Aplicativo de Mineração de Dados Aplicado em Bases de Dados Acadêmicas

Elisa Maria Vissotto[1], Adriane Barbosa Camargo[2]

[1]Universidade Regional Integrada do Alto Uruguai e das Missões (URI)
CEP– 98.400-000 – Frederico Westphalen – RS – Brasil

[2]Universidade Regional Integrada do Alto Uruguai e das Missões (URI)
CEP– 98.400-000 – Frederico Westphalen – RS – Brasil

elisavissotto@hotmail.com, adrianec@uri.edu.br

Abstract. *With technological advances, data storage has taken large proportions, making the discovery of useful knowledge. Thus, it became necessary to find ways to extract knowledge in large databases. For this it is necessary to apply data mining techniques allowing to know new information, whether academic, commercial or business. To represent the techniques of Data Mining an application was developed to extract relevant data from the access of students to the Central Library of the URI and the access to the System UriNet. The objective of this work is the application of data mining techniques to extract relevant information in large databases.*

Resumo. *Com os avanços tecnológicos, o armazenamento de dados tomou grandes proporções, dificultando a descoberta de conhecimento útil. Assim, tornou-se necessário encontrar formas de extrair conhecimento nos grandes bancos de dados. Para isso é necessário aplicar técnicas de Mineração de Dados possibilitando conhecer novas informações, sejam elas acadêmicas, comerciais ou empresariais. Para representar as técnicas de Mineração de Dados foi desenvolvido um aplicativo para extrair dados relevantes dos acessos dos alunos à Biblioteca Central da URI e dos acessos ao Sistema UriNet. O objetivo deste trabalho é a aplicação de técnicas de Mineração de Dados para extrair informações relevantes em grandes bases de dados.*

1. Introdução

Com o avanço da tecnologia, a informação tornou-se importante para a realização de negócios entre organizações e instituições de ensino. Com o objetivo de extrair conhecimentos novos e relevantes das grandes empresas, surgiu a Mineração de Dados, que tem por objetivo buscar padrões e relacionamentos nas informações existentes nas bases de dados, sendo considerada uma das principais tecnologias de descoberta de conhecimento.

Neste contexto, este trabalho abordará o uso da Mineração de Dados, em conjunto com suas técnicas, visando proporcionar aos usuários, sejam de sistemas que gerenciam bibliotecas ou qualquer outro sistema interno em instituições públicas e/ou privadas, um aplicativo *desktop* capaz de auxiliar no processo de agrupamento e contagem de alunos que utilizam a biblioteca ou outros sistemas e gerar relatórios com os principais dados de acessos armazenados no banco de dados.

Sendo assim, foi desenvolvido um aplicativo, que realizará a mineração dos dados do Sistema da Biblioteca Central da URI – Universidade Regional Integrada do Alto Uruguai e das Missões, em conjunto com os dados da base do Sistema UriNet. Este

aplicativo auxilia na visualização e geração de relatórios com a quantidade de acessos aos referidos sistemas, proporcionando ao usuário obter rapidamente os dados. Isso contribui positivamente para que o agrupamento dos cursos com nomes semelhantes seja realizado, mantendo uma melhor organização dos dados e assim, gerando informações mais claras e precisas.

2. Mineração de dados

Com a rápida evolução dos recursos computacionais nos últimos anos houve um crescimento elevado no volume de dados disponível, desafiando a capacidade de armazenamento, seleção e uso. Este crescimento tem gerado uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, *terabytes* de dados em informações significativas e em conhecimento. Através do uso da Mineração de Dados e de suas ferramentas, é possível realizar a extração destes dados transformando-os em informação e conhecimento. (CORRÊA; SFERRA 2003).

Carvalho (2001) define a Mineração de Dados como "[...] o uso de técnicas automáticas de exploração de grandes quantidades de dados, de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam descobertos a olho nu [...]".

As ferramentas da Mineração de Dados são um conjunto de técnicas de Estatística e Inteligência Artificial, que tem como objetivo específico descobrir conhecimento novo, escondido em grandes volumes, armazenado em bancos de dados. (CARVALHO, 2001).

Conforme Amaral (2001), Mineração de Dados é "O processo de busca de relacionamentos e padrões existentes nas bases de dados. Devido à grande quantidade de informações nos sistemas de bancos de dados atuais, esses relacionamentos estão escondidos [...]."

Neste contexto, Amaral (2001) destaca que a Mineração de Dados é um conceito revolucionário, sendo seu principal objetivo encontrar padrões ainda não descobertos em grandes quantidades de dados, dos quais se espera que sejam geradas respostas corretas.

As principais técnicas de Mineração de Dados estudadas durante o desenvolvimento deste trabalho foram a *Classificação, Associação*, *Estimativa, Previsão e Análise de Agrupamentos.*

2.2 As Fases e os Passos do Processo KDD

Há vários anos, a noção de encontrar por padrões úteis, em dados que inicialmente não eram tratados, tem sido denominada descoberta de conhecimento KDD (*Knowledge Discovery in Database)*, Mineração de Dados, extração de conhecimento, descoberta de informação, coleta de informação, dentre outros. (AMARAL, 2001).

Com os avanços tecnológicos, a obtenção de informações relevantes dentro de uma organização tornou-se mais eficiente através da Descoberta de Conhecimento em Banco de Dados (KDD), criada em 1989. Esse processo surgiu como uma alternativa viável de atender parte da demanda de Mineração de Dados úteis e de grande importância para as organizações, e procura, em um nível abstrato, desenvolver métodos e técnicas que ofereçam significado a dados armazenados digitalmente.

Visando uma exemplificação da aplicabilidade do processo de Mineração de Dados no mercado financeiro, Silva (2005) apresenta um exemplo onde a descoberta de conhecimento em bancos de dados pode desempenhar tarefas relevantes, levando resultados positivos ao analista.

O exemplo a seguir, retrata a usabilidade da mineração de dados, exemplificando na prática como pode ser utilizado com o método de KDD. Ex: Realizar empréstimos exige do cliente o fornecimento de dados pessoais e financeiros relevantes. Todas essas informações são utilizadas pelas instituições financeiras como base para decidir se efetuam, ou não, um empréstimo. Inicialmente, métodos estatísticos são utilizados para determinar a aceitação ou rejeição do pedido. São estes os casos que estão no limite e necessitam de análise humana. O processo de KDD pode ser aplicado neste problema da seguinte forma: suponha-se a disponibilidade de um banco de dados histórico sobre clientes da instituição, com aproximadamente 5.000 cadastros, contendo 20 diferentes atributos, tais como: idade, tempo de serviço, vencimentos, bens, status atual de crédito, dentre outros. O tratamento dessas informações, utilizando-se métodos de KDD, geraria automaticamente regras objetivas e claras sobre as principais características dos bons e maus clientes, podendo estas regras também serem aplicadas para aumentar a taxa de sucesso das operações de empréstimo. (SILVA, 2005).

Tomando como base esse exemplo, pode-se verificar que a técnica de agrupamentos foi escolhida para agrupar os nomes dos cursos dispersos no banco de dados da Biblioteca Central e do Sistema UriNet, pois está direcionada aos objetivos de identificação e classificação dos dados.

Este processo envolve três grandes fases, o Pré-Processamento, a Mineração de Dados e o Pós-Processamento conforme mostra a Figura 1. (AMARAL, 2001).

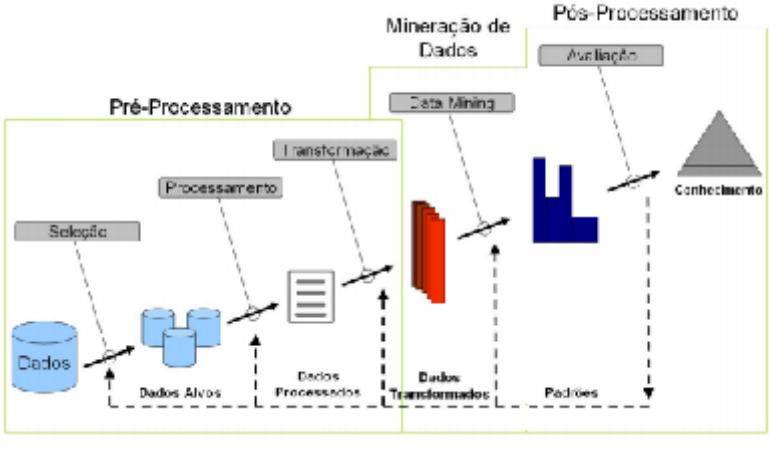

Figura 1. As Fases do Processo KDD

A partir da Figura 1 será feita uma abordagem sobre cada uma das fases do KDD, apresentando seus significados no processo de extração de conhecimento.

Seleção: Seleção dos dados relevantes, escolha da base de dados a ser trabalhada.

Processamento: Realiza-se a redução, limpeza e padronização da base de dados.

Transformação: Possibilita a transformação ou consolidação dos dados no formato apropriado para iniciar o processo de mineração.

Data Mining: Esta é a fase essencial de todo o processo, onde as técnicas escolhidas são aplicadas para análise e extração de padrões de dados.

Avaliação: Nesta última fase é realizada a interpretação dos dados e identificados os padrões que realmente interessam dentre todos os dados analisados. (AMARAL, 2001).

Diante do desenvolvimento do aplicativo SoftMiner, a seleção dos dados ocorreu no momento da escolha da base de dados da Biblioteca Central e do Sistema Urinet. No processamento das informações foram excluídos alguns cursos sem relevância ou com erros de gramática, padronizando os dados somente com informações sobre cursos de graduação. Na transformação os dados foram consolidados, permanecendo somente os cursos de graduação da universidade, excluindo-se os cursos de pós-graduação, extensão, dentre outros. Na fase de *Data Mining* a técnica de agrupamentos foi escolhida para representar os resultados obtidos, sendo que os mesmos foram avaliados com resultado positivo, sendo que, qualquer usuário do sistema poderá visualizá-los rapidamente e gerar relatórios impressos.

O método KDD foi escolhido por ser mais adequado ao objetivo principal deste trabalho, sendo um dos processos mais citados em livros e artigos.

3 Desenvolvimento do Aplicativo SoftMiner para Mineração de Dados

A seguir serão detalhados os Sistemas UriNet e Biblioteca Central, sendo estes sistemas utilizados para a coleta de dados referente aos acessos dos alunos. A partir disso, foi desenvolvido um aplicativo, denominado pelo nome de *SoftMiner* (Aplicativo de Mineração de Dados) que tem o objetivo de extrair conhecimentos relevantes dos *logs* de acesso dos alunos de graduação da URI que acessaram os Sistemas UriNet e Biblioteca Central.

3.1 Sistema da Biblioteca Central da URI

A Biblioteca Central da URI dispõe de acervo atualizado, composto por livros, periódicos, obras de Referência (dicionários, enciclopédias, bibliografias, atlas) monografias, teses e dissertações, abrangendo várias áreas do conhecimento.

Para melhorar as informações contidas no banco de dados do Sistema da Biblioteca também foram aplicadas as três primeiras etapas do processo KDD: Seleção, Processamento e Transformação. Na etapa de Seleção foram escolhidas as bases de dados a serem mineradas e realizadas modificações nos dados da tabela Hist_Biblioteca, que refere-se aos cursos que acessaram o Sistema da Biblioteca e que contém todos os *logs* de acesso dos alunos. Através do Processamento foram realizadas atualizações nos nomes dos cursos, padronizando a base de dados. Na etapa de Transformação, os dados foram transformados e consolidados.

A Figura 2 mostra como estão armazenados os *logs* de acesso ao Sistema da Biblioteca Central, ressaltando-se que os *logs* referentes ao Sistema UriNet também encontram-se da mesma forma.

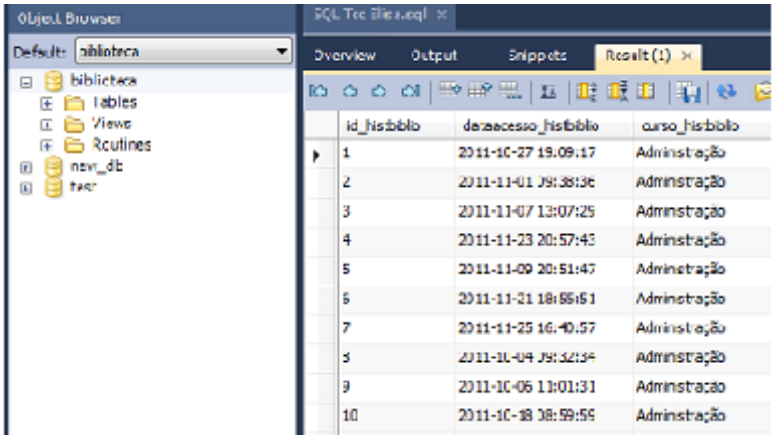

Figura 2. *Logs* de acesso ao Sistema da Biblioteca Central

Observando-se a Figura 2 é possível verificar que o histórico de *logs* apresenta-se na forma de Código, Data (contendo ano, mês, dia, horas, minutos e segundos), além do nome dos cursos oferecidos pela URI. Através desta representação, foi possível agregar a Técnica de Agrupamentos e posteriormente as análises da Mineração de Dados e realizar a descoberta de novas informações. A partir do agrupamento desses atributos, ou seja, dos nomes dos cursos, a quantidade de acessos aos Sistemas UriNet e Biblioteca Central foi encontrada e os nomes dos cursos foram padronizados.

Na Figura 3 estão numeradas as principais funcionalidades da tela de Mineração de Dados do Sistema da Biblioteca Central. **(1) Código:** Mostra o código de cada curso minerado, permitindo visualizar a quantidade de cursos. **(2) Curso:** Lista no *DBGrid* o nome de todos os cursos que restaram após a mineração. **(3) Qtda**: Pode-se visualizar a quantidade de acessos dos alunos de cada curso listado. **(4) Porc_Acesso**: Faz referência à porcentagem de acessos de cada curso, em relação ao total de 56.743 acessos minerados. **(5) Minerar**: Pode-se visualizar o Código, Nome do Curso, Quantidade e Porcentagem de Acessos. **(6) Imprimir Relatório:** Permite gerar um relatório de impressão dos dados minerados. **(7)** refere-se ao número total de acessos ao Sistema da Biblioteca no período de 01/06/11 a 01/07/12.

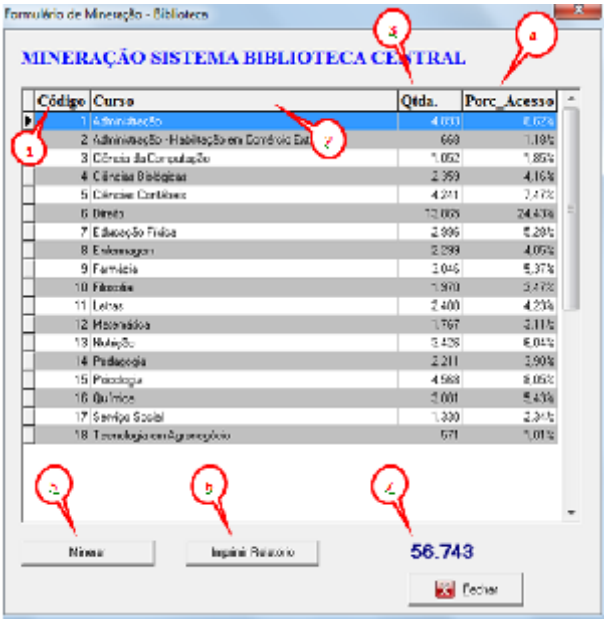

Figura 3. Tela de Mineração da Biblioteca Central

3.2 Sistema UriNet

O UriNet é um sistema desenvolvido para a consulta de informações acadêmicas através da *Internet,* possibilitando que professores e acadêmicos interajam utilizando o espaço para envio de trabalhos, consulta de notas, Fórum de Discussão e Chat.

Os *logs* de acesso ao Sistema UriNet foram disponibilizados pelo setor de Desenvolvimento de Sistemas, contendo 56 nomes de cursos diferentes, deixando o banco de dados com informações irrelevantes e excessivas.

Sendo assim, para padronizar o banco de dados foram renomeados alguns cursos, como: "Química Licenciatura" e "Química Industrial" foram renomeados para "Química". "Direito Diurno" e "Direito Noturno" para "Direito", entre outras modificações.

Nesta etapa inicial onde foram aplicadas as três primeiras fases do processo KDD o banco de dados continha 37 cursos que não possuíam relevância para a realização da Mineração de Dados e através da limpeza e redução desses dados restaram ao final da mineração 22 cursos.

Para padronizar os nomes dos cursos de graduação utilizou-se a ferramenta gráfica *MySQL Workbench* que fornece modelagem de dados e melhor desempenho, permitindo a criação e execução de instruções para atualizar e excluir informações irrelevantes.

Na Figura 4 são destacadas as principais funcionalidades da tela de Mineração de Dados do Sistema UriNet. O item **(1) Código**, mostra o código de cada curso minerado, permitindo visualizar a quantidade de cursos. **(2) Curso**, lista o nome de todos os cursos que restaram ao final da mineração. **(3) Qtda**, é possível visualizar as quantidades de acessos dos alunos de graduação de cada curso listado. **(4) Porc_Acesso**, mostra a porcentagem de acessos de cada curso, em relação ao total de 729.965 acessos minerados. **(5) Minerar**, o analista obterá no *DBGrid* o Código, Nome do Curso, Quantidade e Porcentagem de Acessos. **(6) Imprimir Relatório** permite que seja gerado um relatório de impressão dos dados minerados. O item **(7)** mostra o número total de acessos ao Sistema UriNet no período de 01/06/11 a 01/07/12.

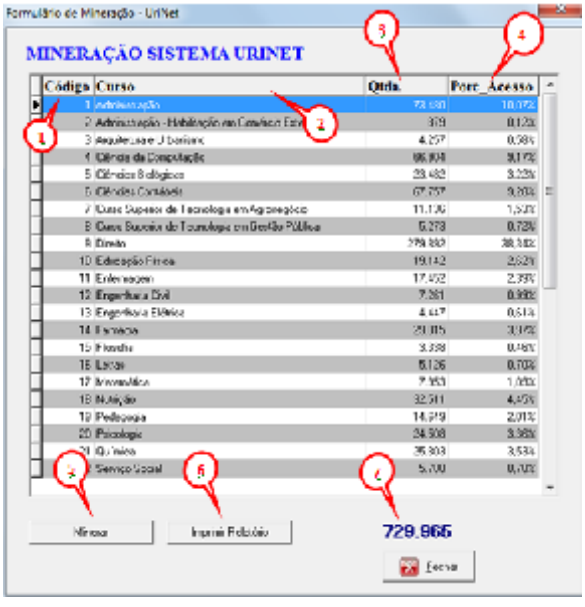

Figura 4. Tela de Mineração do Sistema UriNet

Com a aplicação da Técnica de Análise de Agrupamentos e a lógica de programação, foi possível mostrar de forma precisa o resultado da Mineração de Dados no Sistema UriNet, dando uma visão ampla da quantidade de cursos restantes, bem como seus nomes padronizados e organizados.

Para fazer um comparativo entre os acessos aos dois sistemas, foi realizada a Mineração de Dados sobre a quantidade geral de acessos no período de 01/06/2011 a 01/07/2012, visando analisar as datas com o maior e menor índice de acessos aos Sistemas da Biblioteca e UriNet.

No Sistema da Biblioteca Central também foram aplicados os mesmos métodos para se descobrir os picos de acesso ao Sistema UriNet.

4 Análise dos Resultados e Conclusão

Através do trabalho, foi possível analisar duas bases de dados e fazer a Mineração utilizando-se a técnica mais adequada, permitindo a descoberta de conhecimento novo de forma simples e segura. Assim, instituições de ensino e organizações empresariais podem usufruir desta tecnologia para buscar dados relevantes ainda não descobertos.

Este trabalho buscou identificar através da Mineração de Dados os cursos de graduação da URI que possuem os maiores e menores índices de acesso aos Sistemas UriNet e Biblioteca Central, e constatou-se que o curso de Direito obteve destaque na quantidade de acessos tanto no Urinet, quanto na Biblioteca Central.

Por outro lado, a Mineração de Dados revelou que, os cursos de Tecnologia em Agronegócios e Administração – Habilitação Comércio Exterior obtiveram os menores índices de acesso no Sistema da Biblioteca Central e UriNet, respectivamente.

Com relação aos picos de acesso aos Sistemas UriNet e Biblioteca Central, os resultados revelaram que, no período de fechamento dos semestres os acessos foram mais intensos, possivelmente, devido à consulta de notas e envio de trabalhos.

Os resultados desta análise revelaram que, em relação ao Sistema UriNet, os meses de Abril, Agosto e Setembro mostraram vários dias com poucos registros de

acesso. Este resultado demonstra que, nesse período devem ser intensificados os acessos às atividades de interação no Sistema UriNet.

Considerando-se o período letivo, os resultados da análise de acessos revelaram que, em relação ao Sistema UriNet, os meses de Abril, Agosto e Setembro mostraram vários dias com poucos registros de acesso. Em relação ao Sistema da Biblioteca Central, os meses de Março, Abril, Outubro e Novembro revelaram vários dias com poucos registros de empréstimos de obras.

Portanto, os resultados demonstraram que o curso de Direito obteve maior quantidade de acessos nos dois sistemas analisados. Este resultado é de grande importância para a universidade, visto que, até o momento a mesma não tinha comprovação de tais índices de acesso. A universidade também não possuía conhecimento sobre os cursos que menos acessavam os Sistemas da Biblioteca e Urinet.

Dessa maneira, este trabalho veio agregar conhecimento e disponibilizar informações que ainda eram desconhecidas pela Universidade Regional Integrada, contribuindo positivamente para que os sistemas Urinet e Biblioteca Central pudessem ser aprimorados e adequados para melhor atender às necessidades dos estudantes.

Referências

AMARAL, Fernanda Cristina Naliato do. *Data Mining* - Técnicas e aplicações para o *Marketing* direto. 1ª Ed. São Paulo. Editora Berkeley, 2001, 110 páginas.

CARVALHO, Luís Alfredo Vidal de. *Data Mining*: A Mineração de dados no *Marketing*, Medicina, Economia, Engenharia e Administração. 2ª Ed. São Paulo: Érica, 2001, 234 páginas.

CORRÊA, Ângela M. C. Jorge; SFERRA, Heloisa Helena. Conceitos e Aplicações de *Data Mining*. Revista de Ciência e Tecnologia. Piracicaba, 18 dez. 2003. Disponível em: < http://www.unimep.br/phpg/editora/revistaspdf/rct22art02.pdf>. Acesso em: 12 Abr. 2012.

FAYYAD M. Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY Ramasamy. Advances in Knowledge Discovery and Data Mining. MIT Press, 1st Edition, 1996.

SILVA, Marcelino Pereira dos Santos. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. Mossoró, 2005. Disponível em: <portalsbc.sbc.org.br/download.php?paper=35>. Acesso em: 26 Mar. 2012.