

## O processo da bibliomineração: repositório de dados e mineração de dados para tomada de decisão em bibliotecas<sup>1</sup>

### *The bibliomining process: data warehousing and data mining for library decision-making*

Scott NICHOLSON<sup>2</sup>

#### RESUMO

Apresenta uma metodologia operacional integrada, utilizando técnicas de repositórios de dados (*data warehousing*) e mineração de dados (*data mining*), e combinando ferramentas estatísticas com reconhecimento de padrões. Tendo em vista a utilização de recursos, o objetivo principal é chegar a um entendimento melhor do comportamento de comunidades de usuários. O anonimato dos usuários fica inteiramente assegurado na aplicação do modelo, pelo qual são processados os dados oriundos de arquivos transacionais da biblioteca. O próprio processo da bibliomineração consiste basicamente em coletar, depurar e anonimizar os dados, visando, precipuamente, a descoberta de padrões subjacentes. Os resultados podem ser combinados com diferentes variáveis, antes de sua utilização na tomada de decisão pela Administração da biblioteca.

**Palavras-chave:** bibliomineração, repositório de dados, data warehouse, mineração de dados (datamining), tomada de decisão, administração de biblioteca.

<sup>1</sup> Os termos *bibliomineração*, *repositório de dados* e *mineração de dados* são neologismos aqui utilizados, respectivamente, para traduzir *bibliomining*, *data warehousing* e *data mining* (Nota do tradutor).

<sup>2</sup> Ph.D., Syracuse University School of Information Studies, Center for Science and Technology, Syracuse, NY 13026. E-mail: <smichol@syrr.edu> ou <<http://bibliomining.org>>.

**Tradução, adaptação e resumo:** Prof. Dr. Ulf Gregor BARANOW, Departamento de Ciência e Gestão da Informação, Universidade Federal do Paraná. Av. Prefeito Lothário Meissner, 3400, Campus III, Jardim Botânico, 80210-170, Curitiba, PR, Brasil. E-mail: <[ugb@ufpr.br](mailto:ugb@ufpr.br)>.

Recebido para publicação em 21/9/2004.

## ABSTRACT

*In order to get a better understanding of library user communities and to improve resource utilization, the author proposes an integrated operational application of data warehousing and datamining, combining statistical and pattern recognition tools, both integrated under a new concept named bibliomining. With the application of this model to transactional records of the library, patron privacy is fully respected through the creation of an anonymized version of the operational data, before reformatting them for bibliomining analysis. The bibliomining process itself consists mainly of collecting, cleaning and anonymizing the data, in order to discover underlying behavior patterns. The results can be combined with several variables, before being used to improve decision-making by the library administration.*

**Key words:** *bibliomining, data warehousing, datamining, decision-making, library administration.*

## INTRODUÇÃO

Um dos desafios lançados por Buckland (2003) aos profissionais de Biblioteconomia diz respeito à necessidade de se conhecer melhor as comunidades de usuários. A maior parte das técnicas atualmente utilizadas nesta área é concentrada em medidas de frequência e de agregação, ou seja, em métodos estatísticos que não explicitam os padrões subjacentes. Entretanto, é na descoberta desses padrões que se encontra a chave para conhecer melhor as comunidades de usuários dos serviços oferecidos pelas bibliotecas, com o objetivo de melhor adequar os serviços aos diferentes grupos de usuários. O presente artigo trata do processo da “bibliomineração”, que consiste basicamente na organização de repositórios, juntamente com a mineração dos dados armazenados.

O termo ***bibliomining***<sup>3</sup> (“bibliomineração”) foi utilizado pela primeira vez por Nicholson e Stanton (2003), ao tratar da mineração de dados aplicada à biblioteca. Na literatura especializada, a maior parte das contribuições que contêm os termos **biblioteca** e **mineração de dados** discute o assunto **bibliotecas de**

**software**, visto que, nesta expressão de busca, o termo **mineração de dados** se refere a um determinado tipo de ferramentas disponíveis em bibliotecas especializadas em *software*. Criou-se então o termo **bibliomining** (em inglês), justamente para ir ao encontro dos profissionais interessados na mineração de dados em ambiente de biblioteca, que tentam localizar trabalhos de pesquisa e autores numa especialidade, cujos padrões de comunicação científica se encaixem na Bibliometria.

A bibliomineração, portanto, utiliza ferramentas estatísticas e do reconhecimento de padrões, aplicados a dados produzidos em sistemas de bibliotecas. O processo da bibliomineração consiste em: determinar áreas temáticas de interesse; identificar as fontes de dados internas e externas; coletar, depurar e tornar anônimos os dados destinados a um repositório; selecionar ferramentas de análise apropriadas; descobrir padrões por meio da mineração de dados e pela produção de relatórios com ferramentas tradicionais de análise; analisar e implementar os resultados dessa atividade.

Pelas suas próprias características, esse processo é cíclico. Na medida em que os padrões

<sup>3</sup> Destaque em negrito do tradutor.

forem descobertos, surgem novas questões, reinicializando o processo. Quanto mais áreas da biblioteca forem sendo analisadas, mais completo se tornará o repositório de dados, possibilitando análises com novos enfoques.

#### DETERMINANDO AS ÁREAS DE INTERESSE

O primeiro passo no processo da bibliomineração consiste em determinar a área de interesse para a análise. Este interesse pode decorrer de um problema na biblioteca ou da necessidade de informação para servir nas tomadas de decisão. A primeira determinação refere-se à escolha entre uma mineração de dados direcionada e outra não-direcionada (BERRY; LINOFF 1997). Quando direcionada, a mineração refere-se a problemas. Normalmente, trata-se de um problema específico que motivou a análise (por exemplo, “qual o tipo de procedimento a ser adotado, quando um usuário se atrasa 10 dias na devolução de um livro?”) Por sua vez, a mineração de dados não-direcionada se aplica quando, por exemplo, o diretor de uma biblioteca necessita obter uma visão mais nítida de um assunto de caráter geral (“Como é que as diferentes repartições de bibliotecas e os diversos tipos de usuários estão utilizando as revistas eletrônicas?”

É preciso tomar certos cuidados com a mineração de dados não-direcionada. A mineração como tal é uma técnica probabilística, sendo que algumas técnicas desse tipo fornecerão melhores padrões, mesmo que os dados não sejam muito consistentes. Todos os padrões encontrados devem ser analisados para assegurar sua validade; não serão triviais, mas devem fazer sentido para os profissionais envolvidos. Na modalidade da interação de dados não-direcionada, eventualmente, pode surgir um número excessivo de padrões para a análise. Considerando o tempo necessário para coletar e depurar esses dados, a mineração não-

-direcionada só deveria ser utilizada, quando houver à disposição um repositório de dados bem estruturados.

#### IDENTIFICANDO AS FONTES DOS DADOS

Após determinar as áreas de interesse, o passo seguinte é identificar as fontes de dados apropriadas. No processo da bibliomineração, trabalha-se com dados transacionais, não-agregados, de baixo nível. Isto pode tornar-se muito difícil, senão impossível, nas instituições em que os bibliotecários habitualmente apagam os registros operacionais, no afã de proteger a privacidade dos usuários. O jornal *New York Times* de 4 de abril de 2003 relatou a eliminação diária dos registros referentes à frequência de bibliotecas públicas (“Bibliotecários utilizam picotador de papel para demonstrar sua oposição aos novos poderes do FBI). Entretanto, a técnica do repositório de dados poderia ser utilizada para o próprio histórico da biblioteca, baseado em dados concretos, respeitando-se, ao mesmo tempo, a privacidade dos seus usuários. Considerando que a alocação de verbas para serviços públicos continua a sofrer cortes, é importante aproveitar a possibilidade de comprovar e defender a existência dos serviços prestados pelas bibliotecas.

Existem dois tipos de fontes de dados a serem consideradas: internas e externas. As primeiras são aquelas já integradas ao sistema adotado. Considerando que a maior parte dos sistemas de bibliotecas tem seus dados armazenados em diferentes silos eletrônicos, espalhados pela organização (base de usuários, dados transacionais, servidores de acesso à *Web*), a descoberta e extração de dados internos da instituição pode constituir-se realmente num desafio.

No caso de fontes externas, trata-se de dados não localizados no sistema da biblioteca. Num ambiente acadêmico, isto pode incluir a

informação estatística relacionada a um determinado número de identificação específica, localizado no Centro de Computação ou no sistema de gestão de pessoal. No caso de uma biblioteca pública, pode-se acrescentar as informações em código *zipado*, dos dados do último Censo<sup>4</sup>.

## CRIANDO UM REPOSITÓRIO DE DADOS

Suponhamos que os dados efetivamente já existem disponíveis na biblioteca. Neste caso, apresenta-se o desafio de combinar diferentes fontes de dados num só repositório, sem incluir informações publicamente identificáveis sobre os respectivos usuários.

No presente caso, entende-se por repositório uma base de dados separada dos respectivos sistemas operacionais, contendo uma versão depurada, sob forma anônima, dos dados operacionais, reformatados para análise. Ao criar um repositório de dados, o bibliotecário vai elaborar questões de busca destinadas para extrair dados a partir de fontes identificadas. Ele combina os dados, utilizando campos comuns; a seguir, depura os mesmos, e passa os arquivos resultantes para um arquivo bidimensional de dados ou para uma base de dados relacional, criada especificamente para análise (INMON, 2002). Após testar-se o procedimento, este pode ser automatizado para transferir regularmente os dados dos sistemas operacionais para o referido repositório.

### Protegendo a privacidade dos usuários

À primeira vista, pode parecer que um repositório de dados vai anular a privacidade do usuário, ao combinar diferentes sistemas da biblioteca em uma única fonte. Numa reportagem

do *New York Times*, de 2 de dezembro de 2002 (“Os bibliotecários são orientados sobre a legislação e a privacidade do leitor”), foi citada uma recomendação, de acordo com a qual as bibliotecas deveriam, apagar todos os dados transacionais. Entretanto, desse modo ficaria muito difícil avaliar e justificar os serviços por elas prestados. Em outros casos, não é tomada nenhuma iniciativa, ficando os dados nos sistemas operacionais ou em fitas de segurança à disposição de quem tiver acesso a elas. Ao adotar o processamento do repositório de dados, a biblioteca deverá, antes de tudo, examinar suas fontes de dados, determinar explicitamente o que guardar ou apagar, mas possibilitar a preservação da informação estatística necessária para avaliar as comunidades de usuários. Isto, sem guardar os arquivos individuais, referentes ao membros dessas comunidades.

Os processos de extração e de depuração dos dados são essenciais para proteger a privacidade dos usuários ao criar um repositório. Todos os arquivos são retirados de sistemas internos e externos, e ajustamentos são feitos para conectar os dados, sendo descartadas aquelas informações pelas quais poderia ser identificado algum dado pessoal. A informação pessoal jamais deverá ser repassada para o repositório. Assim, não se correrá perigo de que esses dados venham a ser copiados ou arquivados indevidamente, por pessoa não autorizada. Após a criação do repositório, os dados originais que a ele deram origem podem ser *deletados*. A finalidade precípua do repositório é constituir uma fonte de dados com as informações destinadas à tomada de decisão. Nesta nova fonte, não mais será possível recriar os arquivos transacionais originais.

A seguir, são apresentados dois exemplos de um processo de depuração, onde se mantêm as informações importantes para a tomada de decisão, sem incluir informações que pudessem

<sup>4</sup> O autor se refere à situação nos EUA, onde os dados do Censo são utilizados no âmbito público, com estrita observância da preservação da identidade das pessoas, consoante a legislação em vigor (Nota do tradutor).

servir para a identificação posterior de pessoas. Observa-se que, quando um item emprestado é devolvido, muitas bibliotecas apagam todas as informações referentes a essa transação. Entretanto, dessa maneira são perdidas informações preciosas para tomadas de decisão. Sendo o sistema operacional uma fonte de dados centrada no usuário, o repositório de dados constitui uma fonte sobre os itens de empréstimo. Por isso, antes de *deletar* a informação transacional, deverá ser criado um arquivo no repositório de dados, que combine a informação referente ao item com a informação estatística referente ao usuário. Com esta combinação será captada a informação importante sobre a transação, sem identificar o usuário envolvido. Permitir-se-á a descoberta de padrões acerca dos usuários, mas sem comprometer a informação referente à identificação pessoal (Quadro 1). Os campos sombreados referem-se ao processo da substituição dos dados.

O arquivo de transações do servidor contém informações valiosas sobre a utilização dos serviços da biblioteca baseados na *Web*. Esses arquivos de registro de acesso contêm os protocolos da *Internet* (PI's), que podem ser utilizados para chegar aos computadores

individuais, utilizados para acessar esses serviços. A informação importante do PI, no presente caso, não é de onde vem a pessoa, mas quais as páginas que ela visitou numa determinada sessão de consulta. Em vez de *deletar* os conteúdos desses campos, algum tipo de código deveria ser utilizado para substituir cada endereço de PI acessado. Uma maneira de também preservar a privacidade seria a substituição de todos os endereços de PI idênticos, acessados numa mesma janela, por um código gerado a partir da data e hora de entrada, associado ao PI. Alguns arquivos de transação preservam informações sobre a entrada do usuário. Assim como no caso dos arquivos de circulação, a informação sobre a entrada do usuário deverá ser substituída, no repositório de dados, pela correspondente informação estatística (Quadro 2).

Sempre que forem captadas informações estatísticas, deverá ficar assegurado que nenhuma combinação de variáveis seja capaz de levar à identificação do usuário. No desenvolvimento das ferramentas de captação, os grupos estatísticos devem ser combinados ou omitidos até que todas as combinações de dados estatísticos salvos apontem para os agrupamentos e não para indivíduos (Tabela 1).

**Quadro 1.** Depuração de arquivos transacionais.

Arquivos originais de circulação			Base de dados original dos usuários			
n. de chamada	Assunto	Usuário	Usuário	Nome	Nível de usuário	Departamento
QA 76.9	Ciência da Computação	392-33	373-34	Abby Lavender	Pós-Graduação	Psicologia
PS 159.G8	Literatura Americana	575-49	392-33	Kenneth Moore	Graduação	Matemática
HF 5415.125	Marketing	392-33	575-49	Sophie Richards	Docente	Inglês

  

Repositório de dados – Arquivos de circulação depurados e combinados			
n. de chamada do livro	Assunto	Nível de usuário	Departamento do usuário
QA 76.9	Ciência da Computação	Graduação	Matemática
PS 159.G8	Literatura Americana	Docente	Inglês
HF5415.125	Marketing	Graduação	Matemática

**Quadro 2.** Depuração de arquivos transacionais de servidor da *Web*.

Arquivos original de entradas na <i>Web</i>			
Endereço IP	Hora/data	Página acessada	Instrumento de busca
12.90.201.23	10:32/10-29-02	Index.html	Google.com
98.28.189.49	10:33/10-29-02	Resources/odc.asp	Index.html
12.90.201.23	10:35/10-29-02	Reference.html	Index.html
12.90.201.23	10:36/10-29-02	Databases.html	Reference.html
98.28.189.49	10:37/10-29-02	Resources/oclc.asp	Firstsearch.html

Repositório de dados – Arquivos de transações na <i>Web</i> depurados			
Endereço IP	Hora/data	Página acessada	Instrumento de busca
102902-1032-A	10:32/10-29-02	Index.html	Google.com
102902-1033-A	10:33/10-29-02	Resources/odc.asp	Index.html
102902-3032-A	10:35/10-29-02	Reference.html	Index.html
102902-1032-A	10:36/10-29-02	Databases.html	Reference.html
102902-1033-A	10:37/10-29-02	Resources/oclc.asp	Firstsearch.html

**Tabela 1.** Consolidação de variáveis estatísticas.

Dados Estáticos Originais			
	Graduação	Pós-Graduação	Docentes
Inglês	27	5	8
Computação	14	3	7
Matemática	33	1	6
Psicologia	24	6	7
Administração	24	14	5
Dados Estáticos Depurados			
Inglês	27	5	8
Computação/Matemática	47	4	13
Psicologia	24	6	7
Administração	24	14	5

O profissional deverá buscar orientação jurídica antes de iniciar um projeto de repositório de dados. Nos EUA, a privacidade de arquivos de usuários, atualmente, está sujeita à legislação de cada Estado da

Federação<sup>5</sup>. As bibliotecas vinculadas a instituições de pesquisa deverão buscar orientação junto à respectiva Diretoria, antes de iniciar qualquer projeto de repositório de dados.

<sup>5</sup> A correspondente legislação brasileira é baseada na Constituição de 1988 (Nota do tradutor).

### Construindo um repositório de dados

A criação de um repositório de dados pode absorver até 80% do tempo necessário ao processo da bibliomineração (BERRY; LINOFF, 1997). Recomenda-se que os profissionais envolvidos no processo encarem esta etapa, na verdade, como investimento. Uma vez concluídos os programas referentes à coleta e à depuração dos dados, o sistema vai exigir pouca manutenção, enquanto a biblioteca estiver utilizando as mesmas rotinas. Por causa do elevado investimento inicial, em termos de tempo, sugere-se começar com um tópico de bibliomineração bem delimitado, mas rodando o programa por inteiro. Sendo de natureza iterativa, o processamento oferece aos desenvolvedores do repositório de dados a possibilidade de aperfeiçoar os algoritmos subjacentes à coleta e depuração, ainda na fase inicial do projeto.

### SELEÇÃO DE FERRAMENTAS DE ANÁLISE APROPRIADAS

Uma vez desenvolvido o repositório de dados, pode-se passar para a análise dos mesmos. Conforme mencionado anteriormente, agregações de relações podem ser facilmente calculadas para criar relatórios em formato tradicional. Entretanto, padrões interessantes e úteis podem estar encobertos nesses resultados agregados. E são esses padrões que podem oferecer aos responsáveis pela biblioteca um conhecimento melhor dos seus grupos de usuários.

### Sistemas de gestão da informação (SGIs)

Os SGIs permitem responder às questões básicas referentes aos dados armazenados. Muitos pacotes de sistemas de informação para bibliotecas incluem algum tipo de SGI. Todavia,

os dados que entram nesses sistemas provêm de arquivos operacionais, que não foram depurados nem relacionados a quaisquer dados externos. Por outro lado, a exemplo da maioria das bases de dados de bibliotecas, essas informações são tratadas como silos de dados, onde as bases não são interligadas entre si. Por isso, um SGI direcionado para um repositório de dados específicos e criado para uma determinada biblioteca será muito mais poderoso, fornecendo aquelas informações de que essa biblioteca realmente necessita.

Outra vantagem do SGI consiste no fato de poder vir a ser um sistema de alerta para fatores críticos. Como os gestores e administradores trabalham com relatórios baseados em dados, poderão identificar as variáveis-chave responsáveis pelo ritmo de funcionamento da biblioteca. Poderão ser instalados programas de notificação automática, em relação a essas variáveis, de modo que, se elas extrapolarem um campo previamente especificado, os responsáveis serão automaticamente notificados de uma solução potencialmente problemática. Por exemplo, se a circulação de usuários por hora ficar abaixo ou acima de um determinado nível, os responsáveis pela administração da biblioteca, imediatamente, poderão ser notificados para tomar as providências adequadas.

### Processamento analítico online (OLAP)

Este tipo de processamento, *OLAP*<sup>6</sup>, permite uma visão interativa dos dados ao nível dos responsáveis pela tomada de decisão. Com essa ferramenta, já foram processadas milhares de questões de busca dirigidas à base de dados, combinando todas as variáveis com as medidas selecionadas. Os respectivos relatórios são apresentados em um sistema de fácil acesso por intermédio de um *menu* amigável.

<sup>6</sup> OLAP é a sigla inglesa para *On-Line analytical processing system*, que acabou por ser adotada também entre profissionais brasileiros (Nota do tradutor).

O usuário escolherá, por exemplo, uma das muitas variáveis constantes de uma lista, digamos acerca do uso de periódicos eletrônicos. Podem-se, então, selecionar dimensões básicas, tais como data/hora e assunto. O sistema OLAP vai permitir uma visão desses dados, em alto nível, por exemplo, como relatório sob forma tabulada, por ano e classificação. O usuário poderá então clicar uma das dimensões para expandir o relatório. No presente caso, se o usuário clicar o campo “ano”, a ferramenta poderá expandir esse campo em trimestres, sob o mesmo título, mas recalculando os dados. Em seguida, o usuário poderá clicar um outro campo, para analisar sob outro aspecto os dados disponíveis. As ferramentas de OLAP trabalham com dados organizados em repositórios, sendo que todos os campos devem estar definidos, antes que o sistema seja liberado para uso. As respostas oferecidas pela ferramenta ao administrador são instantâneas, favorecendo a análise e durante o processo da análise; administrador poderá captar diferentes aspectos dos dados, incorporando-os num relatório regular.

### Mineração de dados

Outra maneira de analisar os dados, e que deu origem ao termo *bibliomineração*, é por meio da mineração de dados propriamente dita. Trata-se da descoberta de padrões de dados em nível inferior, utilizando ferramentas estatísticas e de inteligência artificial. A mineração de dados surgiu na área empresarial americana, no início da década de 1990, depois da popularização dos repositórios de dados. Estes haviam possibilitado a armazenagem de grandes quantidades de dados, o que motivou as corporações a se interessarem por instrumentos capazes de tirar vantagens desses repositórios.

Existem dois tipos principais de tarefas a serem desempenhadas pela mineração de dados: *descrição* e *prospecção*. Na descrição, trata-se de compreender dados passados e

presentes. Os padrões descobertos são utilizados para formar “agrupamentos de afinidade” e variáveis comuns a diferentes usuários ou agrupamentos estatísticos com determinadas características. A prospecção, por outro lado, pode ser destinada a fundamentar afirmações sobre o desconhecido, baseadas sobre o que é conhecido. Pode ser utilizada para prever o futuro ou para fazer afirmações sobre o presente. A prospecção oferece dois tipos de recursos, a saber: a *classificação*, cuja finalidade é colocar um item numa categoria, e a *estimativa*, cujo objetivo é produzir um valor numérico para uma variável desconhecida. A tarefa inicialmente escolhida determinará a escolha da modalidade do respectivo recurso prospectivo.

### Pacotes de *software* de mineração de dados

Atualmente, existem numerosos pacotes de *software*, possibilitando o uso de diferentes ferramentas de mineração de dados. Na maioria desses pacotes, os dados são importados para um arquivo bidimensional, ou para um arquivo comum de dados de outro formato. Após importar os dados, o analista poderá selecionar as ferramentas e visualizar os resultados. Os pacotes estatísticos SAS e SPSS dispõem ambos de ferramentas opcionais de mineração, mas são bastante caras. A ferramenta de mineração de dados WEKA (WITTEN; FRANK 1999) é um *software* livre, que contém muitas opções idênticas às daqueles pacotes, mas a custo menor; porém, o uso desta ferramenta não é tão amigável como o daquelas.

Apesar da simplicidade de aplicação desses modelos, eles apresentam um inconveniente: muitas das ferramentas disponíveis são baseadas em métodos estatísticos ou probabilísticos. Esses métodos requerem um certo tipo de dados, por exemplo, provenientes de uma distribuição normal, a fim de produzir, por sua vez, dados confiáveis. A inconveniência decorre



do fato de que tais ferramentas não dispõem de nenhum mecanismo para checar se essas condições dos dados a serem processados foram satisfeitas ou não. Assim, um usuário desprevenido poderá aplicar essas ferramentas e, no caso de os dados não terem sido enquadrados em determinadas condições prévias, vir a produzir respostas que não fazem sentido. Para saber mais sobre algoritmos específicos em pacotes de mineração de dados, recomendam-se os trabalhos de Berry e Linoff (1997; 2000) ou o texto de Witten e Frank (1999), este último referente ao já mencionado *software* livre de mineração de dados WEKA.

#### ANÁLISE E IMPLEMENTAÇÃO

Uma vez concluído o modelo, ele deverá ser validado. O primeiro passo é testar os dados numa amostragem (que não tenha sido utilizada na própria construção do modelo), a fim de averiguar a robustez desse modelo em diferentes ambientes. A validação mais importante, contudo, ficaria a cargo de um profissional experiente, familiarizado com o contexto da biblioteca em questão, o qual poderá concordar ou não com os resultados obtidos. Desta maneira, os padrões baseados nos dados seriam validados a partir do conhecimento tácito que este profissional adquiriu no próprio ambiente de trabalho da biblioteca. No caso de ele não

concordar com o padrão encontrado, será preciso re-analisar cuidadosamente os fluxos de dados. Esse tipo de discordância, em geral, se origina de uma distorção dos dados ou da aplicação errônea de uma ferramenta.

O último passo consiste em implementar o modelo ou tipo de relatório, devendo ser monitoradas as variáveis que determinam os modelos utilizados ao longo do tempo. Quando a média de uma variável se afasta demasiadamente por causa de mudanças na biblioteca, eventualmente, o próprio modelo terá de ser reavaliado.

#### CONCLUSÃO

O objetivo principal deste artigo foi explicar um processo de análise que se sugeriu denominar *bibliomineração*. Enfatizou-se a importância da construção de um repositório de dados, sem comprometer a privacidade dos usuários, problema a ser resolvido antes de iniciar qualquer atividade na área. Deve-se, pois, captar os dados dos arquivos da instituição, sem ferir os direitos do usuário. Por meio de um repositório de dados, ambos os objetivos poderão ser alcançados. Uma vez instalado o repositório, a biblioteca terá disponível uma grande quantidade de relatórios e ferramentas analíticas, valiosos subsídios para poder conhecer mais profundamente suas comunidades de usuários e poder melhor adequar a utilização de seus recursos.

#### REFERÊNCIAS

BUCKLAND, M. Five grand challenges for library research. *Library Trends*, v. 51, n.4. Available from: <<http://www.sims.berkeley.edu/~buckland/trends03.pdf>>. Acess: 27 June 2003.

BERRY, M.; LINOFF, G. *Data mining techniques for marketing, sales, and customer support*. New York: John Wiley, 1997.

BERRY, M.; LINOFF, G. *Mastering data mining*. New York: John Wiley, 2000.

INMON, W. *Building the data warehouse*. 3.ed. New York: John Wiley, 2002.

NICHOLSON, S.; STANTON, J. Gaining strategic advantage through bibliominer: data mining for management decisions in corporate, special, digital, and traditional libraries. In: NEMATI, H.; BARKO, C. (Ed.). *Organizational data mining: leveraging enterprise data resources for optimal performance*. Hershey, PA: Idea Group, 2003.

WITTEN, I.; FRANK, E. *Practical machine learning tools and techniques with Java implementations*. San Francisco, CA: Morgan Kaufmann, 1999.

