Identificação do perfil dos usuários da Biblioteca Central da FURB através de *data mining* para a personalização da recuperação e disseminação de informações

Alberto Pereira de Jesus[1], Evanilde Maria Moser[1], Paulo José Ogliari[2]

[1]Biblioteca Central – Universidade Regional de Blumenau (FURB)
Caixa Postal 15.05 – 89.012-900 – Blumenau – SC – Brasil

[2]Departamento Informática e Estatística – Universidade Federal de Santa Catarina
{albertop,emmoser}@furb.br, ogliari@inf.ufsc.br

Abstract. *This paper describes all data mining deployment stages applied to library user profile identification for information recovery and dissemination WEB systems.*

Resumo. Este artigo descreve todas as etapas de implantação de *data mining* aplicado na identificação do perfil dos usuários de uma biblioteca para a personalização de sistemas WEB de recuperação e disseminação de informações.

1. Introdução

Com o crescimento do volume de publicações e também das necessidades de informações dos clientes, sejam elas em papel ou em formato eletrônico, é importante, que as bibliotecas possuam sistemas de informações capazes de armazenar e indexar informações bibliográficas de forma a facilitar a recuperação e disseminação aos usuários (CARDOSO, 2000).

Neste sentido, dois sistemas têm sido desenvolvidos, sendo eles o sistema de recuperação de informações (SRI) e o sistema de disseminação seletiva de informações (DSI). Enquanto o SRI trata de localizar as informações solicitadas pelo usuário, o DSI tenta prever as necessidades desses usuários, fazendo recomendações e sugestões conforme seu interesse.

Assim, conhecer os usuários é importante e já era uma necessidade no passado, onde o bibliotecário sabia e conseguia lembrar as preferências de cada um de seus usuários para fazer recomendações e ajudá-los na localização de obras. Hoje, devido à grande quantidade de usuários e publicações, precisa-se de ferramentas automatizadas que auxiliem nesse processo.

Sabendo-se que a missão das Bibliotecas, segundo Funaro et al. (2001, p. 1) "é oferecer a seus usuários informações relevantes para a realização de suas pesquisas, facilitando o acesso e localização do material necessário", os sistemas tradicionais de SRI e DSI das Bibliotecas necessitam evoluir e ser inteligentes, a fim de agregar valor ao serviço de referência. Dessa forma é necessário que se conheça o perfil do usuário, delineando suas preferências e seus interesses.

As técnicas de *data mining* permitem a identificação desse perfil, possibilitando a personalização dos processos do SRI e DSI, tornando-os objetivos e seletivos. Esta confluência de acertos caracteriza a relevância da informação. Não adianta o usuário

receber uma comunicação personalizada se ela não for relevante para seus interesses e necessidades.

"O objetivo da personalização de conteúdo é garantir que a pessoa certa receba a informação certa no momento certo" (ARANHA, 2000, p. 10).

Estes sistemas, principalmente o DSI, apesar das facilidades que oferece, apresenta alguns problemas como: o não preenchimento por alguns usuários e as rápidas mudanças que ocorrem em seus interesses. Toma-se como exemplo um professor que lecionava uma disciplina de *data warehouse* e atualmente leciona a disciplina de *data mining*. Como ele preencheu seus dados com antigo perfil, continuará recebendo informações sobre seus interesses preenchidos anteriormente.

Seria prudente que o sistema reconhecesse essas alterações no ambiente e fosse capaz de se adequar às novas características. Isso é possível por meio da aplicação de técnicas de *data mining* sobre os dados contidos nos registros de transações como: empréstimos, reservas e consultas que são armazenados no banco de dados da Biblioteca e servirão para fazer um estudo do perfil do usuário. Estes registros são armazenados diariamente pelas transações de empréstimos, no entanto não são utilizados para tomada de decisão.

Mais especificamente, a aplicação de *data mining* nestes registros permitirá:

a) melhorar o SRI através da personalização das consultas, ao fazer uma busca o retorno da consulta é filtrado segundo o perfil do usuário;
b) facilitar o DSI, recomendando obras de interesse ao usuário.

2. Objetivos

O objetivo geral deste trabalho é desenvolver um sistema de recuperação e disseminação de informações, personalizado segundo o perfil de cada usuário da Biblioteca Central (BC) da Universidade Regional de Blumenau (FURB), por meio da aplicação de técnicas de *data mining*. Como objetivos específicos têm-se:

a) desenvolver um *data warehouse* para dar suporte a aplicação das técnicas de *data mining*, possibilitando também obter informações para tomada de decisões;
b) aplicar técnicas de *data mining* sobre o histórico de empréstimos e reservas dos usuários para identificar o perfil dos mesmos na BC da FURB;
c) desenvolver um sistema WEB de SRI e DSI personalizado dinamicamente para a BC da FURB.

3. Justificativa

O trabalho se justificativa, pois conhecendo as características e preferências do usuário, pode-se assim definir seu perfil que é de elevada importância para o SRI, DSI e para tomada de decisões gerenciais. Possibilitando uma maior satisfação dos usuários, uma melhor utilização e organização da biblioteca, redução de custos com a aquisição de materiais, bem como, facilidade no atendimento dos usuários.

4. Fundamentação Teórica

A quantidade de informações produzidas versus a capacidade de armazenamento dos recursos computacionais a um baixo custo, tem impulsionado o desenvolvimento de novas tecnologias capazes de tratar estes dados, transformá-los em informações úteis e extrair conhecimentos.

Entretanto, o principal objetivo da utilização do computador ainda tem sido o de resolver problemas operacionais das organizações, que coletam e geram grandes volumes de dados que são usados ou obtidos em suas operações diárias e armazenados nos bancos de dados. Porém, os mesmos não são utilizados para tomadas de decisões, ficando retidos em seus bancos de dados, sendo utilizados somente como fonte histórica. Estas organizações têm dificuldades na identificação de formas de exploração desses dados, e principalmente na transformação desses repositórios em conhecimento (BARTOLOMEU, 2002).

Pesquisadores de diferentes áreas estudam e desenvolvem trabalhos para obter informações e extrair conhecimentos a partir de grandes bases de dados, como tópico de pesquisa, com ênfase na técnica conhecida como *data mining*.

Data mining é parte do processo de *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimentos em bancos de dados (DCBD), o qual é responsável pela extração de informações sem conhecimento prévio, de um grande banco de dados, e seu uso para a tomada de decisões (DINIZ; LOUZADA NETO, 2000).

KDD é um processo contínuo e cíclico que permite que os resultados sejam alcançados e melhorados ao longo do tempo. Na figura 1 são apresentados os passos que devem ser executados no processo de KDD. Segundo Diniz; Louzada Neto (2000) embora os passos devam ser seguidos na mesma ordem em que são apresentados, o processo é extremamente interativo e iterativo (com várias decisões sendo feitas pelo próprio usuário e loops podendo ocorrer entre quaisquer dois ou mais passos).

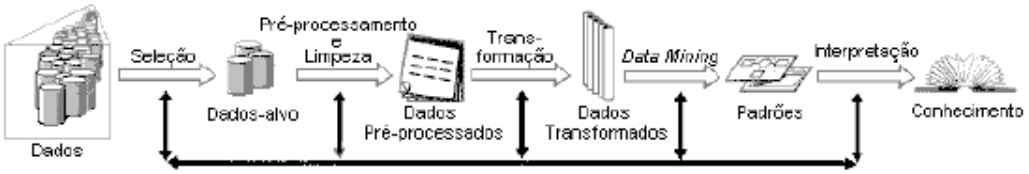

Figura 1 – Passos do processo de KDD (Fonte: FIGUEIRA, 1998, p. 8.)

Para a implantação desta tecnologia é necessário que se conheça a fundo o processo, para que a mesma venha atender às expectativas do usuário. O processo de KDD começa obviamente com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos.

Segundo Harrison (1998, p. 155) "*data mining* é a exploração e análise, por meios automáticos ou semi-automáticos, das grandes quantidades de dados para descobrir modelos e regras significativas".

Deve-se destacar que cada técnica de *data mining* ou cada implementação específica de algoritmos que são utilizados para conduzir as operações *data mining* adapta-se melhor a alguns problemas que a outros, o que impossibilita a existência de um método de *data mining* universalmente melhor. Para cada particular problema tem-se um particular algoritmo. Portanto, o sucesso de uma tarefa de *data mining* está diretamente ligado à experiência e intuição do analista (Diniz; Louzada Neto 2000).

Assim, é importante que se conheçam, as tarefas desempenhadas (Classificação, Estimatição, Previsão, Agrupamento por afinidade, Segmentação e Descrição) e suas técnicas (Análise de seleção estatística, Raciocínio baseado em casos, Algoritmos genéticos, Detecção de agrupamentos, Análise de vínculos, Árvores de decisão e indução de regras, Redes neurais) a fim de dar suporte a sua escolha.

5. Metodologia

Para o processo de extração de conhecimento nos dados da biblioteca sobre o perfil dos usuários, será utilizada a metodologia referenciada por Berry; Linoff (1997). A Figura 2 apresenta o modelo proposto. A mesma será aplicada na BC da FURB.

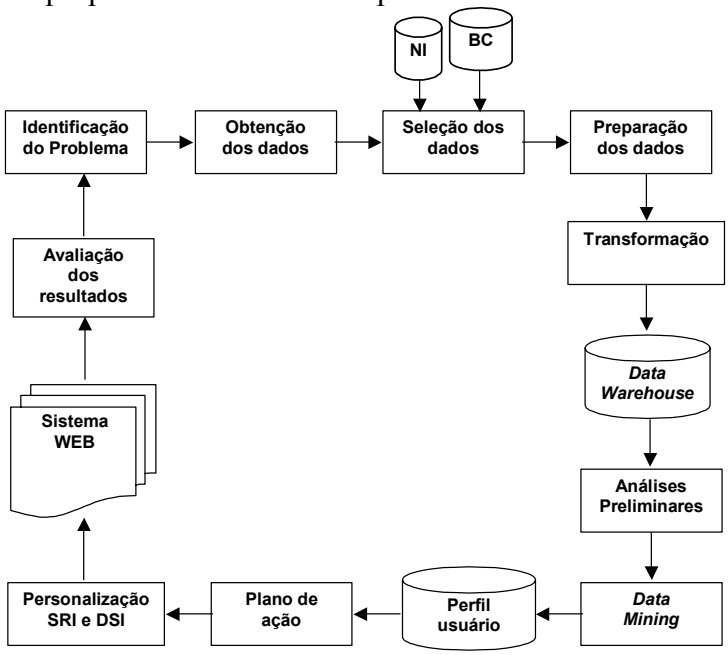

Figura 2. Modelo proposto para aplicação de *data mining* em bibliotecas

5.1. Identificação do problema

A maioria das bibliotecas não possui sistemas de recuperação e disseminação de informações capazes de ajudar no processo de localização das obras de interesse dos usuários. O mesmo é feito pelo serviço de referência com o auxílio de bibliotecários (a) ou especialistas na área.

A BC da FURB atualmente não apresenta nenhum sistema informatizado de DSI aos usuários. Este ainda é feito de forma manual pelo serviço de referência. O SRI não identifica o usuário para tratá-lo de forma seletiva e personalizada. Quando é feita uma consulta, a pesquisa retorna uma grande quantidade de informações (dados) a maioria sem relevância e nenhuma ordenação, o que caracteriza uma alta revocação, mas baixa precisão (CARDOSO, 2000, p. 2).

Definido o problema, elegem-se as variáveis que serão utilizadas na investigação para a resolução do mesmo. As variáveis que tem relacionamento direto para identificação do perfil dos usuários são: usuários; obras da Biblioteca; CDD (classificação decimal dewey do assunto principal da obra); transações (empréstimos, reservas).

5.2. Obtenção dos dados

Mediante a identificação das variáveis que serão utilizadas no processo de extração de conhecimento sobre o perfil do usuário, parte-se para o reconhecimento e a obtenção das mesmas nas fontes de dados. A principal fonte dos dados são os sistemas legados da BC mantidos pela seção de automação (cadastro e a circulação obras). Outra fonte é o sistema de identificação única de pessoas com vínculo na instituição, mantido pelo NI (usuários). A ligação entre o usuário e suas transações é feita através do identificador único do usuário.

5.3. Seleção dos dados

Através de uma engenharia reversa das tabelas de interesse armazenadas nos bancos de dados, torna-se possível reconhecer as variáveis de interesse para assim fazer a seleção.

Foram selecionados na amostra professores e alunos de pós-graduação da FURB. Com a integração das bases, excluem-se alguns dados como CPF, Endereço, etc, por serem usadas com finalidades operacionais que não se aplicam a esta pesquisa.

5.4. Pré-processamento dos dados

Após a seleção dos dados, faz-se a verificação da existência de inconsistências e/ou erros nas variáveis: A data de aquisição continha dados fora do formato padrão e o código CDD em alguns casos estava fora do padrão de catalogação.

5.5. Extração, transformação e carga dos dados

Os dados são estruturados para facilitar e agilizar o processo de mineração. A partir daí, foi gerado um *data mart*, que é parte de um *data warehouse*. Após identificar as variáveis de interesse, chega-se a um modelo que trata da circulação das obras.

A tabela fato é a de circulação de empréstimos, onde cada registro corresponde a uma transação que pode ser dos tipos: empréstimo e reserva. As dimensões encontradas são: a obra e o usuário que a emprestou. A partir do modelo de *data mart* foram criadas as tabelas e rotinas para carga dos dados.

Na área de biblioteconomia já foram institucionalizados alguns códigos para determinados domínios de uma variável, como é o caso da classificação dos livros. Existe uma codificação internacional, conhecida como Classificação Decimal Dewey (CDD), que é usada por diferentes órgãos da área de biblioteconomia, a fim de organizar o acervo e facilitar a localização das obras. Assim foram criados cinco níveis da CDD. A partir da qual foram gerados os assuntos significativos (AS) através da totalização das transações segundo a CDD, reduzindo as mesmas até um nível mínimo de significância.

5.6. Análises preliminares

Em qualquer investigação é fundamental para o pesquisador ter uma visão global dos dados que estão sendo pesquisados, a seguir apresenta-se uma análise descritiva dos dados da amostra, envolvidos neste estudo.

A amostra é composta por 17421 títulos que totalizam 51011 volumes, 3906 usuários, 821 da categoria professores e 3085 da categoria de pós-graduação.

Estes usuários realizaram 68543 transações, 66769 de empréstimo e 1775 de reservas. Tivemos uma média de 17,54 transações por usuários.

5.7. Mineração dos dados

Caracteriza-se pela transformação dos dados em conhecimento. Para encontrar o perfil do usuário utilizam-se as seguintes etapas:

5.7.1 Análise de conglomerados de assuntos significativos

A metodologia de análise de conglomerados (*cluster analysis*) é uma descoberta indireta de conhecimento a partir de algoritmos para encontrar registros de dados que são semelhantes entre si. Estes conjuntos de registros similares são conhecidos como clusters.

Segundo Velasquez et al. (2001, p. 2) “Todos os algoritmos de análise de conglomerados são baseados em uma medida de similaridade ou, ao contrário de

distância, que procuram expressar o grau de semelhança entre os objetos". Uma medida de distância muito utilizada quando os atributos são de natureza quantitativa é a distância euclidiana.

Formam-se agrupamentos das obras em grandes áreas de conhecimento, ou seja, grupos de livros os quais são utilizados por usuários para estudo de determinando assunto ou área. Assim foram analisados alguns métodos estatísticos de agrupamento hierárquico, como o do vizinho mais próximo, do vizinho mais distante, e de Ward. Optou-se pelo método Ward com distância euclidiana, pois o mesmo apresentou melhores resultados e por ser indicado por Aranha (2000) em seu trabalho.

Afirma Velasquez et al. (2001, p. 2) que "Nos métodos hierárquicos o número de classes não é fixado a priori, mas resulta da visualização do dendrograma, um gráfico que mostra a seqüência das fusões ou divisões ao longo do processo iterativo".

Foi aplicado esta técnica as dados de transações dos usuários segundo suas transações por AS, contidos no *data mart* resultando no dendograma apresentado na Figura 3.

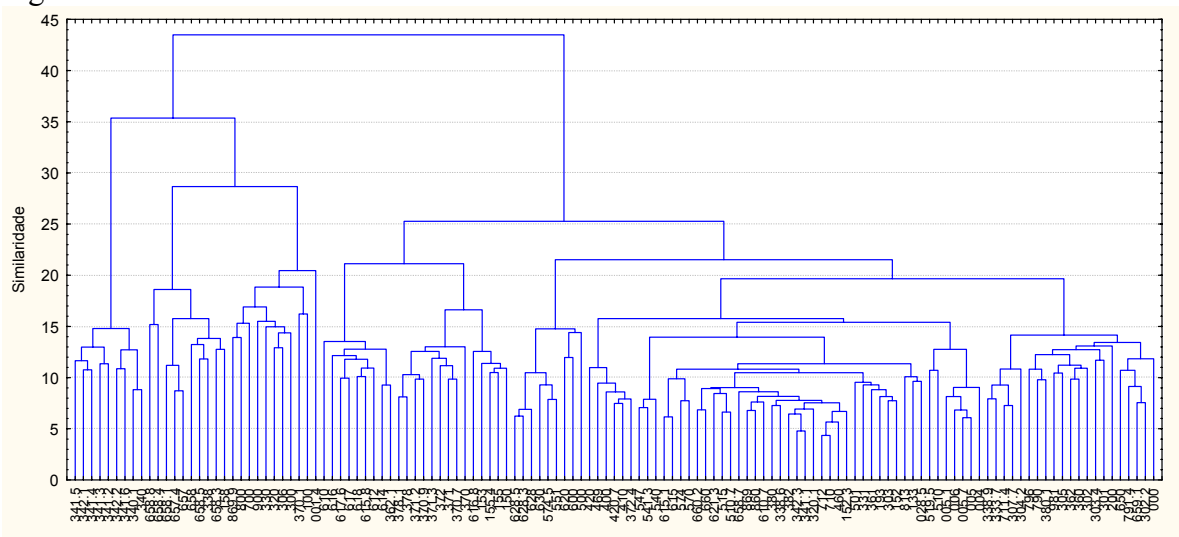

Figura 3. Dendograma com transações dos usuários por AS

Através da análise do dendograma foram gerados 34 grupos de grandes áreas como o apresentado na Tabela 1.

Tabela 1. Exemplo da tabela de grupos de grandes áreas de interesse

Grupo	Descrição do Grupo	AS	Descrição do AS
1	Direito	341.5	Direito penal
		342.1	Direito civil
		341.2	Direto constitucional
		342.2	Direito comercial
		341.6	Direito do trabalho
		340.1	Filosofia do Direito
		340	Direito

5.7.2 Classificação do acervo em grandes áreas

A classificação é uma tarefa muito utilizada em *data mining*. Consiste em examinar os aspectos de um objeto e atribuí-lo a um dos conjuntos de classes existentes. Assim,

classificam-se as obras do acervo da biblioteca em grandes áreas do conhecimento segundo a tabela gerada através da analise de *cluster* apresentada anteriormente.

5.7.3 Descrição do perfil dos usuários

Segundo Harrison (1998 p.181) "ás vezes o propósito de executar *data mining* é simplesmente descrever o que está acontecendo em um banco de dados complicado de maneira a aumentar o conhecimento das pessoas, produtos ou dos processos que produziram os dados".

A descrição do comportamento do usuário da biblioteca, através da análise de suas transações, objetiva identificar seu perfil de utilização de obras, podendo interagir com o mesmo através dos sistemas de SRI e DSI de forma personalizada.

No estudo do perfil, o primeiro nível de descrição seria a maior grande área de interesse, assim determinando a grande área de interesse do usuário. O próximo nível de descrição seria formado por três subáreas de interesse, no quarto nível da CDD, identificados através de uma análise das três principais áreas de transações do usuário. Tomamos como exemplo as transações de um usuário segundo as grandes áreas (Figura 3) e segundo CDD (Figura 4).

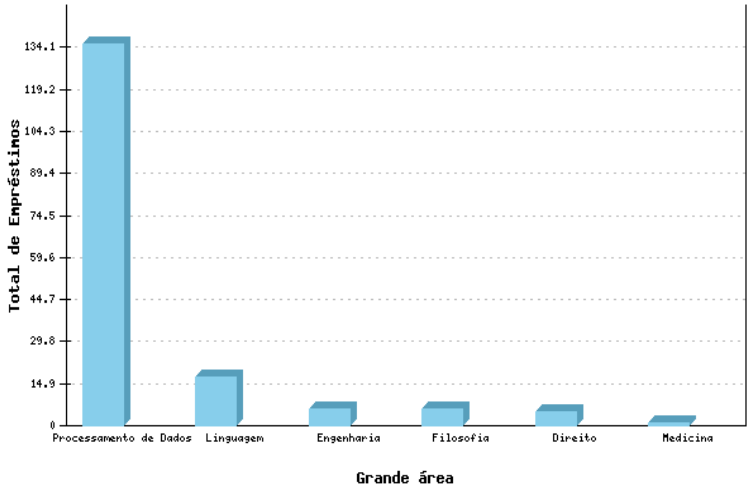

Gráfico 1. Transações usuários por grandes áreas

Pode-se verificar no gráfico 1 que o primeiro nível de descrição apresenta "processamento de dados" como a grande área de interesse do usuário em estudo.

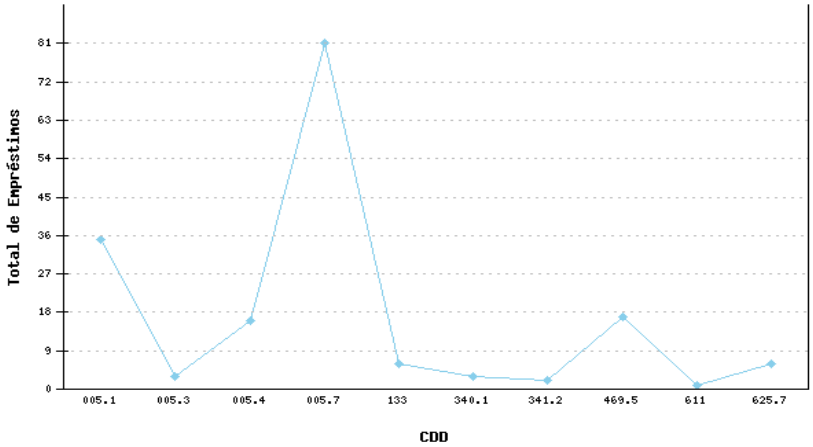

Gráfico 2. Transações usuários por CDD nível quatro

Como pode ser observado no gráfico 2, o segundo nível de descrição apresenta as três principais subáreas de interesse do usuário que são: 005.1, 005.7 e 469.5.

5.8. Plano de ação

Depois de identificado o perfil do usuário, torna-se possível personalizar os sistemas de recuperação e disseminação de informações. Para tanto, utiliza-se um sistema WEB (de SRI e DSI) que foi desenvolvido e personalizado dinamicamente ao perfil de cada usuário.

O sistema desenvolvido fica esperando requisições do servidor WEB quando a recebe processa e retorna páginas HTML com o conteúdo ao usuário personalizado. A arquitetura do sistema desenvolvido pode ser visto na Figura 4.

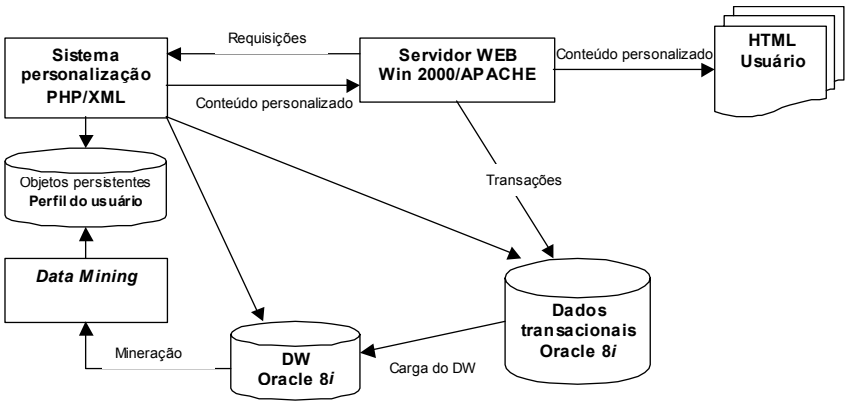

Figura 4. Arquitetura do sistema de personalização

O sistema conta com um banco de dados onde estão contidos os dados transacionais sobre as obras, usuários e suas transações, com um *data warehouse* onde estão os dados que serviram de fonte para a aplicação do *data mining*, e um objeto persistente o qual recebe os dados sobre o perfil do usuário. Quando o usuário faz uma requisição ao servidor WEB este recebe e a repassa para o sistema de personalização que recebe a requisição processa e envia a resposta de volta ao usuário personalizada.

5.9. Sistema WEB

Ao entrar no sistema é apresentada a tela de *login* onde devem ser informados o código e senha do usuário na biblioteca, após validação são carregados os dados do perfil do usuário para uma sessão no servidor, que funciona como objeto persistente ficando ativo até que o usuário saia do sistema.

A tela principal do sistema (Figura 5) é divida em três partes: menu superior, menu lateral, e corpo principal.

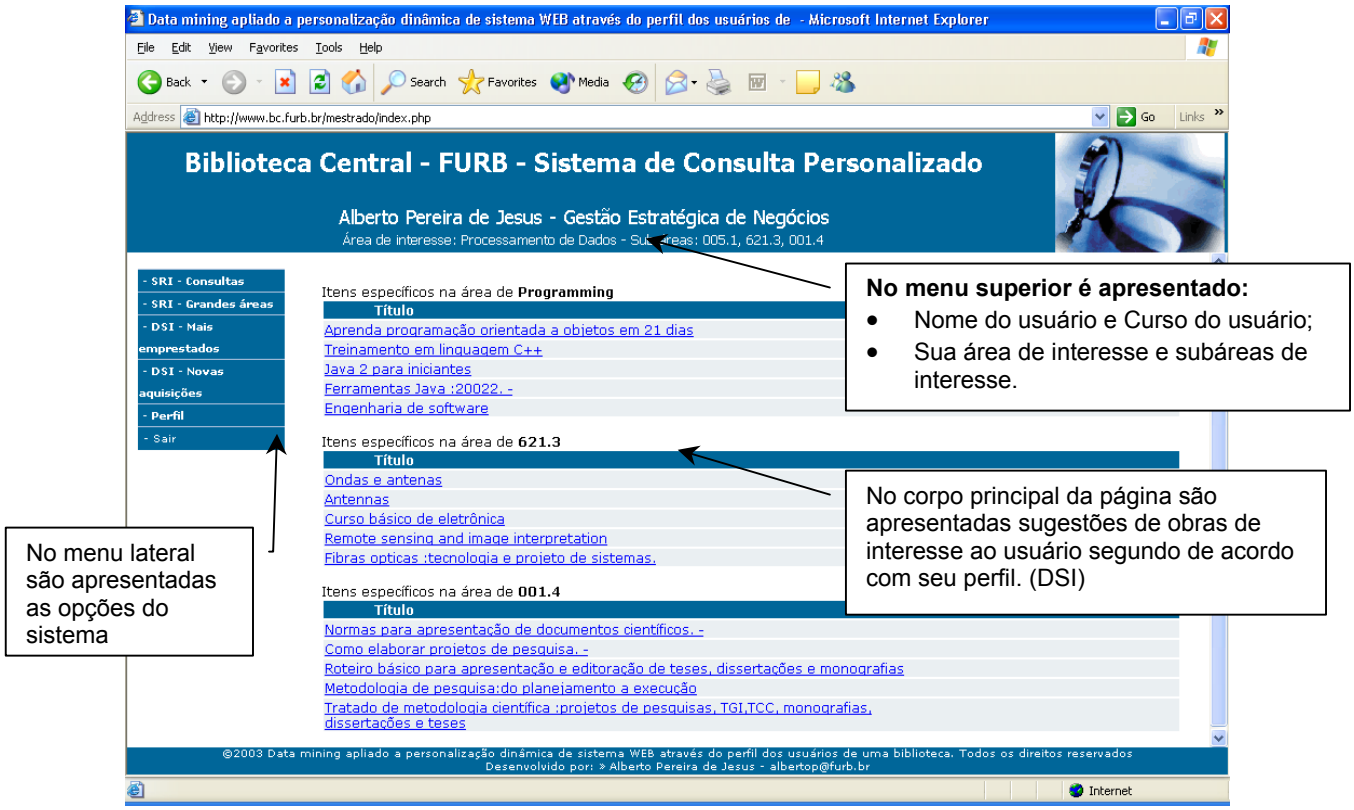

Figura 5. Tela principal do sistema

A tela de resultado da consulta (Figura 6) retornará os títulos encontrados no acervo segundo a expressão de busca determinada, ordenados conforme o perfil do usuário.

Resultado da consulta

Título	CDD	Relevância	Proximidade
Projeto & engenharia de software :teste de software	005.1	1	0
Qualidade & teste de software :engenharia de software, qualidade de software, qualidade de produtos de software, teste de software, formalização do processo de teste, aplicação prática dos testes	005.1	1	0
Guia completo ao teste de software	005.14	1	0.04
O teste gestáltico Bender para crianças	150.1982	999	145.0982
Tecnicas de exame psicologico e suas aplicacoes no Brasil :testes de aptidoes	155.28	999	150.18
Testes para admissão em empresas e empregos públicos	155.28	999	150.18

Figura 6. Tela resultado da consulta

5.10. Avaliação dos resultados

Através do modelo proposto e do protótipo desenvolvido foi possível melhorar o processo de recuperação e recomendações de obras através da identificação da relevância da mesma ao usuário.

6. Conclusão

Este trabalho propôs um modelo para extração automática do conhecimento sobre o perfil de usuários em bibliotecas. O modelo desenvolvido usa técnicas de *cluster*, classificação e descrição, fáceis de serem implementadas e interpretadas.

Os objetivos do trabalho foram alcançados. O *data warehouse* foi desenvolvido, o perfil dos usuários foi identificado com a aplicação de técnicas de *data mining* o sistema proposto implementado.

Quanto à tecnologia envolvida, acredita-se que está apenas nascendo e passará a fazer parte do nosso dia-a-dia. O mercado está em ampla expansão e com possibilidades de grandes negócios, pois a maioria das empresas possui grandes bancos de dados sem nenhuma utilização dos mesmos para tomada de decisões.

7. Referências

ARANHA, Francisco. **Análise de redes em procedimentos de cooperação indireta**: utilização no sistema de recomendações da Biblioteca Karl A. Boedecker. São Paulo: EAESP/FGV/NPP, 2000. 71p.

BARTOLOMEU, Tereza Angélica. **Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento**. 2002. 302f. Tese (Doutorado em Engenharia de Produção) – EPS. Universidade Federal de Santa Catarina, Florianópolis, 2002.

BERRY, Michael J. A, LINOFF, Gordon. **Data mining techniques**: for marketing, sales, and customer support. New York : J. Wiley E Sons, 1997. 454 p.

CARDOSO, Olinda Nogueira. Paes. Recuperação de Informação. **INFOCOMP Revista de Computação da UFLA**, Lavras, v.1, 2000. Disponível em: <http://www.comp.ufla.br/infocomp/e-docs/a2v1/olinda.pdf> Acesso em: 23 out. 2003.

DINIZ, Carlos Alberto R., LOUZADA NETO, Francisco. **Data mining**: uma introdução. São Paulo: ABE, 2000. 123p.

FUNARO, Vânia Martins B. O., CARVALHO, Telma de, RAMOS, Lúcia Maria S. V. Costa. **Inserindo a disseminação seletiva da informação na era eletrônica**. São Paulo: Serviço de Documentação Odontológica de Faculdade de Odontologia da USP. 17p.

HARRISON, T. H. **Intranet data warehouse**: ferramentas e técnicas para a utilização do data warehouse na intranet. Berkeley Brasil: São Paulo, 1998.

VELASQUEZ,Roberto M.G. et. al. Técnicas de Classificação para Caracterização da Curva de Carga de Empresas de Distribuição de Energia - Um Estudo Comparativo. **V Congresso Brasileiro de Redes Neurais**, 2001, Rio de Janeiro. Disponível em: <http://bioinfo.cpgei.cefetpr.br/anais/CBRN2001/5cbrn-6ern/artigos-5cbrn/>.