DATAMINING: UMA TÉCNICA ESSENCIAL PARA TRATAMENTO DE INDÍCIOS DE INFORMAÇÕES[1]

INTRODUÇÃO

As organizações existem para cumprir o importante papel de coletivizar as ações sociais (Chiavenato, 1990). A fim de tornar exeqüível tal determinação, apoiam-se fortemente na compilação, tratamento e uso da informação para realizações internas e formulação e implementação de estratégias.

Historicamente, para conseguir realizar com sucesso a compilação e tratamento dos dados, aquelas organizações estruturaram sistemas de informação de acordo com as plataformas disponíveis à época, considerando um mercado competitivo e de ampla concorrência (Ein-Dor e Seveg, 1997). Em verdade, a introdução da computação em larga escala nas organizações fez que estas, independentemente do seu porte, se vissem compelidas a requerer apoio em todos os seus níveis de funcionamento: tático, estratégico e, maciçamente, operacional (Furlan et al, 1995).

A evolução desta função organizacional está derivando, contudo, para um deslocamento de ênfase em sua implementação hodierna. De fato, presencia-se atualmente, uma gradual migração do espectro técnico, onde se busca eficiência de especificação e de performance, para um espectro mais estratégico da interpretação e uso da informação.

E nesta nova rota, valoriza-se, sobremaneira, a análise de informações qualitativas, as quais, durante longo tempo, foram relegadas à dimensões menores, em prol do uso de dados quantitativos (Van Maanen, 1995). Obtém-se com isso, novas interpretações e extraem-se conclusões importantes de onde, à priori, existiriam apenas dados históricos ou descritivos.

[1] Uma versão preliminar deste trabalho foi submetida ao Encontro Nacional de Engenharia de Produção, edição 99.

Este trabalho pretende justamente evidenciar como a partir de informações aparentemente banais, de domínio público, podem ser extraídas valiosas informações para uso estratégico, procedendo uma "mineração" de dados (*datamining*). Sua execução reafirma a idéia de que é possível extrair, a partir de indícios informacionais, importantes relações que os sistemas de tratamento estruturados não vislumbram, na maioria das vezes por falhas de modelagem e especificação.

Como espaço para exercitar a viabilidade desta técnica de tratamento de dados, em associação com uma espécie de monitoração ambiental, o pesquisador a aplicaram no processo de extração de significado adicional de um conjunto de dados, a proposição *datamining*, aos registros computadorizados de uso das redes locais de duas grandes universidades brasileiras, que permitiram, explicitamente, que se fizesse este acompanhamento rotineiro, durante 30 dias.

O resultado que se apresentará, pretender exortar as possibilidades da técnica e apesar de estar aplicado a um contexto bem particular, com algumas facilidades adicionais e regalias de observação não comuns a este tipo de função informacional, indicará um largo horizonte para sua aplicação.

Para desenvolver este intento, este relato será estruturado em quatro grandes divisões. A primeira, mais conceitual, procurará enquadrar os grandes temas da tecnologia da informação (TI) grafados como importantes neste tópico introdutório, quais sejam: sistemas de informação (características e modelagem); algumas tecnologias específicas e importantes no contexto de TI e deste exercício (*groupware* e *datamining*); e aplicações correntes ligada às técnicas mencionadas (monitoração ambiental e inteligência competitiva).

A segunda parte consistirá no relato do caso estudado (procedimentos, resultados e discussão), e justificará as ligações estipuladas entre os conceitos e a prática. A parte final discutirá as possibilidades e reafirmará o *datamining como uma técnica essencial o tratamento de indícios informacionais,* amealhando conclusões e indicando tendências de uso e mesmo possibilidades de pesquisas derivadas da presente proposição.

1. TECNOLOGIA DA INFORMAÇÃO

Não se pode negar a disseminação e a importância que a informática atinge nos dias de hoje, em todos os campos de aplicação das atividades humanas. Do caráter eminentemente técnico dos primeiros tempos à revolução provocada pela microcomputação e pelas redes de computadores,

que mimetizam um quase anarquismo, a computação enraizou-se de tal forma nas atividades empresariais que é impensável se idealizar hoje em dia uma tarefa de gestão sem o apoio tecnológico (Dornelas e Hoppen, 1999).

O largo escopo de cobertura deste campo de conhecimentos, que compreende desde elementos de comunicação formal entre autômatos ao reconhecimento automático de voz; que vai da precisão micro-eletrônica às expansionistas provas automáticas de teoremas; que encampa desde facilidades para ensino à distância até às redes neuronais especialistas em deduções inteligentes; fez brotar um novo segmento de estudo que se denomina tecnologia da informação.

Este campo tem forte representatividade no contexto científico-acadêmico atual e comporta diversas correntes e disciplinas de estudo. Algumas delas serão brevemente comentadas a seguir, particularmente aquelas que mais interessam para a composição do trabalho.

1.1. SISTEMAS DE INFORMAÇÃO: ALGUMAS REFERÊNCIAS

Para Davis (1974), sistema de informação é um composto integrado homem-máquina, que fornece informação em apoio às atividades de operação, administração e de tomada de decisão, utilizando-se de *hardware* e *software* de computação, procedimentos manuais, modelos de decisão e banco de dados. Esta visão integra elementos técnicos e também ressalta a utilidade dos sistemas de informação para estender as capacidades organizacionais. Mediante ela, há seguras evidências que muitas das modernas técnicas de gerenciamento e suporte à decisão seriam impraticáveis sem o apoio computacional

Ainda em acordo com aquela definição, o desenvolvimento dos sistemas de informação quer em termos de técnicas, quer em termos de especificação, quer em atendimento às necessidades dos usuários, foi sendo aprimorado ao longo das duas últimas décadas, migrando da cobertura das necessidades operacionais para as funções estratégicas (Meirelles, 1994), passando ao longo desta evolução por grandes dificuldades de entendimento, até chegar ao estágio atual em termos de modelagem, especificação e implementação.

Os indícios desta nova postura são evidentes. Se antes a tecnologia da informação servia a utilizadores relativamente passivos, acostumados ao uso dos sistemas de informação em observância a um rígido repertório de operações definidas por especialistas (Riex, 1995), hoje,

em decorrência da difusão de microcomputadores e redes, como a *internet*, fornece a base para uma modificação fundamental no *status quo* do utilizador.

Agora, este indivíduo equipado com um arsenal tecnológico que lhe disponibiliza *softwares* que ele pode dominar, para, inclusive, ensaiar soluções locais para os seus próprios problemas, e o caso do *datamining* em exame é absolutamente enquadrável como uma solução deste tipo, consegue concretamente uma real capacidade de realizar soluções pessoais, autônomas.

Esta autonomia agora vivenciada pelo usuário criativo (Tapscott e Caston, 1995), ao lado de uma informática tradicional gerida por especialistas, recrudesce hoje em dia no seio de muitas organizações. Ela precisa ser, entretanto, coordenada de modo rápido, a fim de não obstacular o avanço da função informacional, que assume importante papel neste final de século e que sofre o impacto do surgimento das organizações virtuais (Malone et al, 1996), empresas expandidas e em redes (Ives et al, 1991) e geografia de competências (Favier, 1998).

Uma solução aparente para estabelecer um mínimo de controle naquela autonomia, desejável, necessária, mas de algum risco, configura-se na modelagem e na especificação de sistemas de informação, feitas em termos de prototipação e em trabalhos de forma cooperativa. Esta configuração tem por base conceitos de fácil entendimento, larga representação e grande difusão no ramo da tecnologia da informação. Tais conceitos permitem um uso específico, como o *datamining,* preservando padrões globais de administração de dados nas instalações. Presentemente, o trabalho tecerá comentários sobre itens técnicos citados neste último parágrafo.

1.1.1. MODELAGEM DE SISTEMAS DE INFORMAÇÃO

Tardieu et al (Léonard, 1995) definem sistemas de informação como um artefato tecnológico vinculado à noção de objetos de dados, em duas feições: natural e artificial. Segundo aquela definição, um sistema de informação seria um objeto que parece natural, "pois captura a realidade da organização em transformação, comunicante e gerando sua própria memória de informações" (Léonard, 1995, p 11). Todavia, devido à sua construção "feito pelo homem para representar as ações e a memória das organizações" (Léonard, 1995, p. 3), um sistema de informação se assemelharia a um objeto artificial, justamente por imitar a aparência dos objetos naturais.

A fim de emprestar um senso efetivo ao mapeamento e representação do mundo natural, pleno e rico de relações, mediante um sistema artificial, em computador, onde pululam restrições e limites, há

que se recorrer à construção de modelos e a especificação de regras que delimitem esta simbolização.

Este cotejo é apresentado na figura 1. Nela se destaca que há um conjunto de conhecimentos e regras que devem ser modelados para delimitar o espaço de uso de um sistema de informação, aí inclusos os processos de funcionamento e os limites de representação. Este conjunto admite a existência de um domínio especifico para as necessidades informacionais, que são passíveis de solução, ou de melhor tratamento, via aplicativos informatizados, mas salienta o conceito informação como principal elemento a cobrir, destacando a sua utilização para resolver problemas de forma pertinente e com eficiência , justo a essência do *datamining*.

Ressalte-se que a representação apresentada na figura 1, ela é válida dentro de um universo fechado, onde se procura sempre uma solução satisfatória, dentro de um limite de alternativas disponíveis. Esta plataforma de busca de solução, que serve de guia para implementar sistemas, é, certamente, a proposta mais conhecida em termos de tomada de decisão, a proposta da racionalidade limitada (Simon, 1961) e conjuga elementos técnicos e elementos da esfera organizacional.

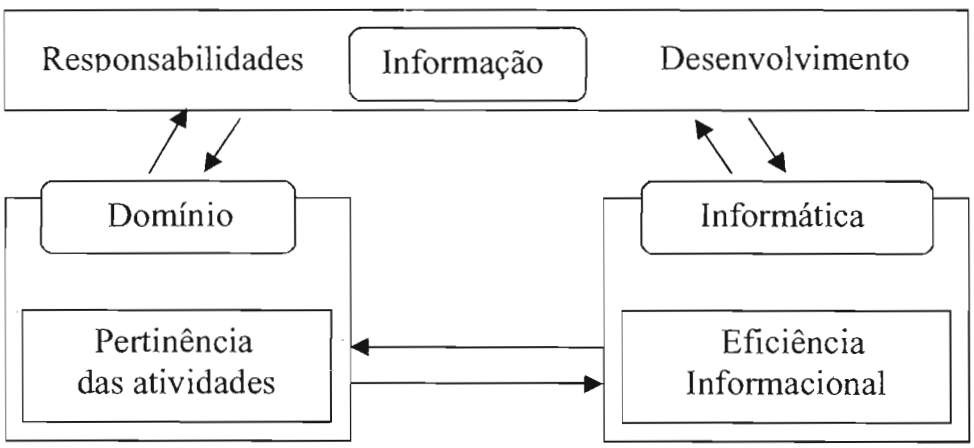

Figura 1 – Uma proposta de modelagem de sistemas de informação e comunicação
Fonte: adaptada de Léonard, 1999

Pelo lado técnico, dois modelos, e seus posteriores desdobramentos, podem ser vistos como fundamentais, para a atual compreensão de mais importância do conceito informacional que da função de otimizar desempenho de sistemas. Estes modelos são o modelo relacional de Codd e o modelo entidade-relacionamento de Shen (Date, 1995).

Referências obrigatórias em quaisquer escritos sobre modelagem e especificação de sistemas nas duas últimas décadas, estes dois modelos popularizaram conceitos largamente difundidos na atualidade como: entidades, classes, relacionamentos, atributos, generalização, herança, especialização, etc. Este conjunto de enunciados é hoje vinculado indelevelmente às noções de tecnologia cliente-servidor e orientação à objeto de dados (Morejon, 1994).

Mais recentemente, admite-se como plenamente viável agregar-se a este conjunto de enunciados, o conceito de ciclo de vida objeto, um protocolo que impõe uma seqüência de chamadas a um objeto, precisando quais os estados que ele pode transitar durante sua existência semântica. Sua introdução em muito contribui para o melhor entendimento da concepção de um sistema e melhor adequação à proposição de se construir conhecimento com a exploração dos dados.

Já pelo ângulo dos fatores organizacionais, percebe-se a inclusão de espaços de responsabilidades de desenvolvimento. Esta nova noção é decisiva para a aceitação da modelagem de sistemas.

De fato, a proposição de modelar sistemas de informação é uma tentativa de resolver os problemas de ambigüidade de representação, no ambiente artificial dos computadores, das funcionalidades de um sistema em seu ambiente natural de execução, qual seja o mundo real, rico em informações e relações normalmente não capturadas.

Esta proposição visa ainda reduzir, pela instituição de regras para assegurar compatibilidade e precisão de funcionamento, os descompassos de concepção e implementação oriundos das diferentes interpretações das atividades existentes e a seu suporte pelo sistema.

Acentua-se então, que a modelagem de sistemas representa um esforço no sentido de minorar as distorções de entendimento, carreadas para a fase de implementação dos mesmos, pelas distintas visões das pessoas envolvidas no trabalho de especificação, seja por conta de suas experiências anteriores, seja por conta de suas construções cognitivas e crenças em distintos paradigmas de conhecimento. Esta esforço leva, incontinente, à visualização de uso de propostas cooperativas.

A problemática dos sistemas de informação se situa em "uma encruzilhada onde não somente ciências informáticas e de gestão são levadas a cooperar, mas igualmente as ciências humanas e sociais, em especial aquelas relacionadas aos processos de aprendizagem coletiva" (Léonard, 1994, p. 11).

A cooperação aqui dita, é tida como uma das tendências em voga na área de especificação, exatamente graças ao uso da modelagem cooperativa de sistemas de informação, uma forma de reduzir as distâncias percebidas entre percepções da mesma realidade, fazendo uso do citado espaço de iniciativas e responsabilidades e utilizando-se de técnicas de prototipação de sistemas.

Neste limiar, aparece a necessidade de uso de outra grande vertente da tecnologia da informação, a qual será rapidamente comentada neste artigo, e que como se verá nos resultados práticos colhidos na monitoração efetuada, conquista, com seus componentes, significativas menções de uso corrente.

1.1.2. A AFIRMAÇÃO DE *GROUPWARE*

Groupware é uma tecnologia de informação voltada à colaboração que afeta a comunicação entre pessoas e as estruturas organizacionais. Em termos técnicos, *groupware* é uma ferramenta computacional utilizada para trabalhos em grupo de uma maneira cooperativa (Coleman, 1995).

O impulso para a adoção desta tecnologia advém dos processos de reestruturação organizacional (reengenharia e *downsizing*) típicos do início dos anos 90 (Courbon, 1998).

As diversas definições de *groupware* buscam uma síntese conceitual melhor referida como tecnologia de *groupware,* que se sugere em linhas gerais como: "uma tecnologia que integra sistemas de computação e facilidades de comunicação e que oferece suporte às atividade de grupos engajados em alcançar um objetivo comum" (Chen e Liou, 1991, p. 333). O quadro 1 mostra uma síntese geral das informações ligadas à tecnologia de *groupware.* As fontes dos dados estão citadas por coluna do quadro.

Categorias de serviços (funcionalidades)	**Modelagem orientada ao**	**Formas usuais de referência (campo de utilização)**
Pacotes para integração de plataformas	Processo	*Group Support Systems (GSS)*
Pacotes para aplicações institucionais	Dado	*Eletronic Meetings Systems (EMS)*
Ferramentas para geração de aplicações	Usuário	*Negotiation Support Systems (NSS)*
Fluxo de documentos (*workflow)* e manipulação de documentos *(handler)*		*Computer Support Cooperative Work (CSCW)*
Apoio à decisão em grupos		
Videoconferências		
Correio eletrônico (*e-mail*)		
Adaptado de Coleman, 1995	**Adaptado de Dennis, 1996**	**Adaptado de Chen e Liou, 1991**

Quadro 1 – Visão geral de informações sobre a tecnologia *groupware*

Dentro de uma perspectiva geral é possível rotular como categorias de serviços vinculadas à *groupware,* em termos de funcionalidades das rotinas empresariais, uma variada gama de aplicações. De fato, desde o tradicional correio eletrônico (*e-mail*), ao sofisticado *software* que integra plataformas distintas de computação, com efetivo destaque para a utilização associada às transações de fluxo de documentos, amplamente difundida pelo *Lotus Notes*®, há um vasto leque de rotinas nas quais se pode encontrar características típicas de um aplicativo *groupware.*

Ainda a título geral, reconhecem-se distinções quanto a o processo de orientação de modelagem e uso do *software*, enraizadas nas formas de ativação e controle do mesmo, as quais podem variar quanto ao processo, aos dados e aos usuários. Também se constatam variações terminológicas para referir *groupware,* relacionadas, sobretudo, à área de inserção e uso do *software.* Estas terminologias são citadas na terceira coluna do quadro 1.

Porém, sem sombra de dúvida, a principal razão que pode ser associada ao sucesso e proliferação de ferramentas *groupware* é a sua capacidade de romper com o paradigma temporal para realização de trabalhos.

Esta nova perspectiva para a realização de trabalhos cooperativos com compartilhamento de recursos informacionais e competências individuais, independentemente de restrições de tempo espaço: *não importa onde, não importa quando,* parece ser o diferencial competitivo desta nova frente da tecnologia da informação.

Ela vem corroborar a noção de usuário criativo, trabalhando em sistemas concebidos cooperativamente à luz do saber fazer de cada indivíduo, a fim de melhorar o desempenho das organizações. O quadro 2 permite antever as possibilidades de suporte associadas às noções de tempo-espaço pré-faladas, sendo já de formulação clássica em textos relacionados a *groupware.*

Temporaneidade ▼	**Localização** ▼	▼
▼	**Mesmo local**	**Locais diferentes**
Mesmo tempo	Reuniões face a face, Sala de decisão	Rede Local
Tempos distintos	Fóruns	*E-mail*, Reuniões virtuais

Quadro 2 - Matriz tempo-espaço para realização de trabalhos em grupos
Fonte: adaptado de DeSanctis e Gallupe, 1987

Agora se faz mister compreender como estes elementos podem conviver harmonicamente em ambientes concorrenciais, auxiliando a captar através de boas modelagens, cooperação explícita,

coordenação efetiva, comunicação adequada, análise e interpretação criativa, o sumidouro inesgotável de dados que se multiplica à uma velocidade incomensurável, carregando diversos conteúdos informacionais inexplorados.

1.1.3. DATAMINING

Apostila-se que a informação está se transformando na maior vantagem estratégica das organizações (Jakobiak, 1991; Rouach 1996) e que a exploração dos recursos e ligações informacionais, está se convertendo na fonte de conhecimento mais relevante e útil àquelas (Lesca e Belkatir 1994, Dou 1995). Aflora, então, a necessidade de se estabelecer um procedimento sistemático de análise intensiva, deste recurso capital para a estratégia dos empreendimentos.

O conceito de *datamining* está justamente associado à esta necessidade de se buscar dados de forma temporânea e coerente, em fontes externas e não estruturadas de informações. Isto é feito mediante o uso de métodos e *softwares* que permitam uma adaptação imediata ao caráter fugaz e mutante do contexto informativo do mercado.

Corresponde na prática, a explorar extensivamente dados externos e internos, que muitas vezes, são relegados à um patamar inferior de importância, através de uma análise minuciosa de seus inter-relacionamentos, buscando extrair conhecimento útil .

Para concretizar tal função, há inúmeras ferramentas. As zonas de congruência destas ferramentas situam-se, via de regra, na opção pela rica análise qualitativa de dados (Moscarola, 1998). Mediante a conjugação de técnicas e ferramentas, obtém-se, através de um ciclo interativo de aprendizagem dentro do contexto dos próprios dados, a informação relevante que estaria obscurecida em função do volume e formato não adequado dos dados.

A potencialidade do *datamining* advém exatamente das pré-faladas vinculações de modelagem e alicerça-se na disseminação de *softwares* amigáveis, robustos em capacidade de processamento e poder de tratamento de dados. Reforça-se ainda, fortemente, na popularização de acessos baseados em padrões às redes de computadores e no uso de navegadores da teia (*web browsers*) e à implementação baseada em uma sólida cooperação sistema-usuário, em função da utilização da modelagem visual interativa. Em tese, afirma-se que o conjunto método-utilitário progrediu nesta direção desde as suas origens na inteligência artificial (Moscarola, 1999).

Apesar dos custos associados ao processo de "mineração de dados" serem, de certo modo, elevados, e isto ser um fator proibitivo ao seu uso em larga escala, há possibilidades, como se quer mostrar com este exemplo, de se criar alternativas metodológicas e computacionais de aplicar o processo *datamining* em bases mais modestas.

Tal adequação exige, primeiramente, um esforço adicional em termos de formulação de acesso aos grandes repositórios de dados (banco de patentes, *internet*, bases de dados comerciais, publicações de domínio público). Em segundo lugar desencadeia uma árdua tarefa de estruturação e tratamento de dados, como se mostrará na seção de procedimentos. Contudo, essa tarefa conduz à extração de informações reveladoras e em alguns casos de natureza estratégica. Este é um dos grandes argumentos a utilizar para se justificar a adoção de *datamining* entre as rotinas relevantes das empresas.

O grande diferencial deste tipo de método de análise de dados é que, à priori, não se tem, como nos casos clássicos dos métodos quantitativos, a premissa de se encontrar uma resposta a uma pergunta predefinida. Há que se extrair ligações baseando-se em relações percebidas em um ciclo de análise-retorno ao exame dos dados. Assim, esta prática só se mostra efetiva e adequada quando aplicada à um repertório de dados muito grande, a fim de configurar uma real tendência nos mesmos.

Esta tendência se articula a partir da ação do investigador e de um processo que se quer taxar como análise léxica (Weber, 1990). A análise léxica leva o usuário a uma ação interativa de aprendizagem com os próprios dados. As intuições preliminares oriundas de indícios que se tenham ou de sinais fracos que se percebam no ambiente, são refinadas do acordo com o resultado que emerge do "léxico" (o conjunto de vocábulos remetidos a tratamento estatístico).

Através de decisões de eliminar, compor e recodificar dados no conjunto parcial obtido, estabelecem-se ciclos de interpretação e busca de novas relações, a fim de reorientar a investigação, que como a prática de uso tem mostrado, rapidamente, converge para aspectos reveladores e significativos dos dados (Moscarola, 1999).

Estes ciclos de aprendizado com os dados são bastante úteis em tratamento de documentos e em pesquisas que se baseiam em *grounded theory* (Strauss, 1987; Lincoln e Guba, 1987). A grande representatividade deste esquema de análise é permitir identificar relações não previstas. Os processos de codificação utilizados: codificação aberta, codificação axial, codificação seletiva,

levam o pesquisador, seja ele um estudante, seja ele um investigador prático em uma empresa, a criar as chamadas modalidades de dados e estas modalidades levam à uma maior sistematização de informações.

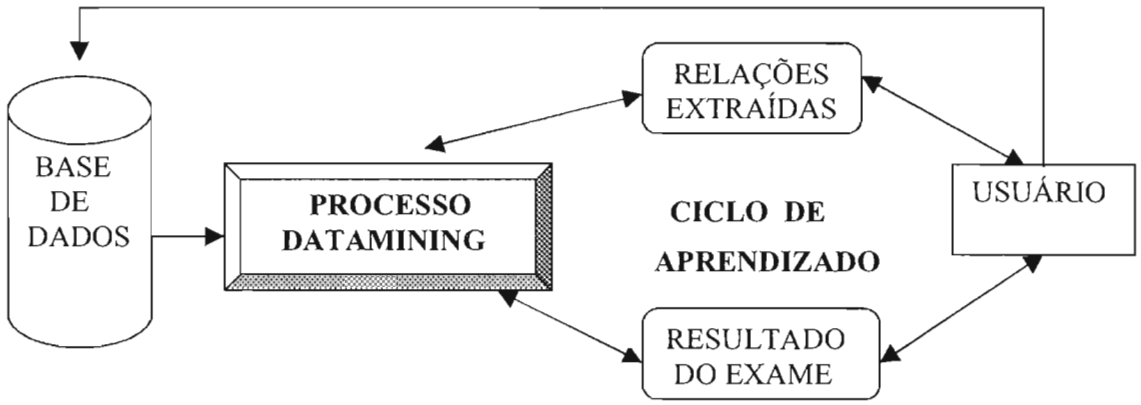

Figura 2 - Ciclo de funcionamento de um processo *datamining*
Fonte: adaptado de Moscarola (1999)

A aliança entre as idéias exploratórias do usuário e as possibilidades inferenciais das ferramentas *datamining*, faz surgir um ciclo de aprendizado recursivo, posto que retorna a si próprio, fato este que é um rótulo indelével do processo. Na figura 2 está sendo considerada a análise léxica de textos, que será parcialmente utilizado no exemplo prático deste texto. Ver-se-á, por fim, antes da descrição do caso em detalhes, um rápido exame do campo de aplicação *datamining*.

2. APLICAÇÕES DE *DATAMINING*

Neste ínterim, antes de apresentar o caso prático realizado para exemplificar a técnica, convém destacar duas das áreas potencialmente destinadas ao uso de um método *datamining* no cotidiano empresarial. As duas aplicações têm conotações vinculadas à estratégia concorrencial e servirão de elo para a retomada desta noção, que será feita na parte final deste texto.

2.1. ANÁLISE DE AMBIENTE

Se faz presente em várias correntes de estudo sobre estratégia empresarial, a menção que é de extrema importância o conhecimento do ambiente em torno das organizações, pois o mesmo

define desde o nível de competição até as escolhas e definições estratégicas (Hunger e Wheelen, 1996).

Porém, apesar de se ter consciência da importância do conhecimento do ambiente para o processo de formulação estratégica, observa-se que um dos maiores entraves à sua implementação é exatamente a falta de rotinas sistemáticas para obter informações úteis e com a agilidade necessária sobre os cenários de inserção da empresa (Furlan et al, 1995; Kotler, 1987). Aquelas rotinas, quando existem, são na maioria dos casos negligenciadas.

Incoerentemente com a necessidade sobejamente reconhecida, pouco tem sido feito para melhorar a atividade de coleta, tratamento e análise de dados ambientais (Swamidass e Newell, 1987). As poucas referências literárias e práticas constatadas neste sentido, apoiam-se em modelos tradicionais, como o modelo de Porter (1991), míopes para constatar a importância das forças políticas que influenciam as vantagens competitivas (Ichikawa, 1997).

De acordo com Hunger e Wheelen (1996), os principais grupos de variáveis ambientais a serem monitoradas com interesse estratégico são: variáveis econômicas, variáveis tecnológicas, variáveis político-legais e variáveis socioculturais. Via de regra estas variáveis apresentam-se sob a forma de agregados informacionais, disponíveis em formato não-estruturado e com diversas nuanças, as quais são relegadas à condição de elementos supérfluos. Ordinariamente estes agregados também não estão registrados em sistemas de informação convencionais (estruturados) e transacionais.

Neste contexto parece que *datamining* prenuncia-se como uma solução adequada para viabilizar este requisito operacional.

2.2. INTELIGÊNCIA COMPETITIVA E MONITORAÇÃO AMBIENTAL

Considerando-se questões de inserção em ambientes concorrenciais, observa-se que capacidade de reagir e o tempo de reação são qualidades fundamentais para a definição de estratégias das organizações, a fim de que as mesmas possam se tornar claramente orientadas para o mercado e para as oportunidades que estão surgindo.

Neste cenário de aceleradas transformações e intensificação da concorrência, a tecnologia de informação é apontada como uma ferramenta decisiva para obter ganhos de qualidade e

produtividade (Tapscott e Caston, 1995). Em relação à informação, contudo, existe um grande potencial a ser explorado e mesmo descoberto. Esta constatação é particularmente verdadeira, quando se olha a necessidade de obter informações externas, do ambiente de negócios, e incorporá-las ao processo de tomada de decisão (Constantineau, 1993), a fim de estabelecer relações e possibilidades em um senso ulterior, não percebido explicita e mediatamente.

Nesta configuração acentua-se o espaço para a consolidação de uma área de atividades dentro das empresas, cuja missão básica é suprir as carências informacionais citadas antes, bem como balizar o trabalho de análise de dados: a inteligência competitiva.

Em perfeita sintonia com as colocações de ambientação e formulação de estratégia antes colocadas, é correto conceituar inteligência competitiva como a "coleção e análise das informações de mercado, informações tecnológicas, informações sobre clientes e concorrentes e a análise e interpretação das tendências externas, políticas e sócio-econômicas (Evaristo, 1994).

O processo de inteligência compreende três fases principais de igual valia para as decisões empresariais e processos de planejamento em geral: procura de dados, fortemente apoiada em uma subatividade conhecida como monitoração ambiental (*environmental scanning*), processamento das informações e disseminação das informações (Evaristo, 1994).

O processo de monitoração visa coletar, compilar, tratar e sumarizar informações para à tomada de decisão e formulação estratégica. Esta técnica baseia-se, sobretudo, no processamento de informações de natureza qualitativa e de fontes externas (Jennings e Lumpkin, 1992).

No âmbito de processamento, onde diversas "intuições" devem ser testadas, enquanto outras são formuladas, a fim de detectar movimentos subliminares do mercado concorrencial, percebe-se que há um nítido espaço para se adotar um método de análise exaustivo de dados, buscando inferir e expandir relações. Claramente após a coleta e antes da difusão dos resultados, as diretrizes dos processos de inteligência competitiva são exclusivamente voltadas à exploração de dados e de seus inter-relacionamentos, especialmente aqueles tácitos.

Novamente aqui, se se quiser enunciar uma questão de aplicabilidade, obter-se-á como resposta, certamente, aquela mesma da seção anterior: ao que parece *datamining* é uma grande solução.

Em virtude de tantas respostas explicitas e positivas, descreva-se o procedimento e a aplicação contextualizada do processo *datamining,* a fim de avaliar seus propalados méritos..

3. UMA APLICAÇÃO DE *DATAMINING*

O objetivo das próximas seções é exibir como foi aplicado o processo de *datamining* a um caso simples. O objetivo deste relato é, mais que evidenciar um exercício da técnica, que pode derivar eventualmente para encaminhamentos e usos personalizados, demonstrar a sua aplicabilidade com algum esforço e a reduzidos custos.

Serão descritos os procedimentos para seleção de evento, a negociação para monitoração dos mesmos, a técnica de monitoração, o procedimento para estruturação e tratamento dos dados. Em seção posterior se procederá o efetivo *datamining*.

3.1. PROCEDIMENTOS ADOTADOS

Como informado previamente, agora será exibido o cenário de execução e os procedimentos estabelecidos para forjá-lo.

3.1.1. A TAREFA

O trabalho optou por observar e acompanhar sistematicamente o resumo de atividades dos usuários das redes de computadores de duas universidades brasileiras, que têm em comum a experiência de rede de dados centralizadas em ambientes computacionais cuja máquinas principais, *host* da rede, são equipamentos de mesmo fabricante e utilizam mesmos sistema operacional.

A forma mais elementar de investigar esta intenção de pesquisa foi utilizar um comando de sistema operacional, *finger*, para verificar a atividade dos usuários. Esta atitude elementar em termos de uso de sistema, emissão de um comando, configurar-se-ia como banal, todavia por ter que ser empregada em moldes cíclicos, levou o pesquisador a solicitar a autorização explícita dos administradores das respectivas redes.

Em uma das universidades, EP, não houve nenhuma restrição de acesso e uso e o próprio período de exploração inicial para desenvolver a sistemática de pesquisa, foi liberado, desde o início, para uso automático. Na segunda universidade, SR, houve que se estabelecer um compromisso através da troca de *e-mails* e bastante negociação. Este já foi um primeiro indício a guardar para a análise.

3.1.2 A SISTEMÁTICA

Foi definido um período de rastreamento em torno de 30 dias, por se entender que em época de atividades normais, este prazo seria suficiente para caracterizar usos e aplicações. Um ciclo preliminar de monitoração semi-automática, foi então estabelecido, com o intuito de identificar as horas de pique para monitoração. Decidiu-se, após observações preliminares, que seria interessante efetuar a "aferição" das atividades quatro vezes ao dia.

Novamente foi negociado com as instituições alvo, a aplicação da sistemática, desta vez em moldes automáticos nos horários definidos. Apesar de alguma reserva, em especial de SR, foi acertada, a monitoração nos prazos e instantes desejados.

A única restrição foi excluir da rotina a monitoração das sextas-feiras, pois, coincidentemente, é o dia de efetivação de cópias de segurança (*back-up*) das instalações e pelo menos em dois horários haveria forte tendência à postergação da monitoração, o que, por tratar-se de uma rotina automática, viria a comprometer toda a seqüência de atividades.

Também foram excluídos da monitoração de final de semana, posto que no período escolhido houve dois grandes "feriadões" e as equipes de operação envolvidas nas duas universidades monitoradas, nem sempre faziam eventuais recargas do sistema. Como o pesquisador não tinha, nem tem, acesso irrestrito aos recursos dos sistemas, esta prática, monitoração de fim de semana, foi abandonada após duas tentativas de implementação.

Os resultados obtidos das observações correspondem, de forma geral, aos registros de guarda (*logs*) ou de contabilização (*accounting*). Nesta configuração eles têm peculiaridades que precisam ser melhor compreendidas. De fato, o conjunto de dados obtidos, como em qualquer processo de análise de dados, precisa passar por um processo de restruturação antes de ser fornecido aos analisadores *datamining*.

Esses registros são ditos banais em termos dos sistemas de computação implantados, dado que não se tem rotinas explícitas para se proceder à uma exploração dos dados registrados, a fim de se lhes estimar qualquer significado adicional. Os arquivos são simplesmente gerados e ganham uma existência média estimada em três meses.

3.1.3 PREPARAÇÃO DE DADOS

Os dados coletados em 112 sessões de monitoração foram inicialmente incorporados a uma planilha eletrônica, onde se procedeu a formatação adequada à entrada de dados, especialmente à codificação de campos para tratamentos estatísticos. Aplicativos *datamining,* como por exemplo o *software* Eureka®, podem operacionalizar grande parte do esforço de codificação e aqui listados.

As variáveis textuais que configuraram até 50 classes nesta fase, foram, desde já, convertidas em variáveis codificadas, caso específico do item programa utilizado, ao passo que aquelas classes que excederam este patamar, foram divididos em três agregados de dados (alto, médio e baixo) de acordo com o número de citações no léxico.

Os grandes agregados de dados colhidos na fase de monitoração para posterior tratamento e análise foram, a saber: data e hora da execução; tipo de programa utilizado; local de utilização do programa, usuário que estava trabalhando (código e nome).

Dois *softwares* de análise de dados Sphinx®, Atlas®, auxiliaram na redução e tratamento de dados. De fato o grande de volume de dados colhido e tratado, especialmente os nomes das máquinas (*servers),* nomes de usuários *(username),* e nomes dos locais de utilização (*domain name)*, visto que continham, já em si próprios, uma codificação. Outra dificuldade foi a tentativa de uniformizar (em vera idealizar) uma tabela para este três categorias, dado que o pesquisador não conhecia integralmente as plantas físicas das redes locais monitoradas.

4. RESULTADOS (ANÁLISE E INTERPRETAÇÃO DOS DADOS)

Os quadros a seguir dão uma amostragem geral dos dados que foram trabalhados. A conjugação destes dados permitiu uma série de inferências, sendo que aquelas taxadas de mais relevantes serão argüidas aqui.

IFES	Citações		Dia	Utilização			Região	
EP	59,5		Segunda	21,2		Dia	EP	SR
SR	40,5		Terça	25,6		Segunda	60,0	40,0
			Quarta	25,3		Terça	60,7	39,3
			Quinta	28,0		Quarta	63,0	37,0
						Quinta	54,9	45,1

Quadro 3 - Percentuais analisados por região e dia
Fonte: Dados da Pesquisa

Hora	Freq.			Região	
Manhã	30,9		Hora	EP	SR
Tarde	40,0		Manhã	60,6	39,4
Noite	24,4		Tarde	53,5	46,5
Madrugada	4,7		Noite	64,3	35,7
			Madrugada	78,4	21,6

Quadro 4 - quantitativos por região e dia
Fonte: Dados da Pesquisa

	Região			Região	
	EP	SR		EP	SR
TELNET	98,2	1,8	CDU	-	100
PINE	21,4	78,6	PROFILE	-	100
MAIL	34,7	65,3	FINGER	-	100
DCL	27,8	72,2	SET	-	100
SEND	3,9	96,1	MWAIT	-	100
LYNX	-	100	CSWING	-	100
BOSS	100	-	GLOGIN	100	-
TPU	11,8	88,2	LOOK	-	100
PHONE	57,1	42,9	NEWS	-	100
FTP	60,0	40,0	SHOW	-	100
Multinet	-	100	TYPE	100	-
TALK	-	100	VMSHELP	100	-

Quadro 5 - Distribuição de uso dos programas por região
Fonte: Dados da Pesquisa

Em tempo ressalve-se que até meados de março serão enviados aos usuários que tiveram os seus perfis avaliados aqui, uma rápida descrição deste intento de pesquisa, e do perfil que lhes foi imputado. O objetivo é arrebanhar opiniões que permitam falar em validade e/ou confirmação das observações efetuadas.

As principais perguntas/inquisições que foram efetuadas ao conjunto dos dados compilados foram as seguintes, considerando que o aspecto de monitoração revelara participação no conjunto de dados equivalentes à 60% de observações registradas para EP e 40% para SR.

Qual o ritmo de trabalho da instituição ? Tomando por base os dados compilados, vê-se que há uma distribuição uniforme entre os dias centrais da semana e uma tendência a um menor ritmo às segundas-feiras (é lícito afirmar que esta queda deva se verificar também às sextas-feiras). A concentração de trabalho é diversa, uma instituição opera mais pela manhã ao passo que outra opera mais à tarde.

Esta última relação está associada ao tipo de trabalho que se executa na rede da instituição. Esta foi a segunda grande pergunta feita aos dados: qual a grande vocação da rede de dados destas instituições. Mormente se idealizar como resposta fundamental à este questionamento, a parte de comunicação, *groupware* e particularmente o *e-mail,* verificou-se uma surpresa. Na região EP, o quesito mais constante na rede é o utilitário de *login* remoto TELNET, que não configura uma aplicação *groupware*, apesar de executar uim processamento cooperativo entre máquinas de uma rede.

Graças à interação com o ambiente prevista ao nível do *datamining*, e antes que se tirassem conclusões apressadas, verificou-se que, na prática, ao invés de conectar pontos remotos de redes em modalidade individual, este uso intenso de TELNET, devia-se à utilização no antigo estilo de *front-ends* para o computador central da rede. Indo mais a fundo constatou-se que o peso da circulação de mensagens na rede EP era creditado ao uso por pacotes administrativos da universidade.

O segunda função da rede EP foi aquela que se esperava e que foi destaque incontestável na rede SR, qual seja a função de comunicação de dados (*mail, phone, send, talk*). Observou-se também que há uma forte diversificação de uso de *softwares* na rede SR, o quê se nota pelos diversos percentuais de 100% na coluna de tipo de programa, ao passo que há uma forte concentração na rede EP.

Esta dado permite inferir distinções nos tipos de gestão de informática nas duas instituições. De fato pela concentração de uso detectada é lícito afirmar que a rede EP tem uma operação centralizada, com uma gestão voltada para a utilização do equipamento monitorado, em regime de *mainframe* para conexões de aplicativos. Esta constatação é ratificada pelo exame dos locais

de utilização, que normalmente estão conectados à portas de comunicação vinculadas à um mesmo IP.

Já a rede SR , em que pese a sua forte concentração em aspectos de comunicação, permite inferir uma informatização mais descentralizada, com muitos mais usuários acessando-a para realizar funções diversas. É significativo o número de ocorrências de uso de sistemas operacionais, proprietários ou não, e de utilitários de gestão de diretórios de dados.

Um outro dado associado ao exame da localização de uso e da forma de ativação dos programas é o consumo de conexão. Grande parte das portentosas referências aos utilitários centralizados e baseados em *mainframe* verificados na rede EP, deixam o registro do tempo de conexão à atividade. Percebe-se, no exame dos dados, que boa parte das conexões permanecem ativas dois ou mais dias, sem nenhuma utilização efetiva.

Em que pese ser mais constante na rede EP, em especial com os usuários administrativos e de pontos concentrados, esta faceta também se percebe nas altas taxas de referência de programas utilitários como *Boss e Glogin* na rede EP e *DCL, Set* e *Mwait* na rede SR. Um exame dos usuários que colaboram na obtenção deste índices, revela uma grande concentração deste fatos nos gestores das redes e nas pessoas responsáveis pelo suporte. Em que pese a justificativa de monitoração constante, tal prática deve ser abolida, pois pode vir a configurar um fator de insegurança na operação das duas redes.

O tratamento intensivo dos dados não se esgotaria nas análises aqui colocadas. Todavia, o programa de inferência utilizado, não permite grupamento em *arrays* léxicos com mais de cinqüenta ocorrências, o quê veio a dificultar uma análise mais ampla.

5. CONCLUSÃO

O presente trabalho teve a intenção de combinar elementos que justificassem a adoção de uma técnica de pesquisa intensiva nos dados, chamada *datamining*, nas rotinas de tratamento de dados das organizações. Ele procurou mostrar como é possível com algum esforço metodológico e uso de recursos computacionais, colher informações de dados taxados habitualmente de irrelevantes.

Para tal, enumerou quais as principais idéias que se aninham atrás de um grande arquétipo chamado tecnologia da informação e em seus componentes mais destacados e exibiu como pertinente a correlação, em termos de organizações modernas, destes elementos.

Justificou o direcionamento a este método especial de coleta e análise de dados por razões de formulação de estratégias competitivas em um ambiente de constantes mudanças e dinâmicas de atuação concorrencial.

Exibiu, com um singelo exemplo, como se pode inferir resultados de uma massa aparentemente trivial de dados e objetiva, em conseqüência destes resultados preliminares, ainda durante o corrente ano, após a validação dos perfis "deduzidos" pelos usuários reais, transpor, com codificação adequada os dados aqui compilados, para o utilitário *Atlas*®.

A idéia final é apresentar aos dirigentes das áreas de computação das universidades monitoradas, um conjunto de idéias úteis para avaliação do potencial de uso e dinamização desta vital atividade nas empresas. Este item vem reforçar a idéia de modelagem de sistemas e partilha de responsabilidades, desenvolvida na primeira parte deste trabalho, como forma de obter mais potência no uso de recursos de informática nas organizações.

Também olhando as tarefas que foram executadas e aqui descritas é possível estabelecer um paralelo com as técnicas de pesquisa empregadas nos engenhos de busca de *sites web*. De fato, os dos modelos, tanto a idéia *datamining*, quanto o cerco de informações por redução de informações acessórias para se centrar em um foco específico de pesquisa, obtido nos engenho de pesquisa em *sites* como o altavista.com, visam ampliar a especificidade de recuperação de informações em um vasto arsenal disponível. Ao ver o pesquisador esta é uma atividade absolutamente compatível com o trabalho cotidiano de bibliotecas. Assim sendo a técnica descrita neste artigo tem excelentes oportunidades de ser aplicada na busca e recuperação de informações, em apoio às idéias implícitas deste modelo, que hoje já são largamente empregadas quiçá de modo intuitivo.

Por fim, em que pese o restrito espaço para o presente desenvolvimento, o conjunto de dados obtido deverá ser exaustivamente examinado, a fim de consubstanciar o relatório final da pesquisa onde o presente tema se insere, e que antevê a aplicação de técnicas *datamining* e de monitoração ambiental para formulação de estratégias empresariais, a partir da modelagem cooperativa de sistemas de informação.

6. REFERÊNCIAS BIBLIOGRÁFICAS

CONSTANTINEAU, L. Making competitive intelligence actionable. *Marketing Reserach*, v.7, n. 1, p. 46-47. 1993,

CHEN, M.. LIOU, Y I. The design of an integrated group support environment. *IEEE Systems Science*. v IV, p 333-342. 1991.

CHIAVENATO, I. *Teoria Geral da Administração*. São Paulo, MacGraw Hill, 1990. 2v.

COLEMAN, D. *Groupware technology and applications*. New Jersey: Prentice Hall. 1995.

COURBON, J. *Le travail en groupe à l'âge des réseaux*. Paris: economica. 1998.

DATE, C. J. *Introdução a Banco de dados*. 4a ed. Rio de Janeiro: Campus. 1995.

DeSANCTIS, G. GALLUPE, R. B. A foundation for the study of group decision support systems. *Management Science*. Vol 33, n 5. p 589-609. 1987.

DORNELAS, J. S. HOPPEN, N. Inovações ligadas ao processo de gestão participativa e ao uso de sistemas de apoio à decisão em grupo, na direção de novas formas de estruturas organizacionais. *Anais XXIII ENANPAD*. Foz do Iguaçu, setembro de 1999. Cd-rom.

DOU, H, *Veille technologique et compétitivité*. Dunod, 1995

EIN-DOR, Philip. SEVEG, Eli. *Administração de Sistemas de Informação*. 3a ed. Rio de Janeiro: Campus, 1997.

EVARISTO, R. CI UNIT: Implementation problems and solutions. In: *Competitive Inteligence Review*. v 5, n. 4. 1994. p. 15-18.

FAVIER, M. *Le travail en groupe à l'âge des réseaux*. Paris :Economica. 1998.

FURLAN, J. D. *Sistemas de Informações Executivas*. São Paulo, Makron Books, 1995;

HAMEL G. & PRAHALAD C.K. Strategy as stretch and leverage. In: *Harvard Business Review*, Mar-Apr, 1993. p. 75-84.

HUNGER, J.D. & WHEELEN, T.L. *Strategic Management*. Addison-Wesley Publising, 5ed, 1996.

ICHIKAWA, E.Y. Considerações críticas sobre o planejamento estratégico. *Anais do ENEGEP 97*. Gramado / RS. out, 1997, Cd-rom

IVES, B. JARVENPAA, S. L. MASON, R. O. Global business drivers: aligning information technology to global business strategy. *IBM Systems Journal*. v. 32, n. 1. 1993. p. 143-161.

JENNINGS, D. F. LUMPKIN, J. R. Insights between environmental scanning activities and Porter's generic strategies: an empirical analysis. *Journal of Management*. v. 18. n. 4, 1992, p. 791-803;

JAKOBIAK, F. *Exemples commentés de veille technologique*. Organisation.1991

KOTLER, P. *Marketing Management: analysis, planning, implementation and control*. Prentice-Hall, 1987.

LÉONARD, M. Introduction aux SI. *Photocopies des transparences de cours*. SES-CUI-UNIGE. Genève, mars, 1995.

LÉONARD, M. Fondements et exigences d'une modélisation de SIC. *Photocopies de transparences de cours*. Genève: Université de Genève, juin, 1998.

LESCA H., BELKATIR, M. *Pertinence: un instrument pour évaluer le besoin de veille stratégie de l'entreprise*. ESA, 1994 jul/aou. 1996. p. 7-12.

LINCOLN, Y.S GUBA, E. *Naturalistic Inquiry*. London: Sage 3a (ed). 1985.

MOREJON, José. *Merise: vers une modélisation orientée objet.* Paris: Les éditions des organisations. 1994.

MOSCAROLA, J. Les actes de langage - Protocoles d'enquêtes et analyse des données textuelles. *Colloque Consensus Ex-Machina*, La Sorbonne, Paris, 1998

MALONE, T W. MORTON, M A.S. HALPERIN, R R. Organizing for the 21st century. *Strategy & Leadership*. June. 1996. p 56-63.

PORTER, M. E. *Vantagem competitiva: criando e sustentando um desempenho superior*. Rio de Janeiro: Campus, 1991;

STRAUSS, A L. *Qualitative analysis for social scientists*. London: Cambridge Press 1987.

SWAMIDASS, P.M. & NEWELL, W.T. Manufacturing strategy, environmental uncertainty and perfomance: a path analytic model. *Management Science,* v.33, n.4. p.509-524. 1987.

SIMON, H. *Comportamento Administrativo*. São Paulo: Atlas, 1991.

TAPSCOTT, D. CASTON, A. *Mudança de paradigma.* São Paulo: Makron Books, 1995.

VAN MAANEN, J. Style as Theory. *Organization Science*. v. 6, n. 1, p. 133-143. 1995.

WEBER, R. P. *Basic content analysis*. London: Sage, 1990.