas de casos de uso do verificador e analisador

O verificador é a parte do sistema responsável pela verificação de consistência das informações de tabelas e valores dos atributos. O analisador é responsável pela análise e montagem dos relatórios. Na Figura 3 é apresentado o diagrama de casos de uso do módulo do sistema, do ator verificador e do ator analisador.

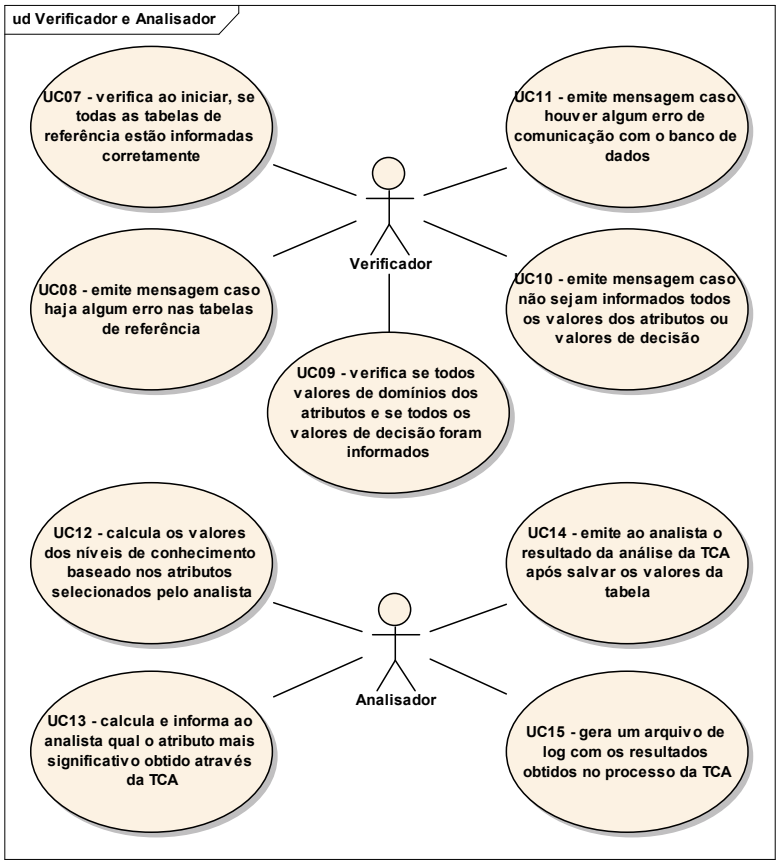


Figura 3 – Diagrama de casos de uso do analisador e verificador

3.2.1.3 Diagrama de casos de uso do DBA

O DBA é o responsável pela importação das tabelas de dados do ERP[3] para o bando de dados do sistema e também responsável em ajustar os campos destas tabelas para compatibilizar as referências de dados para a análise. Na Figura 4, é apresentado o diagrama de casos de uso do DBA.

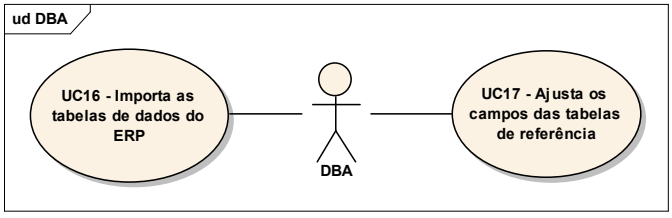


Figura 4 – Diagrama de casos de uso do DBA

[3] ERP é uma ferramenta integrada de gestão, que procura integrar as diversas áreas da empresa como forma de aumentar a produtividade das organizações.

3.2.2 Diagrama de Atividades

Os diagramas de atividade capturam ações e seus resultados, focando o trabalho executado na implementação de uma operação. É uma variação do diagrama de estados da UML possuindo um propósito um pouco diferente, pois é uma maneira alternativa de se mostrar interações, com a possibilidade de expressar como as ações são executadas, o que elas fazem, quando elas são executadas e onde elas acontecem (UML, 2002, p. 48).

Na Figura 5, é apresentado o diagrama de atividades, no modelo raias de natação, onde é apresentado o fluxo de atividades da análise de dados da TCA.

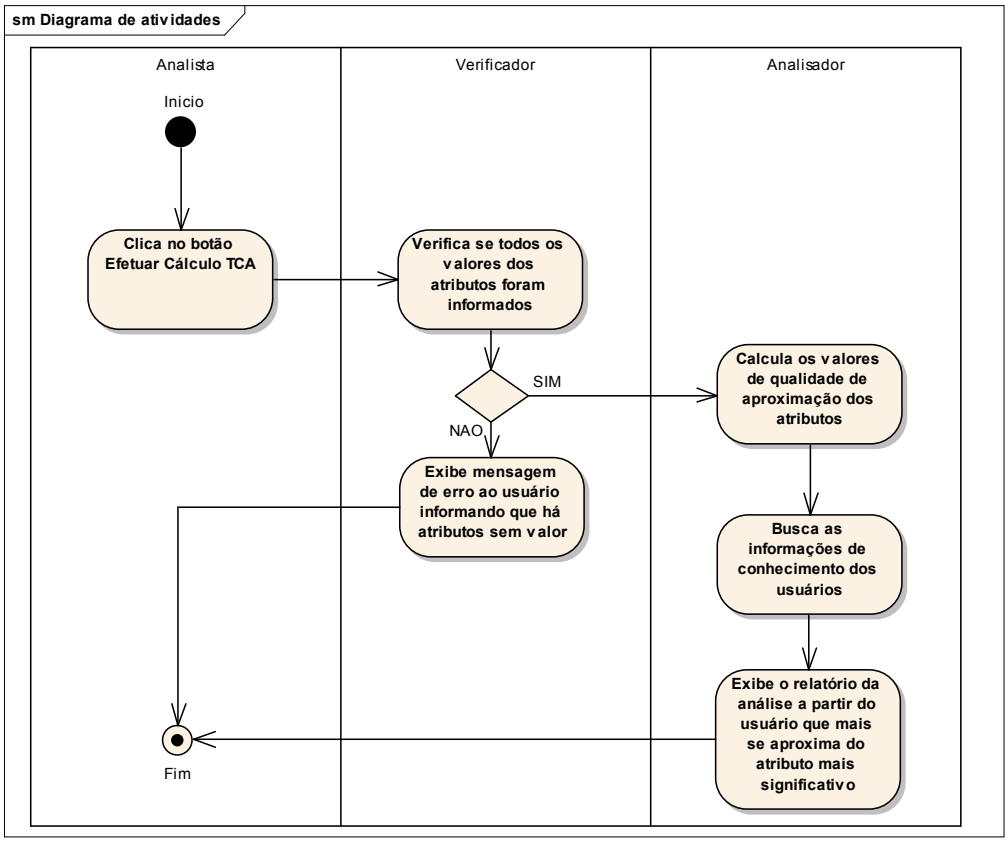


Figura 5 – Diagrama de atividades do fluxo da análise de dados da TCA

3.2.3 Diagramas de Classes

O diagrama de classes demonstra a estrutura das classes de um sistema que representam os objetos gerenciados pela aplicação modelada. Estas classes possuem vários tipos de relacionamentos, como por exemplo, o de associação onde as classes são conectadas entre si, o de dependência onde uma classe depende ou usa outra classe, o de especialização

onde uma classe é uma especialização de outra classe, ou em pacotes onde as classes são agrupadas por características similares. Todos estes relacionamentos são mostrados no diagrama de classes juntamente com as suas estruturas internas, que são os atributos e operações. O diagrama de classes é considerado estático já que a estrutura descrita é sempre válida em qualquer ponto do ciclo de vida do sistema (UML, 2002, p. 42). Na Figura 6 é apresentado o diagrama de classes modelado durante a especificação do sistema.

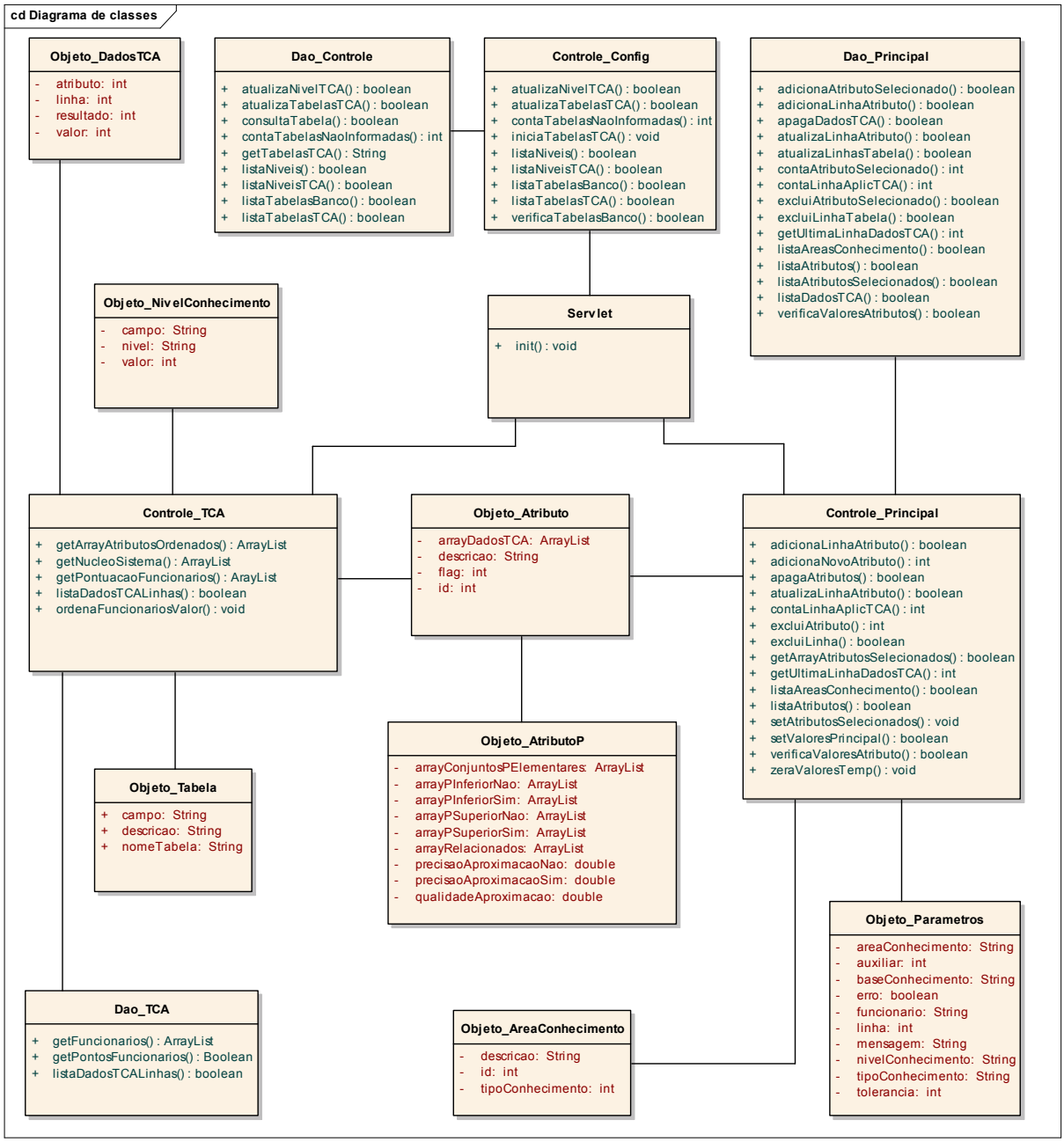


Figura 6 – Diagrama de classes

3.2.4 Exemplo de cálculo da TCA

O presente exemplo de especificação e cálculo da TCA possibilita obter um melhor entendimento de como funciona as bases desta teoria. O texto e fórmulas a seguir são transcritos a partir de Pawlak (1982) com tradução do autor deste trabalho.

O autor inicialmente solicita que seja considerada a tabela de informação, descrita na Tabela 2.

Paciente	Dor de cabeça (C)	Dor Muscular (M)	Temperatura (T)	Gripe (G)
1	Não	Sim	Alta	Sim
2	Sim	Não	Alta	Sim
3	Sim	Sim	Muito Alta	Sim
4	Não	Sim	Normal	Não
5	Sim	Não	Alta	Não
6	Não	Sim	Muito Alta	Sim

Fonte: Pawlak (1982).

Tabela 2 – Tabela de informação de diagnóstico de gripe

Diante do exposto tem-se a definição dos conjuntos U, C, D e Q descritos conforme demonstrado no Quadro 12.

U = {1,2,3,4,5,6} é o conjunto universo de objetos (pacientes);
C = {C,M,T} é o conjunto de atributos de condição;
D = {G} é o conjunto de atributos de decisão;
Q = C ∪ D = {C,M,T,G} é o conjunto de todos os atributos.

Fonte: Pawlak (1982).

Quadro 12 – Definição dos conjuntos da tabela de informação

A função de informação assume os valores descritos no Quadro 13.

f(1,C) = Não; f(1,M) = Sim; f(1,T) = Alta; f(1,G) = Sim;
f(2,C) = Sim; f(2,M) = Não; f(2,T) = Alta; f(2,G) = Sim;
f(3,C) = Sim; f(3,M) = Sim; f(3,T) = Muito Alta; f(3,G) = Sim;
f(4,C) = Não; f(4,M) = Sim; f(4,T) = Normal; f(4,G) = Não;
f(5,C) = Sim; f(5,M) = Não; f(5,T) = Alta; f(5,G) = Não;
f(6,C) = Não; f(6,M) = Sim; f(6,T) = Muito Alta; f(6,G) = Sim.

Fonte: Pawlak (1982).

Quadro 13 – Definição dos conjuntos da tabela de informação

Com os domínios de atributos definidos no Quadro 14.

VD = VM = VG = {Sim,Não};
DT = {Normal,Alta,Muito Alta}.

Fonte: Pawlak (1982).

Quadro 14 – Definição dos domínios dos atributos

Neste sistema de informação, particiona-se U de forma a diagnosticar a gripe, portanto de acordo com a relação de indiscernibilidade ID sobre o atributo de decisão G, obtém-se Y conforme exposto no Quadro 15.

$Y = U / ID = \{Y_1, Y_2\}$
onde
$Y_1 = \{1,2,3,6\}$ é o conjunto de pacientes que apresentam gripe;
$Y_2 = \{4,5\}$ é o conjunto de pacientes que não apresentam gripe.

Fonte: Pawlak (1982).

Quadro 15 – Relação de indiscernibilidade ID

Seja o conjunto de atributos P = {M, T}, a sua relação de indiscernibilidade I_P e a classe de conjuntos P-elementares U/ I_P são descritos no Quadro 16.

IP = {(1,1),(2,2),(2,5),(3,3),(3,6),(4,4),(5,2),(5,5),(6,3),(6,6)};
$U / I_P = \{\{1\},\{2,5\},\{3,6\},\{4\}\}$.

Fonte: Pawlak (1982).

Quadro 16 – Relação IP e classe de conjuntos P-elementares

Os conjuntos P-elementares em U / I_P possuem suas descrições demonstradas no Quadro 17.

$Des_P(\{1\}) = \{(M,Sim), (T,Alta)\}$;
$Des_P(\{2,5\}) = \{(M,Não), (T,Alta)\}$;
$Des_P(\{3,6\}) = \{(M,Sim), (T,Muito Alta)\}$;
$Des_P(\{4\}) = \{(M,Sim), (T,Normal)\}$.

Fonte: Pawlak (1982).

Quadro 17 – Conjuntos P-elementares em U / I_P

Ao considerar como P aproxima ao conjunto Y_1 de pacientes que apresentam gripe e Y_2 de pacientes que não apresentam gripe, tem-se demonstrado no Quadro 18 as definições das aproximações P-inferior P_{Y1}, P-superior P^{Y}_1, o conjunto P-fronteira $Fr_P(Y_1)$ e a precisão de aproximação α_P (Y_1) e também P-inferior P_{Y2}, P-superior P^{Y}_2, o conjunto P-fronteira $Fr_P(Y_2)$ e a precisão de aproximação α_P (Y_2).

$P_{Y1} = \{1\} \cup \{3, 6\} = \{1, 3, 6\};$
$P^{Y}_{1} = \{1\} \cup \{2, 5\} \cup \{3, 6\} = \{1, 2, 3, 5, 6\};$
$Fr_P(Y_1) = P_{Y1} - P^{Y}_{1} = \{2, 5\};$
$\alpha_P(Y_1) = \frac{card(P_{Y1})}{card(P^{Y}_{1})} = \frac{3}{5} = 0{,}6.$
$P_{Y2} = \{4\};$
$P^{Y}_{2} = \{4\} \cup \{2, 5\} = \{2, 4, 5\};$
$Fr_P(Y_2) = P_{Y2} - P^{Y}_{2} = \{2, 5\};$
$\alpha_P(Y_2) = \frac{card(P_{Y2})}{card(P^{Y}_{2})} = \frac{1}{3} = 0{,}3333.$

Fonte: Pawlak (1982).

Quadro 18 – Definições para pacientes que apresentam e que não apresentam gripe

A fórmula da qualidade de aproximação da partição Y pelo conjunto de atributos P é demonstrada no Quadro 19.

$$\gamma_P(Y) = \frac{card(P_{Y1}) + card(P_{Y2})}{card(U)} = \frac{3+1}{6} = 0{,}667$$

Fonte: Pawlak (1982).

Quadro 19 – Definição da qualidade de aproximação da partição Y

Para descobrir as dependências e obter as reduções deve-se, inicialmente, encontrar a qualidade da aproximação sobre todos os possíveis subconjuntos P de atributos de condição. Os resultados podem ser vistos na Tabela 3.

Atributos P	Qualidade de Aproximação $\gamma_P(Y)$	Conjuntos P-elementares em U / I_P
{C,M,T}	0,667	{1},{2,5},{3},{4},{6}
{M,T}	0,667	{1},{2,5},{3,6},{4}
{C,T}	0,667	{1},{2,5},{3},{4},{6}
{C,M}	0,167	{1,4,6},{2,5},{3}
{T}	0,500	{1,2,5},{3,6},{4}
{M}	0,000	{1,3,4,6},{2,5}
{C}	0,000	{1,4,6},{2,3,5}

Fonte: Pawlak (1982).

Tabela 3 – Resultados obtidos dos possíveis subconjuntos P de atributos de condição

Observa-se que as Y-reduções de P = {C,M,T} são {M,T} e {C,T} e o Y-núcleo de P é Coreγ({C,M,T}) = {C,T} $\cap$ {M,T} = {T}. Ou seja, T é o atributo mais significativo de Q, o qual não pode deixar de ser considerado, pois sua eliminação significa obter aproximações de baixa qualidade. Em relação aos atributos C e M, eles são mutuamente intercambiáveis. Assim, fica a critério pessoal trabalhar com {C,T} ou com {M,T}, considerando que ambos os grupos produzem a mesma qualidade de informação em relação à {C,M,T}.

Quanto às dependências, se P e R são dois conjuntos de atributos e se P $\subset$ R, então

IR $\subset$ IP e, portanto, existe a dependência R $\rightarrow$ P. Assim, por exemplo, {M,T} $\rightarrow$ {C,M,T} e, portanto, a dependência {C,M,T} $\rightarrow$ {C,M}.

Da família U / IC = {{1},{2,5},{3},{4},{6}} de classes de condição e da família U / I_C = {{1,2,3,6},{2,5}} de classes de decisão surgem as regras descritas na Tabela 4.

$X_i \in U / I_C$	$Y_j \in U / I_D$	regraDesc $_C(X_i) \Rightarrow Desc_D(Y_j)$	Determinística?
{1}	{1,2,3,6}	{(C,Não),(M,Sim),(T,Alta)} $\Rightarrow$ {(G,Sim)}	Sim
{2,5}	{1,2,3,6}	{(C,Sim),(M, Não),(T,Alta)} $\Rightarrow$ {(G,Sim)}	Não
{3}	{1,2,3,6}	{(C, Sim),(M,Sim),(T,Muito Alta)} $\Rightarrow$ {(G,Sim)}	Sim
{6}	{1,2,3,6}	{(C,Não),(M,Sim),(T,Muito Alta)} $\Rightarrow$ {(G,Sim)}	Sim
{2,5}	{4,5}	{(C, Sim),(M, Não),(T,Alta)} $\Rightarrow$ {(G, Não)}	Não
{4}	{4,5}	{(C,Não),(M,Sim),(T,Normal)} $\Rightarrow$ {(G, Não)}	Sim

Fonte: Pawlak (1982).

Tabela 4 – Resultados de aproximações determinísticas para {C,M,T}

Estas regras simplificam as determinações caso se adotem a Y-redução {M,T} de C, sem perda da qualidade de aproximação da partição. Neste caso, obtém-se a situação mostrada na Tabela 5.

$X_i \in U / I_C$	$Y_j \in U / I_D$	regraDesc $_C(X_i) \Rightarrow Desc_D(Y_j)$	Determinística?
{1}	{1,2,3,6}	{(M,Sim),(T,Alta)} $\Rightarrow$ {(G,Sim)}	Sim
{2,5}	{1,2,3,6}	{(M, Não),(T,Alta)} $\Rightarrow$ {(G,Sim)}	Não
{3,6}	{1,2,3,6}	{(M,Sim),(T,Muito Alta)} $\Rightarrow$ {(G,Sim)}	Sim
{2,5}	{4,5}	{(M, Não),(T,Alta)} $\Rightarrow$ {(G, Não)}	Não
{4}	{4,5}	{(M,Sim),(T,Normal)} $\Rightarrow$ {(G, Não)}	Sim

Fonte: Pawlak (1982).

Tabela 5 – Resultados de aproximações determinísticas para {M,T}

3.3 IMPLEMENTAÇÃO

Neste capítulo são apresentadas algumas informações teóricas sobre técnicas e ferramentas utilizadas para o desenvolvimento do trabalho, detalhando as informações técnicas mais recentes que não sejam de domínio tão comum. Em seguida é apresentado um exemplo da operacionalidade da implementação.

3.3.1 Técnicas e ferramentas utilizadas

A seguir são apresentadas as técnicas e ferramentas utilizadas para a implementação do sistema desenvolvido.

3.3.1.1 Unifield Modeling Language (UML)

Para a criação dos diagramas de casos de uso, de atividades e de classes foi utilizada a ferramenta Enterprise Architect unida à linguagem UML. A UML é uma linguagem visual com o objetivo de modelar sistemas orientados a objetos constituídos de elementos gráficos utilizados na modelagem, que permitem representar os conceitos do paradigma da orientação a objetos. Outra característica da UML é sua independência de linguagem de programação e de processos de desenvolvimento, podendo ser utilizada para a modelagem de sistemas, sem importar qual linguagem de programação será utilizada (BEZERRA, 2002).

3.3.1.2 Java Server Pages (JSP)

Pittella (2006) explica que JSP é uma tecnologia utilizada no desenvolvimento de aplicações para internet. Por ser baseada na linguagem de programação Java, tem a vantagem da portabilidade de plataforma, permitindo a sua execução em vários sistemas operacionais. Esta tecnologia permite ao desenvolvedor de páginas web, produzir aplicações que acessem banco de dados, manipulem arquivos no formato texto, capturem informações a partir de formulários e captem informações sobre o visitante e sobre o servidor.

Uma página criada com a tecnologia JSP depois de instalada em um servidor de aplicação compatível com a tecnologia J2EE, é transformada em um *servlet*[4]. Um exemplo de servidor compatível com a tecnologia JSP é o Tomcat.

O Quadro 20 apresenta a rotina JSP responsável em relacionar os atributos selecionados pelo analista a fim de efetuar todas as combinações possíveis entre estes

[4] *Servlet* é um programa que estende a funcionalidade de um servidor WEB, gerando conteúdo dinâmico. Na requisição de uma página WEB, quando todas as informações necessárias são geradas, os servlets montam a página JSP, que segue os padrões HTML para o aplicativo navegador do cliente.

atributos e enviá-las posteriormente à rotina de cálculo da TCA.

```
public ArrayList relacionaAtributos (ArrayList arrayAtributosSelecionados)
{
  // array de arrays que será devolvido no fim da função
  // encapsulado em um objeto "Atributo P"
  ArrayList arrayAtributosP = new ArrayList();
  Objeto_Atributo oAtributo;
  Objeto_AtributoP oAtribP;
  // o numero de possibilidades deve ser (2 elevado a n) menos 1
  // pois o menos 1 leva primeiro a combinação de todos os atributos
  // começando do último (maior) para o primeiro (menor)
  int possibilidades=(int)Math.pow(2,arrayAtributosSelecionados.size())-1;
  // inicia a contagem decrescente de possibilidades
  for (int i=possibilidades;i>0;i--) {
    // efetua a conversão do número da possibilidade para binário
    String bin = Integer.toString(i, 2);
    // arraySelecionados conterá os atributos que devem ser relacionados
    ArrayList arrayRel = new ArrayList();
    // verifica os bits da String "bin"
    for (int j=0;j<bin.length();j++) {
      // no array de atributos, deve começar do último para o primeiro
      // pois no binário gerado são desconsiderados os zeos a esquerda
      int k = bin.length()-j-1;
      // quando "bin" em J for igual a 1, deve relacionar os atributos
      // acidionando no arrayRel
      if (bin.charAt(j)=='1') {
        oAtributo = (Objeto_Atributo)arrayAtributosSelecionados.get(k);
        arrayRel.add(oAtributo);
      }
    }
    // encapsula o array de atributos em um objeto de atributosP este obj
    // conterá o valor da qualidade de aproximação e atrb. "P Elementares"
    ArrayList conjPElement = new ArrayList();
    oAtribP = new
Objeto_AtributoP(arrayRel,null,null,0.0,null,null,0.0,0.0,conjPElement);
    // por fim, adiciona o objeto de atributosP ao array de atributos
    arrayAtributosP.add(oAtribP);
  }
  return arrayAtributosP;
}
```

Quadro 20 – Rotina JSP que relaciona os atributos selecionados

Nesta rotina, o sistema recebe os atributos selecionados como parâmetro e efetua o cálculo das possibilidades possíveis, elevando 2 à potência do número de atributos, efetuando em seguida, os relacionamentos entre os atributos, utilizando as comparações efetuadas a partir do sistema binário de numeração e retornando a lista dos atributos já relacionados.

O Quadro 21 apresenta a rotina que calcula a qualidade de aproximação dos atributos relacionados com os valores dos objetos, linhas da tabela, a fim de determinar posteriormente, qual a melhor qualidade de aproximação entre os atributos.

```
public void calculaQualidadeAproximacao (ArrayList arrayAtributosP, int
ultimaLinha){
  int inferiorSim, inferiorNao, superior = 0;
  // para cada elemento Atributo P do array
  for (int i=0;i<arrayAtributosP.size();i++) {
    Objeto_AtributoP oAtributoP;
    oAtributoP = (Objeto_AtributoP)arrayAtributosP.get(i);
    // ... captura os arrays P superior e inferior para sim e não
    ArrayList arrayPInfSim = (ArrayList)oAtributoP.getArrayPInferiorSim();
    ArrayList arrayPSupSim = (ArrayList)oAtributoP.getArrayPSuperiorSim();
    ArrayList arrayPInfNao = (ArrayList)oAtributoP.getArrayPInferiorNao();
    ArrayList arrayPSupNao = (ArrayList)oAtributoP.getArrayPSuperiorNao();
    // ... conta em cada array quantos elementos há para SIM
    inferiorSim = calculaQtdElementos (arrayPInferiorSim);
    superior = calculaQtdElementos (arrayPSuperiorSim);
    // ... setando a precisao de aproximação
    oAtributoP.setPrecisaoAproximacaoSim((double)inferiorSim/superior);
    // ... e o mesmo para não
    inferiorNao = calculaQtdElementos (arrayPInfNao);
    superior = calculaQtdElementos (arrayPSupNao);
    // ... setando a precisão de aproximação
    oAtributoP.setPrecisaoAproximacaoNao((double)inferiorNao/superior);
    // por último, com os valores de aproximação inferior para sim e não
    // calcula a qualidade de aproximação para o elemento Atributo P
oAtributoP.setQualidadeAproximacao((double)(inferiorSim+inferiorNao)/ultima
Linha);
  }
}
```

Quadro 21 – Rotina JSP que calcula a qualidade de aproximação

Nesta rotina, o sistema recebe a relação dos atributos relacionados e extrai as informações das aproximações inferior e superior para sim e para não. A partir destas informações, são calculadas as aproximações e em seguida a qualidade de aproximação para o atributo relacionado, conforme as especificações da TCA.

3.3.1.3 JSP Standard Tag Library (JSTL)

JSTL é um conjunto de bibliotecas de *tags* padronizadas que implementam funcionalidades comuns a aplicações web e permitem a substituição de código JSP nas páginas, facilitando a implementação e manutenção do código HTML. As *tags* são estruturas em uma linguagem de marcação, que consistem em breves instruções para instruir o aplicativo navegador sobre a forma de apresentação de textos e gráficos numa página web.

O Quadro 22 apresenta a rotina JSTL responsável em montar a tabela atributos e valores na página principal do sistema. Nesta página, o analista pode visualizar e modificar os valores dos atributos previamente selecionados, bem como inserir e excluir linhas da tabela de informações. Cada linha identifica um objeto de análise, que em cada atributo, possui um

valor correspondente ao domínio do atributo conforme as definições da TCA; estes valores podem ser determinados através dos campos de escolha disponíveis em cada atributo. Quando nenhum valor é selecionado, a mensagem “selecione...” é exibida.

```
<c:forEach var="linha" begin="1" end="${ultimaLinha}" step="1">
  <TR>
  <TD ALIGN="center"><c:out value="${linha}"/></TD>
  <c:forEach items="${arrayAtributosSelecionados}"
var="atributoSelecionado">
    <TD ALIGN="center">
    <SELECT NAME="${linha}-${atributoSelecionado.id}">
      <OPTION VALUE="0" SELECTED>Selecione. . . . . .</OPTION >
      <c:forEach items="${arrayAreasConhecimento}" var="  areaConhec">
        <c:if
test="${areaConhec.tipoConhecimento==atributoSelecionado.id}">
          <c:forEach items="${arrayDadosTCA}" var="dadosTCA">
            <c:if test="${dadosTCA.linha==linha}">
              <c:if test="${dadosTCA.atributo==atributoSelecionado.id}">
                <c:if test="${dadosTCA.valor==  areaConhec.id}">
                  <OPTION VALUE="${areaConhec.id}"
SELECTED>${areaConhec.descricao}</OPTION>
                </c:if>
                <c:if test="${dadosTCA.valor!=  areaConhec.id}">
                  <OPTION
VALUE="${areaConhec.id}">${areaConhec.descricao}</OPTION>
                </c:if>
              </c:if>
              <c:set var="resultado" value="${dadosTCA.resultado}"/>
            </c:if>
          </c:forEach>
        </c:if>
      </c:forEach>
    </SELECT></TD>
  </c:forEach>
  <TD ALIGN="center">
    <SELECT NAME="${linha}-resultado">
      <OPTION VALUE="-1" SELECTED>Selecione</OPTION>
      <c:if test="${resultado==1}">
        <OPTION VALUE="1" SELECTED>Sim</OPTION>
      </c:if>
      <c:if test="${resultado!=1}">
        <OPTION VALUE="1">Sim</OPTION>
      </c:if>
      <c:if test="${resultado==0}">
        <OPTION VALUE="0" SELECTED>Não</OPTION>
      </c:if>
      <c:if test="${resultado!=0}">
        <OPTION VALUE="0">Não</OPTION>
      </c:if>
    </SELECT>
  </TD></TR>
</TD>
  <TD ALIGN="center"><INPUT TYPE="button"VALUE="-><-"ONCLICK="linha.value=
'${linha}';action.value='exlui';frmPrinc.submit()">
</TD></TR>
</c:forEach>
```

Quadro 22 – Rotina JSTL que monta a tabela de dados da tela principal do sistema

3.3.1.4 Tomcat

O Tomcat é um sistema gratuito, de código aberto e desenvolvido para as tecnologias de desenvolvimento Java *servlets* e JSP, compatível com a tecnologia J2EE sob o Projeto Jakarta da empresa Apache Software Foundation. Coutinho (2006) explica que se trata basicamente de um servidor de aplicação onde são instalados os *servlets* para tratar as requisições da página internet que o servidor da aplicação receber.

O autor acrescenta que muitas companhias e desenvolvedores estão contribuindo com o avanço e atualização do Tomcat. A implementação está disponível para qualquer companhia ou desenvolvedor para utilização em servidores, ferramentas de desenvolvimento e criação de sites dinâmicos e interativos para internet.

Existem muitos servidores de aplicação disponíveis no mercado, neste trabalho o Tomcat foi escolhido por ser gratuito e fácil de usar e operar. A versão utilizada foi a 5.5.17.

3.3.1.5 Banco de dados MySQL

O MySQL é um SGBD desenvolvido em linguagem de programação C por dois desenvolvedores suecos e um finlandês que trabalham juntos no projeto desde a década de 80. O sistema utiliza SQL como interface para consulta de dados e possui um sistema de segurança que atende a maioria das aplicações para o qual é utilizado (BELFIGLIO, 2006).

A característica mais marcante do MySQL é a conveniência com o ambiente multiusuário e multitarefa, o que o torna ideal para o desenvolvimento de sites e sistemas para web. As bases de dados podem ser acessadas por diversas linguagens de programação, além de possuir suporte a arquitetura cliente/servidor e ser compatível com diversas plataformas de sistemas operacionais. No presente trabalho foi utilizada a versão 5.0 do MySQL para Windows.

3.3.1.6 Eclipse

D'Ávila (2005) explica que o Eclipse é uma ferramenta de desenvolvimento de software, voltada para o ambiente de desenvolvimento integrado conhecido por IDE. Iniciado

pela IBM, que o utiliza como base na construção de suas ferramentas comerciais para desenvolvimento Java em seus produtos para projeto e desenvolvimento, o Eclipse ganhou popularidade e força na comunidade de desenvolvedores da linguagem Java por diversos motivos, em especial os enumerados a seguir:

a) ser software gratuito, livre e de código aberto;
b) oferecer amplos recursos de produtividade para geração de código, atalhos e automação de desenvolvimento;
c) consistir em um projeto sério e ativo, bem organizado e coordenado, além do amplo apoio da comunidade e de grandes empresas e instituições;
d) ter um ambiente gráfico construído com a biblioteca de componentes própria do projeto Eclipse, combinando grande quantidade de componentes gráficos e interface de usuário com desempenho e leveza;
e) possuir uma arquitetura de software aberta e extensível, permitindo que *plug-ins*[5] sejam criados e facilmente integrados.

O autor acrescenta que como o eclipse é um ambiente de desenvolvimento aberto, o seu uso efetivo em produção corporativa requer agregar *plug-ins* para tecnologias Java especializados, principalmente para a plataforma J2EE. Isto torna o eclipse uma plataforma para a integração de ferramentas de desenvolvimento altamente extensível. Neste trabalho foi utilizada a versão Eclipse SDK 3.2 com suporte ao desenvolvimento web.

3.3.2 Operacionalidade da implementação

A seguir é apresentado um exemplo de funcionamento da implementação, onde são apresentadas algumas telas do sistema, preservando a ordem de execução do aplicativo para demonstrar a operacionalidade da implementação realizada.

Ao ser iniciado, o sistema exibe a tela principal com a tabela de atributos vazia, sugerindo inicialmente ao analista, selecionar quais os atributos deseja utilizar na análise, conforme pode ser visualizado na Figura 7.

---

[5] *Plug-ins* são códigos ou pequenos programas que podem ser implementados através da própria ferramenta Eclipse e adicionados à mesma para prover alguma funcionalidade especial ou muito específica.

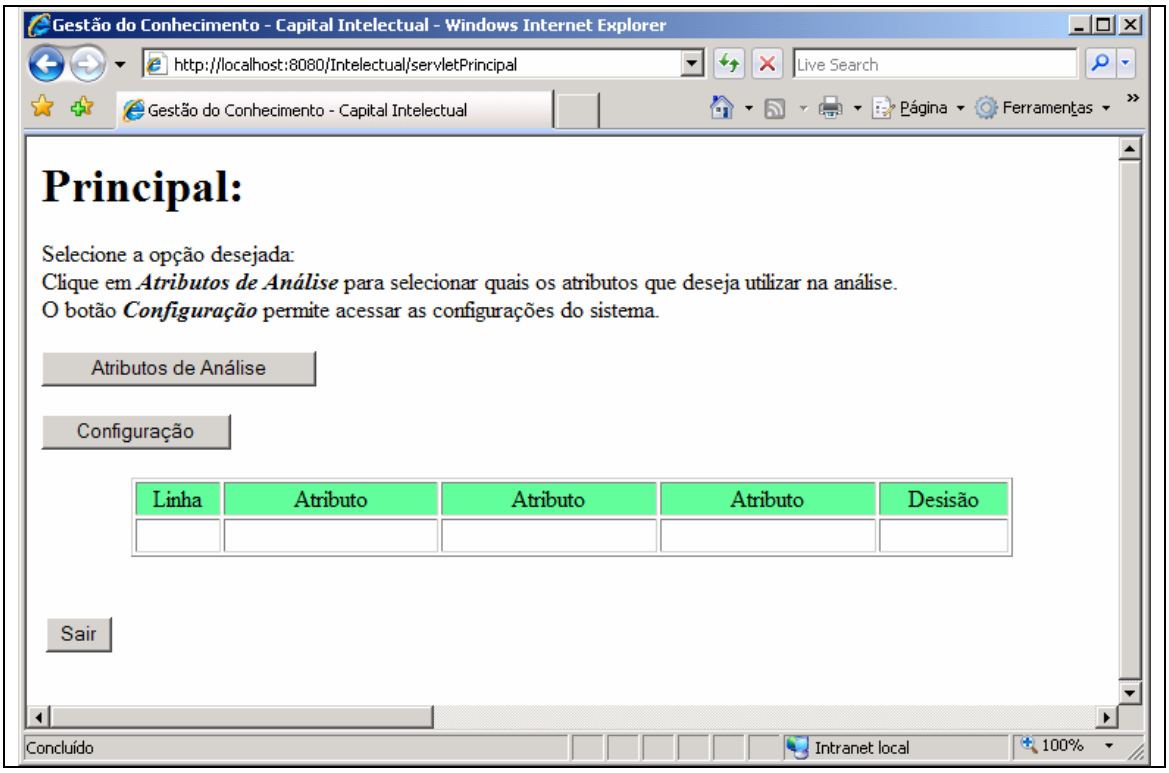


Figura 7 – Tela principal do sistema

Os atributos podem ser selecionados clicando no botão "Atributos de Análise" disponível no canto superior esquerdo, onde o sistema exibe a tela mostrada na Figura 8.

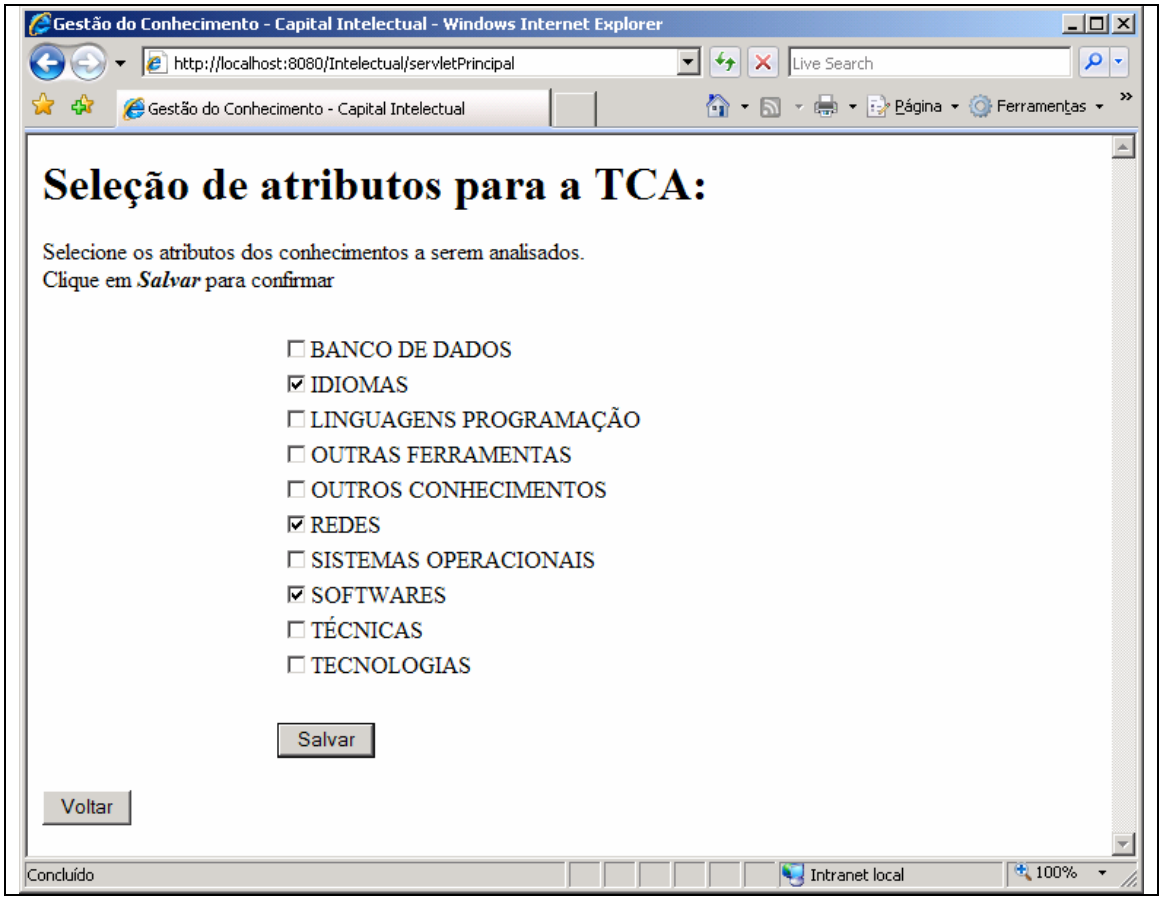


Figura 8 – Tela de seleção de atributos para a TCA

Após selecionar os atributos desejados, o analista deve clicar no botão “Salvar”, disponível abaixo da lista de atributos, o qual conduz o analista novamente para a página principal, onde a tabela de informações é preenchida com os atributos selecionados e com as opções de valores, que combinados, devem satisfazer uma determinada decisão, conforme pode ser visto na Figura 9.

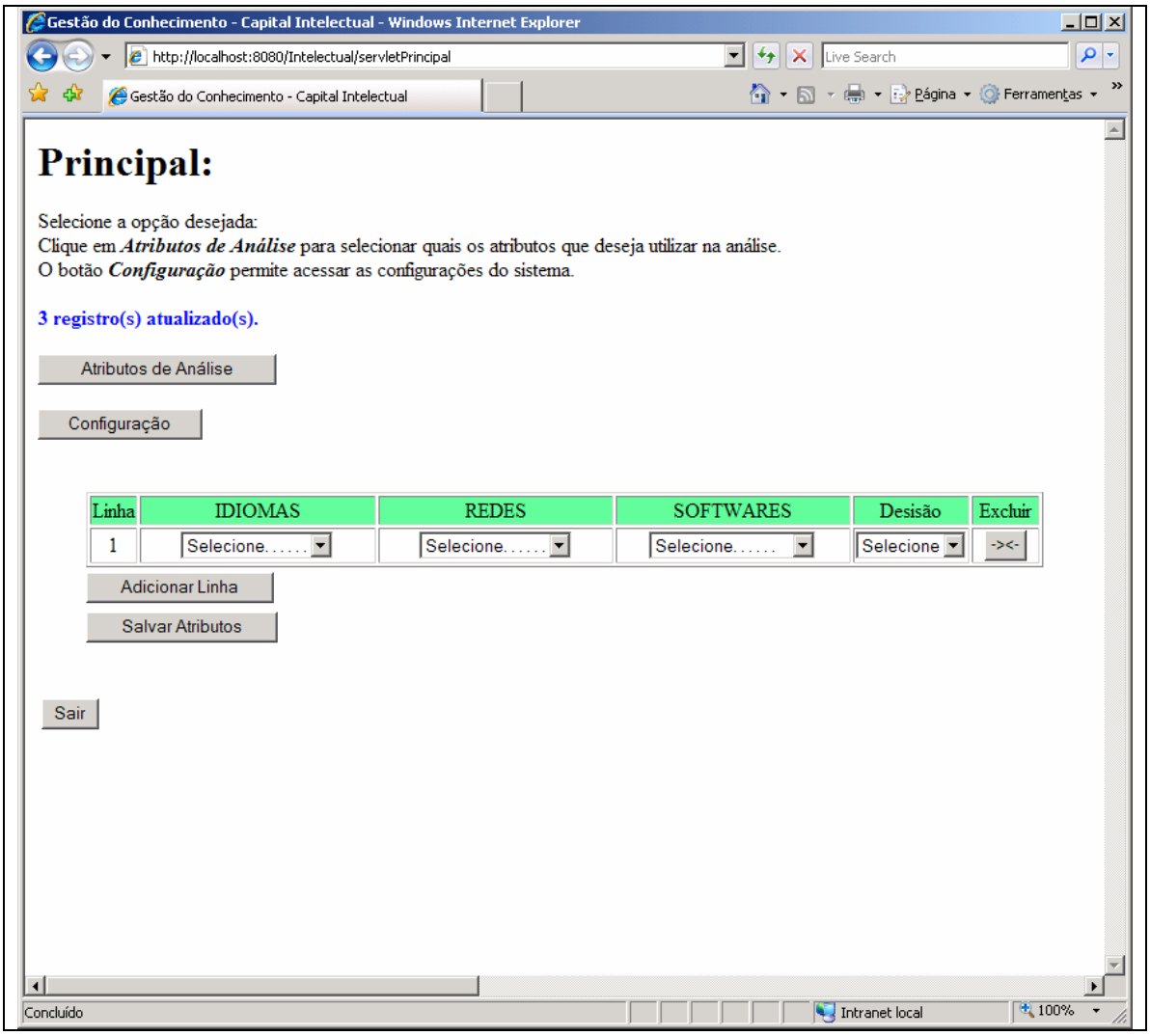


Figura 9 – Tela principal com atributos selecionados

No presente exemplo, constituindo o conjunto de atributos de condição foram selecionados os seguintes atributos:

a) idiomas;

b) redes;

c) softwares.

O botão “Adicionar Linha” permite ao analista inserir outras linhas ao conjunto universo de objetos a ser analisado pelo sistema. Um exemplo de tabela, com os valores UxQ associados com os valores dos domínios informados, pode ser visto na Figura 10.

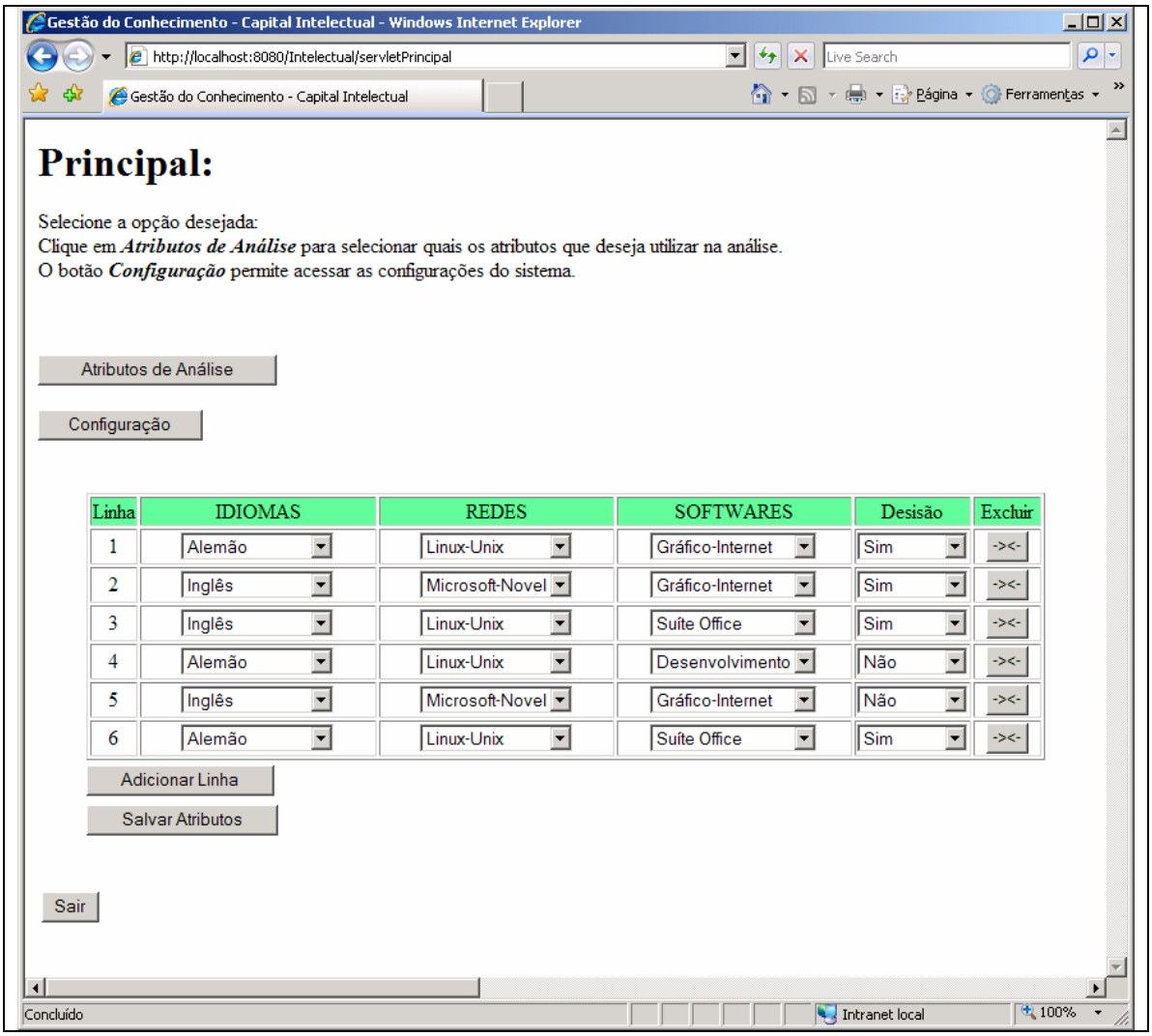


Figura 10 – Tela principal com atributos e valores informados

Nesta tela, o botão "Salvar Atributos" salva os dados informados pelo usuário no banco de dados e inicia o processo de análise da TCA. A partir daí, o sistema efetua os cálculos para determinar qual o atributo mais significativo. Inicialmente são relacionados os atributos formando os conjuntos dos atributos P de condição e em cada um deles, é formado o conjunto dos elementos P-elementares, baseado nos valores indicados pelo analista.

Tomando como base os atributos "Redes" (R) e "Softwares" (S), são obtidos os resultados descritos no Quadro 23.

- $U / I_P = \{\{1\},\{2,5\},\{3,6\},\{4\}\}$
- $P_Y(sim) = \{1,3,6\}$
- $P^Y(sim) = \{1,2,3,5,6\}$
- Precisão de aproximação para sim = 0,67
- $P_Y(não) = \{4\}$
- $P^Y(não) = \{2,4,5\}$
- precisão de aproximação para não = 0,33
- qualidade de aproximação = 0,67

Quadro 23 – Resultados obtidos pela TCA em operacionalidade

Os valores de qualidade de aproximação e os conjuntos P-elementares obtidos para cada subconjunto P de atributo de condição são demonstrados na Tabela 6.

Atributos P	Qualidade de Aproximação $\gamma_P(Y)$	Conjuntos P-elementares em U / I_P
{I,R,S}	0,667	{1},{2,5},{3},{4},{6}
{R,S}	0,667	{1},{2,5},{3,6},{4}
{I,S}	0,667	{1},{2,5},{3},{4},{6}
{I,R}	0,167	{1,4,6},{2,5},{3}
{S}	0,500	{1,2,5},{3,6},{4}
{R}	0,000	{1,3,4,6},{2,5}
{I}	0,000	{1,4,6},{2,3,5}

Tabela 6 – Qualid. de aprox. e conj. P-elementares para os subconjuntos P de condição

Observa-se que as Y-reduções de P são {R,S} e {I,S} e o núcleo de P obtido pela intersecção destes dois conjuntos é {S}, ou seja, o atributo software é o atributo mais significativo e que não pode deixar de ser considerado, pois isto acarretaria em baixa qualidade nas aproximações da TCA.

Sendo S o atributo mais significativo, o sistema inicia a próxima etapa que é a mineração dos dados na base de dados de conhecimento, efetuado os cálculos para cada nível de conhecimento, multiplicando os valores indicados na configuração do sistema, pelos níveis de conhecimento correspondentes informados pelos colaboradores em cada atributo, priorizando na busca, os colaboradores que possuem melhor aproveitamento no atributo mais significativo. A Figura 11 apresenta a tela com os resultados obtidos após o cálculo do atributo mais significativo e da mineração de dados das informações de conhecimento dos colaboradores da organização.

Nesta tela pode ser observado que o sistema relaciona os colaboradores priorizando os valores obtidos no atributo “Softwares”, independente do aproveitamento obtido nos outros atributos. Estes outros atributos são também relacionados, a fim de auxiliar o analista na tomada de decisão, a partir dos valores expostos, possibilitando ao mesmo julgar a necessidade ou não de efetuar uma nova análise dos resultados, modificando os valores dos atributos anteriormente informados e visualizando os resultados obtidos com a nova busca.

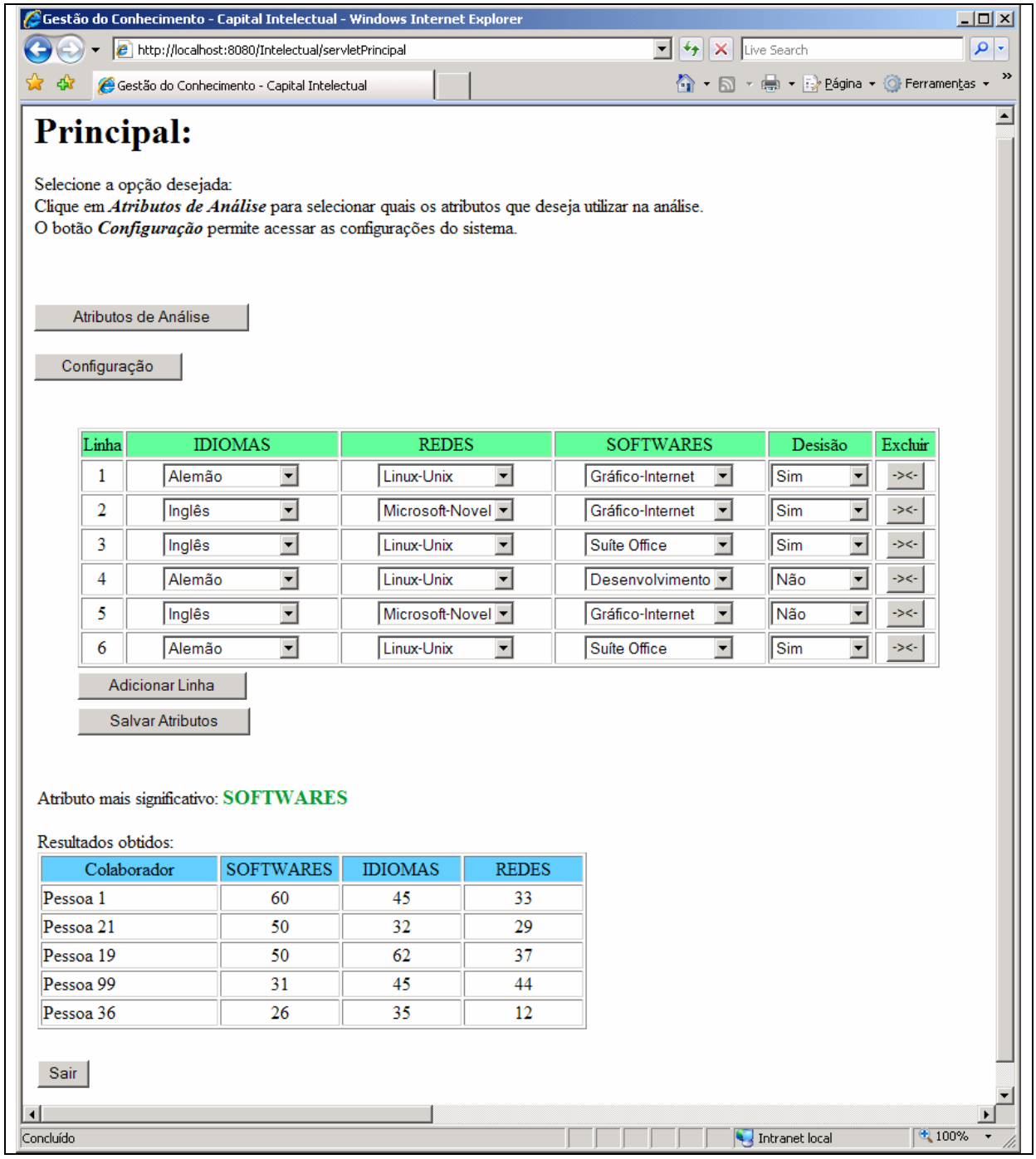


Figura 11 – Tela principal com a apresentação dos valores calculados

3.4 RESULTADOS E DISCUSSÃO

O levantamento e controle do capital intelectual de uma organização pode ser efetuado através do processo manual em papel, ou em formato eletrônico, através de um software ou

ferramenta de apoio ao gerenciamento de TI. Com o objetivo de proporcionar uma maior confiabilidade, segurança e principalmente agilidade no que se refere à manipulação de informações extremamente importantes e com o objetivo de prover e melhorar a gestão do conhecimento, decidiu-se aplicar uma solução automatizando o processo utilizando DM baseado na TCA

Por ser um sistema de gestão do conhecimento, ele procura proporcionar de forma eficiente e eficaz, através de valores fornecidos pelo analista do sistema, o levantamento do capital intelectual dos seus usuários mais preparados e, com os níveis de conhecimento mais adequados à determinadas situações.

Quanto à utilização de *data mining*, o sistema busca implementar todos os conceitos do KDD, desde a seleção e processamento dos dados, passando pela transformação e mineração das informações e exibindo e disponibilizando por fim, a interpretação do conhecimento obtido através do conceito de pontuação dos níveis de conhecimento.

Sobre o conceito da TCA, pode-se afirmar que é uma técnica matemática muito eficaz, pois ela permite a redução das variáveis envolvidas, relacionando o conjunto de atributos selecionados pelo analista com os objetos formados pelos valores destes atributos. A qualidade da aproximação é o ponto principal da análise, pois ela revela o quanto a variável é responsável em gerar algum resultado. Através deste resultado, é obtida a relação do capital intelectual esperado.

A busca por uma distribuição gratuita, orientada a objeto, fácil de ser compreendida e adaptada além de independente de qualquer plataforma de sistema operacional motivou a realização da implementação utilizando o servidor de aplicação Tomcat, utilizando o banco de dados MySQL e a linguagem de programação JSP com as *tags* da JSTL. Esta última escolha facilita bastante o trabalho de um profissional que venha a trabalhar com o design das páginas web do sistema.

Por fim, pode-se afirmar que todos os requisitos funcionais foram contemplados atingindo o resultado final proposto. Em relação aos requisitos não funcionais pode-se afirmar que todos foram atingidos sem maiores dificuldades devido à utilização das tecnologias mencionadas, com o objetivo principal de demonstrar e utilizar a TCA em uma aplicação de gestão de capital intelectual.

4 CONCLUSÕES

A gestão e a divulgação do conhecimento passam constantemente por diversas modificações. A evolução das tecnologias de informação e principalmente, a popularização da Internet, tem contribuído, e muito, para que o conhecimento seja divulgado cada vez mais através de meios eletrônicos, tornando este processo mais ágil e abrangente. Gerir estes conhecimentos de modo eficaz auxilia o crescimento sustentável das empresas. Unir a informação com o processo de decidir que rumo tomar dentro de uma organização é o diferencial competitivo que todas procuram. O presente trabalho foi desenvolvido levando em consideração esta necessidade.

Em relação aos objetivos propostos no início deste trabalho, pode-se afirmar que todos foram alcançados. O objetivo principal, que era demonstrar o potencial do *data mining* para classificação e segmentação de dados baseado na TCA, foi atingido. Com base nos resultados de atributos mais significativos, obtidos através da TCA foi possível aplicar os processos do KDD sobre a base de dados, pesquisando o potencial de cada colaborador através das informações de níveis de conhecimento previamente informadas. Estas informações são disponibilizadas graficamente de forma rápida e objetiva, listando os colaboradores e os resultados de aproveitamento para cada atributo, priorizando na busca o atributo mais significativo a fim de disponibilizar o analista os resultados do capital intelectual obtidos.

O conceito de programação orientada a objeto, facilitou bastante o processo de cálculo dos índices da TCA e permitiu ao sistema alcançar uma maior agilidade, sem depender muito do tempo das respostas das consultas realizadas ao banco de dados, processo este, que em geral, degrada um pouco o desempenho do sistema. A utilização da linguagem de programação JSP utilizando as *tags* JSTL permitiu ao sistema ser disponibilizado na web sem a necessidade de instalação de aplicativos auxiliares para operação e visualização das informações, além de garantir uma excelente portabilidade, tornando-se independente de plataforma de sistema operacional. A utilização do servidor de aplicação Tomcat e o banco de dados MySQL, também facilitou o desenvolvimento do sistema e favorece futuras extensões, que possam ser implementadas, uma vez que, além de serem gratuitos, estes aplicativos também são bastante conhecidos no ambiente de programação.

A maior vantagem da gestão do conhecimento é permitir a organização gerenciar o capital intelectual dos seus colaboradores. Ao concluir este trabalho, notou-se que a aplicação da TCA no processo da gestão do conhecimento, possibilitou entender e aprender como

utilizar o seu potencial a fim de auxiliar a resolução de problemas complexos de uma forma bastante prática e objetiva, bastando apenas selecionar os atributos desejados e seus respectivos valores, para obter a informação almejada, sem a preocupação com o peso numérico dos valores de cada atributo.

Uma das maiores dificuldades encontradas no projeto do sistema foi entender o funcionamento da TCA e adaptá-la no contexto da gestão do conhecimento. Outra dificuldade encontrada durante a implementação do sistema foi possibilitar o fato de os atributos poderem ser livremente selecionados pelo analista, o que torna o sistema bastante dinâmico, necessitou a busca e utilização de técnicas auxiliares para prover esta funcionalidade. Este fato ajudou muito no enriquecimento do conhecimento pessoal, devido às pesquisas realizadas e soluções encontradas. Não foi possível desenvolver a rotina para efetuar a avaliação da eficácia do levantamento dos valores das qualidades de aproximação obtidos pela TCA sobre os atributos selecionados, indicando ao analista se os valores escolhidos podem ou necessitam ser modificados para se obter um melhor resultado do capital intelectual. Também não foi possível desenvolver a rotina que possibilita ao analista, a escolha de um outro método de análise de dados, permitindo efetuar uma comparação de valores e eficácia dos resultados obtidos.

4.1 EXTENSÕES

Este trabalho contempla a gestão do conhecimento no que diz respeito ao levantamento de capital intelectual dos colaboradores de uma organização. Como sugestão para extensão deste trabalho poderiam ser implementadas algumas outras funcionalidades como:

a) implementar uma rotina que alerte o analista se os valores por ele selecionados necessitam ser modificados, podendo assim, obter uma busca de dados mais eficiente e precisa;
b) adicionar outras técnicas de cálculo de informações tais como árvores de decisão ou regressão linear, por exemplo, a fim de possibilitar ao analista, escolher outras formas de obter os resultados da geração do capital intelectual;
c) aplicar este projeto para seleção de pessoal, baseado no capital intelectual.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, Leandro M. et al. **Uma ferramenta para extração de padrões**. Palmas, [2004]. 13 f. Centro Universitário Luterano de Palmas, Palmas. Disponível em: <http://www.sbc.org.br/reic/edicoes/2003e4/cientificos/UmaFerramentaParaExtracaoDePadroes.pdf>. Acesso em: 10 abr. de 2007.

BARTOLOMEU, Tereza. **Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento**. 2002. 282 f. Tese (Doutorado em Engenharia de Produção) - Área de Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

BELFIGLIO, Ricardo. **Manuais de MySQL**. [S.I.], [2006]. Disponível em: <http://my.opera.com/Ricardo%20Belfiglio/blog/index.dml/tag/apostilas>. Acesso em: 13 out. de 2007.

BERNARDES, João N. **Tecnologia da informação para o gerenciamento do conhecimento obtido das bases de dados de uma organização**. 2001. 46 f. Dissertação (Mestrado em Engenharia de Produção) - Área de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina, Florianópolis.

BEZERRA, Eduardo. **Princípios de análise e projeto de sistemas com UML**. Rio de Janeiro: Campus, 2002.

COMPOLT, Geandro L. **Sistemas de informação executiva baseado em um Data Mining utilizando a técnica de árvores de decisão**. 1999. 52 f. Trabalho de Conclusão de Curso (Bacharelado em Ciências da Computação) - Centro de Ciências Exatas e Naturais, Universidade Regional de Blumenau, Blumenau.

COUTINHO, Flávio. **Tutorial de tomcat**. [S.I.], [2006]. Disponível em: <http://www.guj.com.br/java.tutorial.artigo.9.1.guj>. Acesso em: 13 out. de 2007.

D'ÁVILA, Márcio. **O eclipse vai bem obrigado**, 2006. Disponível em: <https://www.mhavila.com.br/topicos/java/eclipse_bem.html>. Acesso em: 12 out. de 2007.

DALFOVO, Oscar. **Modelo de integração de um sistema de inteligência competitiva com um sistema de gestão da informação e de conhecimento**. 2007. 234 f. Tese (Doutorado em Engenharia e Gestão do Conhecimento) - Área de Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis.

FAYYAD, Usama M. et al. **Advances in knowledge discovery and data mining**. Menlo Park: Mit Press, 1996.

GIMENES Eduardo. **Data mining - data warehouse**: a importância da mineração de dados em tomadas de decisões. 2000. 45 f. Trabalho de Conclusão de Curso (Tecnólogo em Processamento de Dados) - Faculdade de Tecnologia de Taquaritinga, Taquaritinga.

GOMES, Carlos F. S.; GOMES Luiz F. A. M. Modelagem de aspectos qualitativos do processo de negociação. **Revista de Administração Mackenzie**, n. 1, p. 83-103, 2004. Disponível em: <http://www.mackenzie.com.br/editoramackenzie/revistas/administracao/adm5n1/83.pdf>. Acesso em: 02 nov. de 2007.

GONÇALVES, Cid F.; GONÇALVES, Carlos A. Desafios e oportunidades para as organizações. **Gerência do conhecimento**, São Paulo, v. 8, n. 1, 2001. Disponível em: <www.ead.fea.usp.br/cad-pesq/arquivos/v08-1art05.pdf>. Acesso em: 10 jun. de 2007.

JESUS, Alberto P. **Data mining aplicado à identificação do perfil dos usuários de uma biblioteca para personalização de sistemas web de recuperação e disseminação de informações**. 2004. 120 f. Dissertação (Mestrado em Ciência da Computação) - Área de Concentração de Sistemas de Computação, Universidade Federal de Santa Catarina, Florianópolis.

MADEIRA, Sara C. **Data mining**. [S.I.], 2003. Disponível em: <www.di.ubi.pt/~smadeira/TALK_DI_UBI_2003.pdf>. Acesso em: 12 abr. de 2007.

MATOS, Hugo. Tarefas e técnicas de data mining. In: _____.**Estado da arte de ferramentas de data mining**. Lisboa, 2004. cap. 4, p. 4-7.

NARDELLI, Biancca **Protótipo de um sistema de informação gerencial aplicado a central de informação dos alunos da FURB utilizando data mining**. 2000. 69 f. Trabalho de Conclusão de Curso (Bacharelado em Ciências da Computação) - Centro de Ciências Exatas e Naturais, Universidade Regional de Blumenau, Blumenau.

NONAKA, Ikujiro; TAKEUCHI, Hirotaca. **Criação de conhecimento na empresa**. Rio de Janeiro: Campus, 1997.

NUNES, Milton S. **Sistemas inteligentes para tomada rápida de decisão nos sistemas elétricos**. [S.I.], [2005]. Disponível em: <http://www.eln.gov.br/setel/dados/arquivos/TrabalhosSelecionados/TecnologiaInformacao/SistemasInteligentesTomadaRapidaDecisaonosSistemasEletricos-MiltonNunesELN.ppt>. Acesso em: 12 abr. de 2007.

PAWLAK, Zdzislaw. Rough Sets. **International Journal of Information & Computer Sciences**. [S.l.], v. 11, p. 341-356, 1982.

PESSOA, Alex S. A.;SIMÕES, José D. S. **Mineração de dados espaço-temporal aplicada a previsão climática utilizando a teoria dos conjuntos aproximativos**. [S.I.], [2003]. Disponível em: <http://hermes2.dpi.inpe.br:1905/col/lac.inpe.br/worcap/2003/10.30.13.19/doc/worcap_versaofinal_alex2003.pdf>. Acesso em: 02 nov. de 2007.

PITTELLA, Felipe. **O que é JSP**. [S.I], [2006]. Disponível em: <http://www.htmlstaff.org/ver.php?id=1592>. Acesso em: 13 out. de 2007.

POLITI, Jaques et al. **Mineração de dados de meteorológicos associados a atividade convectiva empregando dados de descargas elétricas atmosféricas**. Revista Brasileira de Meteorologia, v. 21, n. 2, p. 232-244, 2006. Disponível em: <http://www.sbmet.org.br/>. Acesso em: 28 abr. de 2007.

PRASS, Fernando S. **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining**. 2004. 71 f. Dissertação (Mestrado em Mestrado em Ciência da Computação) - Área de Concentração de Sistemas de Computação, Universidade Federal de Santa Catarina, Florianópolis.

QUONIAM, Luc et al. **Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil**. Brasília, 2001. Disponível em: <http://www.scielo.br/pdf/ci/v30n2/6208.pdf>. Acesso em: 11 abr. de 2007.

RAMOS, Rodrigo R.; MACHADO, Alander O.;COSTA, Helder G. Determinação do grau de coerência aplicado a um sistema de classificação para qualidade em serviços. **Teoria dos Conjuntos Aproximativos**. In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 10., 2003, Bauru. **Anais...** Bauru: Unesp, 2003. p. 2-9.

ROMÃO, Wesley; PACHECO, Roberto; NIEDERAUER Carlos A. P. **Planejamento em C&T:** uma abordagem para descoberta de conhecimento relevante em banco de dados de grupos de pesquisa. [S.I.], [2003?]. Disponível em: <http://www.din.uem.br/wesley/Planejamento.pdf >. Acesso em: 11 abr. de 2007.

UML. **Linguagem de modelagem unificada**. [S.l.], 2002. Disponível em: <http://inf.unisul.br/~osmarjr/download/apostila/app_uml2.zip>. Acesso em: 30 mar. de 2007.