relato de pesquisa

APLICAÇÃO EM DATA MINING UTILIZANDO A TEORIA DOS CONJUNTOS APROXIMATIVOS PARA GERAÇÃO DO CAPITAL INTELECTUAL NAS ORGANIZAÇÕES

Oscar Dalfovo[*]
Sidnei Schmitt[**]
Henrique Raboch[***]

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 conhecimento com 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: Data Mining. Teoria dos Conjuntos Aproximativos. Capital Intelectual.

[*] Universidade Regional de Blumenau. E-mail: odalfovo@gmail.com

[**] Universidade Regional de Blumenau. E-mail: sidnei.schmitt@gmail.com

[***] Universidade Regional de Blumenau. E-mail: hraboch@gmail.com

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. Os autores 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 [...] (PESSOA; SIMÕES, 2003, p. 3)

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 apresenta uma aplicação para efetuar a classificação e segmentação de dados, através da mineração de dados utilizando a técnica da Teoria dos Conjuntos Aproximativos (TCA). O objetivo geral deste trabalho foi 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. Mais especificamente pretende-se com este trabalho demonstrar o potencial do DM para classificação e segmentação de dados baseado na TCA; selecionar os perfis mais adequados dos profissionais cadastrados como capital intelectual; e demonstrar graficamente o resultado da mineração dos dados.

2 A IMPORTÂNCIA DA INFORMAÇÃO PARA AS ORGANIZAÇÕES NO USO DE DATA MINING

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; 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.

2.1 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.

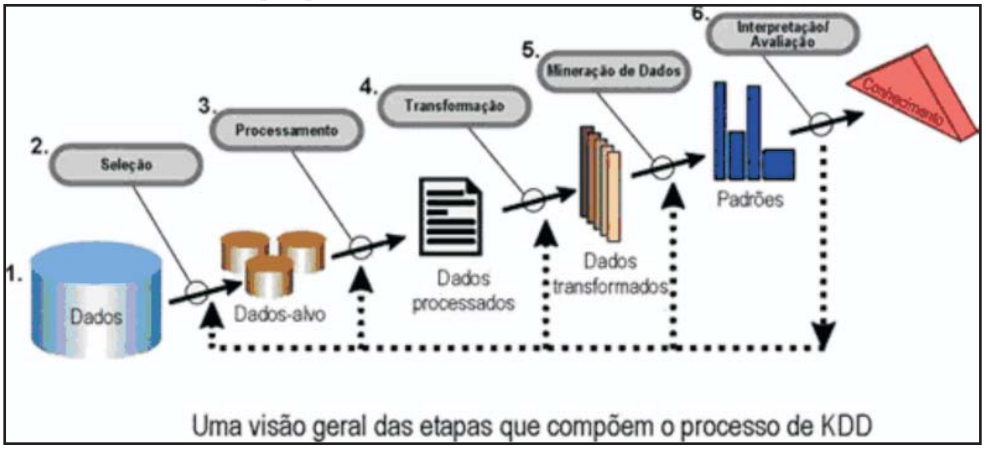

Figura 1: Etapas do processo KDD

Fonte: Fayyad (1996, p. 10).

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.

2.2 A teoria dos conjuntos aproximativos

Os conceitos e a teoria da Teoria dos Conjuntos Aproximativos (TCA) apresentado neste trabalho foi baseado e transcrito em partes a partir de Pawlak (1982) com tradução do autor deste trabalho.

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".

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.

Tabela 1 – Estrutura de uma tabela de informação

	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).

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 ∈ 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.

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

Quadro 1: Domínio do atributo

Fonte: Pawlak (1982).

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

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

Quadro 2: Função de associação do valor ao domínio

Fonte: Pawlak (1982).

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 ⊂ 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.

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

Quadro 3: Função de indiscernibilidade

Fonte: Pawlak (1982).

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.

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

Quadro 4: Descrição do conjunto P elementar

Fonte: Pawlak (1982).

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.

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

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

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

Quadro 5: Definição da aproximação de Y

Fonte: Pawlak (1982).

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 $0 \leq \alpha p(Y) \leq 1$.

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

Quadro 6: Definição da precisão de aproximação

Fonte: Pawlak (1982).

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 $0 \leq \gamma p(Y) \leq 1$.

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

Quadro 7: Definição da qualidade de classificação

Fonte: Pawlak (1982).

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 REDγ(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 COREγ(P). Logo COREγ(P) = $\cap$REDγ(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.

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

Quadro 8: Definição do conjunto Q

Fonte: Pawlak (1982).

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.

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

Quadro 9: Conjunto de regras de decisão para cada classe de decisão Y_j

Fonte: Pawlak (1982).

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.

3 Metodologia e ferramentas

O sistema desenvolvido neste trabalho, como metodologia, foi 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.

Como ferramenta para elaboração deste trabalho, utilizou-se nas especificações os diagramas de casos de uso, diagramas de atividades e diagramas de classes. Já para o desenvolvimento da aplicação utilizou-se a ambiente de programação em JAVA com banco de dados MYSQL.

4 Resultados

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.

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 2.

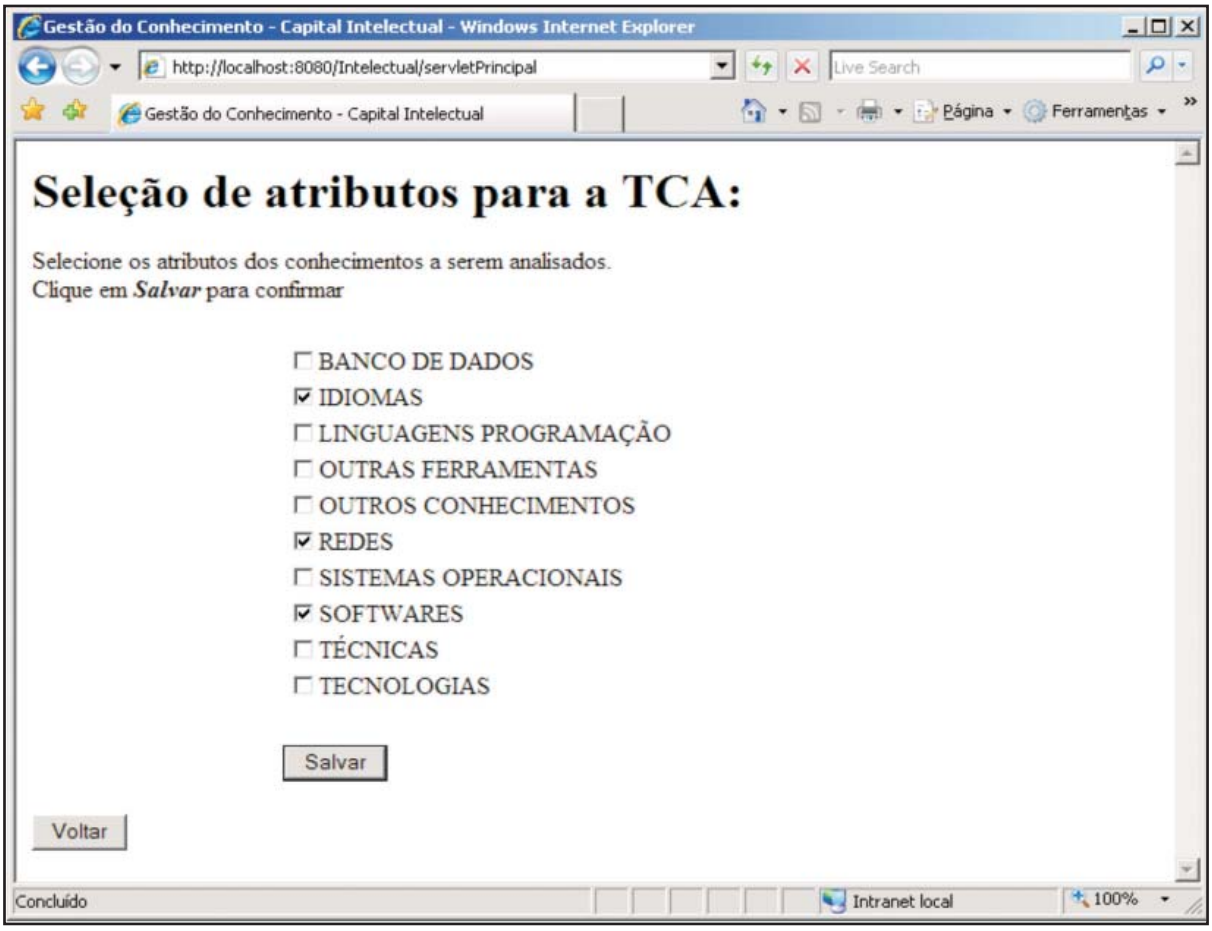

Figura 2: Tela de seleção de atributos para a TCA

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 3.

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.

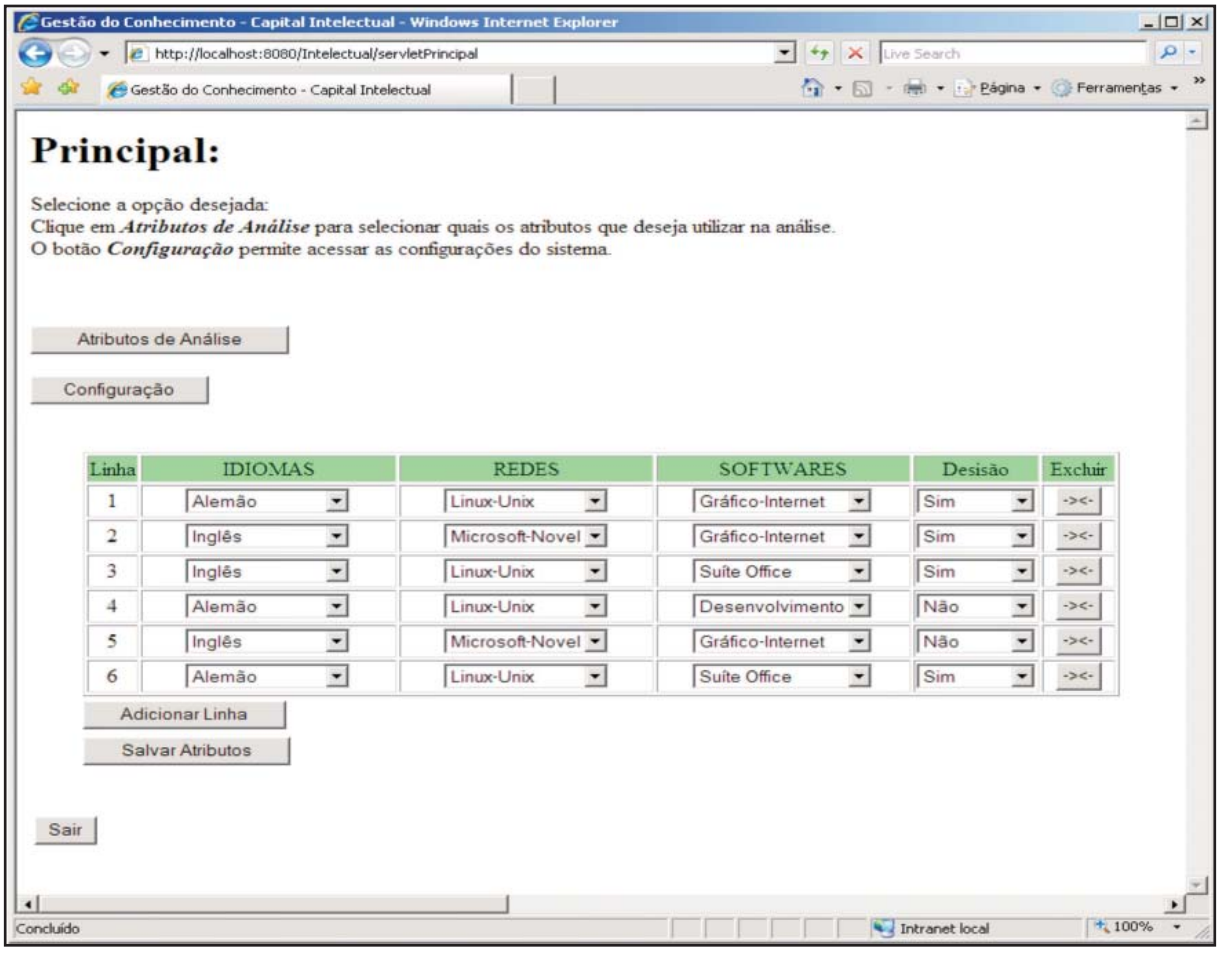

Figura 3: Tela principal com atributos e valores informados

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 10.

- $U / I_P = \{\{1\},\{2,5\},\{3,6\},\{4\}\}$
- $P_Y(sim) = \{1,3,6\}$
- $P^Y(sim) = \{1,2,3,5,6\}$
- 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 10: 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 2.

Tabela 2 – Qualid. de aprox. e conj. P-elementares para os subconjuntos P de condição

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}

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 istto 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 4 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.

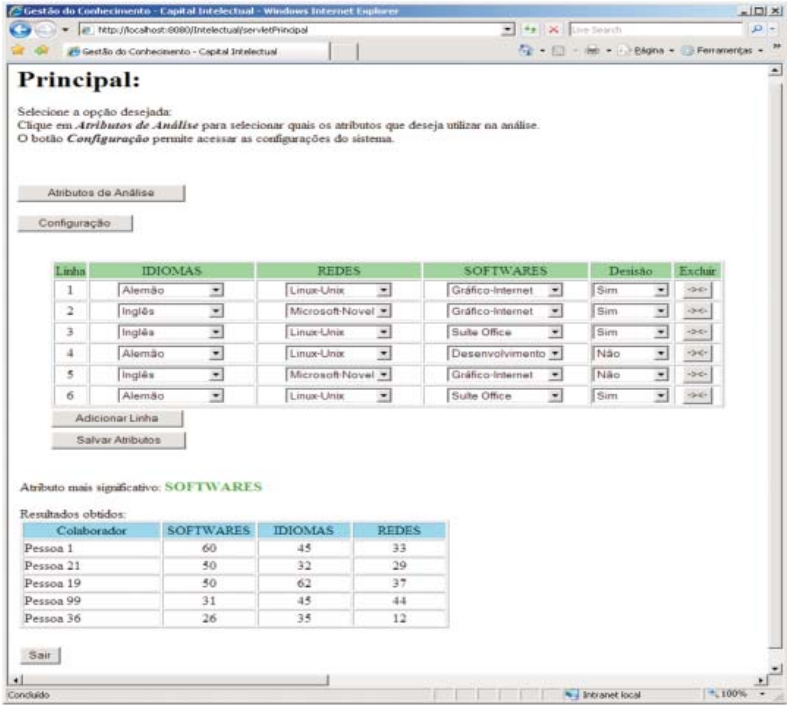

Figura 4: Tela principal com a apresentação dos valores calculados

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.

5 Conclusão

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.

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: 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; 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; aplicar este projeto para seleção de pessoal, baseado no capital intelectual.

DATA MINING APPLICATION USING THE ROUGH SETS THEORY TO GENERATE INTELECTUAL CAPITAL

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.*

Artigo recebido em 23/10/2009 e aceito para publicação em 09/04/2010

Referências

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.

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.

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.

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.I.], 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.

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.