ISSN
0034-7612

RAP
REVISTA DE ADMINISTRAÇÃO PÚBLICA

Gestão do conhecimento usando *data mining*: estudo de caso na Universidade Federal de Lavras*

Olinda Nogueira Paes Cardoso**
Rosa Teresa Moreira Machado***

Palavras-chave: gestão do conhecimento; descoberta de conhecimento em bancos de dados; *data mining*; plataforma Lattes.

Key words: knowledge management; knowledge discovery in database; data mining; Lattes platform.

A gestão do conhecimento abrange toda a forma de gerar, armazenar, distribuir e utilizar o conhecimento, tornando necessária a utilização de tecnologias de informação para facilitar esse processo, devido ao grande aumento no volume de dados. A descoberta de conhecimento em banco de dados é uma metodologia que tenta solucionar esse problema e o *data mining* é uma técnica que faz parte dessa metodologia. Este artigo desenvolve, aplica e analisa uma ferramenta de *data mining*, para extrair conhecimento referente à produção científica das pessoas envolvidas com a pesquisa

* Artigo recebido em fev. 2005 e aceito em mar. 2007.
** Graduada em informática pela Universidade Católica de Salvador (UCSal), mestre em administração pela Universidade Federal de Lavras (Ufla), professora assistente do Departamento de Ciência da Computação da Ufla. Endereço: Caixa Postal, 3037 — CEP 37200-000, Lavras, MG, Brasil. E-mail: olinda@dcc.ufla.br.
*** Economista pela Universidade Federal de Minas Gerais, doutora em administração pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo (FEA/USP). Professora associada do Departamento de Administração e Economia da Universidade Federal de Lavras e editora da revista *Organizações Rurais & Agroindustriais*. Endereço: Caixa Postal, 3037 — CEP 37200-000, Lavras, MG, Brasil. E-mail: rosaflor@ufla.br.

FGV
EBAPE

na Universidade Federal de Lavras. A metodologia utilizada envolveu a pesquisa bibliográfica, a pesquisa documental e o método do estudo de caso. As limitações encontradas na análise dos resultados indicam que ainda é preciso padronizar o modo do preenchimento dos currículos Lattes para refinar as análises e, com isso, estabelecer indicadores. A contribuição foi gerar um banco de dados estruturado, que faz parte de um processo maior de desenvolvimento de indicadores de ciência e tecnologia, para auxiliar na elaboração de novas políticas de gestão científica e tecnológica e aperfeiçoamento do sistema de ensino superior brasileiro.

Knowledge management using data mining: a case study of the Federal University of Lavras

The management of knowledge embraces every form of production, storage, distribution and use of the knowledge, making necessary the use of information technologies to facilitate the process, due to the great increase in the volume of data. An emergent methodology that tries to solve the problem of the analysis of great amounts of data is the knowledge discovery in database (KDD) and data mining, a technique that is part of this methodology. This article aims to develop, apply and analyze a tool of data mining, to extract knowledge regarding people's scientific production involved with the research at the Federal University of Lavras (Ufla). The methodology used involved bibliographical research, documental research, and method of case study. Once it was just used referring data to the scientific production of Ufla. The limitations found in the analysis of the results indicate that it is still necessary to standardize the completion of the Lattes curricula to refine the analyses, and establish indicators. The result was the creation of a structured database, which is part of a larger process of development of science and technology indicators, with the objective of aiding the elaboration of new policies of scientific and technological management and improvement of the superior education system in Brazil.

1. Introdução

O conhecimento tem sido reconhecido como um dos mais importantes recursos de uma organização, tornando possíveis ações inteligentes nos planos organizacional e individual, induzindo a inovações e capacidade de continuamente criar produtos e serviços excelentes em termos de complexidade, flexibilidade e criatividade. O processo de gestão do conhecimento abrange toda a forma de gerar, armazenar, distribuir e utilizar o conhecimento, tornando necessária a utilização de tecnologias de informação para facilitar o processo, devido ao grande aumento no volume de dados.

Ao longo do tempo, percebeu-se que a velocidade de coleta de informações era muito maior do que a velocidade de processamento ou análise das mesmas, o que gera um problema e uma contradição, pois as organizações,

por possuírem uma grande quantidade de dados, possuem uma falsa sensação de que estão bem informadas; porém essas informações de nada servem se não forem analisadas de forma correta e em tempo hábil.

Em outras palavras, a coleta e o armazenamento de dados, por si só, não contribuem para melhorar a estratégia da organização. É necessário que sejam feitas análises sobre essa grande quantidade de dados, estabelecendo-se indicadores para descobrir padrões de comportamento implícitos nos dados, assim como relações de causa e efeito. Processar e analisar as informações geradas pelas enormes bases de dados atuais de forma correta estão entre os requisitos essenciais para uma boa tomada de decisão.

Num ambiente extremamente mutável, como o das organizações na atualidade, torna-se necessária a aplicação de técnicas e ferramentas automáticas que agilizem o processo de extração de informações relevantes de grandes volumes de dados. Uma metodologia emergente, que tenta solucionar o problema da análise de grandes quantidades de dados e ultrapassa a habilidade e a capacidade humanas, é a descoberta de conhecimento em banco de dados.

Data mining, ou mineração de dados, é uma técnica que faz parte de uma das etapas da descoberta de conhecimento em banco de dados. Ela é capaz de revelar, automaticamente, o conhecimento que está implícito em grandes quantidades de informações armazenadas nos bancos de dados de uma organização. Essa técnica pode fazer, entre outras, uma análise antecipada dos eventos, possibilitando prever tendências e comportamentos futuros, permitindo aos gestores a tomada de decisões baseada em fatos e não em suposições.

É possível extrair, por exemplo, um grande número de informações úteis a partir da análise da produção científica, tecnológica e bibliográfica desenvolvida na Universidade Federal de Lavras (Ufla). Para isso, foi criado um banco de dados gerado a partir de arquivos extraídos da plataforma Lattes e, posteriormente, foi desenvolvida uma ferramenta de *data mining*, utilizando os recursos de um sistema gerenciador de banco de dados, para identificar padrões e tendências, gerando base para a gestão do conhecimento na instituição.

As instituições de ensino superior (IES) são organizações voltadas para o conhecimento. Ao longo dos últimos anos, diversos autores vêm discutindo como avaliar a qualidade dos serviços prestados por essas instituições e nunca se questionou tanto a qualidade e os valores cobrados por esses serviços. Tem-se acentuado a necessidade de reflexão sobre a gestão das IES, preparando-as para as transformações que estão ocorrendo no ambiente em que operam. Cabe às próprias IES gerarem soluções para gestão de políticas de ciência, tecnologia e inovação, que tenham um horizonte maior de planejamento a partir dessa enorme massa de dados ainda subutilizados.

Levando-se em consideração os problemas enfrentados pelas universidades com o gerenciamento dos dados, além de diversas limitações encontradas na gestão dos sistemas de informação, o presente trabalho utilizou a técnica de *data mining*, extraindo conhecimento e contribuindo para a melhoria do preenchimento dos dados na plataforma Lattes.

Como parte do processo de descoberta de conhecimento em banco de dados, este artigo tem como objetivo geral desenvolver, aplicar e analisar uma ferramenta de *data mining*, para extrair conhecimento referente à produção científica dos professores da Ufla. Como objetivos específicos, temos:

- selecionar e tratar os dados disponíveis na plataforma Lattes referentes à pesquisa científica na Ufla;
- implementar um programa para transformar os dados selecionados num banco de dados;
- desenvolver uma ferramenta automática de descoberta de conhecimento, utilizando a técnica de *data mining* e descrevê-la;
- descrever as informações geradas e analisá-las.

Este artigo é uma etapa do processo de desenvolvimento do conhecimento, que pode servir de apoio à tomada de decisão, possibilitando, no futuro, a criação de indicadores para efeito comparativo entre instituições de ensino superior e de apoio à gestão da política científica e tecnológica e aperfeiçoamento do sistema de ensino superior do país.

2. Gestão do conhecimento

De acordo com Tarapanoff (2001), as mudanças que vêm ocorrendo nas organizações atualmente convergem para a quebra de um paradigma histórico e, por meio dele, entramos na era sociedade da informação e do conhecimento. A informação como principal matéria-prima das organizações é um insumo comparável à energia que alimenta um sistema; o conhecimento é utilizado na agregação de valor a produtos e serviços; a tecnologia constitui um elemento vital para as mudanças, em especial o emprego da tecnologia sobre acervos de informação. A rapidez, a efetividade e a qualidade constituem fatores decisivos de competitividade.

As organizações estão buscando alguma vantagem sustentável que as diferencie das outras em seu ambiente de negócio, utilizando para isso seu

conhecimento, que é considerado um dos mais importantes recursos de uma organização. O conceito de conhecimento, com base em inúmeras definições, envolve estruturas cognitivas que representam determinada realidade. Segundo Krogh, Ichijo e Nonaka (2001), citados por Alvarenga e colaboradores (2002), conhecimento é como uma crença verdadeira e justificada que significa que as pessoas interpretam as informações conforme sua visão de mundo, também pode ser visto como a experiência, o entendimento e o *know-how* prático que o ser humano possui e que guiam suas decisões e ações.

Assim, a gestão do conhecimento é a área que estuda o modo como as organizações entendem o que elas conhecem, o que elas necessitam conhecer e como elas podem tirar o máximo proveito do conhecimento (Carvalho, 2000). Como o processo de gestão do conhecimento é abrangente e complexo, torna-se necessária a utilização de tecnologias da informação, principalmente no que se refere à análise da grande quantidade de informação que é armazenada.

Antes de chegar a uma definição do que seja gerenciar o conhecimento, é necessário conceituar conhecimento. Diversos autores (Adriaans e Zantinge, 1996; Fayyad et al., 1996; Elmasri e Navathe, 2002; Navega, 2002; Amo, 2003; Moxton, 2004) fazem uma distinção ascendente entre os termos: dado, informação e conhecimento. Dados são fatos, imagens ou sons que podem ou não ser úteis ou pertinentes para uma atividade particular. São abstrações formais quantificadas, que podem ser armazenadas e processadas por computador. Informações são dados contextualizados, com forma e conteúdo apropriados para um uso particular. São abstrações informais (não podem ser formalizadas segundo uma teoria matemática ou lógica) que representam, por meio de palavras, sons ou imagens, algum significado para alguém. Conhecimento é uma combinação de instintos, idéias, informações, regras e procedimentos que guiam ações e decisões; tem embutido em si valores como sabedoria e *insights*. É a inteligência obtida pela experiência. Como exemplo, pode-se citar a experiência que um funcionário possui por ter trabalhado em determinadas atividades numa organização por muito tempo.

Como um organismo vivo, as organizações recebem informação do meio ambiente e também atuam sobre ele. Segundo Navega (2002), durante essas atividades, é necessário distinguir vários níveis de informação. O fundamental a se perceber nesse processo de transformação dos dados ao conhecimento é a sensível redução de volume de dados que ocorre cada vez que se sobe de nível. Essa redução de volume é uma natural conseqüência do processo de abstração. Abstrair, aqui, é representar uma informação por meio de correspondentes simbólicos e genéricos. A importância disso é perceber que, para

ser genérico, é necessário "perder" um pouco dos dados, para só conservar a "essência" da informação.

Tipos de conhecimento

Segundo Tarapanoff (2001), o conhecimento organizacional pode ser classificado em dois tipos. O primeiro é o conhecimento explícito, que pode ser articulado na linguagem formal, sobretudo em afirmações gramaticais, expressões matemáticas, especificações, manuais e assim por diante. Esse tipo de conhecimento pode ser então transmitido, formal e facilmente, entre os indivíduos.

O segundo tipo, o conhecimento tácito, é difícil de ser articulado na linguagem formal. É o conhecimento pessoal, incorporado à experiência individual e envolve fatores intangíveis como, por exemplo, crenças pessoais, perspectivas e sistemas de valor. O conhecimento tácito foi deixado de lado como componente crítico do comportamento humano coletivo. A dimensão cognitiva do conhecimento tácito reflete nossa imagem da realidade — o que é — e nossa visão do futuro — o que deveria ser. Apesar de não poderem ser articulados muito facilmente, esses modelos implícitos moldam a forma com que percebemos o mundo à nossa volta (Tarapanoff, 2001).

Considera-se os conhecimentos explícito e o tácito unidades estruturais básicas que se complementam. Mais importante, a interação entre essas duas formas de conhecimento é a principal dinâmica da criação do conhecimento em uma organização. A criação do conhecimento organizacional é um processo em espiral em que a interação ocorre repetidamente (Tarapanoff, 2001).

Na medida em que o conhecimento, tanto o tácito quanto o explícito, se torna um ativo central, produtivo e estratégico, o sucesso da organização depende cada vez mais da sua habilidade em coletar, produzir, manter e distribuir conhecimento.

Desenvolver procedimentos e rotinas para otimizar a criação, o fluxo, o aprendizado e o compartilhamento de conhecimento e informação numa organização torna-se uma responsabilidade gerencial central. O processo de, ativa e sistematicamente, gerenciar e alavancar o armazenamento de conhecimento numa organização é chamado de gestão do conhecimento (Laudon e Jane, 1999).

A gestão do conhecimento pode ser vista, então, como o conjunto de atividades que busca desenvolver e controlar todo tipo de conhecimento em uma organização, visando à utilização na consecução de seus objetivos. Esse con-

junto de atividades deve ter, como principal meta, o apoio ao processo decisório em todos os níveis. Para isso, é preciso estabelecer políticas, procedimentos e tecnologias que sejam capazes de coletar, distribuir e utilizar efetivamente o conhecimento, bem como representar fator de mudança no comportamento organizacional (Tarapanoff, 2001).

Criando conhecimento

De acordo com Tarapanoff (2001), a criação de conhecimento organizacional pode ser definida como a capacidade que uma instituição tem de criar conhecimento, disseminá-lo na organização e incorporá-lo a produtos, serviços e sistemas. Criar novos conhecimentos também não é apenas uma questão de aprender com os outros ou adquiri-los externamente. O conhecimento deve ser construído por si mesmo, muitas vezes exigindo uma interação intensiva e laboriosa entre diversos membros da organização. Assim, diz respeito também tanto aos ideais como às idéias. Ele também pode ser definido na hora com base na experiência direta e por meio da tentativa e erro, o que exige intensa e trabalhosa interação entre os membros da equipe (Tarapanoff, 2001).

As formas de interação entre o conhecimento tácito e o explícito, e entre o indivíduo e a organização, acontecem por meio de quatro processos principais da conversão do conhecimento que, juntos, constituem a criação do conhecimento, segundo a afirmação de Tarapanoff (2001), ao citar Nonaka e Takeuchi (1997). São quatro processos:

- do tácito para o explícito (externalização), que é um processo de articulação do conhecimento tácito em conceitos explícitos, ou seja, de criação do conhecimento perfeito, à medida que o conhecimento tácito se torna explícito, expresso na forma de analogias, conceitos, hipóteses ou modelos;
- do explícito para o explícito (combinação), cujo modo de conversão do conhecimento envolve a combinação de conjuntos diferentes de conhecimento explícito;
- do explícito para o tácito (internalização), que é o processo de incorporação do conhecimento explícito no conhecimento tácito;
- do tácito para o tácito (socialização), que é um processo de compartilhamento de experiências e, a partir daí, de criação do conhecimento tácito, como modelos mentais ou habilidades técnicas compartilhadas.

Para a criação de conhecimento explícito, diversas técnicas de descoberta de conhecimento podem ser utilizadas pelas organizações. Um dos maiores problemas enfrentados atualmente é o grande volume das bases de dados que as organizações possuem. A descoberta de conhecimento em banco de dados pode ser utilizada como solução para este problema.

Descoberta de conhecimento em banco de dados

A necessidade de informações disponíveis vem crescendo assustadoramente nos últimos anos e vários fatores contribuíram para esse incrível aumento. O baixo custo de armazenagem pode ser visto como a principal causa do surgimento dessas enormes bases de dados. Outro fator é a disponibilidade de computadores de alto desempenho a um custo razoável. Como conseqüência, bancos de dados passam a conter verdadeiros tesouros de informação e, devido ao seu volume, ultrapassam a habilidade técnica e a capacidade humana na sua captação e interpretação.

O sucesso das organizações depende basicamente das decisões tomadas por seus gestores, antes mesmo de apresentar ao mercado seus produtos ou serviços. Tais decisões têm se tornado necessárias em prazos cada vez mais curtos, exigindo dos gestores responsáveis uma atenção redobrada aos ambientes interno e externo da organização. Muitas vezes, más decisões são definidas, não pela inexistência do conhecimento para se escolher melhor, e sim porque o conhecimento não estava disponível para ser utilizado no tempo e lugares certos.

Para que o conhecimento seja extraído de forma eficiente, é realizado um processo chamado descoberta de conhecimento em banco de dados (DCBD ou KDD do inglês *knowledge discovery in databases*), processo este que possui o *data mining* como principal etapa (Amo, 2003). Ou seja, para que o conhecimento seja descoberto, técnicas de *data mining* (mineração de dados) devem ser aplicadas.

Uma definição formal é que DCBD é o processo não trivial de identificação de padrões em um conjunto de dados com as seguintes características:

- validade — a descoberta de padrões deve ser válida em novos dados com algum grau de certeza ou probabilidade;
- novidade — os padrões são novos, ou seja, ainda não foram detectados por nenhuma abordagem;
- utilidade potencial — os padrões devem poder ser utilizados para a tomada de decisões úteis, medidas por alguma função;

- assimiláveis — um dos objetivos do DCBD é tornar os padrões assimiláveis ao conhecimento humano.

De acordo com Adriaans e Zantinge (1996), existe uma confusão entre os termos *data mining* e descoberta de conhecimento em banco de dados. O termo DCBD é empregado para descrever o processo de extração de conhecimento de um conjunto de dados. Nesse contexto, conhecimento significa relações e padrões entre os elementos dos conjuntos de dados. O termo *data mining*, segundo os autores, deve ser usado exclusivamente para o estágio de descoberta do processo de DCBD, que se divide em sete estágios: (1) definição do problema; (2) seleção dos dados; (3) eliminação de incongruências/erros dos dados (filtragem dos dados); (4) enriquecimento dos dados; (5) codificação dos dados; (6) *data mining*; e (7) relatórios. Em outras palavras, a mineração de dados seria uma etapa do processo de DCBD.

Data mining

Talvez a definição mais importante de *data mining* tenha sido elaborada por Fayyad e colaboradores (1996:4): "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Data mining, ou mineração de dados, é uma área de pesquisa multidisciplinar, incluindo principalmente as tecnologias de bancos de dados, inteligência artificial, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados. Embora muita informação já exista sobre o tema, não existe uma padronização e classificação universalmente aceita sobre o assunto, de maneira a facilitar os interessados da área na condução de seus projetos de pesquisa. Uma das justificativas é justamente essa dimensão de novidade do tema e sua relevância na solução para análise de grandes volumes de dados. Além disso, o material existente sobre *data mining* possui abordagens heterogêneas, dependendo da origem ou do público-alvo a que se destina. O tema é estudado e abordado por profissionais de diversas áreas e cada uma possui abordagens específicas, adequadas para as suas necessidades.

Os seguintes pontos são algumas das razões pelas quais o *data mining* vem se tornando necessário para uma boa gestão organizacional: os volumes de dados são muito importantes para um tratamento utilizando somente técnicas clássicas de análise; o usuário final não é necessariamente um estatísti-

co; e a intensificação do tráfego de dados (navegação na internet, catálogos online etc.) aumenta a possibilidade de acesso aos dados.

Segundo Amo (2003), vale ressaltar que é importante distinguir o que é uma tarefa e o que é uma técnica de mineração de dados. A tarefa consiste na especificação do "que" se deseja buscar nos dados, que tipo de regularidades ou categorias de padrões, ou que tipo de padrões poderiam surpreender. Já a técnica de mineração consiste na especificação de métodos que garantam "como" descobrir os padrões que interessam. Entre as principais técnicas utilizadas em mineração de dados estão as técnicas estatísticas e as de inteligência artificial.

Segundo King (2003), *data mining* é um modo de procurar relações interessantes escondidas em um grande conjunto de dados, tais como padrões de *clustering* (agrupamentos) e aproximações de funções. Raramente é um processo completamente automatizado, com uma grande intervenção do analista que conduz o estudo. A aplicação típica de *data mining* começa com um grande conjunto de dados e poucas definições. A maioria dos algoritmos trata os dados iniciais como uma "caixa-preta", com nenhuma informação disponível sobre o que os dados descrevem, quais relações existem entre os dados e se contêm erros. Ao examinar os dados, um algoritmo pode explorar milhares de prováveis regras, utilizando diversas técnicas para escolher entre elas.

Decker e Focardi (1995) definem *data mining* como uma metodologia que procura uma descrição lógica ou matemática, eventualmente de natureza complexa, de padrões e regularidades em um conjunto de dados. Grossman, Hornick e Meyer (2002) definem *data mining* como a descoberta de padrões, associações, mudanças, anomalias e estruturas estatísticas e eventos em dados. A análise de dados tradicional é baseada na suposição, em que uma hipótese é formada e validada por meio dos dados. Por outro lado, as técnicas de *data mining* são baseadas na descoberta, na medida em que os padrões são automaticamente extraídos do conjunto de dados.

De acordo com Moxton (2004), *data mining* é um conjunto de técnicas utilizadas para explorar exaustivamente e trazer à superfície relações complexas em um conjunto grande de dados. Uma diferença significante entre as técnicas de *data mining* e outras ferramentas analíticas é a abordagem utilizada para explorar as inter-relações entre os dados, semelhante à abordagem dada por Grossman, Hornick e Meyer (2002), que também diferenciam as técnicas de *data mining* com relação às técnicas analíticas entre as abordagens de suposição e de descoberta. Segundo esses autores, discordando de outros pesquisadores, as técnicas de *data mining* não pressupõem que as relações

entre os dados devam ser conhecidas *a priori*. Ao ser aplicada a técnica, novas relações entre os dados irão surgir.

A análise automatizada e antecipada oferecida pelo *data mining* vai muito além da simples análise de eventos passados, que é fornecida pelas ferramentas de retrospectiva típicas de sistemas de apoio à decisão. Com a utilização da técnica, novas informações de cunho explícito podem ser geradas e podem fazer parte do conjunto de conhecimentos explícitos de uma organização, podendo servir de subsídio para gerar *insights* e elementos para conhecimento tácito.

O objetivo do *data mining* é descobrir, de forma automática ou semi-automática, o conhecimento que está "escondido" nas grandes quantidades de informações armazenadas nos bancos de dados da organização, permitindo agilidade na tomada de decisão. Uma organização que emprega o *data mining* é capaz de: criar parâmetros para entender o comportamento dos dados, que podem ser referentes a pessoas envolvidas com a organização; identificar afinidades entre dados que podem ser, por exemplo, entre pessoas e produtos e ou serviços; prever hábitos ou comportamentos das pessoas e analisar hábitos para se detectar comportamentos fora do padrão entre outros.

Em termos gerais, segundo Elmasri e Navathe (2002), a técnica de *data mining* compreende os seguintes propósitos:

- previsão — pode mostrar como certos atributos dentro dos dados irão comportar-se no futuro;
- identificação — padrões de dados podem ser utilizados para identificar a existência de um item, um evento ou uma atividade;
- classificação — pode repartir os dados de modo que diferentes classes ou categorias possam ser identificadas com base em combinações de parâmetros;
- otimização do uso de recursos limitados, como tempo, espaço, dinheiro ou matéria-prima e maximizar variáveis de resultado como vendas ou lucros sob um determinado conjunto de restrições.

Segundo Tarapanoff (2001), Elmasri e Navathe (2002) e Amo (2003), o conhecimento descoberto durante a fase de *data mining* pode ser descrito de acordo com cinco tarefas:

- análise de regras de associação — uma regra de associação é um padrão da forma $X \rightarrow Y$, em que X e Y são conjuntos de valores, ou seja, encontrar itens que determinem a presença de outros em uma mesma transação e estabelecer regras que correlacionam a presença de um conjunto de itens

com outro intervalo de valores para outro conjunto de variáveis. Exemplo: sempre que se orienta um aluno de doutorado, é publicado algum documento; descobrir regras de associação entre alunos de doutorado e número de publicações pode ser útil para melhorar a distribuição de orientados por professor;

- classificação e predição — é o processo de criar modelos (funções) que descrevem e distinguem classes ou conceitos, baseados em dados conhecidos, com o propósito de utilizar esse modelo para predizer a classe de objetos que ainda não foram classificados. O modelo construído baseia-se na análise prévia de um conjunto de dados de amostragem ou de treinamento, contendo objetos corretamente classificados. Exemplo: grupos de pesquisas já definidos contendo alguns professores e, a partir da análise de dados das pesquisas de outros professores que não pertencem a esses grupos, sugerir a sua entrada;
- análise de padrões seqüenciais — um padrão seqüencial é uma expressão da forma $<I_1, ...I_n>$, em que cada I_i é um conjunto de itens. A ordem em que estão alinhados os conjuntos reflete a cronologia com que aconteceram os fatos representados por eles. Encontrar padrões previsíveis em um período de tempo significa que um comportamento particular em um dado momento pode ter como conseqüência outro comportamento ou seqüência de comportamentos dentro de um mesmo período de tempo. Exemplo: uma pessoa que cursou mestrado provavelmente fará doutorado em um certo período de tempo;
- análise de *clusters* (agrupamentos) — diferentemente da classificação e predição, em que os dados estão previamente classificados, a análise de *clusters* trabalha sobre dados em que as classes não estão definidas. A tarefa consiste em identificar novos agrupamentos, que contenham características similares e agrupar os registros, ou seja, particionar (segmentar) uma dada população de eventos ou itens em conjuntos. Exemplo: professores de departamentos diferentes, que trabalham em grupos de pesquisas distintos, poderiam estar trabalhando com o mesmo objeto e, dessa forma, seria sugerida a formação de um novo agrupamento dessas pessoas, podendo surgir assim um novo grupo de pesquisa ou reclassificá-lo;
- análise de *outliers* — um banco de dados pode conter dados que não apresentam o comportamento geral da maioria. Eles são denominados *outliers* (exceções). Muitos métodos de mineração descartam esses *outliers* como ruído indesejado. Entretanto, em algumas aplicações, tais eventos raros

podem ser mais interessantes do que os que ocorrem regularmente. Exemplo: descobrir padrões de comportamento de professores que publicam um número muito grande de artigos e que fogem ao padrão dos demais professores.

O *data mining* usa ferramentas de análise estatística, assim como técnicas da área de inteligência artificial, ou técnicas baseadas em regras e outras técnicas inteligentes. A mineração dos dados pode dar-se sobre um banco de dados operacional, ou sobre um *data warehouse*, constituindo um sistema de suporte à decisão.

3. Gestão de ciência, tecnologia e inovação e sua importância

A gestão de ciência, tecnologia e inovação (CT&I) diz respeito à administração e desenvolvimento de estratégias e instrumentos organizacionais, envolvendo aspectos estruturais, culturais, políticos, tecnológicos, gerenciais e de serviços, de forma a promover a pesquisa viável e relevante (Hayashi et al., 2004).

A tomada de decisões no campo da CT&I é uma tarefa complexa, que tem sido simplificada a partir do desenvolvimento de indicadores de ciência e tecnologia (C&T), propostos como ferramentas para auxiliar no planejamento, monitoramento e avaliação de resultados científicos das nações.

Hayashi (2002) afirma que analisar atividades de CT&I é um desafio para a definição de políticas públicas. O avanço do conhecimento produzido por pesquisadores tem de ser transformado em informação acessível para a sociedade, o que coloca os indicadores das atividades de CT&I no centro dos debates.

Na gestão de C&T devem ser consideradas: a escolha de linhas de pesquisa prioritárias quanto à relevância para o desenvolvimento socioeconômico e cultural; e a execução mais eficiente das pesquisas e a conversão mais rápida de resultados obtidos em contribuições para a comunidade. Tais aspectos devem ser considerados em três níveis de gestão: o das políticas públicas, o institucional (universidades, institutos de pesquisa, empresas etc.) e o de programas e projetos específicos de pesquisa (Coelho, 2002).

No entanto, Hayashi (2002) afirma que as principais questões envolvidas nesse âmbito dizem respeito à caracterização e à construção de indicadores que devem ser discutidos e analisados a partir do contexto de produção das atividades científicas, sem deixar de considerar as limitações e dificuldades

para o seu desenvolvimento. O objetivo do trabalho dessa autora foi desenvolver uma metodologia de produção de indicadores para a análise de atividades de CT&I na Universidade Federal de São Carlos (UFSCar). Dessa forma, tais indicadores podem constituir instrumentos para a definição de políticas de C&T nas instituições federais de ensino superior, uma vez que retratam a estrutura, a situação e a performance das atividades de pesquisa científica e tecnológica, tanto para reprodução e geração de conhecimentos, quanto para criação de novos produtos e processos.

A sua metodologia inclui: revisão de literatura em CT&I e sociedade da informação; caracterização do local; coleta de dados na plataforma Lattes; e produção de indicadores de CT&I do local, com o auxílio de ferramentas estatísticas automatizadas (Hayashi, 2002).

A pesquisa da autora indica que, se acompanhados ao longo dos anos, os indicadores de C&T permitirão às instituições: desenvolver mecanismos para planejar, monitorar e avaliar as atividades de pesquisa institucional; estabelecer diretrizes para o desenvolvimento de uma política de C&T sintonizada com os avanços do conhecimento na sociedade da informação; servir de instrumento para conhecimento do perfil do pesquisador, dos programas de pós-graduação e dos grupos de pesquisas institucionais; estabelecer critérios sobre a alocação de recursos humanos, físicos, de equipamentos e material, financeiros e orçamentários, disponíveis e ou mobilizados pela instituição; preservar a memória da atividade científica e tecnológica desenvolvida na instituição; analisar os padrões de publicação científica e tecnológica da instituição; fortalecer e direcionar as ações de organismos de fomento à pós-graduação e pesquisa, entre outros (Hayashi, 2002).

Indicadores de ciência, tecnologia e inovação

Um modelo linear tem sido utilizado para explicar o vínculo entre conhecimento e desempenho econômico e, a partir dele, os governos começaram a articular políticas públicas em relação à ciência. Essa visão deu origem ao modelo linear de C&T ou modelo linear de inovação, desenhado a partir de dois aforismos: a pesquisa básica (o conhecimento geral e um entendimento da natureza e de suas leis) deve ser conduzida sem a preocupação com fins práticos; e a pesquisa aplicada — converte as descobertas da pesquisa básica em inovações tecnológicas que vão ao encontro das necessidades da sociedade. Ao longo de vários anos, esse modelo influenciou largamente universidades, porém, atualmente vem sendo questionado.

Um novo modelo atribui às pesquisas duas coordenadas: uma que dimensiona o avanço do conhecimento que a pesquisa propicia e outra que dimensiona a aplicação que dela decorre. Assim, uma pesquisa pode, ao mesmo tempo, contribuir significativamente para o avanço do conhecimento e ter grandes perspectivas de aplicações práticas.

Segundo Hayashi e colaboradores (2004), existe uma relação entre a capacidade de produzir indicadores de C&T e a de realizar investimentos em C&T por parte de governos e instituições do setor público e privado. Nos últimos anos, o desenvolvimento de políticas e estratégias para execução de metas institucionais conduziu os organismos de ciência e tecnologia e setores públicos a elaborarem instrumentos de medição que possibilitem uma gestão otimizada e racional de seus recursos.

A temática e a produção de indicadores de CT&I fazem parte da agenda científica de organismos e instituições, demonstrando a importância do tema. O uso desses indicadores como subsídio para a construção de políticas em C&T, com foco na informação, é um dos exemplos da importância de trabalhos nessa área (Ferraz e Basso, 2003; Brisolla, 1998, citados por Hayashi et al., 2004).

No contexto nacional, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), criado em 1951, foi a primeira instituição que realizou esforços para gerar indicadores de C&T. Outras iniciativas de construção de indicadores provêm do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), do Ministério da Ciência e Tecnologia (MCT) e, no campo do ensino superior, da Capes. Segundo informações divulgadas pelo MCT em seu site, esse ministério passou a assumir, de forma centralizada, a responsabilidade pela organização e divulgação das informações de C&T do país.

Os indicadores construídos pelo MCT passaram por duas fases: no início, concentravam-se no que passou a se denominar indicadores de insumo, isto é, no dimensionamento dos recursos financeiros e humanos investidos em C&T. A mensuração se limitava à identificação dos recursos aplicados à pesquisa, permitindo a construção do que se chamou de dispêndio interno em P&D, e aos recursos humanos — e sua capacitação — dedicados a tais atividades. Esses indicadores de insumo, seguindo a tendência daqueles dos demais países, possuem as séries mais longas e detalhadas (MCT, 2001).

Como menciona o MCT na apresentação dos indicadores de C&T, tradicionalmente esses indicadores de insumo são desagregados segundo três dimensões:

- a natureza da pesquisa — básica, aplicada e atividades científicas e técnicas correlatas;

- os setores que executam ou financiam essas atividades — governo, instituições de ensino superior e empresas;
- a classificação dos recursos de cada um destes setores, obedecendo a critérios específicos para o governo (segundo objetivos socioeconômicos), as instituições de ensino superior (segundo áreas de conhecimento) e as empresas (segundo setores de atividade econômica).

Mais recentemente, foram desenvolvidos os chamados indicadores de resultados, de início limitados à produção científica e, posteriormente, incorporados à produção de patentes e a transferência de tecnologia entre países.

A constituição e a implantação da plataforma Lattes foram iniciativas conjuntas do MCT, CNPq, Capes e Finep. A plataforma é integrada pelos sistemas currículo Lattes e diretório de grupos de pesquisa no Brasil, que apresentam a opção indicadores de produção de C&T e fornecem uma visão quantitativa dos itens de produção científica e tecnológica cadastrados no currículo e diretório, permitindo consultar as distribuições das diferentes variáveis cadastradas.

A plataforma Lattes

É um conjunto de sistemas de informação, bases de dados e portais da internet, concebido para integrar os sistemas de informação das agências federais, racionalizando o processo de gestão de C&T. Lançada em 16 de agosto de 1999, proporcionou um aumento significativo do número de currículos enviados ao CNPq, que chegou a mais de 100 por dia. Segundo dados do Grupo Stela (2002), a plataforma Lattes possui aproximadamente 480 mil currículos cadastrados.

Os investimentos feitos pelo CNPq são direcionados para a formação e absorção de recursos humanos e financiamento de projetos de pesquisa que contribuem para o aumento da produção de conhecimento e geração de novas oportunidades de crescimento para o país. A função de fomento é a principal ação desenvolvida pelo CNPq, com vistas à promoção do desenvolvimento científico e tecnológico do país. Como linha de trabalho mais tradicional e identificadora da missão do órgão, o fomento é dirigido essencialmente para a formação de recursos humanos e para o apoio à realização de pesquisas.

Para que esses objetivos possam ser alcançados de forma plena, o CNPq decidiu que, a partir de 2002, todos os bolsistas de pesquisa, de mestrado, de

doutorado e de iniciação científica, orientadores credenciados e outros clientes do conselho teriam de ter seu currículo cadastrado na plataforma Lattes do CNPq. A inexistência do currículo impediria pagamentos e renovações. O currículo também seria obrigatório para todos os pesquisadores e estudantes que participam do diretório de grupos de pesquisa no Brasil. Apesar disso, a obrigatoriedade não se estabeleceu até os dias atuais, mas, a qualquer momento, os interessados (bolsistas, pesquisadores e estudantes) podem criar ou atualizar seus currículos e enviá-los ao CNPq.

A plataforma Lattes integra, atualmente, quatro sistemas: o primeiro deles se refere a um sistema eletrônico de currículos, que registra a vida pregressa e atual dos pesquisadores. O segundo sistema é o diretório dos grupos de pesquisa no Brasil, uma base de dados que registra todos os grupos de pesquisa em atividade no país. O terceiro sistema é o diretório de instituições, estas demandam fomento ao CNPq e, finalmente, o quarto sistema chama-se sistema gerencial de fomento, cujo objetivo é possibilitar uma gestão estratégica para dar mais qualidade às atividades de fomento do CNPq. Esses quatro sistemas de informação integrados, articulados com outras bases de dados, localizadas fora da agência — a base de patentes do Instituto Nacional de Propriedade Industrial (Inpi), os bancos de dissertações e teses das universidades — constituem a plataforma Lattes.

O Lattes extrator é o instrumento de extração das informações disponibilizadas na plataforma Lattes. Inicialmente, está sendo disponibilizada a extração dos currículos Lattes e, posteriormente, das demais unidades de análise da plataforma. Atualmente, as instituições licenciadas podem extrair diretamente do banco de currículos Lattes do CNPq os dados curriculares de seus pesquisadores, professores, alunos e colaboradores. O Lattes extrator está limitado a extrair do banco de dados do CNPq os currículos de interesse da instituição, por meio de arquivos no formato XML. Com isto, as instituições podem criar seu próprio banco de currículos Lattes e, para tal, podem contar com o modelo e dicionário disponibilizados pelo CNPq (Grupo Stela, 2002).

A hierarquização dos grupos de pesquisa realizada pelo CNPq coloca em evidência as concentrações geográfica e institucional da pesquisa desenvolvida no âmbito das IES; ordena as instituições sob a ótica da pesquisa científica por grande área de conhecimento, tendo em conta os quantitativos de grupos de pesquisa classificados nos diferentes estratos, em termos absolutos e relativos; e, ao final, confere a existência de correlação entre o grau de qualificação e a produtividade técnico-científica de tais grupos. O indicador de produtividade considera a produção de C&T (artigos, livros e capítulos de livros publicados, produção tecnológica desenvolvida, teses e dissertações defendidas sob orien-

tação de pesquisadores pertencentes ao grupo) dos pesquisadores doutores, cadastrada com o auxílio do sistema de currículo Lattes.

Segundo Macias-Chapula (1998), o foco da produção de indicadores de CT&I esteve, por muitos anos, voltado para a medição dos insumos e, apenas recentemente, aumentou o interesse em medir os resultados das atividades científicas e tecnológicas. A produção de indicadores também tem se concentrado em âmbito nacional, institucional ou com enfoque em áreas do conhecimento específicas e ainda são escassos os indicadores das atividades de CT&I em níveis regionais e locais. Ainda segundo os autores, essa é uma lacuna que precisa ser preenchida.

A partir dessa realidade, Hayashi (2002) optou por construir os indicadores de produção científica institucionais, divididos basicamente em dois grupos: os indicadores de produção científica e tecnológica associada à pós-graduação (que envolve as produções caracterizadas como bibliográficas, as formas de divulgação restrita da produção científica e trabalhos publicados em eventos científicos, entre outros) e os indicadores de produção científica e tecnológica associada aos grupos de pesquisa (além da produção bibliográfica, inclui a produção técnica e as orientações concluídas).

4. Gestão de universidades

A gestão de uma instituição de ensino típica é formada por um conjunto de decisões assumidas a fim de obter um equilíbrio dinâmico entre missão, objetivos, meios e atividades acadêmicas e administrativas (Tachizawa e Andrade, 2002). O trabalho desses autores visa estabelecer um modelo de gestão aplicável às instituições de ensino superior (IES).

Segundo Alvarenga e colaboradores (2002), o foco da gestão estratégica do conhecimento em IES está pautado em:

- diferenciação — diferencia os produtos e os serviços ofertados pela organização, visando criar algo que seja considerado único no setor de atuação;
- concentração — capacidade de satisfazer o público-alvo, com o estabelecimento de uma política funcional voltada para o segmento.

O trabalho desenvolvido por Alvarenga e colaboradores (2002) apresenta uma visão do modelo de gestão do conhecimento proposto para ensino e pesquisa na Universidade Católica de Brasília (UCB), elaborado para suportar necessidades de informação e orientar o processo de gestão das atividades da

universidade, por meio da administração do conhecimento gerado internamente. Tem o propósito de apropriar conhecimento, disseminá-lo e garantir sua incorporação aos serviços e processos de decisão, com foco no desenvolvimento humano.

A compreensão da instituição de ensino e da sua inter-relação com os demais agentes do ramo de atividades, o setor educacional ao qual pertence, é essencial para se desenvolver uma proposta de ferramenta de auxílio à gestão do conhecimento, objetivo deste artigo. Faz-se necessário analisar finalidades e missão, bem como identificar produtos, mercados, fornecedores, concorrentes e órgãos normativos oficiais.

Tal compreensão permitirá estabelecer traços comuns a uma IES e delinear estratégias genéricas inerentes a uma instituição de ensino típica. Tachizawa e Andrade (2002) fazem um questionamento acerca da visão das IES. Citando Fernandes (1998), a universidade é uma organização prestadora de serviço que oferece produtos, que são os profissionais formados, capazes de se inserir no âmbito de trabalho e na sociedade em geral.

Vale ressaltar que cada instituição do sistema deve acoplar-se em nível de gestão das políticas públicas e, para que isso ocorra, é necessário que cada uma defina sua política própria e clara quanto a projetos científica e tecnologicamente viáveis e relevantes (Hayashi et al., 2004). Para isso, é necessário identificar suas capacidades específicas e combinações de recursos e competências, aproveitando bem suas características próprias, além de contextuais e estabelecer formas de parcerias com outras instituições do sistema de C&T.

Por parceiros, entendem-se as entidades/agentes que fornecem recursos às IES na forma de bens, capital, materiais, equipamentos e demais recursos que, por sua natureza, constituem os insumos necessários às atividades internas das instituições de ensino. Nesse contexto, a figura do professor surge como o principal parceiro (colaborador ou fornecedor) da IES (Tachizawa e Andrade, 2002).

Considerando que o produto final de uma IES é o aluno formado, capacitado e habilitado a exercer a profissão para a qual se preparou, o cliente é a organização empregadora desse profissional colocado no mercado. Mercado compreende o conjunto de clientes, constituído das organizações que potencialmente irão absorver os profissionais formados e colocados disponíveis pelas instituições de ensino.

À medida que o gestor de IES tem êxito em integrar o cliente e unir os interesses deste aos objetivos preestabelecidos no plano estratégico (projeto pedagógico) da instituição de ensino, refluiriam os resultados que assegura-

riam o cumprimento da missão e, sobretudo, a sobrevivência (continuidade). São esses resultados que de fato importam à comunidade como um todo e ao gestor da IES em particular (Tachizawa e Andrade, 2002).

Gestão do conhecimento nas relações universidade x empresa: prioridades distintas

Apesar da existência de uma analogia entre universidades e organizações mercadológicas (empresas), elas possuem algumas diferenças que devem ser consideradas. As universidades estão voltadas para a criação e a disseminação do conhecimento. Algumas metas existem, porém, raramente são feitos projetos de pesquisas onde se definem claramente prazos finais. Já com respeito às empresas, há a preocupação com cronogramas, com o cumprimento de metas e outras atividades em curto prazo, no contexto de um ambiente altamente competitivo.

As universidades e as empresas empregam linguagens distintas; enquanto a primeira se preocupa com a codificação do conhecimento, a segunda está voltada ao conhecimento direcionado à geração de produtos. Por exemplo: hipóteses, modelos e variáveis, termos importantes no idioma dos pesquisadores da universidade não possuem a menor importância no vocabulário da maior parte dos representantes das empresas.

Os ambientes de trabalho na universidade e na empresa são bastante diferentes. Para os pesquisadores da universidade, a reputação no meio intelectual é a maior força motivacional, ficando assim o foco de referência situado do lado de fora da organização, em seu grupo de referência profissional.

A universidade não entende as forças de mercado, as demandas de tempo e as estruturas de incentivo da empresa. Já na empresa, para a maioria dos gerentes envolvidos com pesquisa e desenvolvimento, o superior hierárquico é o referencial crítico. As avaliações de desempenho vêm desta fonte e levam em conta resultados específicos provenientes de sua atuação no trabalho. Da mesma forma, a empresa não entende como tal o trabalho realizado nas universidades, nem são familiarizados com os investimentos em recursos humanos e capital físico, que precederam sua relação com a universidade (Alvarenga et al., 2002).

Outro ponto crucial é que os interesses dos pesquisadores da universidade podem mudar e a universidade os deixa relativamente livres para abandonar determinados projetos e ingressar em outros mais motivadores.

Os objetivos das duas organizações mercadológicas são bastante diferentes. A maioria das empresas quer aplicações concretas, quando estabelecem parcerias ou convênios com universidades, visam ao acesso a procedimentos inovadores, soluções de seus problemas, novo conhecimento científico, novas ferramentas e metodologias e novos produtos e serviços. A natureza da pesquisa tecnológica, porém, é complexa, ambígua e abstrata. Muito do conhecimento gerado pode ser tácito, significando que seus princípios subjacentes são difíceis de identificar e articular. Além disso, provavelmente existirão longos espaços de tempo entre o início do projeto e a criação de produtos. Todas essas características podem criar crises, enganos e dificuldades na transferência do conhecimento.

Já as universidades trabalham para a obtenção de um produto muito diferente, que pode ser caracterizado a partir de contribuições para o conhecimento, na forma de novos conceitos, modelos, soluções empíricas, técnicas de medidas e outras contribuições tecnológicas.

5. Metodologia

Para este artigo foram utilizadas as pesquisas bibliográfica e documental e a metodologia de estudo de caso. Além disso, foi aplicado todo o processo de descoberta de conhecimento em bancos de dados.

A pesquisa bibliográfica deu base para a aquisição de conhecimento acerca dos temas envolvidos no projeto, como gestão do conhecimento, mecanismos de descoberta de conhecimento em bancos de dados e técnicas para a construção do sistema de mineração de dados. Envolveu, basicamente, consultas a livros de referência, teses e artigos científicos.

A pesquisa documental foi realizada em documentos referentes à pesquisa científica na Ufla, obtidos a partir do Lattes extrator, que proporcionaram informações úteis para as análises, comparações e para o desenvolvimento da ferramenta de *data mining*. Também foram pesquisados documentos da Ufla referentes às políticas de incentivo ao desenvolvimento de CT&I.

O método do estudo de caso é considerado um tipo de análise qualitativa. Não é uma técnica específica; é um meio de organizar dados sociais preservando o caráter unitário do objeto social estudado (Goode e Hatt, 1969). Bonoma (1985) coloca que o estudo de caso é uma descrição de uma situação gerencial. Esse método, assim como os métodos qualitativos, são úteis quando o fenômeno a ser estudado é amplo e complexo, quando o corpo de conhecimentos existente é insuficiente para suportar a proposição de questões causais

e nos casos em que o fenômeno não pode ser estudado fora do contexto onde naturalmente ocorre.

Yin (1989) afirma que o estudo de caso é uma inquirição empírica que investiga um fenômeno contemporâneo dentro de um contexto da vida real. De acordo com Yin (1989), a preferência pelo uso do estudo de caso deve ser dada quando do estudo de eventos contemporâneos, em situações nas quais os comportamentos relevantes não podem ser manipulados, mas é possível se fazer observações diretas e sistemáticas.

O estudo de caso de que trata este artigo foi realizado na Universidade Federal de Lavras (Ufla), mais especificamente nos setores envolvidos com o desenvolvimento de pesquisa científica. O estudo utilizou dados de fontes secundárias como base para as análises, extraídos dos currículos de pessoas ligadas, de forma direta e indireta, à pesquisa científica da Ufla. Os dados foram disponibilizados pelo uso da ferramenta Lattes extrator, que faz parte da plataforma Lattes.

Entre as etapas predefinidas da técnica de descoberta de conhecimento em bancos de dados (DCBD) foram realizadas:

- seleção dos dados — por meio do Lattes extrator, foram selecionados e extraídos, inicialmente, mais de mil documentos da plataforma Lattes, que continham os registros de toda a produção científica dos docentes, de alunos, ex-alunos, mestrandos e doutorandos da Ufla, entre outras pessoas. Em seguida, foram selecionados 575 currículos que continham dados específicos referentes às produções científica, tecnológica e bibliográfica dos mesmos, principalmente dos professores;
- pré-processamento dos dados — realizado a partir da eliminação de incongruências e/ou erros dos dados (filtragem). Os dados selecionados na etapa anterior ainda continham algumas inconsistências, como ausência de especificação de campos importantes e duplicação de outras especificações. Filtrando-se essas informações, o banco de dados resultante passou a conter 28.389 linhas. Nessa etapa do processo de DCBD não foi realizado o enriquecimento dos dados pelo fato de eles serem referentes a outras pessoas, extraídos dos documentos disponíveis na plataforma Lattes, que já continha as informações necessárias à descoberta de conhecimento proposta;
- transformação dos dados — foram feitos dois tipos de codificação de dados. O primeiro consistiu na transformação dos documentos obtidos no formato XML (dados semi-estruturados) em documentos SQL (BD relacional), contendo o código de inserção e os dados a serem inseridos no banco de dados. O segundo tipo foi, basicamente, a execução desses códigos SQL, gerados

na codificação anterior, no sistema gerenciador de bancos de dados (SGBD) da Oracle;

- *data mining* — a etapa consistiu na elaboração de algumas tarefas de *data mining*, pela implementação de técnicas específicas para esse fim, realizando-se o cruzamento e a comparação de consultas e funções definidas na linguagem de programação PL/SQL, própria do SGBD Oracle;
- interpretação — a interpretação dos resultados obtidos, que gera o conhecimento, é demonstrada a partir da criação de relatórios. O principal relatório desenvolvido foi uma dissertação de mestrado apresentada ao Departamento de Administração e Economia da Ufla, que contém, além de todo o referencial teórico acerca do tema, os resultados apresentados de diversas formas, desde gráficos resumidos até a descrição dos principais resultados.

6. O estudo empírico: gestão de ciência, tecnologia e inovação na Ufla

A Ufla é uma instituição federal de ensino superior, localizada na cidade de Lavras, ao sul do estado de Minas Gerais. É uma universidade com 95 anos de história dedicada à manutenção da alta qualidade do ensino, da pesquisa e da extensão. Atualmente, oferece 10 cursos de graduação e 28 cursos de pós-graduação presenciais. Diretamente ligados às atividades de pesquisa da Ufla estão 302 professores, 2.342 estudantes de graduação e 786 pós-graduandos (PRP, 2004).

Os mais de 200 doutores pesquisadores da Ufla, além de inúmeros mestres, pós-graduados, bolsistas de iniciação científica e técnicos de laboratório desenvolvem suas pesquisas em cerca de 60 laboratórios especializados, bem equipados e estruturados para pesquisa científica e ou tecnológica, além de contarem com vários setores temáticos. Desenvolvem, em parcerias com empresas estatais e privadas, inúmeros projetos e programas de cooperação técnico-científico (PRP, 2004).

A Ufla conta com aproximadamente 65 grupos, que desenvolvem 350 linhas de pesquisa, que compõem os projetos isolados e programas especiais. A universidade é bastante competitiva na captação de recursos nas agências de fomento para as atividades de C&T e disponibiliza seus recursos humanos e infra-estrutura para projetos em cooperação e consultorias nas mais diversas áreas de atuação. Em seu planejamento estratégico, ações estão sendo implementadas para viabilizar um modelo de gestão eficiente da pesquisa, visando

maximizar recursos materiais, humanos e financeiros, de modo a ampliar essa atividade e aumentar sua aplicabilidade e inserção na sociedade.

Desenvolver pesquisa é a grande motivação e incentivo dos docentes, devido à valorização pessoal e profissional; à complementaridade da atividade universitária, uma vez que a pesquisa é parte de sua missão; à contribuição à atividade didático-pedagógica, pois evita repasse copiado de informações; à progressão funcional da carreira do docente; ao incentivo financeiro; às possibilidades de assessoria/consultoria, como tarefas de extensão; ao reforço financeiro para o sistema, advindo de auxílios externos; e à facilitação de inserção na comunidade, que é missão social da universidade.

A contribuição científica e tecnológica da Ufla tem como principais objetivos resgatar os principais resultados da pesquisa na universidade, fazer uma análise crítica da contribuição e do impacto destes para C&T, difundir e ampliar a sua participação no discurso científico e tecnológico nacional. Diversas ações estão sendo implementadas nesse sentido. O controle das atividades de pesquisa é feito pela Pró-Reitoria de Pesquisa, que verifica se os projetos estão sendo apreciados e aprovados em assembléia departamental, se os departamentos estão estabelecendo um banco de projetos, entre outros.

Resultados e discussões

O pressuposto inicial de que há uma grande quantidade de informação e conhecimento "escondidos" nos registros da pesquisa científica da Ufla é bastante válido, uma vez que a riqueza de informações obtidas a partir das respostas alcançadas com as consultas poderia ser mais aproveitada pelos órgãos de direção da universidade envolvidos na pesquisa científica.

Verificou-se que os dados presentes nos currículos extraídos da plataforma não estavam atualizados, o que foi uma limitação para este artigo. Até o presente momento, as informações disponíveis no site oficial do CNPq são de que a versão do Lattes extrator que está disponível extrai apenas currículos atualizados até julho de 2002. De acordo com o site, está sendo desenvolvida uma nova versão que permitirá a extração de currículos mais atualizados, logo que estiver disponível (Grupo Stela, 2002). Logo, apesar da limitação, uma vez disponibilizados novos dados, o mesmo trabalho poderá ser realizado, apenas executando-se as funções já criadas para gerar conhecimento mais atualizado.

Um dos grandes problemas encontrados para realizar a análise dos dados é a falta de padronização dos valores cadastrados. Por exemplo, existem 72

cargos diferentes, 46 órgãos diferentes e 172 unidades distintas cadastrados nos currículos de pessoas ligadas à Ufla. Muitos desses dados, na realidade, representam um mesmo objeto, tal como o Departamento de Administração e Economia que pode, ao mesmo tempo, ser cadastrado como um órgão ou unidade. E mais, esse mesmo departamento poderia ser novamente cadastrado pela sigla DAE. Todas as diferentes formas de cadastrar esse objeto deveriam ser representadas de forma única. Há também casos em que um mesmo objeto é cadastrado de forma redundante em tabelas diferentes, como é o caso de órgãos e unidades.

Outro problema refere-se ao próprio formato do currículo Lattes, que não deixa claro qual é a função de cada pessoa ligada à Ufla. Por exemplo, os dados referentes ao vínculo profissional das pessoas podem ser de seis tipos: celetista, colaborador, livre, outro, professor visitante e servidor público. Observa-se que não há o vínculo definido como "professor", o que torna difícil afirmar com segurança quais são os professores da Ufla, pois existem professores cadastrados como servidor público, livre ou outro. Considerou-se que uma pessoa é professor na Ufla quando possui atividades de ensino cadastradas em cursos oferecidos pela instituição. Porém, não se pode afirmar com exatidão quem são as pessoas que não são professores na Ufla, pois podem existir casos de professores que não cadastraram suas atividades de ensino em seus currículos.

Além dessas limitações, outro fato que prejudicou a análise dos resultados gerados é que poucas pessoas atualizam seus currículos Lattes periodicamente e, quando o fazem, a maioria o faz de forma parcial. Um resultado crítico que advém desse fato é que, dos 575 currículos inseridos no banco de dados, mais de 90% não contêm atividades cadastradas. As atividades podem ser de ensino, pesquisa, direção e extensão, além de serviços técnicos e treinamentos ministrados, que ocorreram ao longo dos anos, ou seja, uma só pessoa pode possuir, por exemplo, diversas atividades de direção cadastradas ao longo de toda a sua carreira. Os menos de 10% das pessoas, exatamente 55 pessoas, que incluíram suas atuações profissionais em seus currículos, têm entre duas e 61 atuações, demonstrando uma variedade muito grande de número de atividades, chegando ao número total de 792 atuações distintas.

O que se pôde observar é que apenas 39 pessoas, aproximadamente 6% do total de currículos cadastrados no banco de dados, realizaram entre uma e 16 atividades de ensino, de um total de 119 atividades cadastradas. Uma observação importante é que essas atividades de ensino erroneamente incluíam atividades de direção como, por exemplo, a gerência de organizações. Dos primeiros resultados, apenas observando-se os números de atividades cadas-

tradas pelas pessoas, é interessante analisar que as mesmas, ao preencherem seus currículos na plataforma Lattes, dão maior prioridade às atividades de ensino e pesquisa do que às demais.

Por outro lado, analisando-se as produções bibliográficas, observou-se que foram publicados 573 artigos de 1968 até o princípio de 2004, a maior parte deles publicada em 2001. Vale ressaltar que como o banco de dados é oficialmente atualizado até julho de 2002, estranha-se o fato de haver publicações cadastradas até o princípio de 2004. Entre esses artigos, 6,4% foram publicados no exterior e a maioria possui de três a cinco autores, com alguns possuindo até oito autores. No caso da Ufla, dos 573 artigos publicados, 77% pertencem à área de ciências agrárias; 13% à de ciências biológicas; 2,3% à de ciências da saúde; 5,4% à de ciências exatas; 0,3% à de ciências humanas; 1,7% à de ciências sociais aplicadas; e 0,3% às áreas de engenharias. Esses foram alguns dos primeiros resultados obtidos com a aplicação do processo de descoberta de conhecimento em banco de dados.

Com a utilização das técnicas de *data mining*, foram criadas funções específicas para descobrir padrões de comportamento mais relevantes nos dados disponíveis. Esses resultados são enquadrados nas categorias de conhecimento que podem ser geradas pela técnica de *data mining*.

Análises de regras de associação

Um primeiro exemplo mostra a associação entre a quantidade de publicações realizadas por pessoas que trabalham na Ufla e as que não trabalham. Essa função envolveu um total de 11 tabelas do banco de dados, das quais sete são relacionadas às atuações e quatro relacionadas às diversas formas de publicações. Foram obtidas 1.977 publicações; destas, 55% foram publicadas por pessoas que não estavam atuando na Ufla na época da publicação e 45% por pessoas que atuavam na Ufla na época da publicação. Vale analisar nesse exemplo que uma pessoa, ao receber afastamento total para treinamento, fazer pós-graduação, por exemplo, não está atuando na Ufla durante o período do afastamento. Isso poderia explicar o resultado encontrado, já que no mestrado e ou doutorado, realiza-se mais pesquisa e publica-se mais. Isto também poderia refletir o fato de que, ao estar atuando na Ufla em atividades de ensino e direção, as pessoas podem ficar com a sua carga horária sobrecarregada e, conseqüentemente, acabem por realizar um número menor de pesquisas e publicações.

Outro exemplo explora um pouco mais os resultados obtidos no exemplo anterior. Refere-se aos 55% das pessoas que não estavam atuando na Ufla,

associados à quantidade de suas publicações nesse período de ausência. Essa função envolveu um total de sete tabelas do banco de dados, sendo três relacionadas às atuações e quatro relacionadas às diversas formas de publicações. No total foram realizadas 1.062 publicações por pessoas que não estavam atuando na Ufla no momento da publicação.

Mais um exemplo de regra de associação mostra a relação entre todas as publicações cadastradas e o tempo de serviço de seus autores na Ufla. Essa função envolveu um total de 11 tabelas do banco de dados, sendo sete delas relacionadas às atuações e quatro tabelas relacionadas às diversas formas de publicações. No total, foram obtidas 915 publicações relacionadas ao tempo de serviço de seus autores com a Ufla. Analisando-se o exemplo, percebe-se que a maioria das publicações feitas por pessoas que atuam na Ufla foi realizada depois que elas começaram a trabalhar na universidade.

Os dois próximos exemplos de regras de associação buscam mostrar a relação existente entre o fato das pessoas terem realizado pós-graduação no exterior ou no Brasil e o fato de essas pessoas terem publicado no exterior.

A relação entre o local onde foi realizada a pós-graduação e o número de publicações no exterior envolveu duas tabelas relacionadas à pós-graduação e quatro relacionadas aos tipos de publicações. No total, foram 74 publicações realizadas no exterior por 42 pessoas, com a maioria delas escrita por pessoas que fizeram pós-graduação no Brasil.

Esse resultado deve-se ao fato de que, nesse banco de dados, o número de pessoas que cursaram pós-graduação no Brasil (34 pessoas) ser muito maior do que o das que cursaram no exterior (oito pessoas). Assim, é natural que o número de publicações no exterior seja maior para o grupo de 34 pessoas do que para o outro. Porém, esse resultado está ligado a outra medida que trata da média de publicações no exterior por cada pessoa. A média de publicações no exterior de pessoas que cursaram a pós-graduação fora do Brasil é maior numa razão de 2,71 com relação às pessoas que cursaram pós-graduação no Brasil. A função que chegou a esse resultado envolveu um total de seis tabelas, sendo duas relacionadas a pós-graduação e quatro relacionadas a publicações. Essa relação indica que quem faz pós-graduação no exterior tende a ter maior visibilidade fora do Brasil, em termos de publicações, do que quem faz pós-graduação no Brasil.

Análises de regras de associação e outliers

Um exemplo tentou verificar a relação existente entre as atividades de pesquisa e o número de linhas de pesquisa nela envolvidas. Nos resultados obtidos

foram analisadas a regra de associação e a ocorrência de *outlier*. Pelo resultado percebe-se a presença de três pessoas com um número muito superior de linhas de pesquisa para suas atividades de pesquisa, podendo ser considerados *outliers*. Essa função envolveu tabelas relacionadas às linhas de pesquisa, grande área, área e subárea, e tabelas relacionadas às atividades de pesquisa desempenhadas. No total, foram obtidas 84 pesquisas e 186 linhas de pesquisa.

Vale esclarecer, para esse banco de dados, a distinção que existe entre os termos "linha de pesquisa" e "atividade de pesquisa". No currículo Lattes, cada atividade de pesquisa na qual uma pessoa está envolvida durante um certo período (por exemplo, qualquer projeto de pesquisa envolvendo um grupo de pessoas ou isolado) pode estar ligada a uma ou mais linhas de pesquisa. As linhas de pesquisa para cada atividade são definidas pelas pessoas ao preencherem seu currículo.

O mesmo fato ocorre com as grandes áreas, áreas e subáreas ligadas às atividades de pesquisa. Uma atividade de pesquisa deve possuir uma grande área e pode possuir uma ou mais áreas e subáreas associadas a ela. Uma pessoa não pode criar uma nova grande área e incluí-la em seu currículo. Porém, as áreas e subáreas não são predefinidas, ou seja, uma pessoa pode criar uma nova área ou subárea para enquadrar sua atividade de pesquisa. Como esses campos são abertos no banco de dados, a tarefa de comparar esses dados é bastante complexa.

Análises de regras de associação e de padrão seqüencial

O objetivo da consulta era avaliar se havia uma relação entre o tempo de conclusão do mestrado e o tempo de início do doutorado. Pela imagem percebe-se um padrão de comportamento, pois a maioria das pessoas leva entre zero e três anos de intervalo entre esses dois tipos de pós-graduação. Nessa mesma consulta pôde-se observar a presença de *outliers* como pessoas que levaram mais de 20 anos entre o mestrado e o doutorado. Essa função envolveu a tabela contendo dados gerais das pessoas e duas tabelas sobre pós-graduação. No total, o resultado envolve 483 pessoas do banco de dados que cursaram mestrado e doutorado.

Análises de padrões seqüenciais

Os exemplos a seguir mostram padrões de comportamento seqüencial dos dados com relação ao tempo. Uma consulta avalia se há uma relação entre o

tempo de cadastramento do currículo e o tempo de vínculo profissional com a Ufla. Pelo resultado, percebe-se um padrão de comportamento, pois a grande maioria das pessoas cadastrou seu vínculo profissional com a Ufla a partir dos anos 1990. Nessa consulta, a função elaborada envolveu as tabelas de dados gerais das pessoas, as tabelas de atuações e a de vínculo profissional, num total de 82 ocorrências.

Outra consulta avalia se há uma relação temporal entre o tempo de serviço das pessoas ligadas à Ufla e o ano de início de suas pesquisas cadastradas. Pelo resultado percebe-se um padrão de comportamento, pois a maioria das pessoas cadastrou suas pesquisas mais recentes nos seus currículos. A função elaborada envolveu as tabelas de dados gerais das pessoas, as de atuações e a de atividades de pesquisa, num total de 79 pesquisas.

Análises de clusters

O exemplo a seguir faz a análise de um agrupamento (*cluster*) que inicialmente era desconhecido e surgiu a partir da consulta para verificar a duração, em anos, das pesquisas realizadas por pessoas da Ufla. Além das pesquisas que estão em andamento e não se pode afirmar a sua duração exata, a maioria das pesquisas dura entre dois e três anos.

Análise de classificação e predição

A análise de classificação difere do agrupamento porque parte de grupos predefinidos dos dados. Como as características dos dados extraídos da plataforma Lattes não têm padrão definido, a tarefa de analisar os grupos já existentes tornou-se muito complexa, uma vez que faltava conhecimento da pesquisadora em agrupar, por exemplo, linhas ou áreas de pesquisa. Por isso, apenas um exemplo será apresentado.

A consulta dividiu as atividades realizadas pelas pessoas da Ufla em três grupos: pesquisa, ensino e direção. O objetivo foi observar, entre os currículos cadastrados, como foi a distribuição das publicações realizadas por pessoas enquanto estavam exercendo cada uma dessas atividades. De um total de 101 publicações. Essa função envolveu três tabelas relacionadas às atividades e quatro tabelas relacionadas aos diversos tipos de publicações.

O resultado mostra que a maioria das publicações foi realizada enquanto as pessoas exerciam atividades de pesquisa; outra parte do total foi quan-

do as pessoas exerciam atividades de ensino e, em menor número, enquanto exerciam atividades de direção. Porém, os agrupamentos não são disjuntos, ou seja, uma pessoa poderia estar ao mesmo tempo realizando diferentes tipos de atividades no momento da publicação. Mesmo assim, esse é um resultado significativo, pois mostra claramente que, dependendo do tipo de atividade em que a pessoa está envolvida, a quantidade de publicações que ela irá realizar sofrerá influência.

7. Conclusão

O objetivo deste artigo foi construir e analisar uma ferramenta de *data mining*, como parte do processo de descoberta de conhecimento em banco de dados, para extrair conhecimento referente à produção científica das pessoas envolvidas com a Ufla, por meio dos dados extraídos da plataforma Lattes. Para tanto, foi implementado um programa para transformar os dados semi-estruturados selecionados dessa plataforma num banco de dados estruturado criado no Oracle. A partir daí, foi desenvolvida uma ferramenta automática de descoberta de conhecimento, utilizando a técnica de *data mining*, cujos resultados gerados foram analisados. Entende-se, portanto, que os objetivos foram alcançados.

Os resultados considerados mais expressivos e sua análise podem ser assim sintetizados. Com relação às limitações e aos problemas envolvendo os dados extraídos da plataforma Lattes:

- um dos grandes problemas encontrados para realizar a análise dos dados é a falta de padronização dos valores cadastrados;
- outro problema refere-se ao próprio formato do currículo Lattes, que não deixa claro qual é a função de cada pessoa ligada à instituição;
- poucas pessoas atualizam seus currículos Lattes periodicamente e, quando atualizam, a maioria dos currículos é preenchida de forma parcial.

Dos primeiros resultados apresentados observando-se os números de atividades cadastradas, é interessante perceber que, ao preencherem seus currículos na plataforma Lattes, dá-se maior prioridade às atividades de ensino e pesquisa do que às demais.

Com relação às publicações:

- percebe-se que a grande maioria delas pertence à grande área de ciências agrárias;

- pessoas que não estão atuando na Ufla publicam mais do que quando estão; o fato de não estar atuando pode significar que possa estar fazendo pós-graduação e, por isso, tende a uma maior quantidade de produção e, conseqüentemente, de publicação. Por outro lado, ao estarem atuando na Ufla em atividades de ensino e direção, as pessoas têm menor disponibilidade de tempo para a produção de trabalhos em pesquisa, conseqüentemente, um número menor de pesquisas e publicações;
- a média de publicações no exterior por pessoa é maior para aquelas que cursaram pós-graduação fora do Brasil;
- a maioria das publicações foi realizada enquanto as pessoas exerciam atividades de pesquisa, seguidas pelas pessoas que exerciam atividades de ensino e, por fim, enquanto exerciam atividades de direção.

É clara a importância dos indicadores de CT&I nas IES. Um esforço deve ser realizado para criar tais indicadores para a Ufla.

A plataforma Lattes, uma vez devidamente atualizada, é uma enorme fonte de informação para a geração de conhecimento útil para a gestão das IES.

Diante dos resultados apresentados, pode-se perceber que, com essa ferramenta, é possível obter-se uma visão mais abrangente dos dados institucionais, pelo fato de ter sido disponibilizada uma grande quantidade de informações sobre a pesquisa científica da Ufla. Portanto, é possível iniciar uma melhoria na gestão do conhecimento dessa instituição fazendo uso dessas informações, pois é exatamente essa a base da gestão do conhecimento: dados integrados, gerando informações analíticas e abrangentes.

Alguns exemplos práticos da aplicabilidade desses resultados na Ufla poderiam ser:

- a partir da verificação da distribuição das atividades de ensino, de pesquisa e de direção, decisões poderiam ser tomadas para tentar não sobrecarregar as pessoas alocadas em determinados órgãos ou unidades, em detrimento de outros;
- analisar os diversos casos de pessoas que fogem ao padrão (*outliers*) dos demais, tentando verificar se esse é ou não um bom comportamento, e se esse deveria ser seguido, formando um novo padrão ou, ao contrário, ser evitado;
- a partir dos agrupamentos de pessoas que inicialmente não estão diretamente ligadas a nenhum departamento ou grupo de pesquisa, criar novas linhas ou áreas de pesquisas, que poderiam ser potencialmente melhor aproveitadas;

- a partir dos diversos padrões de comportamento observados nas informações que foram apresentadas, decisões podem ser tomadas não somente a curto prazo, mas também a longo prazo, pois é possível prever de forma segura prováveis comportamentos futuros;
- as diversas regras de associação que foram apresentadas mostram que dados que aparentemente não estão relacionados, na realidade, possuem aspectos em comum, que podem ser explorados etc.

Apesar de ter sido aplicada em uma área específica, a pesquisa científica na Ufla, o trabalho demonstrou como é possível também utilizar tecnologias da informação para auxiliar na gestão de conhecimento disponível nas instituições de ensino superior. Diversos padrões e associações foram identificados por meio da aplicação da descoberta de conhecimento em banco de dados; porém, há muitas outras descobertas que ainda podem ser feitas aproveitando-se o banco de dados criado.

Por fim, pode-se dizer que o projeto foi apenas um passo para o desenvolvimento de um grande trabalho de mudança na gestão do conhecimento nas atividades gerenciais da Ufla e, quem sabe, futuramente, de outras universidades. O sistema desenvolvido poderá ser incrementado e utilizado em trabalhos futuros, como: atualização da base de dados a partir da nova versão do Lattes extrator; entrevistas com pessoas-chave para estabelecer novos critérios de exploração dos dados, gerando descoberta de novas informações e novo conhecimento, trazendo melhorias para a ferramenta desenvolvida; criação de uma comissão que elabore normas para o preenchimento e atualização dos currículos Lattes das pessoas envolvidas com a pesquisa científica na Ufla; criação de indicadores de CT&I para a Ufla, com o objetivo de auxiliar a elaboração de novas políticas de gestão; a aplicação da ferramenta desenvolvida nos currículos atualizados, assim que eles estejam disponíveis na plataforma Lattes, e comparação dos novos resultados obtidos com os resultados obtidos neste artigo; e aplicação dessa ferramenta em outras instituições de ensino superior, com o objetivo de comparar seus resultados com os obtidos na Ufla.

Referências bibliográficas

ADRIAANS, P.; ZANTINGE, D. *Data mining*. Harlow: Addison-Wesley, 1996. 158p.

ALVARENGA, R. et al. Gestão de conhecimento para ensino e pesquisa: o modelo da UCB. In: CONGRESSO ANUAL DA SOCIEDADE BRASILEIRA DE GESTÃO DO

CONHECIMENTO. *Anais*... São Paulo, 2002. Disponível em: <www.cori.rei.unicamp.br>. Acesso em: 10 out. 2004.

AMO, S. *Curso de data mining*: programa de mestrado em ciência da computação. Uberlândia: Universidade Federal de Uberlândia, 2003. Disponível em: <www.deamo.prof.ufu.br/CursoDM.html>. Acesso em: 5 jul. 2004.

BONOMA, T. V. Case research in marketing: opportunities, problems, and process. *Journal of Marketing Research*, v. 22, maio 1985.

CARVALHO, R. B. *Aplicações de softwares de gestão do conhecimento:* tipologia e usos. 2000. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Minas Gerais, Belo Horizonte.

COELHO, M. I .M. Gestão de C&T: o que é. In: _____. *Gestão de C&T*: planejamento de pesquisa e captação de recursos. 2002. Disponível em: <http://netpage.em.com.br/mines>. Acesso em: 4 out. 2004.

DECKER, K.; FOCARDI, S. *Technological overview*: a report on data mining. CSCS — Swiss National Supercomputing Center, Technical Report, Zurique, 1995. Disponível em: <ftp://ftp.cscs.ch/pub/CSCS/>. Acesso em: 17 mar. 2004.

ELMASRI, R.; NAVATHE, S. B. *Sistemas de banco de dados:* fundamentos e aplicações. 3. ed. Rio de Janeiro: LTC, 2002.

FAYYAD, U. M. et al. From data mining to knowledge discovery: an overview. In: *Advances in knowledge discovery and data mining*. California: AAAI/The MIT, 1996. p.1-34.

FERNANDES, C. V. *Qualidade total no ensino superior*. Rio de Janeiro: Universidade Gama Filho, 1998.

GOODE, W. J.; HATT, P. K. *Métodos em pesquisa social*. 3. ed. São Paulo: Cia. Editora Nacional, 1969.

GRUPO STELA. *Lattes extrator*. Florianópolis: Universidade Federal de Santa Catarina, 2002. Disponível em: <http://lattes.cnpq.br/lattesextrator/>. Acesso em: 7 out. 2004.

GROSSMAN, R. L.; HORNICK, M.; MEYER G. Emerging KDD Standards. In: *Communications of the ACM*, 2002. (Special issue on data mining).

HAYASHI, M. C. P. I. Os indicadores de C&T como ferramenta de gestão da informação científica e tecnológica no contexto universitário. In: CONGRESSO ANUAL DA SOCIEDADE BRASILEIRA DE GESTÃO DO CONHECIMENTO. São Paulo, 2002. *Anais*... São Paulo: SBGC, 2002. 16p.

_____ et al. Ciência, tecnologia e inovação no pólo tecnológico de São Carlos. São Carlos: Universidade Federal de São Carlos/Departamento de Ciência da Informação, 2004. Disponível em: <www.cori.rei.unicamp.br/IAU>. Acesso em: 2 out. 2004.

KING, D. *Numerical machine learning*. Georgia: Tech College of Computing, 2003. Disponível em: <www.cc.gatech.edu/kingd/datamine/datamine.html>. Acesso em: 22 mar. 2004.

KROGH, G. V.; ICHIJO, K.; NONAKA, I. *Facilitando a criação de conhecimento*. Rio de Janeiro: Campus, 2001.

LAUDON, K. C.; JANE, P. *Gerenciamento de sistema de informação*. 3. ed. Rio de Janeiro: LTC, 1999.

MACIAS-CHAPULA, C. A. O papel da infometria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

MOXTON, B. *Defining data mining*. DBMS Data warehouse supplement site, 2004. Disponível em: <www.dbms.mfi.com/9608d53.html>. Acesso em: 20 mar. 2004.

NAVEGA, S. Princípios essenciais do data mining. In: INFOIMAGEM. 2002. *Anais*... Cenadem, nov. 2002. Disponível em: <www.intelliwise.com/snavega>. Acesso em: 14 mar. 2004.

UNIVERSIDADE FEDERAL DE LAVRAS. *Pró-Reitoria de Pesquisa da Ufla apresenta informações sobre a pós-graduação da Ufla*. Disponível em: <www.prp.ufla.br>. Acesso em: 20 mar. 2004.

TACHIZAWA, T.; ANDRADE, R. O. B. *Gestão de instituições de ensino*. 3. ed. Rio de Janeiro: FGV, 2002.

TARAPANOFF, K. (Org.). *Inteligência organizacional e competitiva*. Brasília: Universidade de Brasília, 2001.

YIN, R. K. *Case study research*: design and methods. USA: Sage, 1989.