

APLICAÇÃO DE TESTE DE HIPÓTESE E INFERÊNCIA DE REDES BAYESIANAS PELA UTILIZAÇÃO DA GQM EM SISTEMAS DE INFORMAÇÃO

APPLICATION OF BAYESIAN NETWORKS HYPOTHESIS TESTING AND INFERENCE THROUGH THE UTILIZATION OF GQM IN INFORMATION SYSTEMS

Denis Ávila MONTINI

Doutorando em Ciências da Computação (ITA)
denisavilamontini@yahoo.com.br

Mauri Aparecido de OLIVEIRA

Doutor – Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP), São Paulo – SP – Brasil
mauriao@usp.br

Alessandra de Ávila MONTINI

Doutora – Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP), São Paulo – SP – Brasil
amontini@usp.br

Recebido em 07/2011 – Segundo Recebimento em 11/2011 – Aprovado em 04/2012

RESUMO

O objetivo deste trabalho é realizar a aplicação de uma Rede Bayesiana para controle de qualidade de *softwares* em uma Linha de Produção de *Software*. Para restringir o alcance desta modelagem, foram utilizados conceitos da técnica de Objetivos Questões e Métricas (Goal, Question, Metric – GQM). Dessa forma, o conceito de Rede Bayesiana pôde ser relacionado ao caso particular de um Grafo Acíclico Direcionado – GAD, no qual cada um dos nós representam variáveis discretas aleatórias e os arcos direcionados as relações de causa e efeito entre tais variáveis. Considerando uma Rede Bayesiana como um instrumento gráfico que se restringe a uma estrutura de dados relacionados, selecionou-se o aplicativo de computador Netica como *software* para implementação das Redes Bayesianas. Os resultados, apesar de preliminares, remetem a uma corroboração com as premissas de planejamento apresentadas pelas técnicas propostas por Vasilis (2000), pela GQM e pelas Redes Bayesianas, compondo assim um sistema de informação (SI) que comporta as múltiplas hipóteses e múltiplos controles necessários na Fábrica de *Software*.

Palavras-chave: Rede Bayesiana – Sistema de Informação – *Goal Question and Metrics* – Fábrica de *Software* – *Capability Maturity Model*.

ABSTRACT

The aim of this work is to make an application of a Bayesian Network to control the quality of software in a Software Production Line. To restrict the scope of this modeling technique, concepts of the technique Goal, Question, Metric – GQM - were used. Thus, the concept of Bayesian Network could be related to the particular case of a directed acyclic graph - DAG, in which each of the nodes represents discrete random variables and the directed arcs, the

relations of cause and effect among these variables. Considering a Bayesian Network as a graphic tool that is restricted to a structure of related data, the computer application Netica was selected as the software for the implementation of the Bayesian Networks. The results, although preliminary, point to corroboration of the planning assumptions made by the techniques proposed by Vasilis (2000), by the GQM and by the Bayesian networks, thus making up an information system (IS) that encompasses the multiple hypotheses and multiple controls needed in the Software Factory.

Keywords: Bayesian Network – Information System – Goal, Question, Metric – Software Factory – Capability Maturity Model.

INTRODUÇÃO

Os elementos conceituais de Redes Bayesianas (RB) (NEAPOL, 1990) são apresentados nesta pesquisa, de forma a aplicar a teoria em um contexto de Fábrica de *Software*, em uma das Fases de Teste de produção, assim como exemplificar um processo de confecção de um aplicativo de *software* utilizando a ferramenta Netica (NETICA, 1990). A identificação da rede assim como os fundamentos envolvidos foram resolvidos por intermédio do emprego da técnica de Objetivos Questões e Métricas (*Goal Question and Metrics – GQM*) (BASILI, 1988).

A capacidade de representação gráfica da Teoria

de Redes Bayesianas tem sido cada vez mais explorada em estudos práticos, como em diversas áreas de Engenharia de *Software* e produção devido à aplicabilidade da solução codificada pela visão Bayesianiana. Nas atividades de manufatura, existe a noção de previsão do comportamento da produção, assim como o cálculo de probabilidades que tornam possíveis manipulações estatísticas. Neste trabalho, o cálculo das probabilidades é realizado por meio do Teorema de Bayes para o uso da *GQM* e auxílio ao entendimento das Redes Bayesianas (RBs).

Para se representar simbolicamente um sistema de informação sobre o conhecimento de um fenômeno (LAUNDON, 2007 e O'BRIEN, 2006), esquemas de árvores são utilizados pelas RBs. A modelagem do problema de pesquisa é desenvolvida por processos de formalização matemática, nos quais são incorporados elementos do mundo real.

As variáveis selecionadas foram obtidas por meio de um sistema especialista (STEIN, 2000) e apresentam dependência estocástica, resultando em uma base de dados especialista para as RBs, organizada por fatos e regras específicas. A organização destes conceitos fundamenta a formação de sistemas especialistas probabilísticos contidos em bases de dados classificados *bayesianamente*.

Quadro 1 – Um caso prático de aplicação da *GQM*

GQM							
Objetivo de Negócio	Assegurar que todos os defeitos são corrigidos antes de o software ser liberado para uso.						
Objetivo da Medição	Remover todos os defeitos do software pelo processo de teste.						
Perguntas	7-8 Qual é o número Planejado e Realizado de requisitos testados?		5-6 Qual é a cobertura Planejado e Realizado dos testes?		3-4 Quantos defeitos Unitário, Integrado, Sistema temos atualmente?		1-2 Qual é o número de defeitos no Sistema?
Medidas	8 - Número de requisitos Testados Quantitativos	7 - Número de requisitos Testados Qualitativos	5 - Número de requisitos Testados Quantitativos	6 - Número de requisitos Testados Qualitativos	4 - Número de defeitos Quantitativos	3 - Número de defeitos Qualitativos	1 - Número de defeitos por status Quantitativo 2 - Número de defeitos por status Qualitativo

O quadro 1 representa um estudo prático de uma *GQM*, e a explicação de sua estrutura se dá da seguinte forma:

Objetivo de Negócio: Assegurar que todos os defeitos são corrigidos antes de o *software* ser liberado para uso.

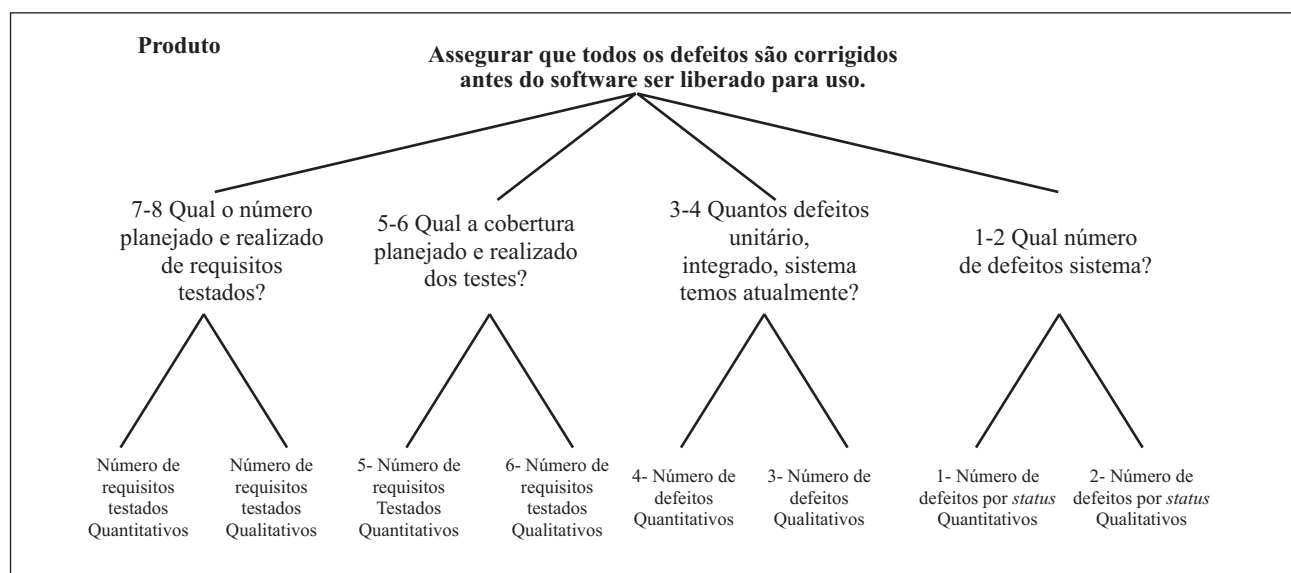
Objetivo da Medição: Remover todos os defeitos do *software* pelo processo de teste.

Perguntas: “7-8 Qual é o número Planejado e Realizado de requisitos testados?”, “5-6 Qual é a cobertura Planejado e Realizado dos testes?”, “3-4

Quantos defeitos Unitário, Integrado, Sistema temos atualmente?” e “1-2 Qual é o número de defeitos Sistema?”.

Medidas: “8 - Número de requisitos Testados Quantitativos”, “7 - Número de requisitos Testados Qualitativos”, “5 - Número de requisitos Testados Quantitativos”, “6 - Número de requisitos Testados Qualitativos”, “4 - Número de defeitos Quantitativos”, “3 - Número de defeitos Qualitativos”, “1 - Número de defeitos por status Quantitativo” e por último “2 - Número de defeitos por status Qualitativo”.

Figura 1 – Um caso prático de aplicação da *GQM* através da Teoria dos grafos.



As RBs e a *GQM* podem ser compostas estruturalmente por duas partes, sendo uma qualitativa e outra quantitativa (SIMÕES, 2001 e BASILI, 1988). A organização do sistema de informações pôde ser dividido em duas partes (quantitativa e qualitativa), e com a aplicação da *GQM* pôde ser visualizado esquematicamente o resultado da implementação desta teoria no quadro 1. Um grafo é um modelo gráfico (BIGGS, 1976), que constitui um caso particular de grafo acíclico direcionado, GAD (PEARL, 1988), no qual as variáveis são representadas por nós, e os arcos que os ligam significam dependências diretas entre essas variáveis. Ele

é acíclico no sentido de que não há, no grafo, arcos ligando a saída de um nó à própria entrada. Na figura 1, apresenta-se a resolução de um Grafo implementado com a solução da *GQM* do quadro 1.

O trabalho apresenta como as Redes Bayesianas foram propostas, definidas e construídas, assim como o processo de decisão e as regras de inferências para a sua utilização. As Redes Bayesianas propostas são mostradas como um exemplo no diagnóstico da Fase de Teste de um produto, sendo que o processo de desenvolvimento deste conceito foi resolvido com a ferramenta Netica (NETICA, 1990).

1 METODOLOGIA

O sistema de apoio à decisão ao gerenciamento de projeto tem a intenção de ser uma ferramenta utilizada no processo de triagem à contingência de problemas no projeto, com o intuito de disciplinar as operações humanas a fim de diminuir os defeitos residuais do processo resultante da atividade de construção de *software*. Para a formulação de um processo de planejamento e controle eficaz, é necessário identificar as formas de raciocínio e mapas mentais que os profissionais de desenvolvimento de *software* precisam ter durante a construção. Esta organização de ações e procedimentos deve visar à viabilização de conhecimentos a respeito do projeto para a tomada de decisão.

O conhecimento sobre o Planejamento e Controle da Produção (PCP), de projetos aplicados a Linhas de Produção de *Software*, por exemplo, Células de Manufatura (MONTINI, 2005, 2006a, 2006b, 2006c, 2006d, 2006e, 2007 e 2008), estão ligados aos aspectos de conhecimento empírico. O conhecimento necessário para o dimensionamento deve estar ligado às evidências disponíveis fundamentadas no conhecimento de engenharia de *software* e no conhecimento formal. Deve-se enfatizar que informações são os resultados do processamento, manipulações e organizações de dados de tal forma que representem um acréscimo ao conhecimento que o usuário final recebe (CUNHA, 2008). O conhecimento empírico pôde ser sistematizado por intermédio de informações constantes em bases de dados históricas.

O conhecimento baseado em evidências, por sua vez, é dado pelas evidências acumuladas ao longo do tempo e sumarizadas por processos estatísticos adequados. O conhecimento formal baseia-se na literatura de engenharia, disponibilizada por periódicos, livros e outras formas de armazenamento, incluindo ainda processos de dedução (O'BRIEN, 2006).

Tomando como premissa a existência de

uma base histórica para a calibração dos dados, contendo informações estruturadas e não estruturadas, pode-se obter, por meio de técnicas estatísticas, as características e a taxa de acerto dos analistas programadores (CMMI, 2006). A restrição do escopo varia de acordo com os objetivos de negócio e a forma com que as atividades de programação têm seguido as diretrizes da empresa. Além disso, o contrato que firma as bases técnicas do projeto (as quais podem ser modeladas por meio de regras de produção da *GQM* e submetidas à inferência Bayesiana) é usado para a formação dos Sistemas de Informações Gerenciais, SIGs (O'BRIEN, 2006).

Neste cenário que se restringe à pesquisa operacional, os defeitos incutidos durante a Fase de Construção podem ser tratados por intermédio de técnicas de descobrimento de conhecimento em base de dados, *Knowledge Discovery in Database* (KDD) (SILVA, 2002), baseadas neste estudo de caso na Modelagem Bayesiana (BARRETO, 2002). A resolução deste problema tem como alicerce as técnicas de Inteligência Artificial, IA, para apoio à tomada de decisão (RUSSEL, 1995).

Uma definição para IA que pode ser utilizada vem da Enciclopédia Britânica: "Habilidade de se adaptar afetivamente ao ambiente, realizando uma mudança em si mesmo ou pela mudança do ambiente ou encontrando um novo". Ou seja, a conclusão é que é praticamente impossível definir exatamente o que é inteligência. Dessa forma, o conceito de IA é considerado tão amplo que alguns pesquisadores dividem o conceito de IA em duas classes: a IA Fraca e a Forte.

A IA Fraca refere-se a um tipo de característica "pensante" que pode ser adicionada aos computadores para torná-los mais úteis aos humanos. Exemplos mais claros de IA Fraca são os Sistemas Especialistas (SE). No caso da IA Forte, os computadores podem mimetizar processos do pensamento humano. Em outras palavras, ela tenta modelar o processo do cérebro. Atualmente, podemos dizer que a IA

Forte é mais adequada aos interesses de pesquisa do que às aplicações em negócios.

De acordo com Sage (1990), o objetivo da inteligência artificial (IA) é o desenvolvimento de paradigmas ou algoritmos que requeiram máquinas para realizar tarefas cognitivas, nas quais os humanos são atualmente melhores.

A cognição pode ser definida como sendo o processo de conhecer e, mais precisamente, o processo de estar consciente, sabendo, pensando, aprendendo e julgando. O estudo da cognição está ligado às áreas da psicologia, linguística, ciência da computação, neurociência, matemática, etologia (estuda o comportamento animal) e filosofia. A partir dessa definição de cognição, parece difícil conceber que seja possível encontrar tais características em algo que não seja um ser humano e muito menos em uma máquina. Isso porque a cognição está dentro da mente humana. Dessa forma, a missão concentra-se em entender a relação da mente com o mundo.

A *GQM*, também conhecida como sendo o conjunto das regras de pesquisa qualitativa definidas por objetivos, questões e métricas, tem como finalidade modelar o raciocínio humano para o desenvolvimento de sistemas computacionais visando à tomada de decisão racional em ambientes em que haja incertezas (BASILI, 1988). Nela, são modelados os conjuntos de indicadores, os quais se constituem de funções. A *GQM* pode comportar qualquer tipo de análise e, por este motivo, testou-se o uso de RBs e foram mapeadas com um valor escalar entre “0” e “1”, a fim de indicar o seu grau de pertinência a este conjunto. São também estabelecidas as regras de inferência, que normalmente são extraídas de especialistas humanos ou valores numéricos. A partir de sinais, sintomas e queixas apresentados pelo paciente, o processo de inferência determina como as regras serão ativadas e combinadas.

Para que se possa implementar a *GQM*, é necessário discutir os métodos de RBs

empregados, desenvolvendo os elementos da Teoria de Bayes. Adotou-se nesta pesquisa a pré-condição que relaciona uma RB à Teoria dos Grafos, restringindo-se aos grafos acíclicos orientados. Esta condição foi adotada para que os nós representassem as variáveis aleatórias, e o arco unindo dois nós representassem a dependência probabilística entre as variáveis quantitativas e qualitativas associadas. Esta ideia é mostrada na figura 1.

A representação gráfica explicita as relações de dependências e é uma ferramenta poderosa no processo de verificação e na aquisição de conhecimentos (CMMI, 2006). Quando uma RB está ligada a uma base de dados, esta pode ser construída de forma a se tornar adaptativa e se atualizar conforme as probabilidades estimadas a partir dos dados armazenados, sendo chamadas de Redes Bayesianas Adaptativas (RBA). Esta capacidade oferece a possibilidade de implementar técnicas de KDD (SILVA, 2002).

Com esta composição de teorias, propõe-se que a Teoria de Bayes é útil e aplicável no tratamento de múltiplas hipóteses e múltiplas evidências restritas por um delimitador organizacional. Esta abordagem é aderente ao conceito probabilístico de independência entre eventos distintos, nos quais os dados sobre um evento não trazem informação adicional sobre outro evento. Assim, dado que um evento ocorreu, isto não traz informação para o evento 2, e vice-versa. Isto, no entanto, é afetado por ocorrências eventuais na estrutura conhecida, corrigindo a probabilidade a partir de cada novo registro de êxito obtido do fenômeno estudado.

2 NOÇÕES DE PROBABILIDADE

Utilizar a Teoria das Redes Bayesianas implica o uso de conceitos básicos de probabilidade. O elemento atômico da linguagem de probabilidade é o “evento”, o qual pode ser definido como um estado de alguma coisa em algum intervalo de tempo. O exemplo clássico de um dado

“evento” é o lançamento de uma moeda no qual se pode esperar somente um resultado entre dois possíveis: um deles “Cara” e o outro “Coroa”. A representação matemática seria “1” evento, entre “2” possíveis; portanto este evento é apresentado por uma fração: 1 de 2 ou 1/2 ou ainda 50% de chances para qualquer resultado. Refletindo sobre o assunto, quando ocorre um evento e o conceito de sua probabilidade tem-se a medida da frequência em que ocorre.

No metaconceito do modelo da moeda, uma probabilidade de 0,5 é atribuída ao evento de que a moeda irá cair cara. Porém, se for conhecido que no mundo real a moeda está viciada em favor da cara, existe a necessidade de se calibrar o modelo a fim de ajustar o modelo real do metamodelo teórico, e redefinir a certeza e a proporção da probabilidade em favor do lado favorecido e testado no mundo real. A notação da probabilidade é descrita como sendo a probabilidade de “e” dado “ ξ ”. O símbolo ξ representa o estado de conhecimento de quem provê a probabilidade. Exemplo: $P(e|\xi)$.

Uma outra interpretação da probabilidade “e” representa o grau de crença de uma pessoa de que o evento irá ocorrer em um experimento único (RUSSEL, 1995). A interpretação de probabilidade como frequência em uma série de experimentos repetidos é referida como interpretação objetiva que calibra o modelo (CMMI, 2006).

Ao contrário, a interpretação de probabilidade como grau de crença é chamado de interpretação subjetiva ou Bayesiana. Na interpretação Bayesiana, a probabilidade sempre dependerá do conhecimento de quem fornece a probabilidade. Em um sistema especialista real, uma pessoa pode avaliar a probabilidade baseada na informação que ela assume ser verdade.

Para a implementação do sistema especialista, precisou-se utilizar mais um conceito da probabilidade, chamado Probabilidade Condicional (PC). No estudo da PC para o caso do lançamento da moeda viciada, a probabilidade

de que a moeda irá mostrar “Cara” no terceiro lançamento, corrigida pela constatação do vício, descreve-se como sendo a probabilidade do evento dois “e2”, dado que o evento um “e1” é conhecido, além da crença da distorção do “ ξ ”. Sendo assim, escreve-se $P(e2|e1, \xi)$.

A outra maneira de expressar a PC é em termos de “variáveis” em vez de “eventos”. Para cada variável é atribuído um valor a partir de uma coleção de estados, na qual cada “estado” corresponde a algum “evento”. Para o estudo das variáveis, sabe-se da estatística que a variável pode ser discreta, tendo um número finito de estados, ou pode ser contínua. Por exemplo, uma variável discreta pode ser usada para representar as possíveis saídas de um lançamento de uma moeda, “Cara” (“0”) ou Coroa (“1”); já o peso da moeda pode ser representado por uma variável contínua, sendo aproximadamente $2,54 + 0,01$ gramas.

Para as variáveis discretas, pode-se ter os eventos “Cara” e “Coroa”, e cada evento pode ser representado pelos estados “0” e “1”. Conhecendo as variáveis e seus estados, pode-se estabelecer seus relacionamentos. Para a primeira simulação, definiu-se que para o evento: “Cara” = “x” atribui-se um valor inteiro “k” = “0”; “Coroa” = “x” atribui-se um valor inteiro “k” = “1”. Das proposições, é possível afirmar a seguinte sentença: se um evento consiste em “x” = “k”, para a variável “x”, então existe um estado discreto “k, sendo que $x=k$ ”.

2.1 Teorema de Bayes

O Teorema de Bayes atribui à parte quantitativa e à qualitativa uma formação constituída pelo conjunto de probabilidades totais e condicionais, *a priori*, associadas aos nós e arcos existentes no grafo das variáveis contínuas e discretas. A utilização de cada tipo de variável é customizada de acordo com a análise do cenário e necessidade de cada nó individualmente selecionado.

Para facilitar a visualização da escolha do tipo

de variável em cada nó dentro do modelo, é utilizado o método de representação gráfica proporcionado pelas RBs, que permite visualizar a influência da coleta da experiência do analista ao constatar o fenômeno e a atribuição de um valor “discreto” no modelo da RB. A pronta correção da probabilidade em tempo real permite a realização de simulações e análises de sensibilidade, envolvendo alocação de recurso ou suspensão de atividades, assim como a identificação de um impacto no produto decorrente desta decisão aplicada a uma Linha de Produção de *Software*.

As referidas distribuições de probabilidades condicionais, que representam o conhecimento adquirido do especialista ou dos dados, devem ser registradas por meio de algum Teorema ou Corolário matemático. Isso tem sido obtido satisfatoriamente pelo Teorema de Bayes, que é a base para a parte quantitativa das RBs. É importante ao usuário conhecer, ao menos em certo grau, os fundamentos desse teorema, já que isso é importante na Fase de Validação da RB em um processo de classificação.

A utilização da “fórmula das probabilidades das causas ou dos antecedentes” apresenta o revés de que uma má avaliação dos $P(H_i)$ pode tornar sua aplicação discutível (RUSSEL, 1995). A equação da “fórmula das probabilidades das causas ou dos antecedentes” é dada por:

$$P(H_i|e) = \frac{P(H \cap e)}{P(e)} = \frac{P(e|H_i) P(H_i)}{\sum_{j=1}^k P(e|H_j) P(H_j)}$$

Esta fórmula consiste na formalização do Teorema de Bayes por meio da atribuição de valores discretos a cada uma das variáveis. Neste caso, “i” varia de $i = “1”, “2” \dots, “k”$; e para as variáveis discretas “H” varia de “H” = “H 1”, “H 2”, ..., “H k” para cada nó do grafo dentro de um espaço amostral “S”.

A calibração manual da PC ocorre quando um evento “e” é associado ao espaço amostral “S”; a probabilidade deste caso particular ocorrer

constitui a hipótese “Hi”, dada a ocorrência do evento “e”. Para o processo de aquisição do conhecimento, os “H’s” são as hipóteses possíveis e mutuamente excludentes do estado de uma variável. Os eventos “e’s” são entendidos como as evidências provocadas pelo estado “H” da variável.

Um resultado do teste constitui uma hipótese para o uso em RB, e é obtido a partir de uma suposição prévia, ou verificação, de que dois ou mais eventos (“e1”, “e2”, ..., “en”) são independentes. No momento em que um dado fenômeno ocorre, levanta-se uma hipótese “H”, ou seja, o conhecimento sobre a ocorrência do evento “e1” não traz informação adicional sobre a ocorrência de “e2” e vice-versa, portanto:

$$P(H_i|e_1 \cap \dots \cap e_n) = \frac{P(H_i) \prod_{k=1}^n P(e_k|H_i)}{P(e_1 \cap \dots \cap e_n)}$$

descreve a equação para o Teorema de Bayes, sob a hipótese de evidências múltiplas e independentes.

2.2 Atualização bayesiana para novas evidências

As RBs estão inseridas em aplicações de sistemas inteligentes e extração de conhecimento, a fim de capacitar o controle e planejamento de projeto com o abastecimento de dados atualizados por evidências obtidas das realidades percebidas a partir do mundo real do projeto, corrigindo o modelo proposto *a priori* pela correção *a priori*.

Essa capacidade de atualização é proporcionada pelo Teorema de Bayes. Para esta tarefa, considera-se a referida hipótese de independência condicional, que simplifica a atualização bayesiana da rede, bastando que seja adaptada a equação do Teorema de Bayes. De fato, considerando-se uma hipótese H_i e denotando-se uma sequência de eventos, supostamente independentes em relação à “Hi”, por “e1”,

“e2”,...,”en” = “en”, a ocorrência de uma nova evidência “en+1” é representada da seguinte forma:

$$P(H_i/e^n \cap e_{n+1}) = \frac{P(H_i/e^n) P(e_{n+1}/e^n \cap H_i)}{P(e_{n+1}/e^n)}$$

representando o Teorema de Bayes, sob a hipótese de evidências múltiplas e independentes para novas evidências obtidas da realidade percebida a partir do mundo real do projeto, corrigindo o modelo proposto *a priori* pela correção *a priori*. A correção da fórmula descrita mostra que, ao ser calculada a probabilidade condicional da hipótese “Hi”, dado um conjunto de evidências “eⁿ”, isto é, $P(H_i/e^n)$ a probabilidade conjunta das evidências passadas $P(e_1 \cap \dots \cap e_n)$ pode ser desprezada na equação, pois $P(H_i/e^n)$ passa a representar completamente a experiência passada, e sua atualização para um novo evento “e_{n+1}” requer apenas sua multiplicação pela razão

$$\frac{P(e_{n+1}/e^n \cap H_i)}{P(e_{n+1}/e^n)},$$

a qual contabiliza o impacto da nova evidência.

2.3 Classificação e descrição bayesiana

Evoluindo a utilização da proposta da aplicação de RB para data-mining na descrição e classificação, identificou-se a necessidade de que a RB aprenda com determinadas características e parâmetros a partir dos dados. Tal fato é essencial para que esta se adapte ao mundo real, e para corrigir o planejamento e o Planejamento e Controle da Produção, PCP, de projetos aplicados às Linhas de Produção de *Softwares* a fim de produzir o conhecimento pela estruturação de classes ou categorias das probabilidades ligadas aos aspectos experimentais.

De certa forma, a IA pode ser representada pela RB para a necessidade da aprendizagem por intermédio do levantamento das informações necessárias na fase de estruturação dos processos (CMMI, 2006), *a priori* associados aos estados

das variáveis aleatórias envolvidas no domínio. Na figura 1, tem-se o Grafo desenvolvido para atender as necessidades simultâneas da Teoria dos Grafos e da *GQM*, a fim de ser pré-requisito de uma RB com a obrigatoriedade de se ter uma estrutura qualitativa e uma parte quantitativa. Este *constructo* visa possibilitar a implementação do conhecimento e futuro aprendizado induzido pelos dados obtidos. Identificado o modelo, pesquisou-se uma ferramenta que permitisse o teste do conceito, selecionando-se a Netica.

O processo de aprendizado foi calibrado pela ferramenta que fornece uma opção de gerar dados estatisticamente e corrigir o modelo, por meio da intervenção do analista. Para que esta hipótese fosse verificável, foi implementado um modelo da *GQM*, mostrado no quadro 01, a partir da ferramenta Netica para testar o conceito do aprendizado na estrutura gráfica. O aprendizado numérico atualmente é considerado relativamente simples, enquanto o aprendizado da estrutura gráfica, que envolve a definição e orientação das relações de dependência condicional, representa uma limitação da ferramenta pela incapacidade de representar o mundo real.

O comprometimento do modelo até pode ser calculado e corrigido; dessa forma, utilizou-se a abordagem ingênua para o aprendizado da estrutura como estratégia de obtenção de uma RB para a Pesquisa Bayesiana. O enfoque ingênuo assume a mencionada hipótese simplificadora de independência condicional entre as variáveis. De forma a aplicar esse conceito à *GQM*, sejam “K”, “X”, “Y”, e “Z” que compõem quatro variáveis aleatórias discretas. Diz-se que “K” é condicionalmente independente de “X” e “Y”, sendo que a distribuição das probabilidades de “K” e “X” são independentes do valor de “Y”, dado o valor de “Z”.

Esta suposição da *GQM* implica, na prática, que os atributos para as variáveis em questão são condicionalmente independentes, dada uma determinada classe “Ci,” resultando que a probabilidade de se observar o conjunto de

atributos: “e1”,..., “en” é dada pela equação,

$P(e_1 \cap \dots \cap e_n | C_i) = \prod_{k=1}^n P(e_k | C_i)$, da suposição da *GQM* aplicada à classe C_i para quatro variáveis.

Com isso, as probabilidades $P(e_1 | C_i)$, ... $P(e_n | C_i)$ podem ser aprendidas facilmente a partir de uma amostra de estimação. Utilizando a fórmula das probabilidades das causas ou dos antecedentes do Teorema de Bayes, atribuindo-se ao caso ou indivíduo a classe de maior probabilidade posterior, implica atribuir-lhe a classe de maior valor dada pela fórmula, $P(C_i) \times \prod_{k=1}^n P(e_k | C_i)$.

Ou seja, a classe de maior valor para a *GQM* aplicado à classe C_i para quatro variáveis. Alguns estudos de comparação foram informados pela literatura e dão conta de que o classificador ingênuo, para certos domínios, é comparável em termos de resultados com as Redes Neurais e as Árvores de Decisão. Contudo, há ocasiões em que a rigidez da hipótese de independência condicional não funciona a contento, e ela tem de ser relaxada, requisitando aprendizado generalizado, de forma a produzir resultados mais próximos da realidade.

O Treinamento de RB eleva bastante o grau de complexidade do aprendizado de estruturas, pois todas as relações de dependência/independência passíveis de ocorrer a partir de uma distribuição de probabilidade conjunta “P”, são, a princípio, passíveis de superposição no grafo, e, portanto devem ser investigadas. Na aprendizagem generalizada, identifica-se a inserção do modelo em um campo em aberto da pesquisa, que requer exaustivos estudos e pesquisas envolvendo elevado nível de complexidade e de processamento computacional. A partir das noções aqui apresentadas das aplicações da RB, e com o objetivo de efetuar um aprendizado automático, foi coletada uma amostra oriunda de um processo formal de uma Fase de Teste do projeto. Uma das regras mais comuns é o Teorema de Bayes, definido pela fórmula,

$$P(X|Y, \xi) = \frac{P(Y|X, \xi)}{P(Y, \xi)} P(X, \xi),$$

para $P(Y, \xi) > 0$, no qual $P(x|\xi)$ é a distribuição da probabilidade de X antes de conhecer Y, e $P(X|Y, \xi)$ é a distribuição de probabilidade de X depois de conhecer Y. Estas distribuições são algumas vezes chamadas respectivamente *a priori* de X e *a posteriori* de X. Este teorema auxilia na alteração da probabilidade a partir do surgimento de uma nova evidência.

3 AS REDES BAYESIANAS RESTRINGIDAS PELA GQM DE TESTE

O universo delimitado pela *GQM* definiu o conjunto de variáveis discretas a serem modeladas. A RB modelada pela *GQM* definiu as relações entre variáveis de interesse do estudo dentro do domínio de PCP de projetos da Linhas de Produção de *Software*. A representação consiste em um conjunto de distribuições de probabilidades locais, combinado em uma estrutura de árvores que permite um conjunto de asserções de independência condicional para a construção de uma distribuição global a partir dessas distribuições locais.

A seguir segue a delimitação das medidas a serem modeladas com base no quadro 1, e pelas regras de “2” e “3” foram estabelecidas as regras de inferência para a construção do quadro 2. Para ilustrar a representação, considera-se o domínio de uma inspeção da Fase de Teste de um componente de *software* embarcado em tempo real. O primeiro passo para construir uma Rede Bayesiana é decidir quais são as variáveis e estados do modelo. Utilizou-se a definição da *GQM* para determinar o domínio, e o resultado da inferência estudada foi: “Assegurar que todos os defeitos são corrigidos antes de o *software* ser liberado para uso”. O segundo passo foi estabelecer para cada uma das perguntas da *GQM* uma árvore em forma de grafo e aplicar os fundamentos do Teorema da Rede Bayesiana para a criação do modelo. Após esta última etapa ser implementada, foram obtidos os resultados apresentados no quadro 2.

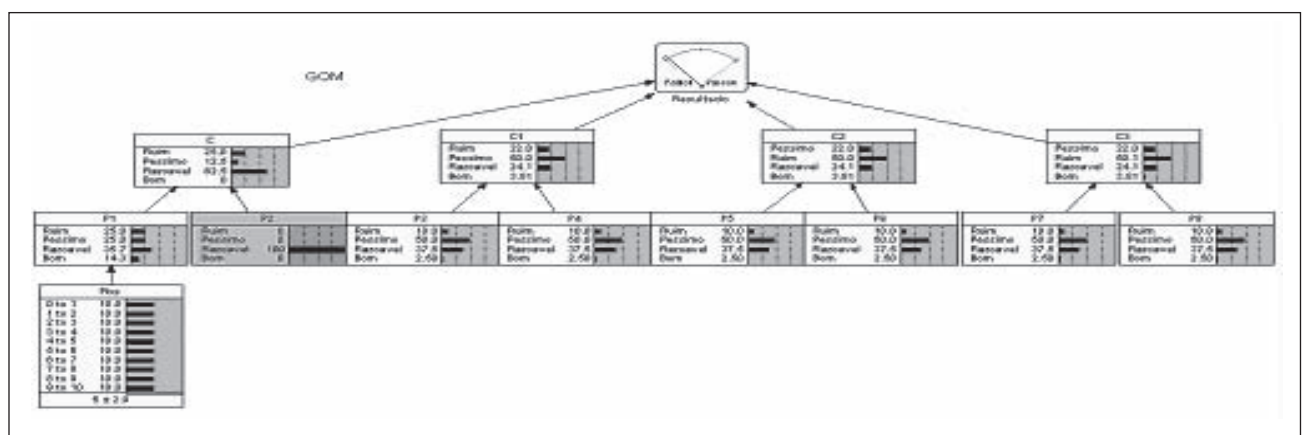
Quadro 2 – Regras de Inferências da *GQM* - Grafo - Bayesiana.

- 1-1 - Qual número de defeitos Sistema? com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-1-1 – Quantitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-1-1-1 – Entradas de Dados 01: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
 - 1-1-2 – Qualitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-1-2-1 – Entradas de Dados 02: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
- 1-2 - Quantos defeitos Unitário, Integrado, Sistema tem-se atualmente? com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-2-1 – Quantitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-2-1-1 – Entradas de Dados 01: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
 - 1-2-2 – Qualitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-2-2-1 – Entradas de Dados 02: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
- 1-3 - Qual a cobertura Planejado e Realizado dos testes? com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-3-1 – Quantitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-3-1-1 – Entradas de Dados 01: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
 - 1-3-2 – Qualitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-3-2-1 – Entradas de Dados 02: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
- 1-4 - Qual o número Planejado e Realizado de requisitos testados? com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-4-1 – Quantitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-4-1-1 – Entradas de Dados 01: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.
 - 1-4-2 – Qualitativo: com os estados Péssimo, Ruim, Razoável e Bom;
 - 1-4-2-1 – Entradas de Dados 02: com os estados 0 to 1, 1 to 2, 3 to 4, 5 to 6, 7 to 8, 9 to 10.

O terceiro passo foi o processo de implementação da modelagem para construir uma Rede Bayesiana por intermédio de uma ferramenta especializada, direcionada para um grafo acíclico para manter as características de inserção de valores para qualquer nó do grafo, respeitando a Teoria da Independência Condicional. Ao implementar

todas estas restrições, tem-se uma estrutura da Rede Bayesiana para novas evidências e o Teorema de Bayes, agora sob a hipótese de evidências múltiplas e independentes percebidas a partir do mundo real do projeto. O passo final foi definir as distribuições locais para cada estado. O resultado pode ser visto na figura 2.

Figura 2 – Exemplo de Redes Bayesianas



4 APLICAÇÕES

A aplicação desta RB foi modelada para as necessidades de controle e planejamento de produção de *software* para a aquisição de conhecimento do comportamento da Linha de Produção de *Software*. A consequência natural deste fato é a melhora do dimensionamento das atividades nos próximos processos. A aquisição de conhecimento tem várias aplicações, tais como: engenharia de produção de *software*, gerenciamento de *software*, planejamento de processos de manufatura, etc.

Em 1991, o SEI (*Software Engineering Institute*) publicou o *Capability Maturity Model* para *Software* (CMM ou SW-CMM), um modelo para julgamento da maturidade dos processos de construção de *software* de uma organização e para identificar as principais práticas que são necessárias para aumentar a maturidade destes processos.

O CMMI, também desenvolvido pela SEI, é um modelo de qualidade criado para integrar engenharia de *software*, engenharia de sistemas e desenvolvimento integrado de produto e processo.

Nesse trabalho, a aplicação resolveu a questão de aderência de relacionamento entre indicadores da *GQM* com os grafos acíclicos da Rede Bayesina para determinar um indicador de tomada de decisão sobre a qualidade dos produtos testados durante a Fase de Homologação na fabricação de *software*. Para isso, foi realizado um estudo de caso de um componente de *software* embarcado em tempo real produzido em C++, no qual havia a suspeita de haver defeitos após a Fase de Testes. A técnica então foi utilizada na Fase de Teste de aceite do usuário, *Users Accept and Test – UAT*.

O *software* Netica também possibilitou a implementação da Teoria da Decisão, incluindo-se as Redes Bayesianas e os nós de decisão, ou seja, as ações a serem tomadas, possibilitando determinar se o planejamento e a execução estavam dentro de um padrão aceitável. O modelo derivado da *GQM* inclui mais de 70 casos de testes distribuídos nos 15 nós Bayesianos, sendo que a ferramenta possibilita a inserção de dados ou variáveis discretas em um nó específico quando um fenômeno ocorre.

5 RESULTADOS OBTIDOS

5.1 Fase inicial

Na fase inicial, foi realizado um estudo teórico sobre a *GQM* e Redes Bayesianas para entender seus conceitos básicos e também a sua aplicação prática, principalmente no campo da confiabilidade. Após feito esse estudo teórico, procedeu-se uma análise das ferramentas existentes para a inferência de Redes Bayesianas, que permitissem o teste da hipótese da restrição do escopo pela inferência da *GQM*.

5.2 Resultado

O resultado da pesquisa foi a calibração da tabela de probabilidades para as Inferências Bayesianas na Linha de Produção pelo *software* Netica. A ferramenta tem grau de usabilidade que permite a visualização tanto da Rede por meio do grafo, figuras 3 e 4, assim como também da tabela de probabilidades e decisão. Dessa forma, as inferências são calculadas rapidamente pelo departamento de Engenharia de Produção da empresa estudada (TCS, 2003).

Figura 3 – Interface gráfica com os Resultados do Netica.

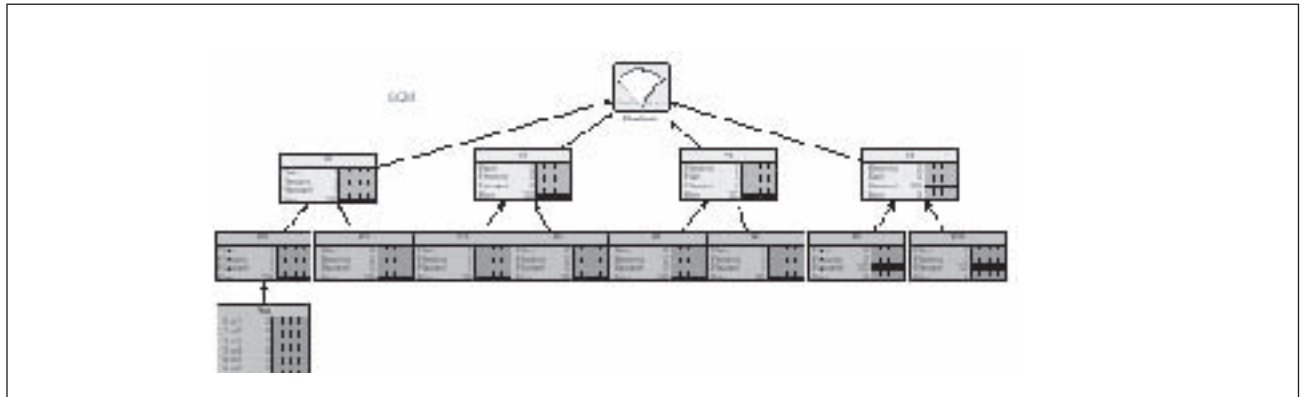


Figura 4 – Interface tabular com os Resultados do Netica.

Netica - [C Table (in net N008_05_08_teste1)]

File Edit Table Window Help

Node: C

Chance % Probability

Apply Okay

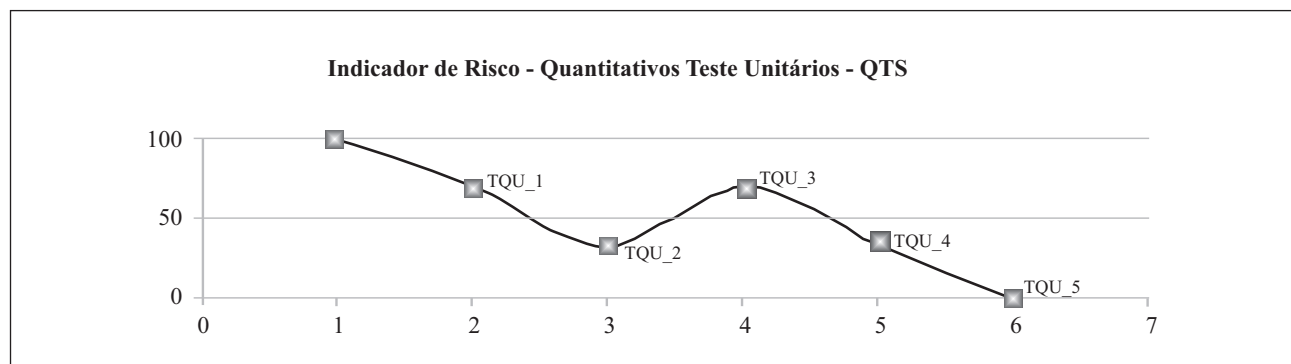
Reset Close

P1	P2	Ruim	Pessimo	Razoavel	Bom
Ruim	Ruim	100.00	0.000	0.000	0.000
Ruim	Pessimo	75.000	25.000	0.000	0.000
Ruim	Razoavel	75.000	0.000	25.000	0.000
Ruim	Bom	75.000	0.000	25.000	0.000
Pessimo	Ruim	75.000	25.000	0.000	0.000
Pessimo	Pessimo	100.00	0.000	0.000	0.000
Pessimo	Razoavel	25.000	50.000	25.000	0.000
Pessimo	Bom	15.000	75.000	10.000	0.000
Razoavel	Ruim	0.000	0.000	100.00	0.000
Razoavel	Pessimo	0.000	75.000	25.000	0.000
Razoavel	Razoavel	0.000	0.000	100.00	0.000
Razoavel	Bom	0.000	0.000	100.00	0.000
Bom	Ruim	75.000	0.000	25.000	0.000
Bom	Pessimo	0.000	25.000	75.000	0.000
Bom	Razoavel	0.000	0.000	100.00	0.000
Bom	Bom	0.000	0.000	0.000	100.00

Este modelo foi submetido às mesmas evidências e foi constatado pelo processo de Verificação e Validação (CMMI, 2006). Logicamente o modelo tem uma forma de lidar com o conhecimento de forma distinta, porém os resultados obtidos foram bastante aproximados da realidade. As

evidências obtidas, suas respectivas formas e resultados, possibilitaram encontrar três erros em testes unitários com alta severidade no produto final, “0” zero erro nos testes de sistemas e “0” zero erro na Fase de Teste integrados em dois ciclos de testes de homologação.

Figura 5 – Evolução da probabilidade do risco, baseada nos erros encontrados em testes: Unitário, Integrado e de Sistema.



Na figura 5 tem-se os valores linguísticos apresentados como evidência e, após o processo de inferência da GQM com a da Rede Bayesiana, os resultados obtidos foram três erros com severidade alta em 70 casos possíveis, sendo a taxa de erro com indicador de risco de 4,29%, o que possibilitou a intervenção e correção dos erros encontrados.

6 DISCUSSÃO E CONCLUSÕES

Este trabalho apresentou a implementação de uma RB baseada em um processo especialista para um sistema de tempo real, utilizado para controle de defeitos em componentes de *software* de computador.

A figura 5 apresenta o controle da probabilidade a partir dos defeitos encontrados no componente do *software* de computador desenvolvido em laboratório, com o monitoramento do teste e da probabilidade de encontrar um erro ao longo do tempo na Fase de UAT. Os dados foram coletados de hora em hora por meio das atividades de inspeções de qualidade, sendo que

estas atividades são rotineiras (CMMI, 2006). A tendência da probabilidade de se encontrar um erro tem que tender a zero até a fase de entrega ao cliente. Foi constatada esta tendência por meio da coleta de dados que possibilitou encontrar os defeitos na Fase de Teste e Aceite de Usuário.

Os resultados, apesar de preliminares, remetem a uma corroboração com as premissas de planejamento apresentadas pelas técnicas propostas por Vasilis, GQM, Teorema de Bayes e RB, compondo, assim, um sistema de informação que comporta as múltiplas hipóteses e múltiplas evidências necessárias na fábrica de *Software*. Algumas divergências podem ocorrer devido ao fato da falta de refinamento do modelo para reproduzir a realidade.

Caso se considere que o modelo, no qual grande parte dos erros de testes esteja acertadamente representada, seja um sistema de apoio à decisão de projeto de *software* baseado na GQM e Bayes, este pode se tornar uma eficiente ferramenta no processo de verificação e validação da Fase de Teste e Construção de *software*. A fundamentação deste sistema de controle

também pode viabilizar a construção de sistemas especialistas de IA, pois constitui uma estrutura estável e reproduzível de conhecimento, a respeito do comportamento de uma Linha de Produção de Célula de Manufatura (MONTINI, 2005). A utilização de ferramentas especializadas, como o Netica, facilita a implementação da teoria bayesiana, e traz a possibilidade da aplicação de conceitos não muito utilizados em computação, principalmente conceitos de definição de critérios integrados à gestão como a *GQM*, aplicados a questões de confiabilidade de testes em fábricas de *softwares*.

REFERÊNCIAS

BARRETO, A.; VIEIRA, R. C.; NASSAR, S. M. *Redes Bayesianas e Produção de Conhecimento: uma abordagem de data-mining em dados de um concurso vestibular*. In: 35º Reunião Regional da ABE/SOBRAPO, Florianópolis-SC, 2003 e V Escola Regional da Sociedade Brasileira de Computação (SBC) - Centro-Oeste, 2002.

BASILI, V. R.; ROMBACH, H. D., Goal-Question-Metrics Paradigm, *IEEE Transactions on Software Engineering*, 1988.

BIGGS, N. L.; LLOYD, E. K.; WILSON, R. J. Graph Theory 1736–1936. *História da Teoria dos Grafos, com reprodução de artigos clássicos*. Clarendon Press, Oxford, 1976.

CMMI® for Development (CMMI-DEV) Version 1.2, *CMU/SEI-2006-TR-008*, Carnegie Mellon University. Disponível no site: <http://www.sei.cmu.edu/publicatons/documents/06.reports/06tr008.html>. Acesso em Agosto de 2006.

CUNHA A.M. *CE-245 Tecnologias da Informação*, ITA – Instituto de Tecnologia da Aeronáutica. Disponível em: <http://www.ita.br/~cunha>. Acesso em Maio de 2008.

LAUNDON, K. C.; LAUNDON, J. P. *Sistemas de Informação Gerenciais*. Pearson Prentice Hall: 7. Ed., 2007.

MONTINI, D. A. *Modelo de indicadores de risco para o orçamento de componentes de software para célula de manufatura*. Denis Ávila Montini. 360p. Dissertação (Mestrado) em Engenharia de Produção – Universidade Paulista (2005).

_____; ALBUQUERQUE, A. R. P. L.; NASCIMENTO, M. R. Strategy for the use of the model of personal software process for the establishment of measurement and analysis for obtain CMMI level 2 in a study of case of a Brazilian software factory. In: *ASEE - 5th ASEE Global Colloquium on Engineering Education*, Rio de Janeiro, Brasil, 2006a.

_____; SPINOLA, M. M.; SACOMANO, J. B.; NASCIMENTO, M. R.; BATTAGLIA, D. Application of model PSP manual and supported by tool in a study of case of Brazilian plant of software. *Revista produção on line*, Florianópolis-SC, Brasil, 2006b.

_____; NASCIMENTO, M. R.; BATTAGLIA, D.; ANDRADE, D. C. Metodologia de Gestão de Vulnerabilidades para Fábrica de Software. In: *WCCSETE'2006 – World Congress On Computer Science, Engineering and Technology Education*, 2006, Itanhaém, São Paulo, Brasil, 2006c.

_____; SEKHAR, C.; NEGRY, T. T.; NASCIMENTO, M. R.; BATTAGLIA, D. Strategy for the use of the model of personal software process for the establishment of measurement and analysis for obtain CMMI level 2 in a study of case of a brazilian software factory. In: *TACTICS Iberoamerica 2006*, Uruguai, Montevideú, 2006d.

_____; MOREIRA, G. S.; VIEIRA, L. A.; BATTAGLIA, D.; GNATIUC, C. E.; CUNHA, A. M. Estudo de caso de uma estratégia de integração de middleware para um serviço SOA de gerenciamento e controle de fábrica de software: TCS – Tata Consultancy Services – Intranet website de Base de Conhecimento Corporativa KnowMax. In: *TACTICS Iberoamerica 2007*, Brasil, São Paulo, 2007.

_____; MOREIRA, G. S.; VIEIRA, L. A.; BATTAGLIA, D.; GNATIUC, C. E.; CUNHA, A. M. Study of case of a strategy of middleware integration for SOA service of administration and control of factory of software: TCS – Tata Consultancy Services – Intranet website Corporative Knowledge Data Base KnowMax. In: *Global TACTiCS 4th Global Conference*, Índia, Kerala, 2008.

NEAPOLITAN, R. *Probabilistic Reasoning in Expert Systems*. John Wiley & Sons, Inc. New York, 1990.

Netica 4.02, Norsys and Netica are trademark of Norsys Software Corp. Copyright © 1990-2007.

O'BRIEN, J. *Sistemas de Informação*. 2. Ed. Editora Saraiva, 2006.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 2. Ed. EUA: Califórnia, 1988.

RUSSEL, S.; NORVIG, P. *Artificial Intelligence - A Modern Approach*. Prentice-Hall, 1995.

SAGE, A.P. *Concise Encyclopedia of Information Processing in Systems and Organizations*. New York: Pergamon, 1990.

SILVA, W. T.; LADEIRA, M. Mineração de Dados com Redes Bayesianas. XXI JAI – Jornada de Atualização em Informática. *Anais do XXII Congresso da Sociedade Brasileira de Computação*, v. 2, p. 235-286, 2002.

SIMÕES, P. W. T. A.; NASSAR, S. M.; PIRES, M. M. S. Sistema de Apoio na Avaliação da Falência do Crescimento Infantil. In: *Congresso Brasileiro de Computação*, Workshop de Informática Aplicada à Saúde, 2001.

STEIN, C.E. *Sistema especialista probabilístico: base de conhecimento dinâmica*. Dissertação (Mestrado em Ciências da Computação) - Programa de Pós-Graduação em Ciências da Computação. UFSC, Florianópolis, 2000.

Tata Consultancy Services, transnacional Indiana no segmento de consultoria e Fábrica de desenvolvimento de software com sede na Índia e atuação no Brasil desde 2003.

VASILIS, S.; SMITHSON, S. Information systems evaluation in practice: a case study of organizational change. *Journal of Information Technology*, v. 15, iss 2, p. 93-105(13). 2000.