ALBERTO PEREIRA DE JESUS

DATA MINING APLICADO À IDENTIFICAÇÃO DO PERFIL DOS USUÁRIOS DE UMA BIBLIOTECA PARA A PERSONALIZAÇÃO DE SISTEMAS WEB DE RECUPERAÇÃO E DISSEMINAÇÃO DE INFORMAÇÕES

FLORIANOPÓLIS - SC

2004

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO
EM CIÊNCIA DA COMPUTAÇÃO

Alberto Pereira de Jesus

DATA MINING APLICADO À IDENTIFICAÇÃO DO PERFIL DOS USUÁRIOS DE UMA BIBLIOTECA PARA A PERSONALIZAÇÃO DE SISTEMAS WEB DE RECUPERAÇÃO E DISSEMINAÇÃO DE INFORMAÇÕES

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Paulo José Ogliari

Florianópolis, maio 2004.

Ficha Catalográfica elaborada pela

Biblioteca Central da FURB

Jesus, Alberto Pereira de
J58d Data Mining aplicado à identificação do perfil dos usuários de uma biblioteca para a personalização de sistemas Web de recuperação e disseminação de informações / Alberto Pereira de Jesus. - Florianópolis, 2004.
119 p. : il.

Orientador: Paulo José Ogliari.
Dissertação (mestrado) - Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Ciência da Computação.

1. Banco de dados. 2. Disseminação seletiva da informação. 3. Serviços de informação. 4. Sistemas de recuperação da informação. 5. World Wide Web (Sistema de recuperação da informação). I. Ogliari, Paulo José. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 025.525

DATA MINING APLICADO À IDENTIFICAÇÃO DO PERFIL DOS USUÁRIOS DE UMA BIBLIOTECA PARA A PERSONALIZAÇÃO DE SISTEMAS WEB DE RECUPERAÇÃO E DISSEMINAÇÃO DE INFORMAÇÕES

Alberto Pereira de Jesus

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Raul S. Waslawick, Dr.
Coordenador do Curso de Pós-Graduação

Banca Examinadora

_____

Paulo José Ogliari, Dr. (orientador), UFSC

_____

Amélia Silveira, Dr.,FURB

_____

Francisco José Espósito Aranha Filho, Dr., FGV

_____

Frank Augusto Siqueira, Dr.,UFSC

Dedico este trabalho aos meus pais, que incentivaram nos momentos difíceis, fazendo com que seguisse em frente, sempre.

AGRADECIMENTOS

Agradeço a minha família, por estarem ao meu lado em todos os momentos, incentivando e auxiliando em mais esta etapa da minha vida.

À Biblioteca Central da Universidade Regional de Blumenau - FURB, por disponibilizar espaço para a realização deste trabalho.

Aos colegas de trabalho Mauro Tessari, Marcos Rogério Cardoso, Evanilde Maria Moser, Izildinha Ramos Accetta e Maria Genoveva Lemos por auxiliarem no desenvolvimento deste estudo.

Aos colegas de mestrado Fernando Prass, Juliano Pacheco, Jones Daros, Jaqueline Uber Silva pela convivência e crescimento mútuo.

Ao Professor Doutor Paulo José Ogliari, pela orientação e por acreditar e confiar no meu potencial e grato pelo apoio prestado para conclusão desta pesquisa.

Ao Professor Doutor Francisco Aranha, por disponibilizar material e pelas valiosas sugestões a este trabalho.

Aos Professores Doutores Amélia Silveira e Frank Augusto Siqueira, pela participação na banca examinadora deste estudo.

A todos aqueles que diretamente ou indiretamente contribuíram para a conclusão deste trabalho.

“Não basta saber,
é preciso também aplicar;
não basta querer,
é preciso também agir.”
(Goethe)

RESUMO

As bibliotecas têm como missão oferecer a seus usuários informações relevantes para a realização de suas pesquisas, facilitando o acesso e localização do material necessário. Desse modo, é importante que se conheça o perfil dos usuários quanto a suas necessidades bibliográficas. Propõe-se um modelo para aplicação das técnicas de *data mining* para identificar o perfil dos usuários de bibliotecas, personalizando os processos de recuperação e disseminação da informação. Para tanto, desenvolve-se um *data warehouse* com os dados históricos de transações de empréstimos, dando suporte à aplicação da tecnologia de extração de conhecimento - *data mining*. Dessa forma, torna-se possível a descoberta de correlações e informações implícitas. Com o resultado dos conhecimentos obtidos sobre o perfil do usuário desenvolve-se um sistema WEB personalizado de recuperação e disseminação seletiva de informações. Na execução desse modelo na Biblioteca Central da Universidade Regional de Blumenau, utilizaram-se as técnicas de análise de conglomerados sobre os assuntos das obras, identificando os grupos de livros que formam grandes áreas de conhecimento. As técnicas de *data mining* possibilitam classificar as obras em grupos e descrever o perfil do usuário. O sistema WEB desenvolvido utilizou o perfil identificado para a personalização da recuperação e disseminação de informações, tornando os processos de recuperação e disseminação de informações eficientes e também seletivos, facilitando a busca por informações, aumentando desta forma a satisfação dos usuários. Cabe destacar que o modelo desenvolvido poderá ser aplicado em outras bibliotecas.

Palavras-chave: *Data Mining*. *Data Warehouse*. Bibliotecas. SRI. DSI. Personalização de Conteúdo WEB.

ABSTRACT

A library has to, as a purpose, offer to its users relevant information for the accomplishment of their research, facilitating the localization and access of the needed material. Therefore, it's important to be known the users' profile, regarding their bibliographical preferences. A "data mining" techniques application model is proposed for library's users profile identification, personalizing the information search and spread proccess. Hence, a "data warehouse" containing historical data from material loans is developed, in order to support the application to knowledge extraction's technology - "data mining". Hence, allowing the discovery of correlations and implicit information. Using the knowledge obtained from the users' profile a information recovery and selective spread personalized WEB system is developed. Executing this model at Biblioteca Central da Universidade Regional de Blumenau (university's library), cluster analysis' techniques were used on the material's subjects, identifying book groups which form big knowledge areas' groups. The "data mining" techniques allow the material classification in groups and the user's profile description. The developed WEB system used the identified profile for the personalization of the information recovery and spread and also making it selective, becoming the the information search easier, increasing the users' satisfaction level. Let alone that the developed model may be applied in other libraries.

Key words: Data mining. Data warehouse. SRI. DSI. Libraries. WEB Personalization.

LISTA DE ABREVIATURAS E SIGLAS

AS – Assuntos Significativos

BC – Biblioteca Central da FURB

BD – Banco de Dados

CDD – Classificação Decimal Dewey

DSI – Disseminação Seletiva da Informação

FURB – Fundação Universidade Regional de Blumenau

KDD – *Knowledge Discovery in Databases*

MBR – *Memory Based Reasoning*

NI – Núcleo de Informática da FURB

SQL – *Structured Query Language*

SRI – Sistema de Recuperação de Informação

UFSC – Universidade Federal de Santa Catarina

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com o crescimento do volume de publicações e também das necessidades de informações dos usuários, 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 nas bibliotecas, sendo eles: o primeiro sistema de recuperação de informações – SRI e o segundo 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.

Funaro et. al. (2000, p. 2) afirmam que "A falta de tempo para realizar suas próprias pesquisas bibliográficas demonstra que a DSI torna-se uma atividade de grande importância e aceitação quando usada como meio suplementar de informação uma vez que permite aos pesquisadores obter maior disponibilidade para dedicarem-se à execução de suas pesquisas propriamente ditas".

Assim, conhecer os interesses bibliográficos dos usuários é importante e já era uma necessidade no passado, quando 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 para realizar esta tarefa.

Sabendo-se que a missão das Bibliotecas, que 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 recuperação e disseminação da informação das Bibliotecas necessitam evoluir e ser inteligentes, a fim de agregar valor a esses serviços de referência. Torna-se 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 assim, a personalização dos processos de recuperação e disseminação da informação, 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 oferecem, apresentam alguns problemas. Ao pedir para que cada usuário preencha um cadastro determinando seus interesses a fim de determinar seu perfil, pode se ter 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 em um semestre lecionava uma disciplina de *data warehouse* e no semestre seguinte leciona a disciplina de *data mining*. Como ele preencheu seus dados com o antigo perfil que era *data warehouse*, continuará recebendo informações sobre seus interesses preenchidos e não sobre os reais interesses atuais que é *data mining*.

Assim, 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 feitos diariamente, estando disponíveis em cada uma destas operações, mas nunca, são utilizados como base para geração de informações para tomada de decisões.

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

a) melhorar o processo de recuperação de informações através da personalização das consultas (o retorno da consulta por expressões é filtrado segundo o perfil do usuário);

b) facilitar o processo de disseminação de informações, recomendando obras de interesse ao usuário.

1.1 PROBLEMA DA PESQUISA

As bibliotecas, em geral, não possuem 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 especialistas na área.

A Biblioteca Central da FURB atualmente não apresenta um sistema informatizado de disseminação de informações aos usuários. Este é feito de forma manual pela seção de referência. O sistema de recuperação de informação não identifica o usuário para tratá-lo de forma seletiva e personalizada. Quando é feita uma consulta, a pesquisa retorna as informações sem nenhuma ordenação; há alta revocação e pouca precisão.

"Precisão é a fração dos documentos já examinados que são relevantes, e revocação é a fração dos documentos relevantes observada dentre os documentos examinados" (CARDOSO, 2000, p. 2).

1.2 QUESTÕES DA PESQUISA

Diante destas considerações, as seguintes questões de pesquisa podem ser levantadas:

a) Que mecanismos poderiam ser utilizados para identificar as preferências de usuários de uma Biblioteca, utilizando-se como fonte de informação as bases de dados relativas aos registros de empréstimos e reservas?

b) Quais são as tarefas e técnicas de *data mining*, que poderiam ser utilizadas para identificar o perfil dos usuários de uma Biblioteca?

c) Como a utilização de técnicas de personalização dinâmica aplicada a sistemas de recuperação e disseminação seletiva de informações de uma Biblioteca ajudaria os usuários a encontrarem livros de seu interesse?

1.3 OBJETIVOS

1.3.1 Objetivo geral

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*.

1.3.2 Objetivos específicos

a) desenvolver um *data warehouse* para dar suporte a aplicação das técnicas de *data mining*, possibilitando também obter informações para tomada de decisões com a criação de relatórios através de ferramentas apropriadas;

b) aplicar técnicas de *data mining* sobre o histórico de empréstimos, reservas e consultas dos usuários para identificar o perfil dos mesmos na Biblioteca Central da FURB;

c) desenvolver um sistema WEB de recuperação e disseminação seletiva de informações personalizado dinamicamente para a Biblioteca Central da FURB.

1.4 JUSTIFICATIVA

Através da aplicação de técnicas de *data mining* em bibliotecas é possível o conhecimento das características e preferências de seus usuários, determinando assim seu perfil que é de elevada importância para os processos de recuperação e disseminação seletiva das informações 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 e facilidade no atendimento dos usuários.

1.5 LIMITAÇÕES DO ESTUDO

O estudo apresenta algumas limitações:

a) serão utilizados somente dados parciais, referentes ao ano de 2003;

b) os usuários explorados são somente os professores e os alunos de pós-graduação.

1.6 RESULTADOS ESPERADOS

Pretende-se identificar o perfil do usuário na biblioteca, agregando valor aos serviços de referência, fazendo com que os sistemas de SRI e DSI sejam personalizados dinamicamente segundo o perfil minerado do usuário, tornando a busca por informações mais fácil.

1.7 ESTRUTURA DO TRABALHO

O conteúdo está organizado da seguinte forma:

O primeiro capítulo traz uma visão geral do trabalho, o problema, seus objetivos, justificativas, limitações, resultados e a estrutura do trabalho.

O segundo capítulo apresenta a fundamentação teórica sobre *data mining* e os serviços de referência.

O terceiro capítulo descreve os métodos que serão adotados para investigação do perfil do usuário da Biblioteca, visando atender ao que está proposto nas questões da pesquisa e nos objetivos.

O quarto capítulo relata a aplicação do modelo proposto e o desenvolvimento de um protótipo.

O quinto capítulo traz as conclusões e recomendação para trabalhos futuros.

As referências, os anexos e os apêndices completam este trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

A quantidade de informações produzidas aliada à 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, sendo utilizados somente como fonte histórica. Estas organizações têm dificuldades na identificação de formas de exploração desses dados, e mais ainda 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, o qual possibilita a extração de informações de um grande banco de dados, e seu uso para a tomada de decisões (DINIZ; NETO, 2000).

Para a implantação da tecnologia de *data mining* é necessário que se conheça todo o processo, para que a mesma venha atender às expectativas do usuário. Assim, o presente capítulo apresenta uma revisão bibliográfica sobre *data mining* e as principais características desta tecnologia.

2.1 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS - DCBD

Descoberta de conhecimentos em bancos de dados - DCBD ou *Knowledge Discovery in Databases* - KDD possibilita a descoberta eficiente de conhecimentos sobre uma grande coleção de dados.

KDD é um processo contínuo 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; Neto (2000) "embora os passos devam ser executados na ordem 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)".

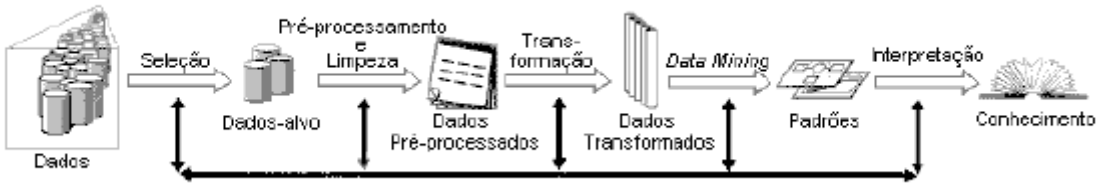

Figura 1 – Passos do processo de KDD

Fonte: FIGUEIRA, 1998, p. 8.

O processo de KDD começa obviamente com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos.

Cada etapa do processo será apresentado nas subseções seguintes, sendo que o passo do *data mining* o qual é objetivo principal deste trabalho, será tratado em destaque na seção 2.2.

2.1.1 Seleção dos dados

Os dados representam a fonte para a descoberta do conhecimento. Geralmente estão armazenados em bancos de dados das organizações ainda não explorados e provenientes dos sistemas legados que através da aplicação do processo de KDD resultarão em conhecimento.

Para entender como os dados podem ser transformados em informação resultando em conhecimento faz-se necessária sua diferenciação. Dados são os componentes básicos, ou seja, qualquer elemento identificado em sua forma bruta, sendo que por si só não conduz em compreensão de determinada situação, e a partir dos quais a informação é criada. Informações são dados inseridos em um contexto, ou seja, uma situação que está sendo analisada. Assim verifica-se que a partir da informação que será gerado o conhecimento, o qual permite tomar decisões adequadas, trazendo assim a tão necessária vantagem competitiva para uma empresa (BARTOLOMEU, 2002).

Assim o objetivo principal da seleção dos dados é identificar a origem dos mesmos, e extrair o subconjunto de dados necessários, através da seleção das variáveis de interesse para a aplicação de *data mining*. A seleção varia de acordo com interesses e resultados esperados. Essas variáveis de interesse podem ser do tipo qualitativa (nominal, ordinal) ou quantitativa (discreta, contínua) (DINIZ, NETO, 2000).

Segundo Diniz; Neto (2000) “as variáveis selecionadas para o *data mining* são denominadas variáveis ativas uma vez que são usadas para distinguir segmentos, fazer predições ou desenvolver outras operações específicas do *data mining*”.

2.1.2 Pré-processamento dos dados

Diniz; Neto (2002, p. 25) afirmam que "a qualidade dos dados é essencial para o obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração".

O pré-processamento dos dados tem como objetivo assegurar a qualidade dos dados selecionados. Para isso é necessário fazer uma limpeza dos dados a fim remover valores desconhecidos, não preenchidos e tratar dados incompletos e com erros. Isto pode ser feito utilizando se uma combinação de métodos estatísticos e técnicas de visualização de dados (Diniz; Neto, 2002).

2.1.3 Transformação dos dados

Os algoritmos de mineração normalmente não podem acessar os dados em seu formato nativo, seja em razão da forma como são armazenados, seja pela normalização adotada na modelagem do banco. É necessária a conversão dos dados para um formato apropriado. Podemos fazer a estruturação dos mesmos a fim de facilitar o processo de mineração gerando um *data warehouse*. Técnicas como discretização e redução de dimensionalidade podem ser utilizadas (Diniz; Neto, 2002).

2.1.4 *Data Mining*

Consiste na efetiva aplicação do algoritmo escolhido sobre os dados a serem analisados com o objetivo de localizar os padrões desejados. A qualidade dos resultados desta etapa dependerá diretamente da correta realização das anteriores.

> "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 ligada à experiência e intuição do analista" (Diniz; Neto, p.28 2000).

Como o objetivo principal deste trabalho é a aplicação do *data mining* na seção 2.2 serão apresentados detalhes sobre o *data mining*, suas técnicas e algoritmos a fim de verificar as alternativas possíveis para utilização neste trabalho.

2.1.5 Interpretação

Nesta etapa, as informações resultantes das etapas anteriores são interpretadas e analisadas de forma que o conhecimento resultante de todo o processo possa ser aplicado um plano de ação para a solução do problema que havia dado origem à aplicação deste processo (DINIZ; NETO, 2000).

2.1.6 Conhecimento

Como resultado de todo o processo temos o conhecimento adquirido sobre os dados minerados. Após a interpretação e a execução do plano de ação podemos avaliar os resultados obtidos.

2.2 *DATA MINING*

Data mining ou mineração de dados é "o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados, e seu uso para tomada de decisões" (DINIZ; NETO, 2000 p. 2).

Faz parte de um processo maior conhecido como *Knowledge Discovery in Databases* - KDD, ou descoberta de conhecimentos em bancos de dados. Mais detalhes sobre o processo de KDD pode ser visto no item 2.1.

Afirmam Diniz e Neto (2000, p. 2) que "vários autores tem tecido definições, muitas vezes conflitantes, sobre *data mining*, o que acrescenta dificuldades para uma definição única". Definições de alguns autores são dadas a seguir:

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

> "Enquanto o banco de dados representa a memória da empresa, registrando diversos tipos de informação sobre os clientes bem como interações entre cliente e empresa, a mineração de dados é a inteligência que, quando aplicada a esta memória, pode identificar padrões, desvios e tendências úteis para o aprimoramento do negócio" (REATEGUI, 2002, p. 140).

Podemos concluir que *data mining* se relaciona com a análise de dados e o uso de meios computacionais (equipamentos e softwares) na busca de conhecimento em um grande conjunto de dados. É extremamente importante a escolha apropriada da ferramenta ou conjunto de ferramentas de *data mining* que será implantada. Assim, é importante que se conheçam, as tarefas desempenhadas e suas técnicas, a fim de dar suporte a sua escolha.

2.3 TAREFAS DESEMPENHADAS PELO *DATA MINING*

Segundo Reategui (2002), as tarefas de *data mining* podem ser divididas em dois grandes grupos:

a) descoberta direta de conhecimento: este tipo de mineração é orientada por objetivo, ou seja, explica o valor de determinado campo (renda, idade, sexo, etc.) a partir de outros. Para tanto se seleciona um campo alvo e solicita-se ao sistema como estimá-lo, classificá-lo ou prevê-lo;

b) descoberta indireta de conhecimento: não há campo alvo. Simplesmente perguntamos ao sistema como identificar padrões significativos nos dados. A partir daí a técnica de mineração trabalha livremente na descoberta de padrões que podem ser úteis.

As próximas subseções descrevem as principais tarefas de *data mining*:

2.3.1 Classificação

A classificação é a tarefa mais comum de *data mining*. Consiste em examinar características de um objeto ou situação e atribuir a ele uma classe pré-definida.

Exemplos de tarefas de classificação:

a) classificar livros de uma biblioteca por área;

b) classificar clientes em baixo, médio ou alto risco de empréstimo bancário;

c) classificar clientes potencialmente consumidores de um determinado produto a julgar pelo seu perfil;

d) discriminar solicitações de cobertura de seguros fraudulentas das não fraudulentas;

e) atribuir palavras-chave a artigos jornalísticos.

Espera-se poder atribuir a cada um dos registros do banco de dados uma determinada classe. As técnicas de redes neurais artificiais, árvores de decisão(CHAID, CART), análise discriminante e regressão logística são apropriadas para a tarefa de classificação.

2.3.2 Estimação

Enquanto a classificação lida com resultados discretos, a estimação trabalha com valores numéricos contínuos. Tendo-se algumas variáveis explicativas usa-se a estimação para retornar (obter) um valor para alguma variável desconhecida, tal como: rendimento, altura, saldo do cartão de crédito, etc.

Exemplos de tarefas de estimação:

a) estimar o número de filhos de uma família;

b) estimar a renda de uma família;

c) estimar a probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames;

d) estimar a probabilidade de um consumidor realizar uma compra.

Estimar é determinar da melhor maneira possível um valor baseando-se em outros valores de situações idênticas. As técnicas de redes neurais artificiais, algoritmos genéticos, estatística (intervalos de confiança e predição) são adequadas para estimação.

2.3.3 Previsão

A previsão é o mesmo que classificação ou estimação, exceto pelo fato que os registros são classificados de acordo com alguma atividade futura prevista ou valor futuro estimado. Qualquer uma das técnicas usadas na classificação ou estimação podem ser adaptadas para o uso na previsão.

Os dados históricos (série temporal) são usados para construir um modelo que explica o comportamento atual observado. Quando este modelo é aplicado a entradas atuais, o resultado é a previsão de atitudes futuras.

Exemplos de tarefas de previsão:

a) determinar se o índice Bovespa subirá ou descerá amanhã;

b) prever qual será a população de uma cidade daqui a 5 anos;

c) prever quais consumidores deixarão de comprar dentro dos próximos seis meses.

A previsão consiste na determinação do futuro de uma grandeza. As regras de associação, raciocínio baseado em casos, redes neurais artificiais, árvores de decisão e séries temporais são todas adequadas. A escolha da técnica dependerá da natureza dos dados.

2.3.4 Agrupamento por afinidade

A tarefa de agrupamento por afinidade tem por objetivo encontrar quais produtos ou serviços os consumidores buscam conjuntamente.

Exemplos de tarefas de agrupamento por afinidade:

a) um mercado de vendas a varejo pode dispor os produtos vendidos conjuntamente no mesmo corredor;

b) um comerciante da web pode usar a análise de afinidade para determinar o layout do seu catálogo;

c) bancos e companhias telefônicas podem usar análise de afinidade para determinar quais novos produtos oferecer para seus consumidores preferenciais.

O agrupamento por afinidade é uma abordagem simples para gerar regras de associação a partir de dados. Se dois livros são emprestados freqüentemente em conjunto, pode ser gerada uma regra: quem empresta o livro A empresta o livro B com probabilidade P1. A técnica mais utilizada é a análise de seleção estatística.

2.3.5 Segmentação

A segmentação é o processo de agrupamento de uma população heterogênea em vários subgrupos ou *clusters* (conglomerados) mais homogêneos. O que distingue a segmentação da classificação é que na primeira existem classes pré-definidas.

Exemplos de tarefas de segmentação:

a) agrupar usuários de uma biblioteca em grupos homogêneos por área de interesse com o seu perfil;

b) segmentação de clientes em um projeto de marketing;

c) colocar em um mesmo conjunto bactérias encontradas com características semelhantes.

A segmentação é normalmente uma técnica preliminar utilizada quando nada ou pouco se sabe sobre os dados, como na metodologia da descoberta não supervisionada de relações. Técnicas utilizadas para segmentação são redes neurais artificiais, estatística (análise de conglomerados) e algoritmos genéticos.

2.3.6 Descrição

O propósito é descrever o que está acontecendo numa base de dados, de forma que aumente o nosso entendimento sobre os consumidores, produtos. A boa descrição do comportamento facilitará encontrar uma explicação para o mesmo.

Exemplos de tarefas de descrição:

a) análise de dados pode demonstrar que “nos EUA, mulheres apóiam democratas mais do que homens”;

b) a maior parte dos pacientes que apresenta o sintoma X, possui a doença Y.

2.4 TÉCNICAS DE *DATA MINING*

Existem várias técnicas de *data mining*. A técnica a ser usada é determinada pelo tipo de informação que se quer obter através dos dados.

Berry; Linoff (1997) afirmam que nenhuma técnica resolve todos os problemas de mineração de dados. A familiaridade com uma variedade de técnicas é necessária para encontrar o melhor caminho para resolver o problema.

Harrison (1998) indica que não há uma técnica que resolva todos os problemas de mineração de dados. A escolha dependerá da tarefa específica a ser executada e dos dados disponíveis para análise.

Chen et. al. (1996) *apud* Bartolomeu (2002) propõem os seguintes esquemas de classificação, baseados em questionamentos que devem ser levantados:

a) "Que tipos de bancos de dados tenho para trabalhar?" Um sistema de descoberta de conhecimento pode ser classificado de acordo com os tipos de bancos de dados sobre os quais técnicas de mineração de dados são aplicadas, tais como: bancos de dados relacionais, bancos de dados de transação, orientados a objetos, dedutivos, espaciais, temporais, de multimídia, heterogêneos, ativos, de herança, banco de informação de Internet e bases textuais.

b) "Que tipo de conhecimento pretende-se explorar?" Vários tipos de conhecimento podem ser descobertos por extração de dados, incluindo regras de associação, regras características, regras de classificação, regras discriminantes, agrupamento, evolução e análise de desvio.

c) "Que tipo de técnica deve-se utilizar?" A extração de dados pode ser categorizada de acordo com as técnicas de mineração de dados subordinadas. Pode ser categorizada, de acordo com a abordagem de mineração de dados subordinada, tal como: extração de dados baseada em generalização, baseada em padrões, baseada em teorias estatísticas ou matemáticas, abordagens integradas etc.

Segundo Diniz; Neto (2000) para cada particular problema tem-se um particular algoritmo ou técnica. Abaixo são relacionados as principais técnicas e algoritmos de mineração de dados:

2.4.1 Análise de Associação - *Market Basket Association Analysis*

A análise de associação gera redes de interações e conexões presentes nos conjuntos de dados usando as associações item a item. Onde por associação item a item entende-se que a presença de um item implica necessariamente na presença de outro item na mesma transação.

O exemplo mais fácil é o do carrinho do supermercado do qual se pode extrair muita informação sobre que produtos os consumidores compram em conjunto com grande chance.

Considere um banco de dados de compras, onde cada compra (transação) consiste de vários artigos (itens) comprados por um consumidor. A aplicação de técnicas de análise de associação neste conjunto de transações pode revelar afinidades entre uma coleção de itens. Estas afinidades entre itens são representadas por regras de associação. Uma regra expõe, em forma textual, quais itens implicam a presença de outros itens

As informações resultantes podem ser usadas para vários objetivos, como planejar a arrumação de lojas, criar "pacotes" de produtos, fazer vendas cruzadas, entre outros. (Harrison, 1998).

2.4.2 Raciocínio baseado em casos

O MBR (*Memory Based Reasoning*) ou raciocínio baseado em casos é uma técnica de *data mining* dirigida, que usa exemplos conhecidos como modelo para fazer previsões sobre exemplos desconhecidos (Harrison, 1998).

Berry; Linoff (1997) afirmam que o MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão.

Segundo Harrison (1998) uma das maiores vantagens do MBR é a habilidade de ser executado em qualquer fonte de dados, mesmo sem modificações. Os dois elementos-chave no MBR são a função de distância usada para encontrar os vizinhos mais próximos e a função de combinação, que combina valores dos vizinhos para fazer uma previsão. Outra vantagem do MBR é sua habilidade de aprender sobre novas classificações simplesmente introduzindo novos exemplos no banco de dados. Uma vez encontradas a função de distância e a função de combinação corretas, tendem a permanecer muito estáveis, mesmo com a incorporação de novos exemplos para novas categorias nos dados conhecidos.

Esta facilidade de incorporar mudanças ao domínio e à extensão separa o MBR da maioria das outras técnicas de *data mining*, que precisam ser reaplicadas para incorporar informações substancialmente novas (KIMBALL, 1998).

2.4.3 Algoritmos genéticos

Os algoritmos genéticos aplicam mecanismos de seleção genéticos e naturais para uma busca usada para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão. Como tal são usados no *data mining* dirigido. Os algoritmos genéticos são semelhantes à estatística, pois também precisam conhecer o modelo em profundidade (BERRY; LINOFF, 1997).

Harrison (1998) diz que os algoritmos genéticos usam os operadores seleção, cruzamento e mutação para desenvolverem sucessivas gerações de soluções. Com a evolução do algoritmo, somente os mais previsíveis sobrevivem, até as funções convergirem em uma

solução ideal. O algoritmo genético tem sido muito usado para aprimorar MBRs e redes neurais. Árvores de decisão e indução de regras

2.4.4 Análise de agrupamentos

Harrison (1998) define esta técnica como a construção de modelos que encontram registros de dados semelhantes. Estas reuniões por semelhança são chamadas grupos (*clusters*).

Segundo Berry; Linoff (1997) trata-se de *data mining* não-direcionado, uma vez que a meta é encontrar similaridades não conhecidas previamente. Existem muitas técnicas para detecção de *clusters*, como métodos de estatística ou redes neurais, sendo que o mais utilizado é o *k-means*.

Agrupar por semelhança pode fornecer o ponto de partida para saber o que há nos dados e descobrir como usá-los melhor (Harrison, 1998).

A identificação de *clusters* é a tarefa de descoberta indireta de conhecimento a partir da construção de modelos para encontrar registros de dados que são semelhantes entre si.

As técnicas de análise de agrupamentos dividem-se em dois tipos: as de agrupamento hierárquico e as de agrupamento não-hierárquico.

2.4.5 Análise de vínculos

Harrison (1998) afirma que a análise de vínculos segue as relações entre registros para desenvolver modelos baseados em padrões nas relações. Este é um aplicativo de construção de teoria gráfica de *data mining*.

Como ferramenta, a técnica de análise de vínculos não é muito compatível com a tecnologia de bancos de dados relacionais. A maior área onde é aplicada é a área policial, onde pistas são ligadas entre si para solucionar os crimes.

2.4.6 Árvores de decisão e indução de regras

As árvores de decisão são usadas para *data mining* dirigido, particularmente a classificação. Dividem os registros do conjunto de dados de treinamento em subconjuntos separados, cada um descrito por uma regra simples em um ou mais campos.

Segundo Kimball (1998), uma das principais vantagens das árvores de decisão é que o modelo é bem explicável, uma vez que tem a forma de regras explícitas. Isso permite às pessoas avaliarem os resultados, identificando atributos-chave no processo. As próprias regras podem ser expressas facilmente como declarações lógicas em uma linguagem como *Structured Query Language* - SQL, de modo que possam ser aplicados diretamente em novos registros.

Harrison (1998) identificou como uma das principais vantagens das árvores de decisão a facilidade de explicação de seu modelo, devido a sua forma de regras explícitas.

2.4.7 Redes neurais

Harrison (1998) diz que as redes neurais são provavelmente a técnica de *data mining* mais comum, talvez sinônimo de *data mining* para alguns. São modelos simples de interconexões neurais no cérebro, adaptados para uso em computadores. Na forma mais comum, aprendem com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão. As redes neurais podem também ser aplicadas ao *data mining* não-dirigido e às previsões em séries temporais.

Uma das principais vantagens das redes neurais é a sua variedade de aplicações. Devido a sua utilidade, as ferramentas que suportam redes neurais são fornecidas por várias empresas para uma variedade de plataformas. As redes neurais são interessantes também porque detectam padrões nos dados de forma analógica ao pensamento humano – um fundamento interessante para uma ferramenta de *data mining* (BARTOLOMEU, 2002).

As redes neurais apresentam duas desvantagens: a dificuldade de compreender os modelos produzidos por elas e a particular sensibilidade ao formato dos dados que as alimentam. Representações de dados diferentes podem produzir resultados diversos, e o ajuste dos dados é uma parte significativa do esforço para utilizá-las.

2.5 ÁREAS DE APLICAÇÃO DE *DATA MINING*

A seguir, são relacionadas algumas áreas que tem aplicado a tecnologia de mineração de dados, *data mining*, segundo Diniz; Neto (2000); Reategui (2002) e Bartolomeu (2002):

Marketing: as técnicas de *data mining* são aplicadas para descobrir preferências do consumidor e padrões de compra, com o objetivo de realizar marketing direto de produtos e ofertas promocionais, de acordo com o perfil do consumidor.

Detecção de fraudes: muitas fraudes óbvias (tais como, a compensação de cheque por pessoas falecidas) podem ser detectadas sem a utilização de tecnologias de mineração de dados. Porém, padrões mais sutis de fraude podem ser difíceis de serem detectados, como, por exemplo, predizer quem se tornará inadimplente em seus pagamentos.

Medicina: caracterizar comportamento de paciente para prever visitas, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças.

Instituições governamentais: descoberta de padrões para melhorar as coletas de taxas e impostos, detectar fraudes, bem como formular políticas públicas.

Ciência: técnicas de mineração de dados podem ajudar cientistas em suas pesquisas, por exemplo, para encontrar padrões em estruturas moleculares, dados genéticos, mudanças globais de clima etc, podendo oferecer, rapidamente, conclusões valiosas.

Controle de processos e controle de qualidade: auxiliar no planejamento da produção e buscar por padrões de condições físicas na embalagem e armazenamento de produtos.

Instituições financeiras: detectar padrões de uso de cartão de crédito fraudulento, identificar clientes "leais", determinar gastos com cartão de crédito por grupos de clientes, encontrar correlações escondidas entre diferentes indicadores financeiros.

Apólice de seguro: análise de reivindicações – determinar quais procedimentos médicos são reivindicados juntos, prever quais clientes comprarão novas apólices, identificar padrões de comportamento de clientes perigosos, identificar comportamento fraudulento.

Transporte: determinar as escalas de distribuição entre distribuidores, analisar padrões de carga etc.

2.6 USO DE *DATA MINING* PARA PERSONALIZAÇÃO DE CONTEÚDO WEB

A personalização de conteúdo é uma característica de sistemas automatizados que permite que usuários distintos sejam tratados de forma diferente. Estes sistemas são capazes de identificar o usuário e direcionar a cada um conteúdo, recomendação de livros e de serviços que sejam relevantes.

Segundo REATEGUI (2002, p. 153) "o termo personalização é utilizado para designar sistemas capazes de reconhecer os usuários, armazenar dados sobre a interação com cada um e personalizar sua "comunicação" para que o usuário seja tratado de acordo com o seu perfil e seus interesses".

> "Estratégias efetivas de personalização demandam a aplicação criteriosa e objetiva de técnicas de descoberta do conhecimento e mineração de dados, determinando padrões de comportamento a partir de variadas fontes de dados transformando esses padrões em serviços personalizados que resultem em aumento de lucratividade ou eficácia desses serviços" (MEIRA JR et al. 2002, p. 179).

Sendo assim, aplicando-se *data mining* é possível uma personalização onde os interesses e necessidades dos usuários serão consideradas. A personalização de conteúdo pode ser utilizada em sistemas de recuperação e disseminação de informações para facilitar o acesso às informações desejadas pelos usuários. Pois segundo Meira JR et al. (2002, p. 179), "a personalização não afeta a semântica dos serviços pois os aspectos funcionais não são alterados".

2.7 *DATA MINING* EM BIBLIOTECAS

Alguns trabalhos já foram realizados aplicando-se *data mining* em bibliotecas, como o realizado por Santos (1998) que implantou *data warehouse* em bibliotecas para dar suporte à aplicação de técnicas de *data mining.* Em seu estudo, Aranha (1999) utilizou *data mining* explorando o perfil de usuários da biblioteca Karl A. Boedecker para geração de valor para pesquisadores por meio de cooperação indireta. Em um segundo trabalho, Aranha (2000) utilizou *data mining* para análise de redes em procedimentos de cooperação indireta para utilização no sistema de recomendações da biblioteca Karl A. Boedecker.

2.8 SERVIÇOS DE REFERÊNCIA NAS BIBLIOTECAS

Podemos observar uma gradativa evolução dos serviços de referência nas bibliotecas, da disponibilidade dos serviços de forma eletrônica e a evolução de sistemas WEB. Diante da crescente quantidade de informações a cada dia, cresce a necessidade do auxílio desse serviço na recuperação de informações pertinentes ao usuário. Estudos sobre as mudanças atuais que vêm ocorrendo nos serviços de referência, geralmente apontam o fator tecnológico como a alavanca para essa nova postura. A exemplo disso, pode-se destacar a pesquisa de Rieh (1999)

que relata os novos meios de fornecer informação aos usuários de bibliotecas sob o ponto de vista principalmente da inovação tecnológica. Enfoca a pesquisa baseada na opinião dos gerentes de bibliotecas, e destas como um todo e dos usuários. Na opinião dos gerentes constatou-se que deve haver treinamento sistemático e educação continuada, tanto do pessoal do atendimento, quanto dos bibliotecários, para que sejam plenamente utilizados os recursos eletrônicos oferecidos atualmente. Na perspectiva das bibliotecas, algumas consideraram que as bases de dados eletrônicas possibilitaram que os produtos fossem oferecidos de forma mais rápida e, em vista disso, já se julgaram satisfeitas. Outras têm sentido o estresse tecnológico, mas ao mesmo tempo acreditam que isso vem otimizar, entusiasmar e revitalizar os serviços.

Segundo Funaro et. al. (2000) para que as bibliotecas possam satisfazer as necessidades de seus clientes é indispensável que a postura do gerente de biblioteca seja pró-ativa. Estudos de usuários, perfis, uso de tecnologias avançadas, aprimoramento profissional e educação continuada são os temas que aparecerão ainda mais na era moderna. O oferecimento de serviços personalizados, individuais, ou em grupos parece que será cada vez maior e a agregação de valor é indispensável para se manter a qualidade e a competitividade dos produtos e serviços oferecidos.

Os principais serviços de referência são: a recuperação de informações através dos sistemas de recuperação de informações - SRI e disseminação seletiva de informações – DSI, auxiliando os usuários no seu processo de pesquisa a fim de encontrar informações pertinentes as seus estudos e trabalhos. Como a proposta deste trabalho é auxiliar nos processos de recuperação e disseminação das informações estes, são descritos nas subseções a seguir.

2.8.1 Sistemas de Recuperação da Informação - SRI

"Recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados geralmente textos" (CARDOSO, 2000, p. 1).

Um sistema de recuperação da informação pode ser estruturado conforme a Figura 2:

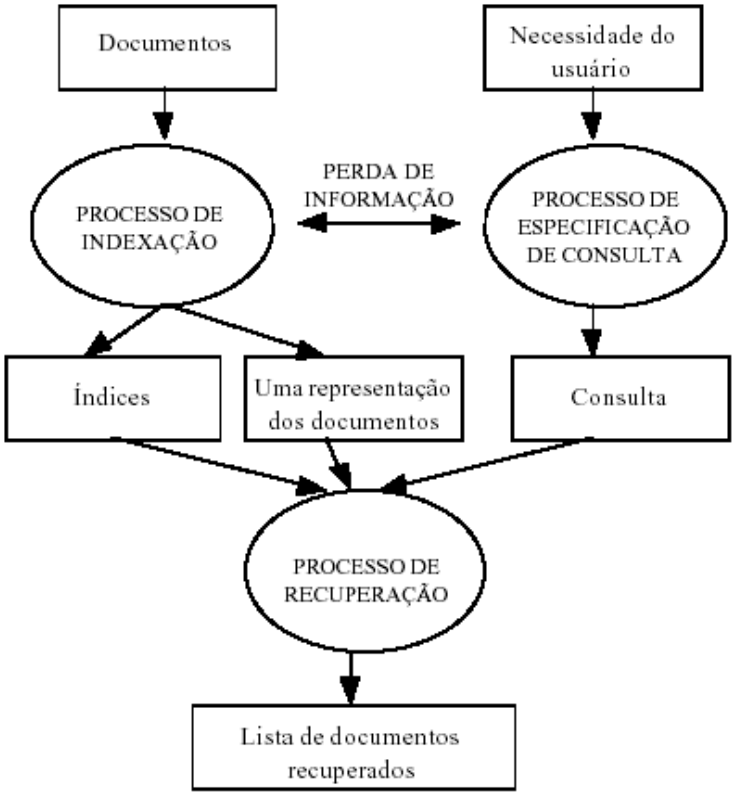

Figura 2 - Componentes de um sistema de recuperação de informações

Fonte: CARDOSO, 2000, p. 1

Os SRI surgiram com o objetivo de minimizar os problemas de recuperação de informação relevante, sendo que a maioria dos sistemas convencionais permite ao usuário encontrar a informação desejada através de uma consulta baseada em palavras. Dessa forma, se um documento possui as palavras que o usuário solicitou, ele é recuperado. Porém, esses SRI geralmente não utilizam técnicas de cálculo de relevância e proximidade dos documentos com os interesses do usuário.

2.8.2 Disseminação Seletiva da Informação - DSI

A DSI compreende serviços do tipo de notificação corrente de informação bibliográfica, extraído de sumários de periódicos ou de índices, constituindo-se em meios eficientes para manter os usuários da biblioteca informados através do perfil de interesse, em sua área de atuação.

Alguns autores como Luhn (1961) entendem a DSI como o "serviço dentro de uma organização que se refere à canalização de novos itens de informação, vindos de quaisquer fontes para aqueles pontos dentro da organização, onde a probabilidade de utilização, em conexão com interesses ou trabalhos carentes, é grande". Nesse aspecto entende-se que essa colocação ainda é muito pertinente, e explora adequadamente a priorização de interesses para um grupo/usuário individual, além de ressaltar a importância da biblioteca nesse contexto. Para o autor, o processo da DSI envolve uma série de fatores que contribuem para que o usuário gaste menos tempo com o exame e a seleção da literatura corrente. Esses processos são categorizados por: coleta da informação produzida, indexação dessa informação e divulgação aos usuários.

Segundo Mondschein (1990), a DSI é um serviço personalizado e atualizado, direcionado a um usuário, fornecendo-lhe listas de publicações mais recentes. Para ele, o que distingue a DSI de outros serviços de alerta é o desenvolvimento do perfil do usuário que pode ser prontamente alterado com a mudança da necessidade de informação. Nos anos 80, a DSI continuou a ser usada pelos cientistas que requisitavam este sistema, mas, havia ainda ajustes a serem realizados quanto às formas de indexação para garantir a relevância da recuperação da informação.

Pode-se considerar, então, que o serviço de DSI continua tendo aceitação pelos pesquisadores mudando, atualmente, não a essência do serviço em si, mas sim, a forma de oferecimento.

2.8.3 Classificação Decimal Dewey - CDD

Os sistemas de classificação bibliográfica foram elaborados com o objetivo de organizar os acervos de bibliotecas e facilitar o acesso dos usuários à informação contida nesses acervos.

Classificar é dividir em grupos ou classes, segundo as diferenças ou semelhanças. É dispor os conceitos, segundo suas semelhanças e diferenças em um certo número de grupos metodicamente distribuídos. É um processo habitual do homem, pois vivemos automaticamente classificando coisas e idéias, a fim de comprendê-las e conhecê-las.

Segundo a OCLC (2004) a classificação bibliográfica é uma linguagem de indexação, cuja função principal é organizar o conhecimento registrado em livros ou em outros documentos para possibilitar:

a) a ordenação dos documentos nas estantes;

b) a ordenação das referências nas bibliografias ou das entradas de assunto nos catálogos.

A primeira classificação bibliográfica importante de caráter universal foi a *Decimal Classification de Melvil Dewey*. Foi o primeiro sistema a utilizar números arábicos decimais simbolizando os assuntos. Uma idéia simples mas revolucionária na época (CARVALHO, 2002).

A Classificação Decimal de Dewey - CDD, é um instrumento de organização do conhecimento continuamente revisto para se manter atualizado. O sistema foi idealizado por

Melvil Dewey em 1873 e publicado pela primeira vez em 1876. É um dos sistemas bibliográficos mais utilizados em todo mundo (OCLC, 2004)

Melvil Dewey, em seu sistema de classificação, apresentou todos os conhecimentos humanos, divididos em 10 grandes grupos, numerados de 0 a 9, sendo que o grupo 0, abrange material miscelâneo ou muito geral para ser incluído em qualquer dos demais grupos (CARVALHO, 2002).

Usando uma base de três algarismos, estabeleceu as classes fundamentais apresentadas na Tabela 1.

Tabela 1– Classificação Decimal Dewey – CDD

Classificação	Descrição
000	Obras Gerais
100	Filosofia
200	Religião
300	Ciências Sociais
400	Filologia
500	Ciências Puras
600	Ciências Aplicadas
700	Belas-Artes
800	Literatura
900	História

Cada classe é por sua vez dividida, sucessivamente, em grupos numerados de 1 a 9. A partir do terceiro algarismo, o sistema passa a desenvolver-se por meio de números decimais, continuando a ser subdividido por nove subclasses, até onde a especialização do assunto mostrar necessário (CARVALHO, 2002). Um exemplo do funcionamento do sistema é apresentado na Tabela 2.

Tabela 2– Exemplo do funcionamento da CDD

Classificação	Descrição
500	Ciências Puras
550	Geologia
552	Petrologia
552.3	Rochas ígneas
552.311	Rochas plutônicas
552.313	Rochas vulcânicas

3 MÉTODO E TÉCNICAS DA PESQUISA

A pesquisa é quantitativa, empregando-se a técnica "*survey*" ou levantamento, com delineamento descritivo.

De acordo com a natureza dos dados, que apresentam variáveis quantitativas a respeito dos usuários e do acervo da BC da FURB, o problema pode ser classificado como pesquisa quantitativa, pois tudo pode ser quantificado, requerendo o uso de recursos e técnicas estatísticas

Utiliza a técnica "*survey"*, à medida que levanta os dados históricos armazenados em base de dados do Núcleo de Informática (NI) da FURB, quanto a professores e alunos de pós-graduação e a base de dados da BC da FURB, quanto às obras e transações de empréstimos realizadas em 2003, por estes usuários.

É descritiva, pois descreve o encontrado, relacionando variáveis. Silva e Menezes (2001, p. 100) afirmam que a pesquisa descritiva "visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis".

A pesquisa é aplicada uma vez que busca gerar novos conhecimentos, por meio da mineração de dados – *data mining*. Apóia-se em dados armazenados no banco de dados da BC da FURB, relacionando-os, em uma aplicação prática dirigida à solução do problema de pesquisa. Este problema refere à necessidade da BC da FURB otimizar o processo de recuperação e disseminação da informação.

Para tanto, busca-se a construção de um modelo para identificação do perfil do usuário, baseado na aplicação da tecnologia de *data mining* para análise de dados relativos aos registros de obras, usuários e transações de empréstimos, bem como a aplicação deste

conhecimento para o desenvolvimento de um sistema WEB de SRI e DSI personalizados dinamicamente.

Segundo Vergara (2000, p.47) “a pesquisa aplicada é fundamentalmente motivada pela necessidade de resolver problemas concretos, mais imediatos, ou não. Tem, portanto, finalidade prática, ao contrário da pesquisa pura, motivada basicamente pela curiosidade intelectual”.

3.1 POPULAÇÃO

A população foi constituída pelos usuários que possuem vínculo com a BC da FURB no segundo semestre de 2003, inseridos na categoria de professores e alunos de pós-graduação. O universo pesquisado foi de 3.906 usuários sendo composto por 821 professores divididos em 29 departamentos e por 3.085 alunos de pós-graduação divididos em 60 cursos mantidos pela FURB em 2004.

3.2 VARIÁVEIS

No banco de dados (BD) do NI, a primeira variável foi o código de identificação único do usuário na FURB. Este código identifica o usuário quanto a: nome, sexo, idade, endereço. A segunda foi a categoria do usuário que identifica o usuário quanto a sua categoria, professor ou aluno de pós-graduação. A terceira, curso/unidade identifica o aluno quanto ao curso que freqüenta, e professor quanto a unidade onde trabalha. Com estas variáveis forma-se a população da pesquisa.

O banco de dados da BC da FURB contém as transações de empréstimo realizadas. A primeira variável utilizada é código da obra, ou seja, o número da obra na biblioteca. O número de classificação do assunto da obra, na classificação decimal dewey (CDD), é a segunda variável, assim permitindo identificar o uso de itens do usuário.

Relaciona-se a primeira variável do NI com a primeira da BC, baseando-se nas transações realizadas, assim gerando uma tabela com os usuários e suas transações de obras.

3.3 CARACTERIZAÇÃO DA BIBLIOTECA CENTRAL DA FURB

A Biblioteca Central Martinho Cardoso da Veiga é um órgão suplementar da Universidade Regional de Blumenau.

Sua missão é desenvolver e colocar à disposição da comunidade universitária um acervo bibliográfico que atenda as necessidades de informação para as atividades de ensino, pesquisa e extensão, adotando modernas tecnologias para o tratamento, recuperação e transferência da informação (ACCETTA, 1998).

Está aberta à comunidade em geral para consultas e permite o empréstimo domiciliar aos usuários vinculados à Instituição, ou seja, corpo docente, discente, e técnico administrativo da FURB.

A Biblioteca Central começou seu processo de automação em 1987, quando através de um convênio com a FGV – Fundação Getúlio Vargas, passou a usar o Sistema Bibliodata Calco. Desde então percorre um caminho de inovações e melhorias no seu processo de automação. O sistema de automação da Biblioteca Central foi utilizado por outras bibliotecas das universidades do estado. Foi uma das primeiras bibliotecas a disponibilizar consulta do seu acervo pela Internet (ACCETTA, 1998).

Os serviços oferecidos atualmente aos usuários da Biblioteca Central são consulta ao acervo tanto na rede local ou via Internet. O usuário pode efetuar reservas, verificar livros em espera ou renová-los sem precisar vir até a seção de empréstimo.

O processo de automação da Biblioteca Central está sempre buscando oferecer a informação de forma rápida e de fácil acesso ao seu usuário. Desse modo, visa oferecer novos

serviços de recuperação e disseminação de informações personalizadas dinamicamente ao usuário.

3.4 MODELO PROPOSTO PARA APLICAÇÃO DE *DATA MINING* EM BIBLIOTECAS

Para o processo de extração de conhecimento nos dados da biblioteca sobre o perfil dos usuários, utiliza-se a metodologia referenciada por Berry & Linoff (1997). A Figura 3 apresenta o modelo proposto de metodologia.

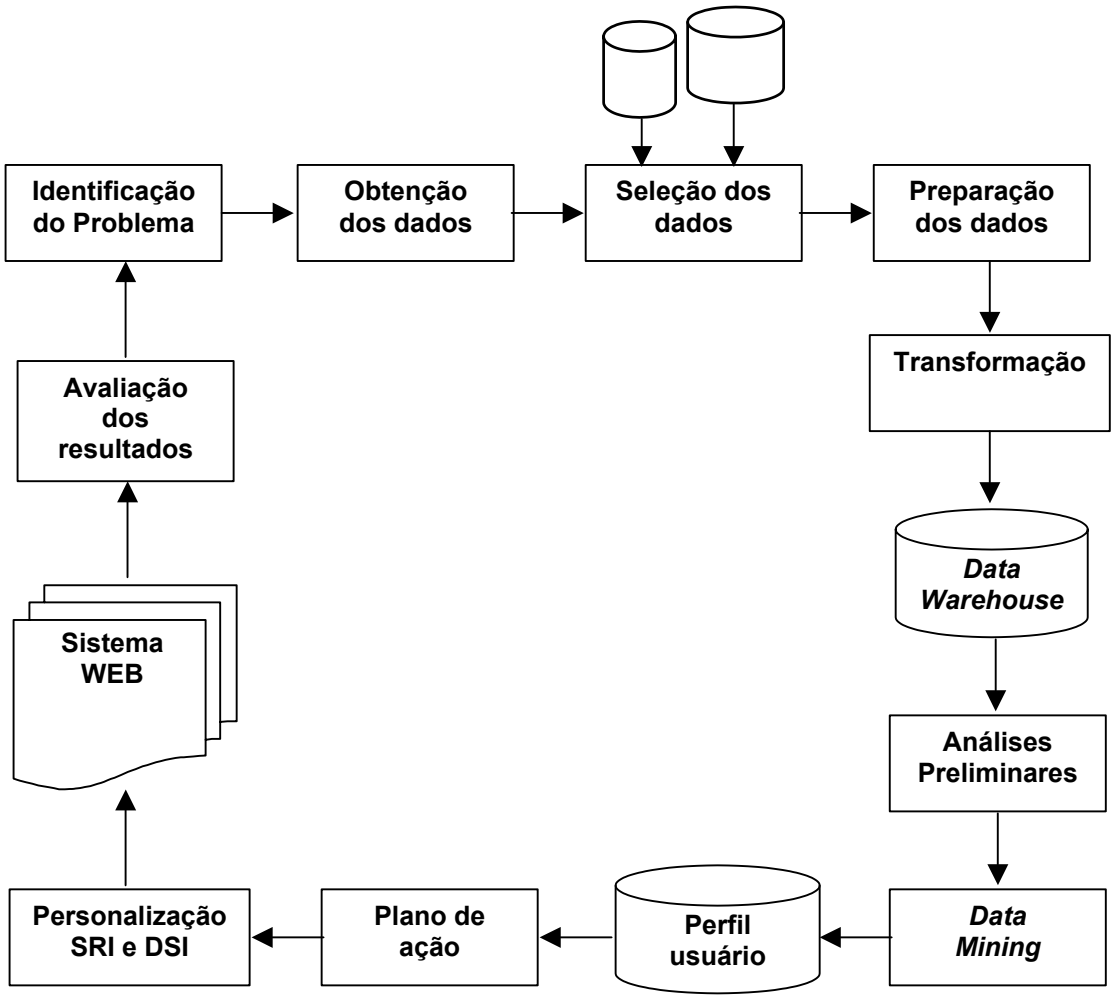

Figura 3 – Modelo do processo de KDD para biblioteca, adaptado de BERRY & LINOFF, 1997

A seguir são descritas as etapas do modelo:

3.4.1 Identificação do problema

A investigação deve ser construída com o objetivo de resolver/ou esclarecer um problema. Este, por sua vez, deve estar diretamente relacionado a um objeto de estudo.

O problema é o ponto de partida da pesquisa. Da sua formulação e delimitação determinará o desenvolvimento da investigação.

3.4.2 Obtenção dos dados

Os dados serão obtidos junto à Biblioteca Central e ao Núcleo de Informática da FURB. Os mesmos estão armazenados de forma relacional no banco de dados *Oracle* versão *8i.*

Os dados serão cuidadosamente analisados a fim de conhecê-los. Será estudado o processo de entrada dos registros de obras, transações de empréstimos e como são organizados em tabelas no banco de dados. Nesta etapa será feita uma engenharia reversa da base, criando um modelo de dados.

3.4.3 Seleção dos dados

Com a integração das bases, excluem-se as variáveis fora da área de pesquisa, por serem usadas com finalidades operacionais ou que não se aplicam ao público alvo.

O conjunto de variáveis selecionadas permitirá extrair uma variedade de informações e revelar o perfil do usuário.

3.4.4 Pré-processamento dos dados

Devem ser retiradas as inconsistências, valores não especificados e erros.

3.4.5 Extração, transformação e carga dos dados

Nesta fase os dados serão estruturados para facilitar e agilizar o processo de mineração. Desta forma, pode ser gerado um *data warehouse* ou *data mart,* este seria uma parte constituinte de um *data warehouse*. Para tanto, devem ser realizadas as seguintes tarefas:

a) identificar a tabela principal e as tabelas secundárias;

b) identificar os códigos chaves que permitirão os relacionamentos entre as tabelas principal e secundárias;

c) criar relacionamentos entre a tabela principal e as tabelas secundárias a partir dos códigos chave.

Desenvolver um modelo dimensional contemplando as variáveis de interesse para aplicação do *data mining*. Faz-se necessário desenvolver rotinas para fazer a carga dos dados.

3.4.6 Análises preliminares

Em qualquer investigação é fundamental para o pesquisador ter uma visão global dos dados. Para isso recomenda-se que sejam feitas análises estatísticas preliminares.

A utilização de representações gráficas também auxilia na interpretação dos dados. Neste caso, deve-se ter o cuidado de escolher o tipo de representação mais adequado para cada tipo de dado em questão, de acordo com a sua natureza.

Serão realizadas análises descritivas a respeito dos usuários, obras e transações entre eles.

Antes de iniciar o processo de *data mining* é importante fazer a reavaliação da seleção das variáveis. A nova avaliação será feita em função dos resultados obtidos a partir

das análises estatísticas realizadas. Posteriormente deverá ser introduzido o conhecimento de especialistas e só então deverá ser iniciado o processo de descoberta de conhecimentos.

3.4.7 *Data Mining*

Caracteriza-se pela transformação dos dados tratados em conhecimento. Para tanto é empregada uma tecnologia conhecida como *data mining*, com a finalidade de realizar a exploração e a análise dos dados por meio automático ou semi-automático, em busca de relacionamentos entre dados, padrões, regras que caracterizam tendências.

Devem ser selecionadas as tarefas a serem realizadas, bem como as técnicas mais apropriadas para a extração de conhecimento, através de uma avaliação e seleção daquelas que apresentem o melhor resultado e que sejam compatíveis na utilização. Após a escolha, opta-se por ferramentas que possibilitem a utilização desta técnica, ou será necessário desenvolver programas que suportem a mesma.

3.4.8 Plano de ação

O objetivo final da aplicação de *data mining* é resolver o problema identificado, com o conhecimento obtido. Neste estudo buscou-se melhorar os sistemas de SRI e DSI, através do desenvolvimento de um sistema WEB personalizado com base no perfil do usuário.

4 APLICAÇÃO DO MODELO PROPOSTO DE *DATA MINING* NA BIBLIOTECA CENTRAL DA FURB

Este capítulo trata especificamente da aplicação do modelo proposto de KDD para bibliotecas. É explanado todo o processo de execução, utilizando-se da apresentação dos passos realizados a fim de facilitar a compreensão, bem como conduzir aqueles que pretendam adotar o modelo.

4.1 IDENTIFICAÇÃO DO PROBLEMA

Conforme o problema definido no item 1.1, que identifica a dificuldade da recuperação de informações, elegem-se as possíveis variáveis que serão utilizadas na investigação para a resolução do mesmo. As possíveis variáveis para identificação do perfil dos usuários são:

a) usuários (dados cadastrais), categoria, curso;

b) obras (dados cadastrais), classificação CDD;

d) transações (empréstimos, reservas).

Objetiva-se, identificar o perfil dos usuários da Biblioteca Central da FURB, através da análise do conjunto de variáveis de interesse, tendo como fonte de dados os sistemas da biblioteca. Para isso aplicam-se técnicas de *data mining*, possibilitanto ampliar o conhecimento sobre o usuário e sobre o acervo.

Com os conhecimentos obtidos, será possível aplicar técnicas de personalização de conteúdo. Com isso elabora-se um sistema WEB personalizado dinamicamente, para recuperação e disseminação de informações.

4.2 OBTENÇÃO DOS DADOS

Mediante a identificação das possíveis 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 junto à Biblioteca Central da Universidade Regional de Blumenau.

As principais fontes dos dados são os sistemas legados da BC, mantidos pela seção de automação da biblioteca, que controlam o cadastro e a circulação de todo o material, bem como o sistema de identificação única das pessoas com vínculo na Universidade, mantido pelo NI, que controla todos os usuários ativos na instituição. Cabe ressaltar que a biblioteca não interfere no cadastro do usuário, somente busca os usuários e seus dados no sistema de identificação único.

4.2.1 Reconhecimento das variáveis

A fim de entender e conhecer os dados foi estudado como os mesmos são cadastrados e armazenados, e qual a importância de cada variável para a identificação do perfil do usuário.

Fez-se uma engenharia reversa do banco de dados do Núcleo de Informática e Biblioteca Central com as tabelas que continham as variáveis de interesse para a pesquisa, bem como um estudo dos dados das mesmas para o reconhecimento dos tipos de dados e sua organização. A Figura 4 apresenta o modelo entidade relacionamento dos dados de usuários armazenados no banco de dados do NI e a Figura 5 apresenta o modelo entidade relacionamento dos dados de obras e circulação armazenados no banco de dados da BC.

Abaixo são descritas as tabelas do NI, apresentadas na Figura 4:

a) PESSOA: tabela que controla os usuários da universidade, contém os dados cadastrais mantidos pelo sistema de identificação única;

b) PESSOA_VINCULO_INSTITUICAO: visão que determina os usuários ativos, gerada através de várias tabelas do sistema acadêmico e de recursos humanos que controlam os vínculos da pessoa com a universidade.

Figura 4 – Modelo entidade relacionamento banco de dados NI

Abaixo são descritas as tabelas da BC, apresentadas na Figura 5:

a) ACERVO_DADOS_MFN: tabela principal da biblioteca, que armazena os dados de todo o acervo; As informações estão estruturadas de acordo com o Formato USMARC (Catalogação legível por computador – Padrão Norte-Americano) que é o padrão adotado na Rede Bibliodata-Calco, da qual a Biblioteca participa desde 1988;

b) CIRCULACAO_HISTORICO: contêm as transações de empréstimos e reservas realizadas pelos usuários contidos na visão PESSOA_VINCULO_INSTITUICAO e as obras contidas na tabela ACERVO_DADOS_MFN;

c) BIBLIOTECA_DEPOSITARIA: descreve a biblioteca onde foi realizado empréstimo;

d) CIRCULACAO_TIPO_MOVIMENTO: descreve o tipo de movimento de circulação.

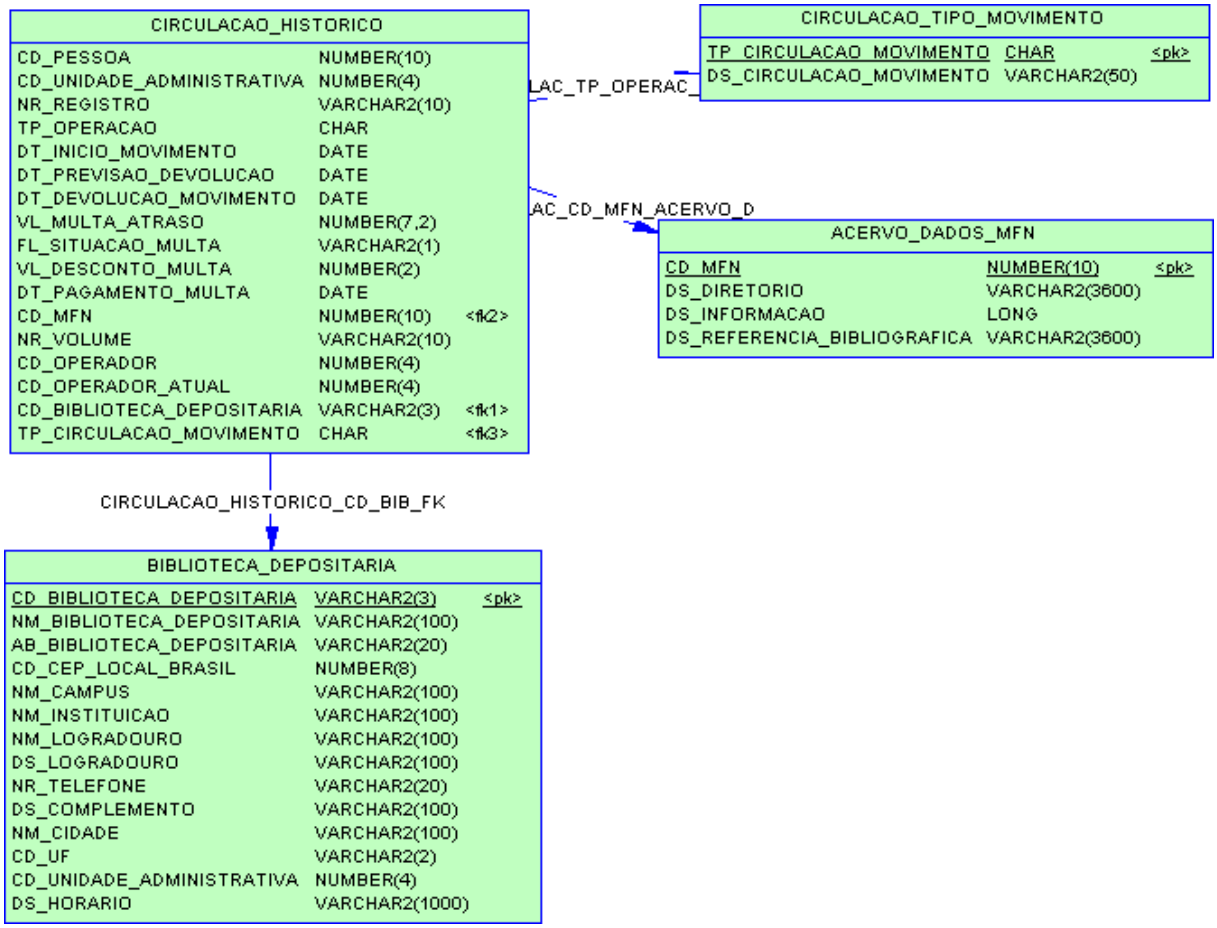

Figura 5 - Modelo entidade relacionamento banco de dados BC

4.3 SELEÇÃO DOS DADOS

Após a obtenção dos dados relativos às variáveis pré-definidas, foram selecionadas todas as variáveis de interesse e descartadas as demais.

As tabelas 1 e 2 apresentam as variáveis selecionadas do sistema de identificação única a respeito do usuário.

Tabela 3– Tabela PESSOA

Nome do campo	Descrição
CD_PESSOA	Código do usuário no Sistema de Identificação
NM_PESSOA	Nome do usuário
DT_NASCIMENTO	Data de nascimento
VL_IDADE	Idade
CD_SEXO	Sexo do usuário
NM_LOGRADOURO	Nome do logradouro
NR_LOGRADOURO	Número do logradouro
DS_COMPLEMENTO_LOGRADOURO	Complemento do logradouro
NM_BAIRRO	Nome do bairro
CD_CEP_LOCAL_BRASIL	CEP
NM_LOCALIDADE	Cidade
CD_UF	Estado
NR_TELEFONE_RESIDENCIAL	Telefone residencial
NR_TELEFONE_CELULAR	Telefone comercial
DS_EMAIL	E-mail

Tabela 4 – Tabela PESSOA_VINCULO_INSTITUICAO

Nome do campo	Descrição
CD_PESSOA	Código do usuário no Sistema de Identificação
DS_CATEGORIA	Categoria
NM_CURSO_UNIDADE	Curso ou departamento

As tabelas 3 a 6 apresentam as variáveis selecionadas dos dados obtidos junto a Biblioteca Central, contemplando as obras e as transações dos usuários.

Tabela 5 – Tabela ACERVO_DADOS_MFN

Nome do campo	Descrição
CD_MFN	Código da obra
DS_DIRETORIO	Contém os parágrafos da obra, sua localização dentro do campo DS_INFORMACAO e o tamanho da informação
DS_INFORMACAO	Contém todas as informações da obra
DS_REFERENCIA_BIBLIOGRAFICA	Referência bibliográfica da obra

Tabela 6 – Tabela CIRCULACAO_HISTORICO

Nome do campo	Descrição
CD_PESSOA	Código do usuário na Biblioteca
NR_REGISTRO	Número do registro emprestado
TP_OPERACAO	Tipo de Operação (Reserva, Empréstimo)
DT_INICIO_MOVIMENTO	Data início da operação
DT_DEVOLUCAO_MOVIMENTO	Data fim da operação
CD_MFN	Código da obra emprestada
NR_VOLUME	Número do volume da obra
CD_OPERADOR	Código do operador que iniciou a transação
CD_OPERADOR_ATUAL	Código do operador que finalizou a transação
CD_BIBLIOTECA_DEPOSITARIA	Biblioteca depositária da obra emprestada

Tabela 7 – Tabela BIBLIOTECA_DEPOSITARIA

Nome do campo	Descrição
CD_BIBLIOTECA_DEPOSITARIA	Código da biblioteca depositária
NM_BIBLIOTECA_DEPOSITARIA	Nome da biblioteca depositária

Tabela 8 – Tabela CIRCULACAO_TIPO_MOVIMENTO

Nome do campo	Descrição
TP_CIRCULACAO_MOVIMENTO	Tipo de circulação
DS_CIRCULACAO_MOVIMENTO	Descrição do tipo de circulação

4.4 PRÉ-PROCESSAMENTO DOS DADOS

Após a seleção dos dados, foi possível fazer a verificação da existência de inconsistências e erros e os mesmos foram corrigidos, conforme descrito a seguir:

a) formato do campo data de aquisição na tabela ACERVO_DADOS_MFN: o mesmo em algumas obras se apresentava em um formato diferenciado do determinado que seria dia/mês/ano apresentava-se em ano/mês/dia. As datas incorretas foram corrigidas;

b) descrição errada do campo classificação CDD na tabela ACERVO_DADOS_MFN: o campo em alguns registros não se encontrava no formato padrão.

4.5 EXTRAÇÃO, TRANSFORMAÇÃO E CARGA DOS DADOS

Para dar suporte à aplicação das técnicas de *data mining* definiu-se um modelo dimensional, gerando um *data warehouse* a partir das variáveis de interesse identificadas nas tabelas 1 até 6, chegando-se a um modelo que trata da circulação.

4.5.1 Definição do *data warehouse*

A tabela fato é a tabela de circulação, onde cada registro corresponde a uma transação que pode ser dos tipos: empréstimo e reserva. As dimensões encontradas tratam-se de informações que estão em volta de um fato de circulação, e são elas: a obra, o usuário que fez o empréstimo e o local. O modelo dimensional é apresentado na Figura 6.

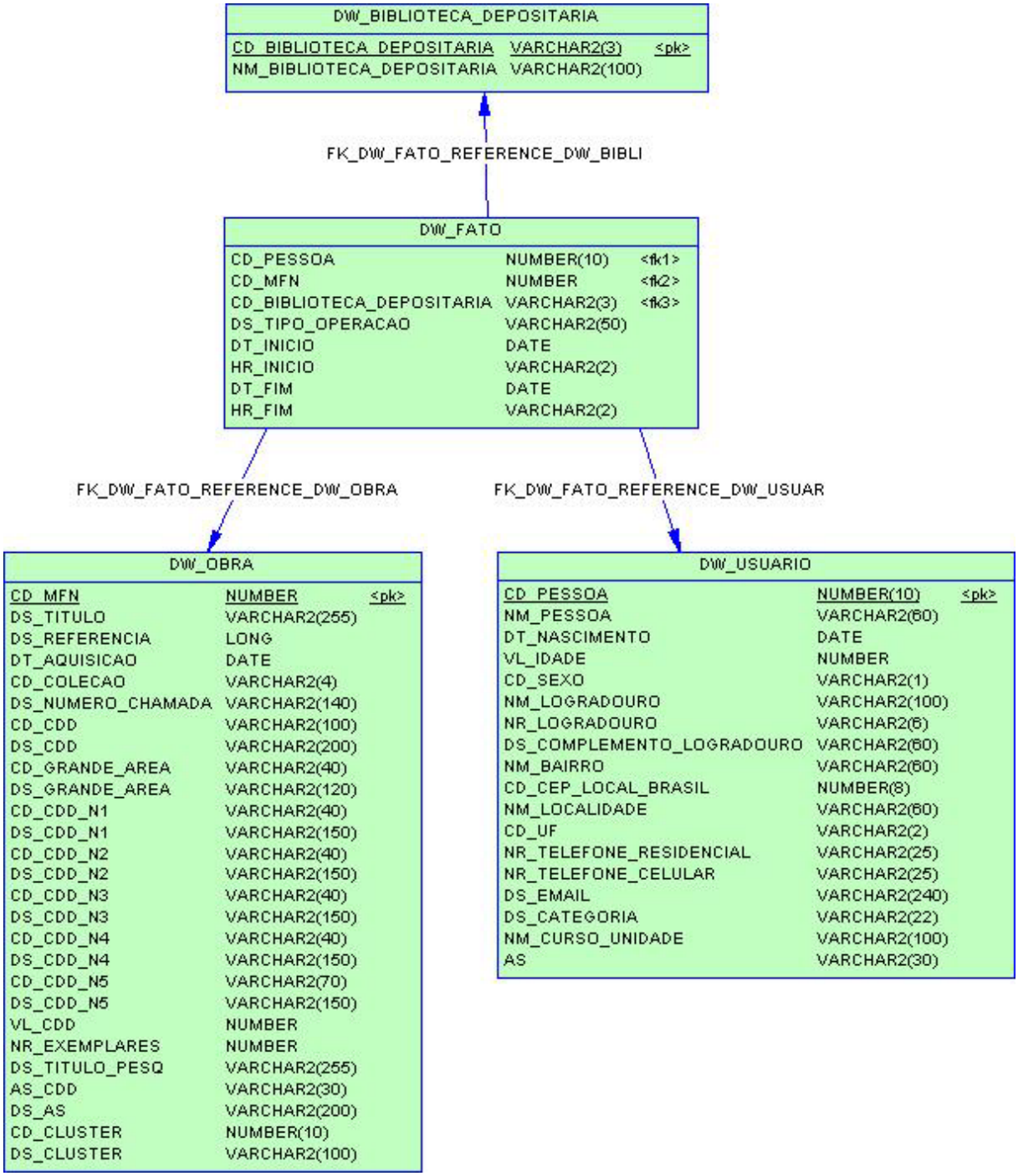

Figura 6 – Modelo *data warehouse*

A partir do modelo de *data warehouse* desenvolvido foram geradas as tabelas e rotinas para carga dos dados. As tabelas geradas a partir do modelo são descritas a seguir:

Tabela 9 – DW_FATO

Nome do campo	Descrição
CD_PESSOA	Código do usuário no Sistema de Identificação
CD_MFN	Código da obra
DS_TIPO_OPERACAO	Tipo de operação
DT_INICIO	Data de início
HR_INICIO	Hora de início
DT_FIM	Data de fim
HR_FIM	Hora de fim
CD_BIBLIOTECA_DEPOSITARIA	Biblioteca depositária

A tabela de fatos (Tabela 9) tem como fontes os dados a tabela HISTORICO_CIRCULACAO (Figura 5).

Tabela 10 – DW_USUARIO

Nome do campo	Descrição
CD_PESSOA	Código do usuário no Sistema de Identificação
NM_PESSOA	Nome do usuário
DT_NASCIMENTO	Data de nascimento
VL_IDADE	Idade
CD_SEXO	Sexo do usuário
NM_LOGRADOURO	Nome do logradouro
NR_LOGRADOURO	Número do logradouro
DS_COMPLEMENTO_LOGRADOURO	Complemento do logradouro
NM_BAIRRO	Nome do bairro
CD_CEP_LOCAL_BRASIL	CEP
NM_LOCALIDADE	Cidade
CD_UF	Estado
NR_TELEFONE_RESIDENCIAL	Telefone residencial
NR_TELEFONE_CELULAR	Telefone comercial
DS_EMAIL	E-mail
DS_CATEGORIA	Categoria
NM_CURSO_UNIDADE	Curso ou departamento

A dimensão usuário (Tabela 10) tem como fontes os dados da tabela PESSOA (dados cadastrais) e PESSOA_VINCULO_INSITUICAO (vínculo do usuário e categoria) (Figura 4).

Tabela 11 – DW_OBRA

Nome do campo	Descrição
CD_MFN	Código da obra
DS_TITULO	Título
DS_REFERENCIA	Referência bibliográfica
DT_AQUISICAO	Data de Aquisição
CD_COLECAO	Tipo de coleção
DS_NUMERO_CHAMADA	Número de chamada
CD_CDD	Código CDD
DS_CDD	Descrição CDD
CD_GRANDE_AREA	Código grande área CDD
DS_GRANDE_AREA	Descrição grande área CDD
CD_CDD_N1	Código CDD nível 1
DS_CDD_N1	Descrição CDD nível 1
CD_CDD_N2	Código CDD nível 2
DS_CDD_N2	Descrição CDD nível 2
CD_CDD_N3	Código CDD nível 3
DS_CDD_N3	Descrição CDD nível 3
CD_CDD_N4	Código CDD nível 4
DS_CDD_N4	Descrição CDD nível 4
CD_CDD_N5	Código CDD nível 5
DS_CDD_N5	Descrição CDD nível 5
VL_CDD	Código CDD sem caracteres alfanuméricos
NR_EXEMPLARES	Número de exemplares do título
DS_TITULO_PESQ	Título para pesquisa
AS_CDD	Assunto significativo
DS_AS	Descrição do assunto significativo
CD_CLUSTER	Código do cluster
DS_CLUSTER	Descrição do cluster

A dimensão obra (Tabela 11) tem como fontes principais as tabelas ACERVO_DADOS_MFN e CIRCULACAO_TIPO_MOVIMENTO (Figura 5). Através das rotinas desenvolvidas para carga dos dados foi possível extrair as informações de interesse.

4.5.2 Verificação e/ou criação das codificações

No caso de existirem codificações para os domínios de uma variável é importante que sejam identificadas e entendidas. E no caso de não existirem deverão ser criadas, pois somente neste formato os dados poderão ser trabalhados posteriormente na extração de conhecimentos com a utilização de tecnologias do tipo *data mining*.

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 CDD – Classificação Dewey, que é usada por diferentes órgãos da área de biblioteconomia no preenchimento de seus respectivos registros que apresentem o campo "CDD".

Exemplos:

- 005.756 é a codificação para o assunto Banco de Dados

- 005.133 SQL é a codificação para o assunto Linguagem de consulta SQL

Esta codificação serve basicamente para facilitar os procedimentos uma vez que consistem numa padronização.

Uma outra grande vantagem de se utilizar uma mesma codificação em registros diferentes é que, a partir deles será possível fazer o cruzamento posterior em diferentes bases de dados.

É importante ressaltar aqui que, dependendo do tipo de codificação instituída será possível fazer agrupamentos posteriores dos domínios das variáveis. Caso não exista, mas seja possível realizar, estas deverão ser criadas. Assim foram criados 5 níveis CDD, armazenados em campos da tabela dimensão obra, para classificar as obras em grupos do mais genérico ao mais especifico. Exemplo:

- 005.73 – classificação da obra;
- 0 - classificação CDD nível 1;
- 00 - classificação CDD nível 2;
- 005 - classificação CDD nível 3;
- 005.7 - classificação CDD nível 4;
- 005.73 - classificação CDD nível 5;

4.5.3 Criação de assuntos significativos

Os assuntos significativos (AS) foram gerados através da totalização das transações segundo a CDD, reduzindo as mesmas até um nível mínimo de significância de 500 transações para nível 2 da CDD, 250 para o nível 3 e de 100 para o nível 4.

Na amostra de 17.421 obras existiam 3.474 áreas de CDD diferentes, das quais foram criados 131 assuntos significativos, conforme tabela contida no Apêndice 1.

A média de obras por AS foi de 132,98 e a média de transações por AS foi de 523,23. Sendo que o assunto significativo com maior número de obras foi o 700 – Artes e com maior número de transações foi 658.4 – Administração.

4.6 ANÁLISE DESCRITIVA DOS DADOS

Esta seção apresenta uma análise descritiva dos dados que foram carregados no *data warehouse*.

4.6.1 Acervo

O acervo da biblioteca analisado é composto por 17.421 títulos (representa a identificação única de uma obra) que totalizam 51.011 volumes (agrupamento de obras com

mesmo título), com uma média de 2,92 volumes por título. Os mesmos estão divididos por tipo de materiais e estes organizados em coleções conforme o tipo de empréstimo. Na tabela a seguir são descritos os materiais versus suas coleções.

Tabela 12 – Divisão acervo por material e coleções

Material			Coleção			
Descrição	**Total**	**%**	**Código**	**Descrição**	**Total**	**%**
Filmes	612	3,51%	DV	DVD	13	0,07%
			FV	Fita de Vídeo	599	3,44%
Literatura Cinzenta	882	5,06%	MO	Monografia	511	2,93%
			RP	Relatório de Pesquisa	6	0,03%
			TE	Tese	365	2,10%
Livros	15850	90,98%	CE	Coleção Especial	237	1,36%
			CG	Coleção Geral	15457	88,73%
			FF	Folheto	87	0,50%
			RF	Coleção Referência	69	0,40%
Outros	77	0,44%	CD	CD-Rom	44	0,25%
			DP	Diapositivo	16	0,09%
			DQ	Disquete	1	0,01%
			FC	Fita Cassete	1	0,01%
			MP	Mapa	8	0,05%
			PR	Partitura	7	0,04%
Total	17421	100%	Total		17421	100%

Conforme tabela anterior os grupos de materiais que se destacam são os livros com 90,98 % do acervo seguido da literatura cinzenta 5,06% do acervo. Abaixo são descritos os grupos de materiais:

a) filmes: compostos pelas Fitas de Vídeo e DVD. Totalizam 612 obras, representando 3,51% do acervo. Destacam se as Fitas de Vídeo com 3,44% do acervo;

b) literatura cinzenta: composto pelas Teses, Monografias e Relatórios Científicos totalizando 882, representado 5,06% do acervo. Destacam-se as Teses com

2,10% e as Monografias com 2,93%, as quais são as principais produções científicas da FURB;

c) livros: composto pela Coleção Especial, Coleção Geral, Coleção Referência e Folhetos. Totalizam 15.859 obras, representado 90,98% do acervo, sendo que a Coleção geral apresenta o maior número de obras 88,73%;

d) outros: composto pelas coleções de CD-Rom, Diapositivo, Disquete, Fita Cassete, Mapa, Partitura. Totalizam 77 obras, representando 0,44 % do acervo.

As bibliotecas utilizam os sistemas de classificação para organizar seus acervos de forma inteligível e sistemática (ARANHA, 1999), detalhes sobre a classificação de obras ver item 2.8.3. A BC da FURB utiliza o sistema CDD - Classificação Decimal *Dewey*. Na amostra de 17.421 obras existiam 3.474 áreas de CDD diferentes, como seria muito onerosa a análise de uma quantidade tão grande de áreas diferentes, foram feitas reduções destas áreas, conforme descrito nos itens 4.5.2 e 4.5.3.

As 3.474 áreas diferentes de CDD do acervo foram divididas em 10 grupos CDD nível 1, 96 grupos CDD nível 2, 490 grupos CDD nível 3, 1.344 grupos CDD nível 4, 2.224 grupos CDD nível 5, e 131 assuntos significativos.

Tabela 13 – Total de títulos por área CDD nível 1

CDD	Descrição	Total Obras	% Obras	Total Exemplares	Média Exemplares
000	Generalidades	919	5,3%	3098	3,37
100	Filosofia & Psicologia	1235	7,1%	2834	2,29
200	Religião	178	1,0%	304	1,71
300	Ciências Sociais	6335	36,4%	17901	2,83
400	Lingüística	501	2,9%	1687	3,37
500	Ciências Naturais & Matemática	1055	6,1%	4393	4,16
600	Ciências Aplicadas	4669	26,8%	15126	3,24
700	Artes	1154	6,6%	2866	2,48
800	Literatura	866	5,0%	1698	1,96
900	Geografia e História	509	2,9%	1104	2,17
Total		17421	100%	51011	2,93

Conforme a Tabela 13 pode-se verificar que a maior concentração de obras está nas áreas das Ciências Sociais e Aplicadas totalizando 63,2% do acervo. O Gráfico 1 demonstra os dados da Tabela 13.

Com relação à média de exemplares, verifica-se uma média alta nas áreas de concentração da Universidade e baixa nas demais, como religião e literatura.

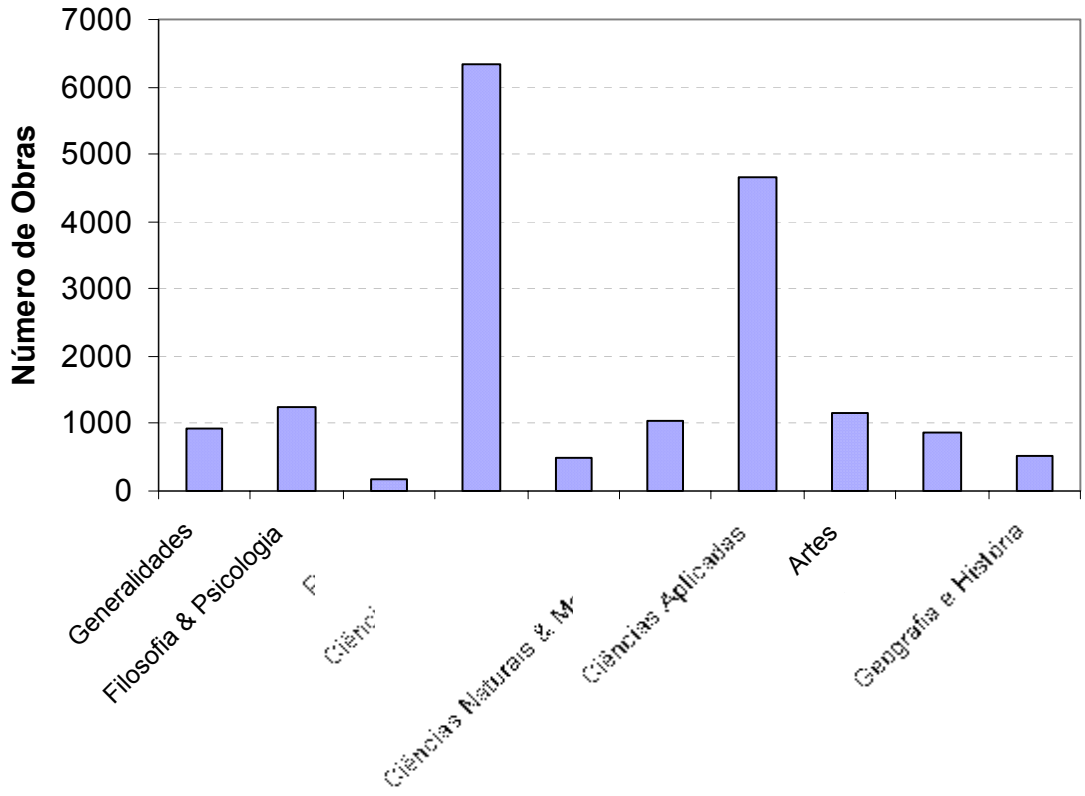

Gráfico 1 – Total de títulos por área CDD nível 1

4.6.2 Usuários

Os usuários da biblioteca são controlados através do sistema de identificação única da Universidade. A Biblioteca é de acesso livre a qualquer pessoa para consulta de material, mas o empréstimo é restrito as pessoas que possuem vínculo com a instituição, ou seja, seus corpos discentes, docentes e técnico administrativo.

Para o estudo foram secionados 3.906 usuários, sendo 3.085 alunos de pós-graduação distribuídos em 60 cursos e 821 professores distribuídos em 29 departamentos.

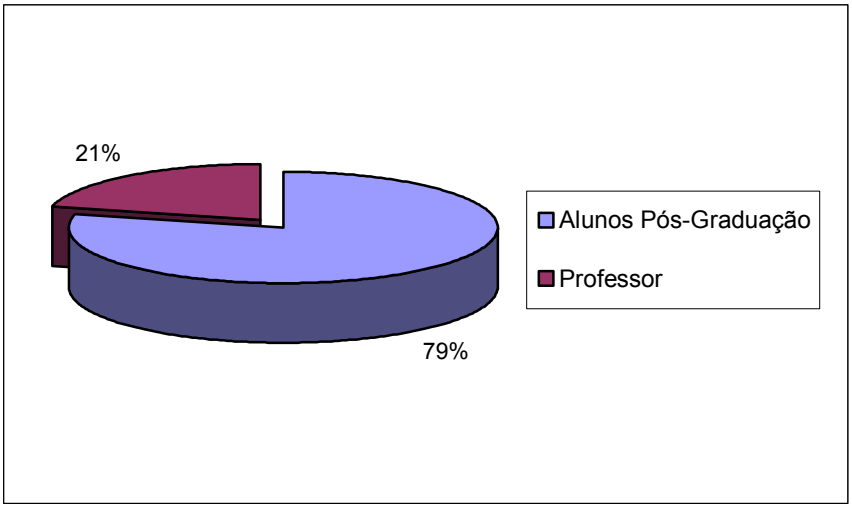

Gráfico 2 – Usuários por categoria, FURB, 2003

O período analisado constitui-se das transações efetuadas no ano de 2003. Dos 3.906 usuários, somente 1.553 realizaram movimentações de empréstimos ou reservas. Na tabela abaixo são apresentados os valores.

Tabela 14 – Usuários por categoria

	Total Usuários		Total c/ Transações		
Categoria	**(TUsu)**	**%**	**(TUsuTran)**	**%**	**TUsuTran/TUsu**
Alunos Pós-Graduação	3085	79%	932	60%	30%
Professor	821	21%	621	40%	76%
Total Global	3906	100%	1553	100%	

Como apresentado na Tabela 14 anterior 79% dos usuários são alunos de Pós-Graduação, mas somente 30% destes realizaram transações e foram responsáveis por 60% delas. Já os Professores são somente 21% dos quais 76% realizaram transações e são responsáveis por 40% delas.

4.6.3 Transações

Foram realizadas 68.543 transações de circulação de itens pelos usuários de interesse, sendo 66.769 transações de empréstimo e 1.775 transações de reservas.

Tabela 15 –Transações por categoria usuário

Categoria	Total de Transações	Total de Usuários	%	Média
Alunos Pós-Graduação	31746	932	60%	34
Professor	36797	621	40%	59
Total Global	68543	1553	100%	44

De acordo com a Tabela 15, a média de transações realizadas pelos usuários é de 44 livros, sendo que para professores a média é de 59 obras por usuário, enquanto que para os alunos de pós-graduação a média é de 34 obras por usuário.

Tabela 16 – Transações por áreas CDD nível 1

CDD	Descrição	Total Exemplares	Empréstimos			Reservas		
			Total	%	Média	Total	%	Média
000	Generalidades	3098	4185	6,3%	1,35	163	9,19%	0,05
100	Filosofia & Psicologia	2834	3873	5,8%	1,37	147	8,29%	0,05
200	Religião	304	482	0,7%	1,59	17	0,96%	0,06
300	Ciências Sociais	17901	24396	36,5%	1,36	674	37,99%	0,04
400	Lingüística	1687	2270	3,4%	1,35	37	2,09%	0,02
500	Ciências Naturais & Matemática	4393	4742	7,1%	1,08	64	3,61%	0,01
600	Ciências Aplicadas	15126	19250	28,8%	1,27	448	25,25%	0,03
700	Artes	2866	3590	5,4%	1,25	41	2,31%	0,01
800	Literatura	1698	2450	3,7%	1,44	156	8,79%	0,09
900	Geografia e História	1104	1531	2,3%	1,39	27	1,52%	0,02
Total		51011	66769	100%	1,31	1774	100%	0,03

De acordo com a Tabela 16, tem-se uma maior freqüência de empréstimo na área de Ciências Sociais com 36,5%, seguidos de Ciências Aplicadas com 28,8%. Como comentado anteriormente, estas são as principais áreas de concentração dos cursos da Universidade.

Quanto à taxa de utilização do acervo temos em média 1,31 empréstimos por exemplar, o que podemos considerar uma taxa baixa. Tem-se uma maior utilização do acervo nas áreas 200 (Religião) e 800 (Literatura), devido à quantidade inferior de material, o que explica as maiores médias de reservas nestas áreas.

Para cada transação associa-se uma data e hora, como se podem verificar nos gráficos 3, 4 e 5 a seguir:

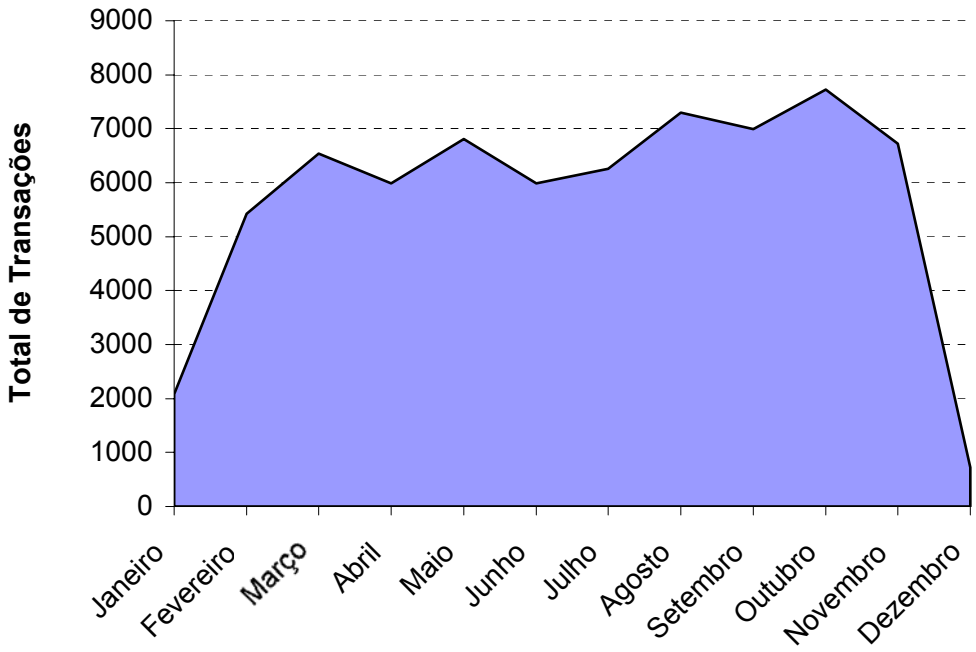

Gráfico 3 –Total de transações por mês

Nas transações por meses (Gráfico 3) pode-se observar um baixo número de empréstimos nos meses de janeiro e dezembro os quais são os meses de recesso. Sempre ao início e final do semestre verifica-se uma maior utilização do material.

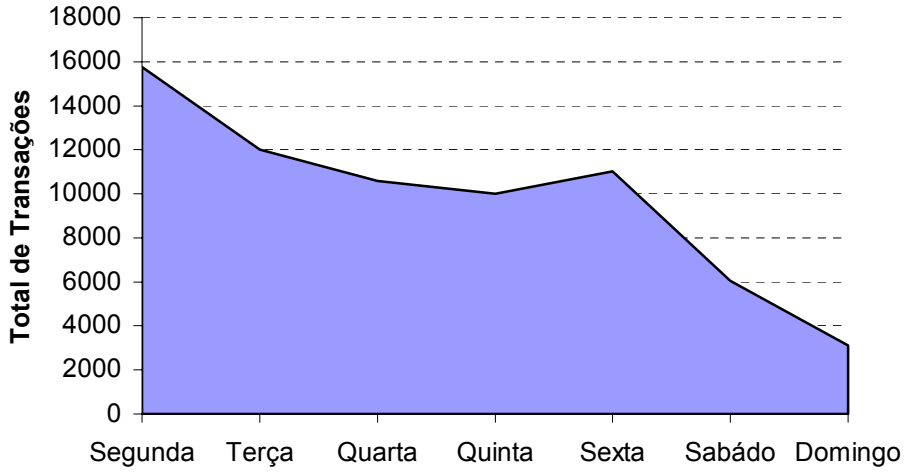

Gráfico 4 – Total de transações por dia da semana

Quanto às transações por dia da semana (Gráfico 4) observa-se uma maior freqüência no início da semana, e uma queda para o final da mesma. Tem-se um aumento na sexta devido aos cursos ministrados somente ao final de semana.

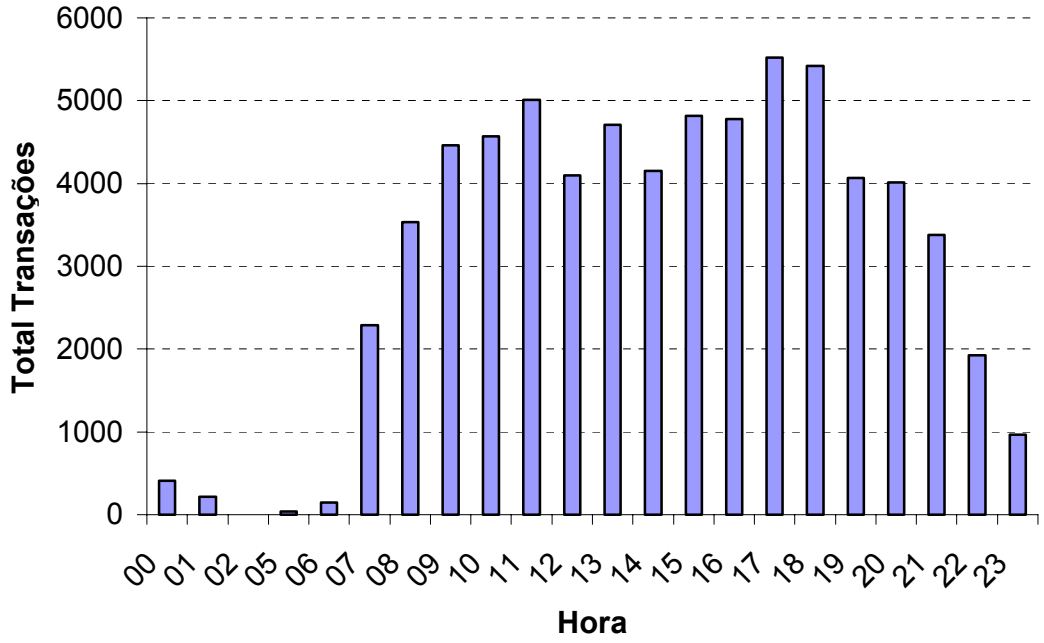

Gráfico 5 – Total de transações por hora

Quanto às transações por hora (Gráfico 5) verifica-se que o horário de pico é ao final da tarde às 17:00 e 18:00 horas, quando iniciam as aulas do período noturno. As transações realizadas após as 22:00 horas até as 7:00 são realizadas pelo usuário via Internet. Ressalta-se que as 3:00 e 4:00 horas não existem transações, pois o sistema está fora do ar para backup.

4.7 *DATA MINING*

O objetivo do trabalho é encontrar o perfil do usuário da biblioteca através de aplicações de técnicas de *data mining*.

4.7.1 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 (1999) 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".

A Tabela 17 apresenta um exemplo dos dados de entrada para a aplicação da técnica de *cluster*, através do software STATISTICA[1].

[1] http://www.statsoft.com

Tabela 17 – Tabela de exemplo de transações usuários por AS

CD_PESSOA	000	001.4	004	005	005.1	005.7	006	028.5	100	133	150	...
82	0	0	0	0	0	0	0	0	0	0	0	...
109	0	1	0	0	0	0	1	0	1	0	0	...
128	0	0	0	0	0	0	0	0	0	0	0	...
181	0	0	1	1	1	1	0	0	0	0	0	...
232	1	0	0	0	0	0	0	0	0	0	0	...
254	0	0	0	0	0	0	0	0	0	0	0	...
283	0	0	0	0	0	0	0	0	0	0	0	...
298	0	0	0	1	1	1	0	0	0	1	0	...
354	0	1	0	0	0	0	0	0	0	0	0	...

A Tabela 17 analisada contempla 1.553 usuários, os quais realizaram 68.543 transações em 131 AS, sendo apresentado uma média de 44,13 transações por AS.

A BC da FURB já possui uma tabela de grandes áreas apresentada no Anexo 1, a qual foi criada de forma manual pelos bibliotecários com base nos cursos da Universidade. Os dados obtidos por esta análise serão posteriormente comparados à tabela existente, para obtenção de um melhor resultado na criação das grandes áreas de interesse do usuário.

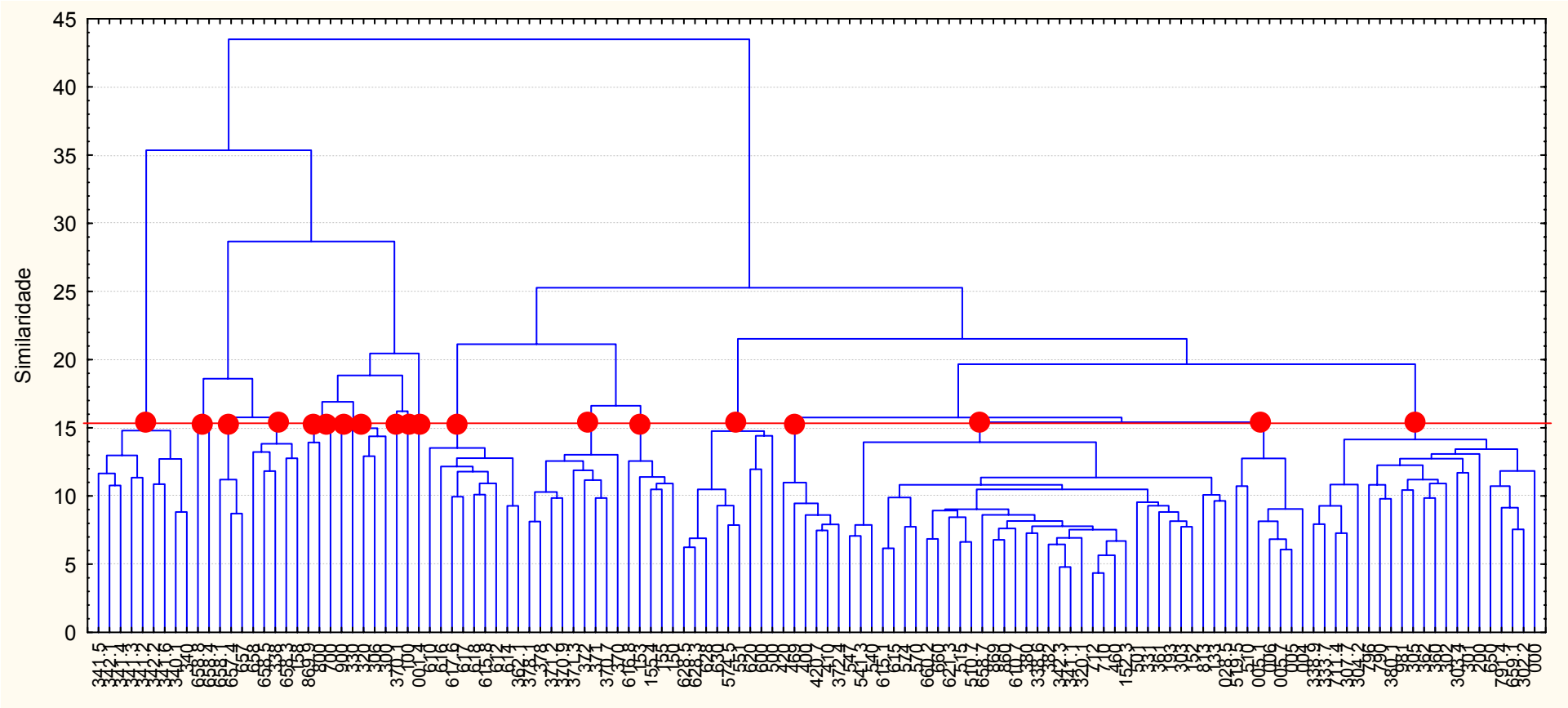

Figura 7 – Dendograma com transações dos usuários por assunto significativo

Através da análise do dendrograma apresentado na Figura 7, que identifica as transações de usuários por assunto significativo, foram definidos 19 grupos que são apresentados na tabela Apêndice 2 e descritos a seguir.

Aranha (1999) sugere a existência dos seguintes grupos na aplicação de *cluster* sobre assuntos significativos:

a) muito especializados – concentram usuários com interesse em poucos assuntos distintos;

b) pouco especializados – apresentam usuários com interesses distribuídos de forma uniforme em vários assuntos;

c) bimodais – contém usuários muito especializados e pouco especializados.

O grupo 1 encontrado é o de Ciências Jurídicas que contempla a grande área de direito, apresenta-se como um grupo especializado, sua principal área é a 340 (Direito) que é formado por 9 AS.

Os grupos 2, 3, 4 e 8 contemplam as de Ciências Sociais e Aplicadas com as respectivas áreas: Marketing, Contabilidade, Administração e Economia/Política, agrupando 14 AS. Os três primeiros são grupos especializados; sua principal área é 650. Já o grupo de Economia/Política é um grupo bimodal. No grupo 2 (Administração) é interessante destacar o AS 158 (Psicologia aplicada), este assunto relaciona-se com Sucesso, Relações Humanas.

O grupo 5 trata da Literatura, um grupo não especializado, sua principal área é a 800 (Literatura), formado por 2 AS. Este material é utilizado pelo público em geral.

Os grupos 6, 9, 13 e 16 contemplam a área de Ciências da Educação com as respectivas áreas de Artes, Filosofia da Educação, Pedagogia e Letras, que são grupos especializados formados por 17 AS.

O grupo 7 reflete a área de História e Geografia, formado somente pelo AS 900 (História,Geografia e Biografia). Apresenta-se como um grupo bem especializado.

O grupo 10 considera a Filosofia, formado somente pelo AS 100 (Filosofia).

O grupo 11 denota a Metodologia Científica, formado somente pelo AS 001.4 (Metodologia da Pesquisa). Este material é utilizado por todos os cursos de pós-graduação para elaboração de seus trabalhos de conclusão de curso.

Os grupos 12 e 14 contemplam a área de Saúde. No primeiro encontram-se as áreas de medicina, odontologia, fisioterapia e enfermagem, sua principal área é a 610. Destaca-se neste grupo o AS 362.1 (Serviço Social – Saúde Pública). O segundo grupo Psicologia tem como principal área 150; neste grupo é interessante destacar a presença do AS 616.8 (Psiquiatria). São formados por 14 AS.

O grupo 15 trata das Ciências e Tecnologia, o qual contempla áreas de Engenharia Ambiental, Ecologia, Agricultura, Tecnologia, Geologia, Ciências Puras e Aplicadas. É um grupo bimodal, formado por 14 AS.

O grupo 17 relaciona as Ciências Tecnológicas, contemplando principalmente as áreas de engenharia, mas encontram-se algumas outras áreas. Este grupo apresentou-se de forma pouco especializado e muito diversificado, além da grande quantidade de AS (35). Assim separa-se este grupo e aplica-se a técnica de *cluster* novamente, a fim de torná-lo mais especializado.

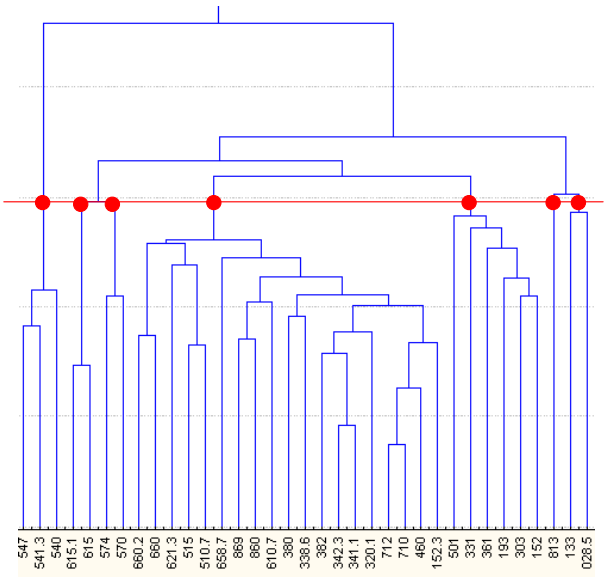

Figura 8 – Dendograma com transações dos usuários do grupo 17

Conforme Figura 8, a nova análise do grupo 17 gerou 7 novos subgrupos: Química, Farmácia, Biologia, Engenharias, Filosofia, Ficção e Outros. Com a nova divisão formaram-se grupos mais especializados, mas ainda apresenta o subgrupo de engenharias com alguns AS nada relacionados a esta área.

O grupo 18 denota as Ciências Exatas, contempla as áreas de Matemática e Tecnologia da Informação, formado por 7 AS.

O grupo 19 trata das Humanas e Comunicação. Trata-se de um grupo pouco especializado, que traz áreas distintas como Educação Física, Turismo, Religião e Ciências sociais. Formado por 20 AS, também apresenta um grande número de assuntos significativos. Assim separa-se este grupo e aplica-se a técnica de *cluster* novamente, a fim de torná-lo mais especializado.

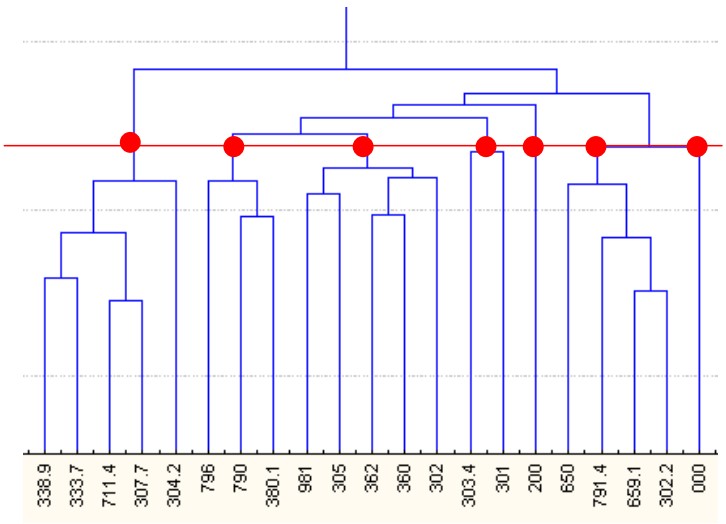

Figura 9 – Dendograma com transações dos usuários do grupo 19

Conforme Figura 9, a nova análise do grupo 19 gerou 7 novos subgrupos: Ciências Sociais, Educação Física /Turismo e Lazer, Serviço Social, Sociologia, Religião, Comunicação Social e Generalidades. Com a nova divisão os grupos tornaram-se mais especializados.

4.7.2 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. Este estudo classifica as obras do acervo da biblioteca em grandes áreas do conhecimento segundo os 34 grupos de grandes áreas identificados no processo de *cluster* apresentado no item 4.7.1 e a tabela pré-definida por especialistas da área de biblioteconomia conforme Anexo 1 a qual contém 105 grupos de grandes áreas.

4.7.2.1 Classificação do acervo em grandes áreas pela tabela *cluster*

Este processo de classificação utilizou-se da tabela constituída através da análise de *cluster*, apresentada no apêndice 2. A Tabela 18, apresenta um exemplo da mesma.

Tabela 18 – Exemplo tabela grandes áreas gerada pelo *cluster*

Descrição do Grupo			
Grupo	**Subgrupo**	**AS**	**Descrição do AS**
1	Direito	341.5	Direito penal
1	Direito	342.1	Direito civil
1	Direito	341.4	Direito processual
1	Direito	341.3	Direto administrativo
1	Direito	341.2	Direto constitucional
1	Direito	342.2	Direito comercial
1	Direito	341.6	Direito do trabalho
1	Direito	340.1	Filosofia do Direito
1	Direito	340	Direito

Conforme descrito anteriormente no processo de criação dos AS cada obra recebeu um AS. Assim, através da tabela de grandes áreas gerada através de *cluster* como o exemplo acima, os AS agrupados receberam um nome de acordo com o assunto ou área dos mesmos.

Através de uma rotina SQL os registros de obras foram classificados e o resultado foi armazenado nos campos CD_CLUSTER e DS_CLUSTER da tabela DW_OBRA (Tabela 11). A rotina para classificação pode ser vista no quadro a seguir.

```
update dw_obra do set do.CD_CLUSTER=(select CD_GRUPO from dw_cluster dc where dc.cd_as=do.AS_CDD);
update dw_obra do set do.DS_CLUSTER=(select DS_GRUPO from dw_cluster dc where dc.cd_as=do.AS_CDD);
```

Quadro 1 – Rotina SQL para classificação de obras em grandes áreas

Com a execução das rotinas descritas no Quadro 1 cada obra do acervo foi classificação em grupo pré-determinado pela análise de *cluster*. O resultado da classificação é apresentado na Tabela 19.

Tabela 19 – Resultado classificação obras pela tabela *cluster*

CD_CLUSTER	DS_CLUSTER	Total livro	% livro	Total transação	% transação
1	Direito	2143	12,3%	8819	12,9%
2	Marketing	889	5,1%	4297	6,3%
3	Contabilidade	368	2,1%	2001	2,9%
4	Administração	799	4,6%	3276	4,8%
5	Literatura	645	3,7%	1985	2,9%
6	Artes	591	3,4%	1661	2,4%
7	História	403	2,3%	1162	1,7%
8	Economia	833	4,8%	2863	4,2%
9	Filosofia da educação	406	2,3%	1773	2,6%
10	Filosofia	387	2,2%	1159	1,7%
11	Metodologia Cientifica	168	1,0%	1497	2,2%
12	Saúde	1288	7,4%	4784	7,0%
13	Pedagogia	956	5,5%	3768	5,5%
14	Psicologia	656	3,8%	2189	3,2%
15	Ciência e Tecnologia	1121	6,4%	4309	6,3%
16	Letras	518	3,0%	2395	3,5%
17	Química	201	1,2%	982	1,4%
18	Farmácia	145	0,8%	578	0,8%
19	Biologia	119	0,7%	433	0,6%
20	Engenharias	1034	5,9%	4812	7,0%
21	Filosofia	350	2,0%	1261	1,8%
22	Ficção	144	0,8%	363	0,5%
23	Outros	227	1,3%	489	0,7%
24	TI / Matemática	647	3,7%	3163	4,6%
25	Ciências Sociais	445	2,6%	2056	3,0%
26	Educação Física / Turismo e Lazer	556	3,2%	1825	2,7%
27	Serviço Social	517	3,0%	1650	2,4%
28	Sociologia	194	1,1%	826	1,2%
29	Religião	178	1,0%	499	0,7%
30	Comunicação Social	320	1,8%	1052	1,5%
31	Generalidades	173	1,0%	616	0,9%
Total		17421	100,0%	68543	100,0%

Conforme tabela anterior foram formados 31 grupos com uma média de 562 obras por grupo e 2.211 transações por grupo. Destaca-se a área de direito com 12,3% das obras e 12,9% das transações.

4.7.2.2 Classificação do acervo em grandes áreas pela tabela biblioteca

Este processo de classificação utilizou-se da tabela definida pelos bibliotecários, apresentada Anexo 1. A Tabela 20 exemplifica a classificação de obras em grandes áreas.

Tabela 20 – Exemplo da tabela de classificação de obras em grandes áreas definida por bibliotecários

Código CDD	Descrição grande área	Código grande área
000	Generalidades	000
001.4	Metodologia Cientifica	001.4
001.5	Generalidades	000
004	Processamento de Dados	004
007	Generalidades	000
020	Biblioteconomia	020
028.5	Literatura Infanto-Juvenil	028.5
028.6	Biblioteconomia	020
040	Generalidades	000

Toma-se como exemplo a área de "Processamento de Dados", interpretando a tabela tem-se o intervalo: [004–007] Processamento de Dados.

Através de uma rotina SQL, os registros de obras foram classificados e o resultado dos dados foi armazenado no campo CD_GRANDE_AREA e DS_GRANDE_AREA da tabela DW_OBRA (Tabela 11). A rotina para classificação desenvolvida pode ser vista no Quadro 2.

```
select aa.cd_area, aa.ds_area  from area_cdd aa, area_cdd ab
where to_number(replace(ab.cd_area,'.',',')) =
        (
        select max(to_number(replace(a2.cd_area,'.',',')))
        from area_cdd a2
        where to_number(replace(a2.cd_area,'.',','))
        <= to_number(replace('&CDD','.',','))
        )
and aa.cd_area = ab.ds_complemento
order by aa.cd_area, aa.ds_area
```

Quadro 2 – Rotina SQL para classificação CDD em grandes área

Na rotina SQL para classificação das obras em grandes áreas segundo sua CDD, temos como parâmetro de entrada "**&CDD"** que é a CDD a ser classificada e como retorno temos a código e descrição da grande área.

A Tabela 21 apresentada o resultado da classificação das obras segundo as grandes áreas.

Tabela 21 – Classificação das obras em grandes áreas pela tabela biblioteca

DS_GRANDE_AREA	Total livros	% livros	Total transações	% transações
Administração e Serviços Auxiliares	1347	7,7%	6774	9,9%
Administração Publica	54	0,3%	150	0,2%
Agricultura	137	0,8%	381	0,6%
Arquitetura e Urbanismo	215	1,2%	912	1,3%
Artes	492	2,8%	1371	2,0%
Astronomia	24	0,1%	78	0,1%
Biblioteconomia	34	0,2%	95	0,1%
Biografia	91	0,5%	197	0,3%
Biologia	119	0,7%	433	0,6%
Botânica	65	0,4%	141	0,2%
Ciência Política	255	1,5%	941	1,4%
Ciências Puras	77	0,4%	281	0,4%
Comercio, Comunicação e Transportes	103	0,6%	377	0,6%
Construção Civil	55	0,3%	221	0,3%
Contabilidade	210	1,2%	992	1,4%
Costumes	41	0,2%	128	0,2%
Dicionários e Enciclopédias	2	0,0%	7	0,0%
Direito	2264	13,0%	9354	13,6%
Ecologia e Meio Ambiente	465	2,7%	2054	3,0%
Economia	561	3,2%	2090	3,0%
Economia Domestica	67	0,4%	215	0,3%
Educação	1435	8,2%	5877	8,6%
Educação Física e Recreação	447	2,6%	1348	2,0%
Enfermagem	61	0,4%	245	0,4%
Engenharia	130	0,7%	572	0,8%
Engenharia Civil	46	0,3%	152	0,2%
Engenharia Elétrica	154	0,9%	609	0,9%
Engenharia Florestal	58	0,3%	191	0,3%
Engenharia Química	109	0,6%	462	0,7%
Engenharia Têxtil	30	0,2%	143	0,2%
Estatística Demográfica	5	0,0%	11	0,0%
Farmácia	138	0,8%	568	0,8%
Filosofia	542	3,1%	1639	2,4%
Física	75	0,4%	356	0,5%
Fisioterapia	157	0,9%	675	1,0%
Generalidades	137	0,8%	514	0,7%

Geografia	90	0,5%	274	0,4%
Geologia	73	0,4%	369	0,5%
Historia	328	1,9%	1087	1,6%
Linguagem	501	2,9%	2307	3,4%
Literatura	866	5,0%	2606	3,8%
Literatura Infanto-Juvenil	127	0,7%	224	0,3%
Marketing e Propaganda	544	3,1%	2252	3,3%
Matemática	285	1,6%	1727	2,5%
Medicina	1010	5,8%	3184	4,6%
Metodologia Cientifica	168	1,0%	1497	2,2%
Moda	38	0,2%	157	0,2%
Odontologia	221	1,3%	1079	1,6%
Paleontologia	7	0,0%	15	0,0%
Processamento de Dados	451	2,6%	2011	2,9%
Psicologia	693	4,0%	2381	3,5%
Química	201	1,2%	982	1,4%
Religião	178	1,0%	499	0,7%
Serviço Social	369	2,1%	1133	1,7%
Sociologia	771	4,4%	3121	4,6%
Tecnologia	20	0,1%	73	0,1%
Tecnologia de Alimentos	47	0,3%	119	0,2%
Turismo	176	1,0%	714	1,0%
Zoologia	55	0,3%	178	0,3%
Total	17421	100,0%	68543	100,0%

Conforme tabela anterior, através da classificação das obras pela tabela definida pelos bibliotecários foram formados 59 grupos, os quais apresentam uma média de 295 livros por grupo e 1.161 transações por grupo. Destaca-se o grupo de Direito com 13% das obras e 13,6% das transações. Verifica-se a existência de alguns grupos com pouca significância como o grupo Dicionários e Enciclopédias que apresenta somente 2 livros e 7 transações.

Nas classificações apresentadas pode-se observar alta freqüência de obras nas áreas dos cursos da instituição e uma baixa freqüência nas demais áreas.

4.7.3 Descrição do perfil dos usuários

Segundo Harrison (1998 p.181) “às 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”.

O objetivo é descrever o comportamento do usuário da biblioteca. Através da análise de suas transações (empréstimos e reservas), foi possível identificar seu perfil de utilização de obras, possibilitando interagir com o mesmo através dos sistemas de SRI e DSI de forma personalizada.

No estudo do perfil do usuário, o primeiro nível de descrição é a maior grande área de interesse, assim determinando a grande área de interesse do usuário. O próximo nível de descrição é formado por três subáreas de interesse, identificados através de uma análise das três principais áreas, segundo classificações das transações do usuário no quarto nível da CDD.

Cada usuário tem uma descrição do seu perfil, assim, pode-se prever seus interesses, possibilitando fazer sugestões, recomendações de livros e filtrar consultas.

Toma-se como exemplo as transações de um usuário segundo a CDD, que são apresentadas na Tabela 22.

Tabela 22 – Exemplo de empréstimos de usuário totalizados pela CDD

CDD	Descrição CDD	Total de Transações
005.74	Processamento de dados - Administração	43
005.1	Software – Desenvolvimento	30
005.756	Projeto de banco de dados	20
005.75	Programação (Computadores) - Gerência	18
469.5	Língua portuguesa – Sintaxe	17
005.43	Sistemas operacionais (Computadores)	16
133	Parapsicologia	6
625.70981	Rodovias – Brasil	6
005.133 SQL	SQL (Linguagem de programação de computador)	5
005.369 DELPHI	Delphi (Programa de computador)	3
340.1	Direito e política	3
341.2	Direito constitucional	2
611	Corpo humano	1
Total		170

As transações apresentadas na Tabela 22, são classificadas em grandes áreas pelo processo descrito no item 4.6.1. Totalizadas estas transações por grandes áreas temos o perfil do usuário segundo a sua grande área de interesse, que pode ser visto no Gráfico 6.

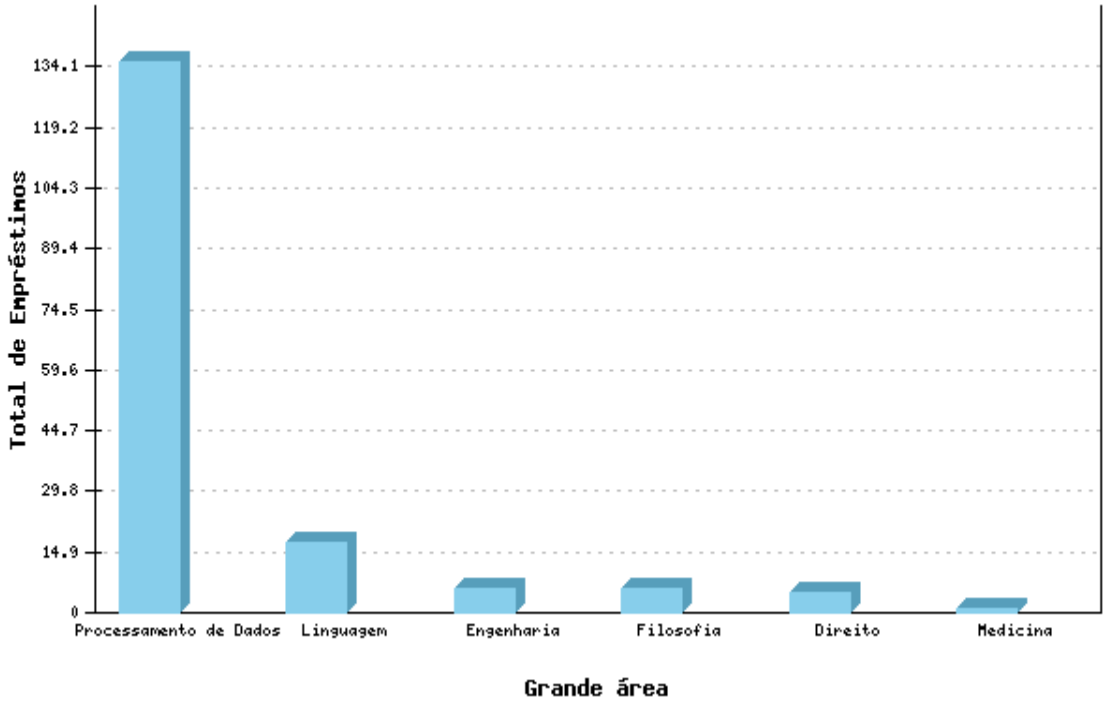

Gráfico 6 – Transações usuário por grandes áreas

Pode-se verificar que o primeiro nível de descrição (Gráfico 6) apresenta Processamento de Dados como a grande área de interesse do usuário em estudo. O próximo passo é identificar as subáreas de interesse do usuário.

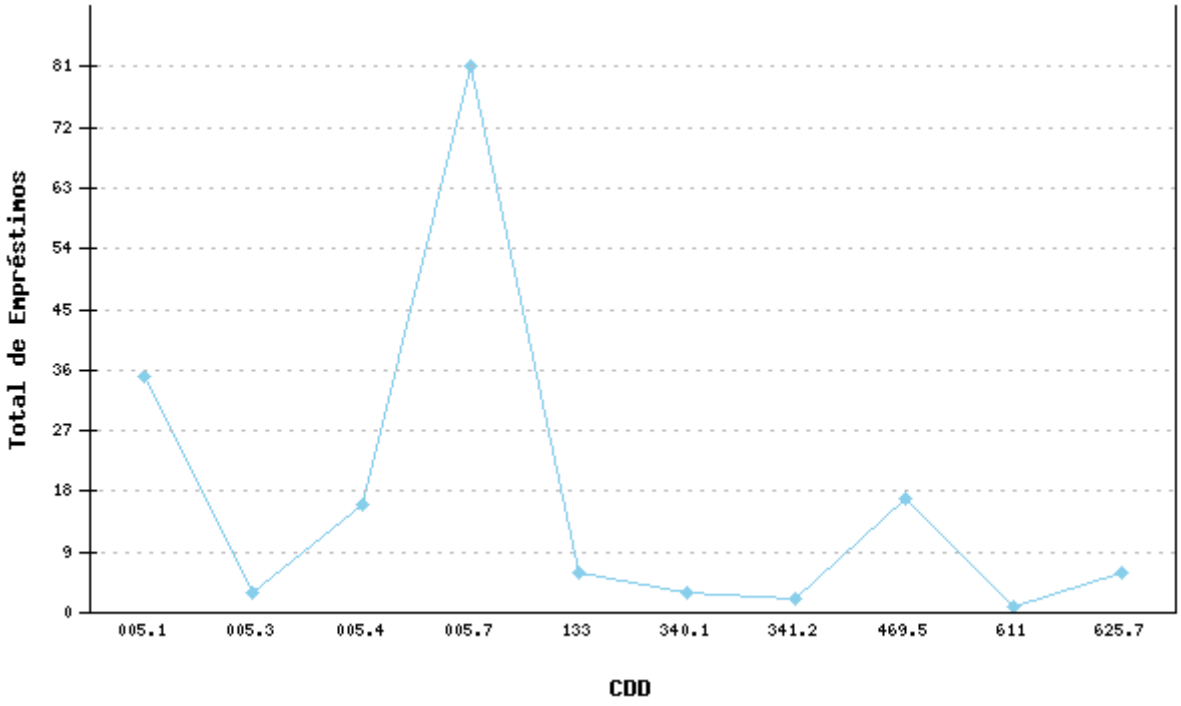

Gráfico 7 – Transações usuário por CDD nível quatro

Como pode ser observado no Gráfico 7, o segundo nível de descrição apresenta as três principais subáreas de interesse do usuário que são:

a) 005.7 - Programação;

b) 005.1 – Banco de dados;

c) 469.5 – Língua Portuguesa.

4.8 PLANO DE AÇÃO

Após a mineração, o plano de ação para solução do problema será executado, isto é, o conhecimento obtido a respeito do usuário será aplicado à personalização dos sistemas de recuperação e disseminação de informações através do desenvolvimento de um sistema WEB.

4.8.1 Personalização do SRI

Os serviços de recuperação de informações são os mais utilizados em bibliotecas, e devem ser além de robustos (com capacidade de gerenciamento de uma grande quantidade de dados), rápidos e eficientes, devido à enorme quantidade de informações. Tendo em vista esses requisitos, o objetivo em personalizar este serviço é de tornar mais fácil aos usuários encontrar o que eles desejam.

A personalização nos serviços de recuperação tem por objetivo facilitar ao usuário a localização de obras de seu interesse. Essa personalização é implementada pela reordenação das obras que são apresentados aos usuários. A personalização destaca as obras de acordo com os interesses do usuário descritos em seu perfil (MEIRA JR et. al., 2002).

O processo de recuperação de informações tradicional, conforme visto no item 2.8.1, foi remodelado conforme Figura 10.

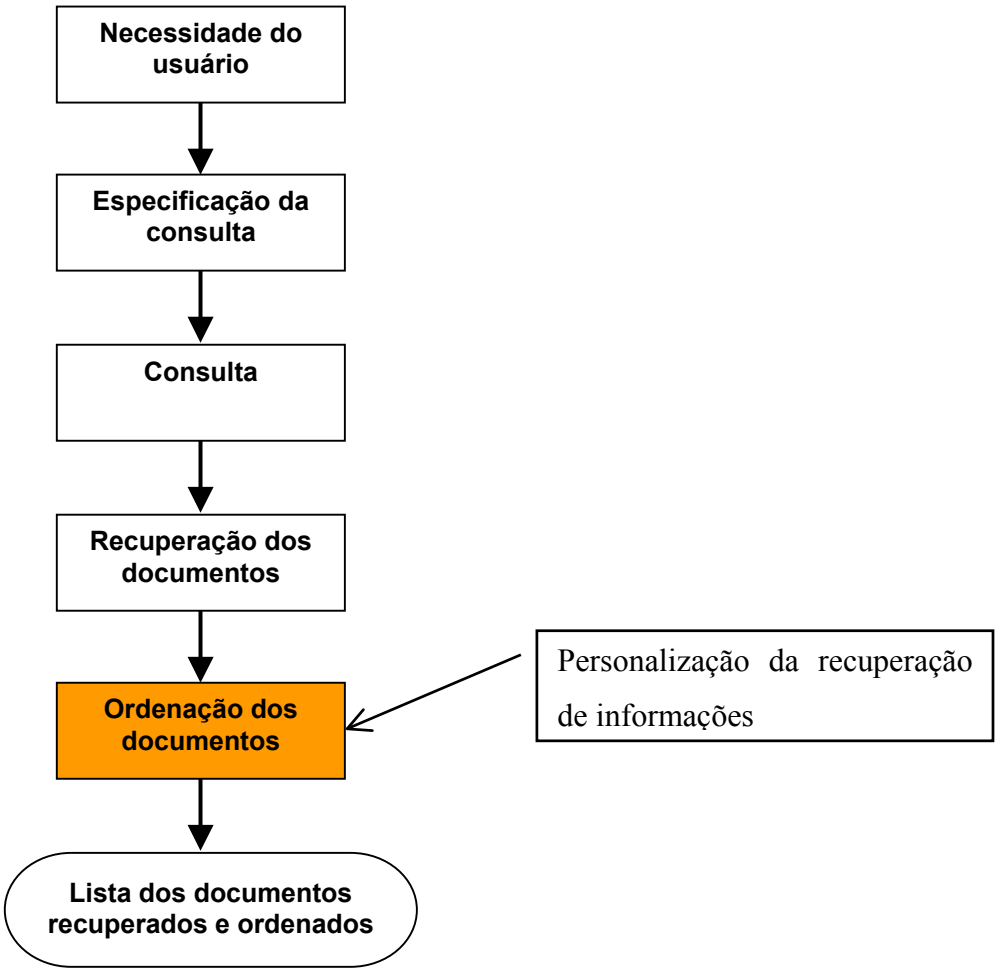

Figura 10 – Modelo de personalização de SRI

4.8.2 Personalização da DSI

Os serviços de disseminação seletiva da informação são responsáveis em tentar prever e fazer recomendações de obras que são potencialmente de interesse dos usuários.

A personalização dos sistemas de disseminação seletiva de informações acontece através da recomendação basendo-se no perfil do usuário. Estas recomendações ou sugestões são feitas para novas obras adquiridas e para obras que são emprestadas com maior freqüência.

O que distingue a DSI de outros serviços de alerta é a personalização segundo o perfil do usuário. Na Figura 11 apresenta-se o modelo de personalização do DSI. Mais detalhes sobre DSI podem ser vistos no item 2.8.2.

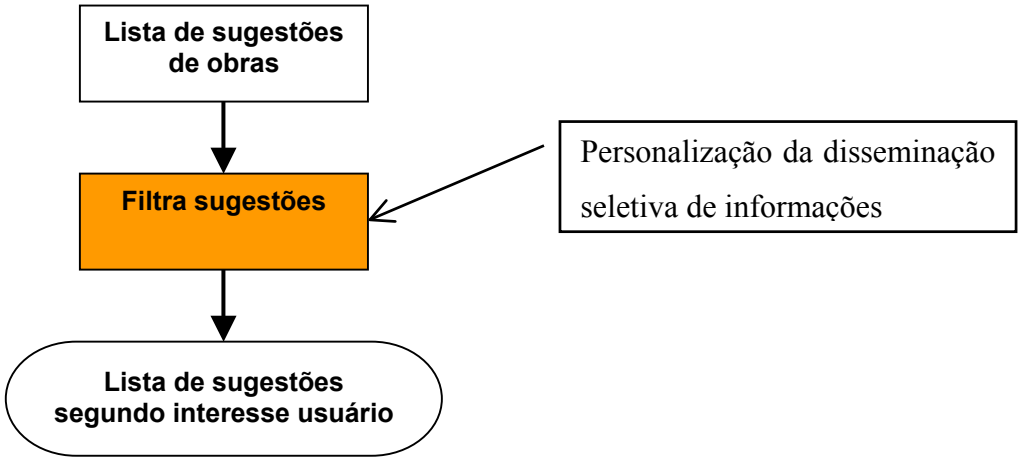

Figura 11 – Modelo de personalização do DSI

4.8.3 Arquitetura do sistema

O sistema desenvolvido disponibiliza a recuperação e sugestão de informações em um ambiente personalizado dinamicamente para cada usuário de acordo com suas preferências. Ao acessar o sistema o usuário solicita uma requisição de conteúdo ao servidor WEB; este a repassa ao sistema de personalização, que a recebe, processa e retorna ao usuário páginas HTML com o conteúdo personalizado. A arquitetura desse sistema pode ser vista na Figura 12.

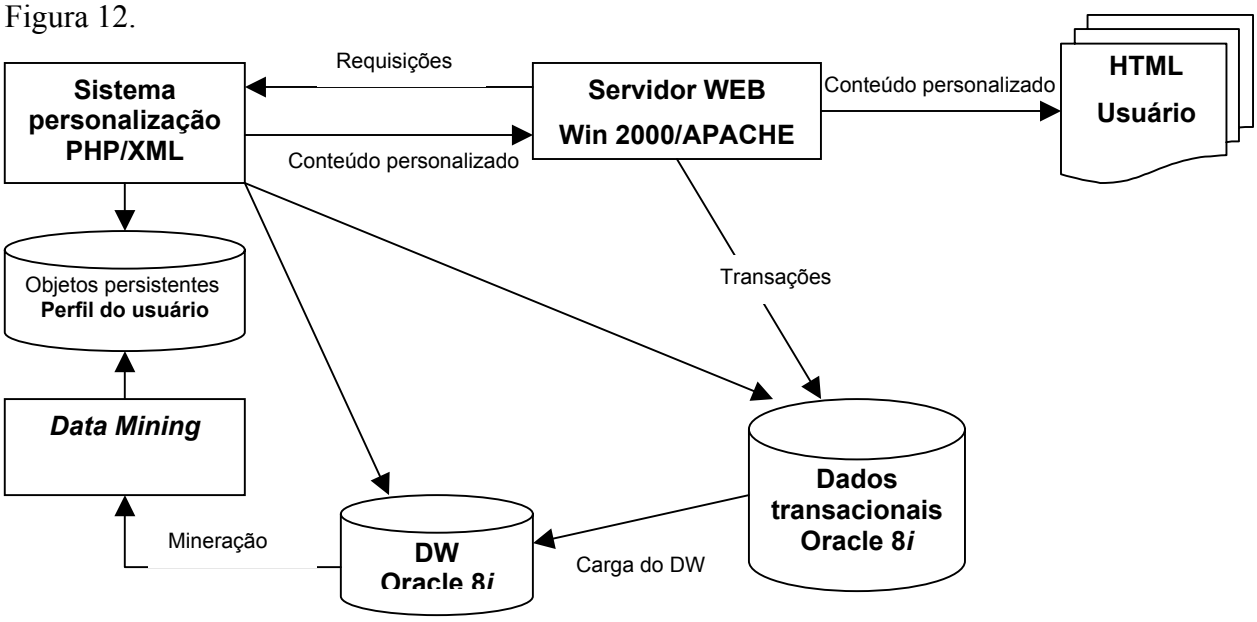

Figura 12 – Arquitetura do sistema de personalização

O sistema conta com uma 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 objetos persistentes o quais recebem os dados sobre o perfil do usuário.

4.8.4 Estrutura do sistema

Este item trata da estrutura do sistema desenvolvido. Na Figura 13 é apresentado um macro fluxo do sistema e sua funcionalidade.

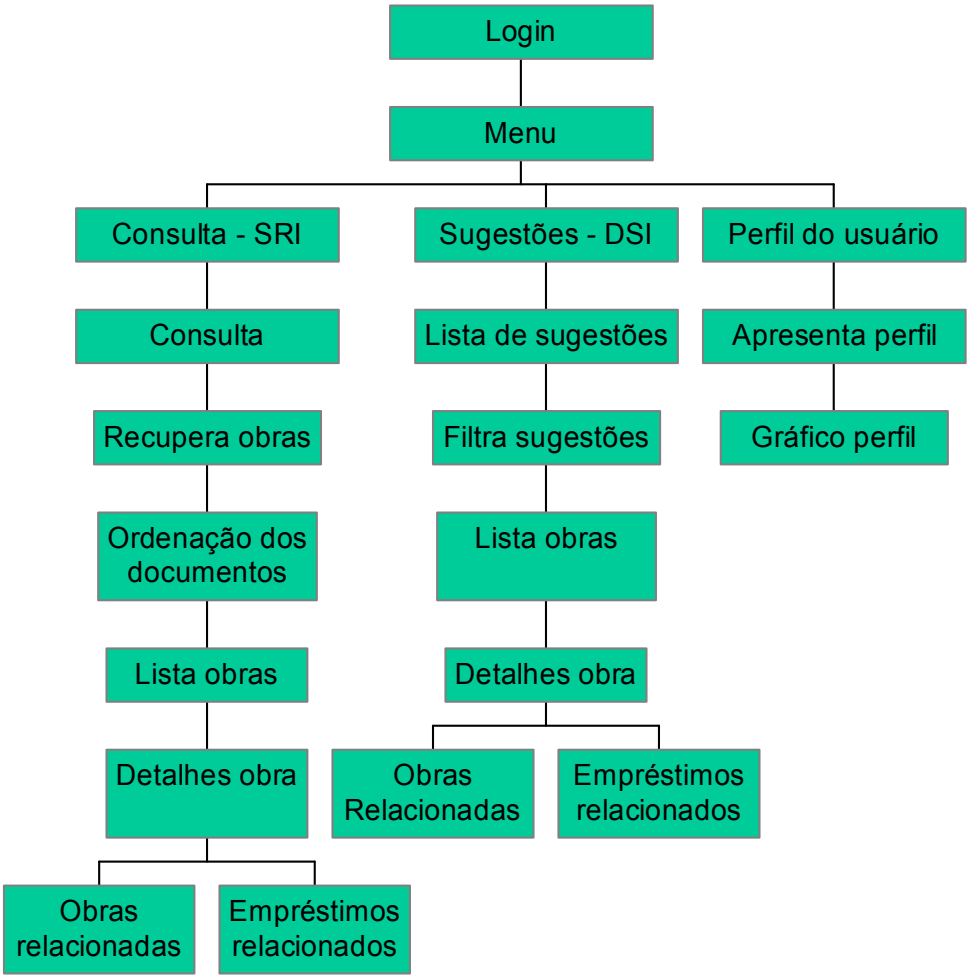

Figura 13 – Macro fluxo do sistema de personalização

Login: tela inicial do sistema que identifica o usuário, através da validade do seu código e senha. Segundo Reategui (2002) o primeiro passo para a personalização de um website é construir um mecanismo para que o site possa identificar os usuários que acessam. Dois mecanismos podem ser utilizados: *cookie* ou *login*. Neste sistema optou-se por *login*,

pois já existem código e senha que identificam o usuário na biblioteca. Assim foi implantada uma rotina para validação do usuário. Após a 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.

Menu sistema: tela menu contém as opções do sistema, é apresentada após a validação do usuário. Faz chamada aos recursos: consulta, sugestões e perfil do usuário.

Consulta: apresenta tela para a especificação da consulta a ser realizada.

Recupera obras: busca obras no banco de dados segundo os termos especificados na tela de consulta, retornando os que contenham os termos pesquisados.

Ordenação das obras: calcula a relevância das obras recuperadas de acordo com os interesses do usuário. Através da diferente entre a CDD da obra e a CDD de interesse do usuário.

Lista obras: recebe uma lista de obras com a relevância e apresenta ao usuário de forma ordenada.

Detalhes da obra: com a seleção da obra de interesse na lista, são mostrados os detalhes da obra (todos os dados sobre a mesma, autores, assuntos, editora, etc.). São apresentados também sugestões correlacionadas a partir dos empréstimos efetuados e sugestões de obras da mesma classificação.

Sugestões: são geradas sugestões de acordo com as novas aquisições de determinado período ou com obras mais emprestadas. Quando o usuário solicita as sugestões, o sistema identifica as áreas de interesse através do seu perfil e sugere as obras que foram recém adquiridas pela BC, nas áreas de seu interesse.

Perfil: apresenta tabelas e gráficos a respeito do perfil do usuário.

4.9 SISTEMA WEB

Esta seção apresenta a visualização do sistema desenvolvido, com uma seqüência de utilização do mesmo passo a passo.

A seguir são apresentadas as telas do sistema e uma breve descrição delas.

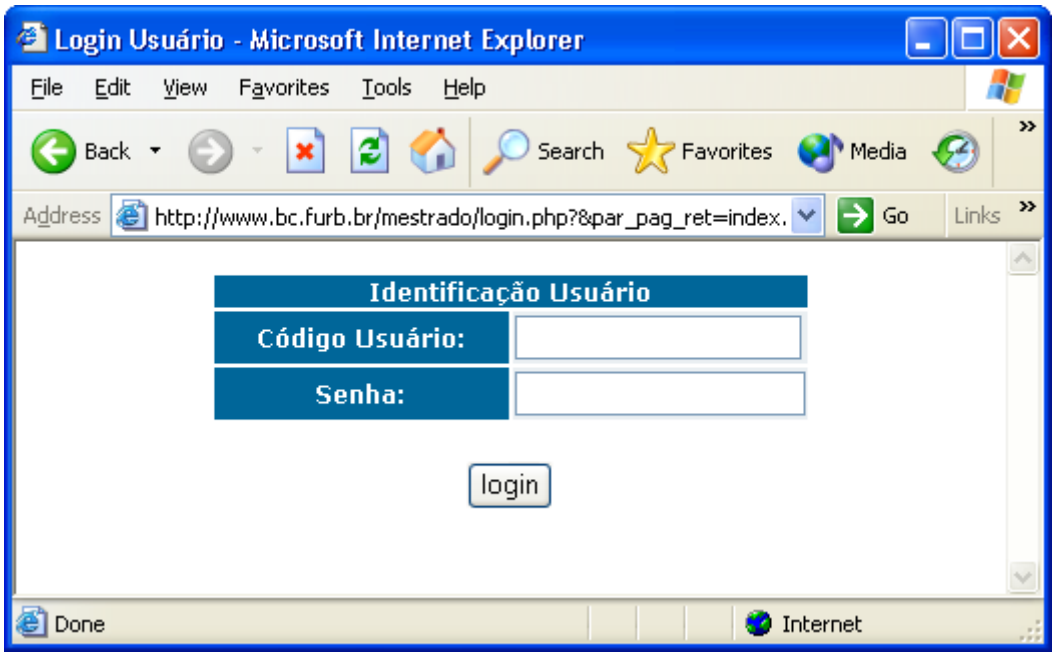

Figura 14 - Tela de *login*

Na tela de *login* (Figura 14) devem ser informados o código e senha do usuário na biblioteca, clicando no botão *login* ocorrerá a validação dos dados. Se os mesmos estiverem corretos será iniciado o sistema, senão será apresentada uma tela de erro.

Figura 15 -Tela principal

A tela principal (Figura 15) do sistema é divida em três partes: menu superior, menu lateral, e corpo principal.

No menu superior é apresentado o nome do usuário, seu departamento ou curso, sua grande área de interesse e subáreas de interesse.

No menu lateral da direita, têm-se as opções:

a) SRI - consulta: realiza a recuperação de informações;

b) SRI - grandes áreas:realiza a pesquisa através de hierarquia;

c) DSI – mais emprestados: faz sugestões de obras mais procuradas;

d) DSI - novas aquisições: faz sugestões de novas obras;

e) perfil do usuário: apresenta o perfil do usuário;

f) sair: sair do sistema.

No corpo principal são apresentados os dados do sistema. Acessando o menu lateral opção DSI – consultas será apresentada a tela de consulta apresentada na Figura 16.

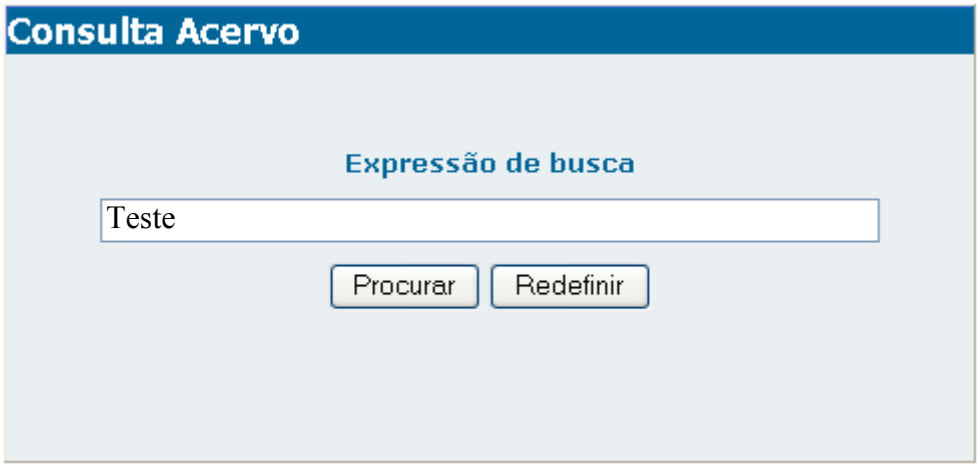

Figura 16 - Tela de consulta

Na tela de consulta (Figura 16) o usuário digita a expressão de busca a ser realizada, e deve clicar em procurar para realizar a consulta no acervo. A expressão de busca pode ser uma palavra ou uma frase. Por exemplo: teste, teste de software.

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 17 - Tela resultado da consulta

A tela de resultado da consulta (Figura 17) retornará os títulos encontrados no acervo ordenados conforme o perfil do usuário. O usuário deve selecionar o título de interesse e clicar no link para apresentar detalhes do título.

Figura 18 - Tela detalhes da obra

A tela detalhes da obra (Figura 18) retorna os dados da obra (classificação do material, tipo do material, autor, título, editora, ano de publicação e situação do material se esta ou não disponível), obras e empréstimos correlacionados.

Além da consulta por palavras e termos, podem ser realizadas consultas através dos grupos criados pelo processo de classificação, acessando a opção do menu DSI - grandes áreas.

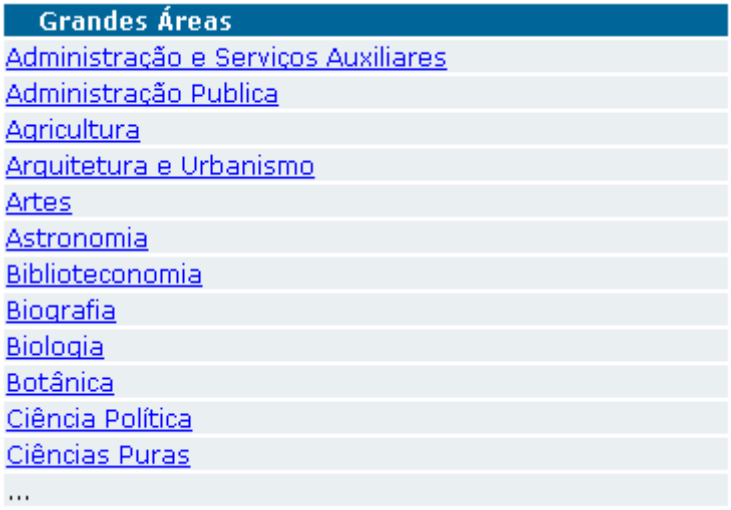

Figura 19 – Tela consulta por grupos

A tela de consulta por grupos (Figura 19) apresenta uma lista com os grupos de grandes áreas criados. Clicando sobre o grupo desejado será apresentada a tela a seguir.

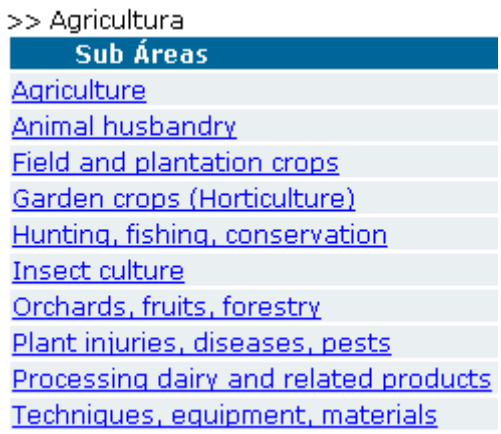

Figura 20 – Tela assuntos por grupo

A tela assuntos por grupo (Figura 20) apresenta os assuntos no nível três da CDD agrupados no grupo selecionado. Após serão listados os títulos encontrados com o assunto selecionado conforme Figura 18.

A DSI tenta prever as necessidades dos usuários fazendo sugestões de obras que sejam de interesse ao usuário. Foram definidas duas formas de sugestões, a primeira de novas aquisições e a segunda de livros mais emprestados. Acessando o menu lateral DSI – Novas aquisições serão apresentadas as ultimas obras adquiridas nas áreas de interesse do usuário.

Itens específicos na área de **Programming**

Título	Data Aquisição
Projeto & engenharia de software :teste de software	06/11/03
Algorithms in Java	16/10/03
PHP :a Bíblia	10/10/03
Introdução à programação orientada a objetos usando JAVA	29/09/03
PHP e MySQL :desenvolvimento Web	15/09/03

Itens específicos na área de **621.3**

Título	Data Aquisição
Dispositivos semicondutores :tiristores : controle de potencia em CC e CA	01/12/03
A revitalização da Usina Hidrelétrica do Gasparinho :uma proposta museológica	07/11/03
Aterramentos elétricos :conceitos básicos, técnicas de medição e instrumentação, filosofias de aterramento	23/10/03
Schaum`s outline of theory and problems of digital signal processing	13/10/03
Digital signal processing technology :essentials of the communications revolution	03/10/03

Itens específicos na área de **001.4**

Título	Data Aquisição
Pesquisa empírica em ciências humanas :(com ênfase em comunicação)	11/12/03
Fundamentos de metodologia científica :teoria da ciência e prática da pesquisa	11/1203
Fundamentos de metodologia cientifica	11/12/03
Elaboração de pesquisa científica	10/12/03
Metodologia do trabalho acadêmico	22/10/03

Figura 21 -Tela sugestões de novas aquisições

A tela de sugestões de novas aquisições (Figura 21) retorna as últimas aquisições nas subáreas de interesse do usuário.

Para acessar sugestões das obras mais emprestadas na área de interesse do usuário deve ser selecionada a opção DSI – mais emprestados.

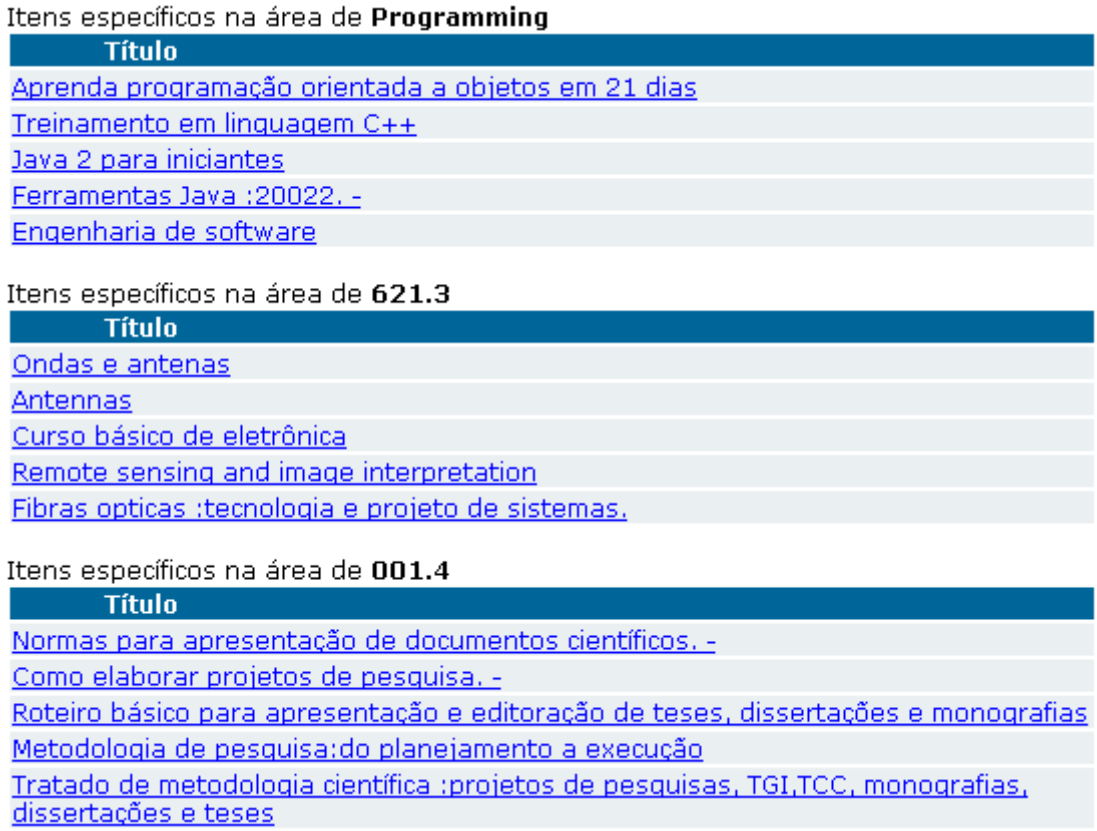

Figura 22 -Tela sugestões obras mais emprestadas

A tela de sugestões de obras mais emprestadas (Figura 22) retorna os títulos com maior freqüência de circulação nas subáreas de interesse do usuário.

Acessando a opção no menu perfil e apresentado a tela a seguir.

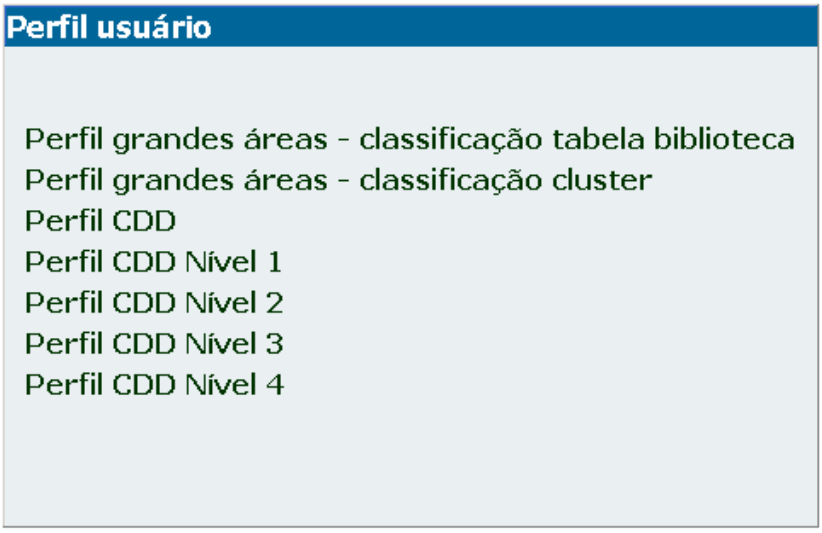

Figura 23 -Tela perfil do usuário

A tela perfil do usuário (Figura 23) apresenta os vários níveis de visualização do perfil do usuário. Selecionando um nível é possível visualizar sua totalização.

Código CDD	Descrição CDD	Total Movimento
005	Computer programming, programs, data	30
621	Applied physics	20
006	Special computer methods	11
001	Knowledge	10
658	General management	10
150	Psychology	3
511	General principles	3
611	Human anatomy, cytology, histology	3
823	English fiction	2
160	Logic	1
516	Geometry	1
519	Probabilities and applied mathematics	1
625	Engineering of railroads, roads	1
659	Advertising and public relations	1
612	Human physiology	1
443	French dictionaries	1

Ver gráfico

Figura 24 -Tela tabela perfil do usuário

A tela tabela perfil do usuário (Figura 24) totaliza os empréstimos dos usuários segundo suas transações conforme o nível do perfil escolhido para visualização.

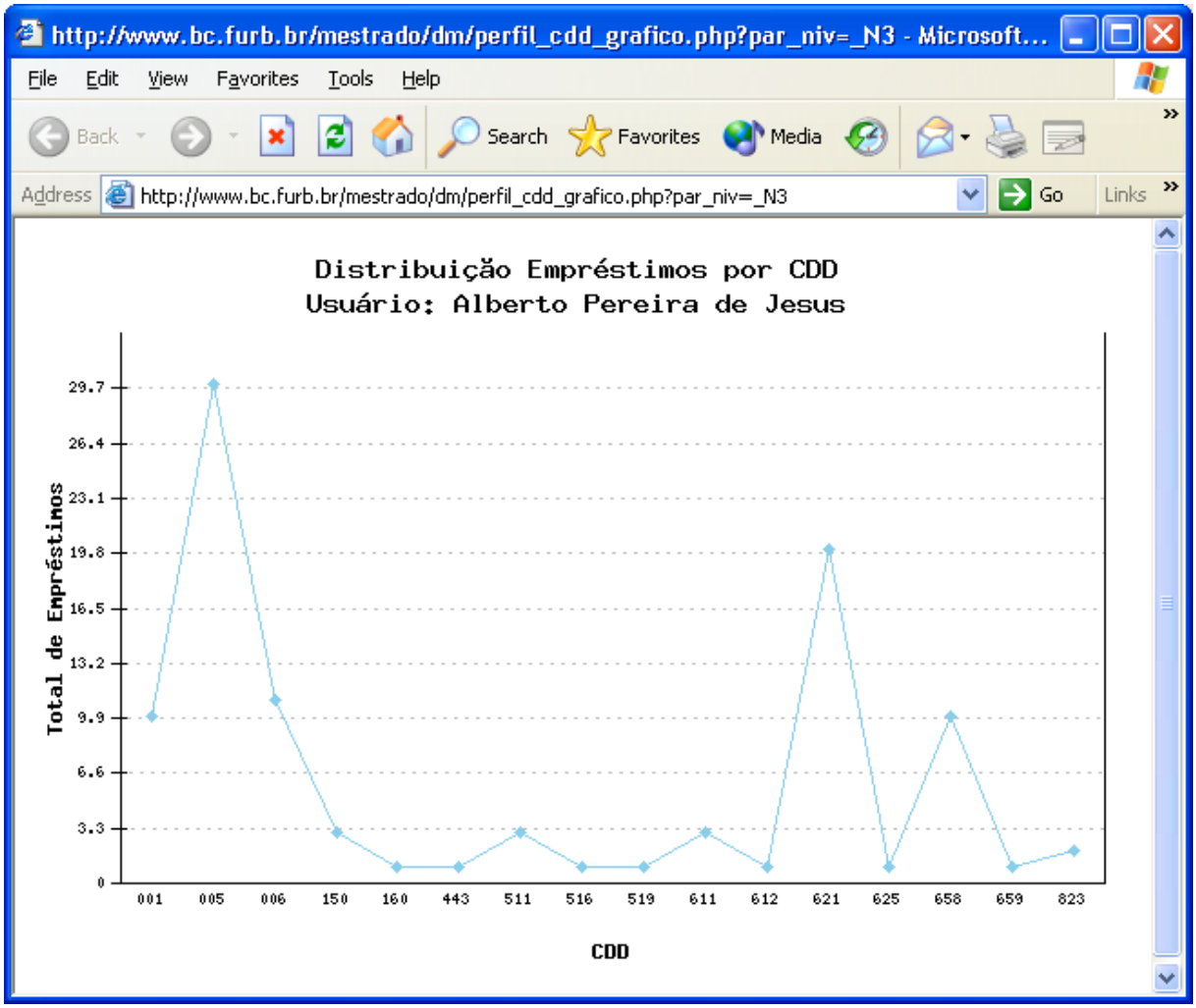

Figura 25 - Tela gráfico perfil do usuário

A tela gráfico perfil do usuário (Figura 25), apresenta um gráfico com o perfil do usuário selecionado.

O sistema apresentado facilita o processo de recuperação e disseminação de informações. Através da ordenação dos resultados da busca, fica facilitado para o usuário o processo de seleção de obras de seu interesse. Com o a disseminação seletiva de informações, possibilita-se ao usuário se manter informações sobre os novos livros adquiridos e livros com grande procura em suas áreas de interesse.

5 CONCLUSÕES E RECOMENDAÇÕES

Com a revisão bibliográfica pôde-se conhecer a tecnologia de *data mining,* necessária para dimensionar a sua aplicação na Biblioteca Central da FURB. A tecnologia de *data mining* ainda é pouco aplicada em bibliotecas. Destacam-se dois trabalhos realizados por Aranha[1999 e 2000], que aplica técnicas de *data mining* gerando listas de recomendações de itens. Diferencia-se dos trabalhos citados anteriormente, pela utilização de *data mining* aliada as técnicas de personalização de sistemas WEB, gerando SRI e DSI mais eficientes.

Cabe ressaltar dois resultados importantes:

a) o modelo para aplicação de *data mining* em bibliotecas;

b) a aplicação do modelo proposto para sua validação.

A definição de um modelo facilitou a realização do estudo, pois o mesmo determinou os passos que deveriam ser realizados para obtenção dos resultados com sucesso.

O modelo proposto baseou-se na aplicação das técnicas de *data mining* sobre a classificação CDD das obras, possibilitando determinar o perfil dos usuários quanto aos seus interesses bibliográficos. A CDD é um padrão utilizado em várias bibliotecas, tornando fácil a aplicação deste modelo em outras bibliotecas que utilizem este padrão.

Através da aplicação do modelo na Biblioteca Central da FURB salienta-se que:

a) os dados são fatores determinantes na aplicação de *data mining*; para tanto devem receber em tratamento minucioso quanto a obtenção e seleção;

b) o conhecimento dos especialistas (bibliotecários) enriqueceu os dados da pesquisa, mostrando-se como elemento muito importante, quanto à determinação das variáveis e análise dos resultados obtidos;

c) os dados devem ser organizados para aplicação de *data mining,* gerando um modelo dimensional;

d) a análise estatística dos dados é uma etapa preliminar à aplicação da tecnologia de *data mining* adequada, uma vez possibilitou obter uma visão global dos dados a partir dela;

e) a escolha da técnica de *data mining* foi fundamental para o sucesso dos resultados;

Os objetivos do trabalho foram alcançados. Quanto ao principal, o desenvolvimento de um sistema WEB de recuperação e disseminação de informações personalizado ao usuário, mostrou-se funcional ao seu propósito, facilitando o processo de recuperação de informações através da ordenação do resultado das pesquisas, disseminando informações de interesse ao usuário.

Quanto aos objetivos específicos:

a) o *data warehouse* desenvolvido foi eficiente para aplicação das técnicas de *data mining*, possibilitando também informações para tomada de decisões através da criação de relatórios com ferramentas apropriadas;

b) a aplicação das técnicas de *data mining* sobre os asssuntos significativos gerou grupos com correlações entre livros implícitas. A classificação dos livros possibilitou descrever o perfil dos usuários, possibilitando conhecer melhor os seus hábitos e interesses na biblioteca;

c) o sistema desenvolvido de SRI e DSI personalizado, facilitou o processo de recuperação de informações e tornando eficiente a disseminação seletiva de informações.

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 gerados a partir de seus sistemas legados, sem nenhuma utilização para tomada de decisões.

5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Como sugestões para trabalhos futuros podemos citar:

a) ampliar os estudo sobre o perfil do usuário, melhorando os processos de identificação das áreas de interesse;

b) testar novas técnicas sobre os dados;

c) buscar novas variáveis que sejam determinantes no perfil do usuário;

d) incorporar mais funcionalidades ao sistema desenvolvido.

REFERÊNCIAS

ACCETTA, Izildinha Ramos. **Serviço de Levantamento e Comutação Bibliográfica da Biblioteca Central da FURB:** a qualidade da comunicação na recuperação da informação. 1998. 47f. Monografia (Curso de Pós-Graduação em nível de especialização em "Qualidade na Comunicação") – FURB, Universidade Regional de Blumenau, Blumenau, 1998.

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.

ARANHA, Francisco. **Perfil de usuários da biblioteca Karl A. Boedecker:** geração de valor para pesquisadores por meio de cooperação indireta. São Paulo: EAESP/FGV/NPP, 1999. 59p.

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.

CARVALHO, Doris de Queiroz. **Classificação decimal de direito**. 4. ed. rev. e atual. Brasília : Presidência da República, 2002. 257 p.

DINIZ, Carlos Alberto R., NETO, Francisco Louzada. **Data Mining:** uma introdução. São Paulo: Associação Brasileira de Estatística, 2000. 123p.

FIGUEIRA, Rafael. **Mineração de dados e bancos de dados orientados a objetos**. 1998. 96f. Dissertação (Mestrado em Ciências da Computação) – UFRJ, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1998.

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**.São Paulo: Berkeley, 1998.

KIMBALL, R. **Data Warehouse Toolkit:** técnicas para construção de data warehouses dimensionais. São Paulo: Makron Books, 1998.

LUHN, H.P. Selective dissemination of nerw scientific information with the aid of electronic processing equipment. **American Documentation Institute, Berkeley,** v.12, p.131-138, Apr. 1961.

MEIRA JR, Wagner et. al. **Sistemas de comércio eletrônico**: projeto e desenvolvimento. Rio de Janeiro: Campus, 2002. 371 p.

MONDSCHEIN, L. G. SDI use and productivity in the corporate research environment. **Special Libraries Association, Washington**, v. 8, n. 4, p. 265-78, Fall 1990.

OCLC- OnLine Computer Library Center. **Introduction to Dewey Decimal Classification**. Disponível em <http://www.oclc.org/dewey/versions/ddc22print/intro.pdf>. Acesso em: 16-mar-2004.

REATEGUI, Eliseo. **Data mining e personalização dinâmica.** Criciúma: X Escola de Informática da SBC-Sul, 2002.

RIEH, S. Y. Changing reference service environment: a review of perspectives from managers, librarians, and users. **Journal of Academic Librarianship**, Cincinnati, v. 25, n. 3, p. 178-86, May 1999.

SANTOS, Érico Resende. **Implantação de tecnologia de data warehouse em bibliotecas com uso da tecnologia adequada**. São Paulo: EAESP/FGV, 1998. 53p.

SILVA, Edna Lúcia da; MENEZES, Estera Muskat. **Metodologia da pesquisa e elaboração de dissertação**. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001. 121p.

TEIXEIRA, Cenidalva Miranda de Souza, SCHIEL, Ulrich. A Internet e seu impacto nos processos de recuperação da informação. **Ciência da Informação**, Brasília, v. 26, n. 1,p. 65-71, jan./abr. 1997.

VASCONCELOS, Eduardo Mourão **Complexidade e pesquisa interdisciplinar**: epistemologia e metodologia operativa. Petrópolis: Vozes, 2002. 343p.

VELASQUEZ,Roberto M.G. et. al. Técnicas de Classificação para Caracterização da Curva de Carga deEmpresas 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/5cbrn_033.pdf.>. Acessado em: 16-mar-2004.

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração**. 3.ed. São Paulo : Atlas, 2000. 92p.

APÊNDICE A – TABELA DE ASSUNTOS SIGNIFICATIVOS

AS	Obras por AS	% Obras por AS	Transações por AS	% Transações por AS
000	173	1,0%	616	0,9%
001.4	168	1,0%	1497	2,2%
004	82	0,5%	291	0,4%
005	63	0,4%	212	0,3%
005.1	179	1,0%	1015	1,5%
005.7	47	0,3%	212	0,3%
006	80	0,5%	281	0,4%
028.5	127	0,7%	224	0,3%
100	387	2,2%	1159	1,7%
133	100	0,6%	265	0,4%
150	126	0,7%	429	0,6%
152	37	0,2%	100	0,1%
152.3	44	0,3%	223	0,3%
153	99	0,6%	366	0,5%
155	112	0,6%	310	0,5%
155.4	146	0,8%	596	0,9%
158	129	0,7%	357	0,5%
193	55	0,3%	215	0,3%
200	178	1,0%	499	0,7%
300	203	1,2%	686	1,0%
301	97	0,6%	439	0,6%
302	94	0,5%	306	0,4%
302.2	76	0,4%	219	0,3%
303	37	0,2%	128	0,2%
303.4	97	0,6%	387	0,6%
304.2	150	0,9%	626	0,9%
305	114	0,7%	476	0,7%
306	159	0,9%	560	0,8%
307.7	92	0,5%	511	0,7%
320	217	1,2%	731	1,1%
320.1	38	0,2%	210	0,3%
330	254	1,5%	886	1,3%
331	88	0,5%	300	0,4%
333.7	52	0,3%	230	0,3%
338	98	0,6%	313	0,5%
338.6	19	0,1%	202	0,3%
338.9	90	0,5%	368	0,5%
340	130	0,7%	414	0,6%
340.1	86	0,5%	256	0,4%
341.1	70	0,4%	247	0,4%
341.2	211	1,2%	1074	1,6%
341.3	301	1,7%	1251	1,8%
341.4	462	2,7%	2274	3,3%

341.5	145	0,8%	555	0,8%
341.6	181	1,0%	582	0,8%
342.1	424	2,4%	1826	2,7%
342.2	203	1,2%	587	0,9%
342.3	51	0,3%	288	0,4%
360	92	0,5%	202	0,3%
361	78	0,4%	285	0,4%
362	111	0,6%	270	0,4%
362.1	101	0,6%	413	0,6%
370	168	1,0%	569	0,8%
370.1	406	2,3%	1773	2,6%
370.7	110	0,6%	768	1,1%
370.9	53	0,3%	277	0,4%
371	120	0,7%	420	0,6%
371.2	87	0,5%	338	0,5%
371.3	144	0,8%	504	0,7%
372	140	0,8%	494	0,7%
372.4	73	0,4%	336	0,5%
378	71	0,4%	176	0,3%
378.1	63	0,4%	222	0,3%
380	37	0,2%	91	0,1%
380.1	177	1,0%	715	1,0%
382	65	0,4%	285	0,4%
400	93	0,5%	442	0,6%
410	83	0,5%	514	0,7%
420	108	0,6%	383	0,6%
420.7	45	0,3%	260	0,4%
460	56	0,3%	248	0,4%
469	116	0,7%	460	0,7%
500	268	1,5%	859	1,3%
501	55	0,3%	233	0,3%
510	106	0,6%	561	0,8%
510.7	41	0,2%	209	0,3%
515	48	0,3%	366	0,5%
519.5	90	0,5%	591	0,9%
540	90	0,5%	326	0,5%
541.3	43	0,2%	243	0,4%
547	68	0,4%	413	0,6%
551	65	0,4%	355	0,5%
570	43	0,2%	197	0,3%
574	76	0,4%	236	0,3%
574.5	62	0,4%	217	0,3%
600	182	1,0%	678	1,0%
610	155	0,9%	532	0,8%
610.7	82	0,5%	314	0,5%
612	126	0,7%	461	0,7%
614	81	0,5%	393	0,6%
615	85	0,5%	230	0,3%
615.1	60	0,3%	348	0,5%
615.8	157	0,9%	675	1,0%

616	241	1,4%	669	1,0%
616.8	173	1,0%	488	0,7%
617	95	0,5%	258	0,4%
617.6	221	1,3%	1079	1,6%
618	111	0,6%	304	0,4%
620	188	1,1%	761	1,1%
621.3	154	0,9%	609	0,9%
628	92	0,5%	405	0,6%
628.3	23	0,1%	211	0,3%
628.5	46	0,3%	251	0,4%
630	195	1,1%	572	0,8%
650	83	0,5%	393	0,6%
657	154	0,9%	704	1,0%
657.4	56	0,3%	288	0,4%
658	211	1,2%	1090	1,6%
658.1	158	0,9%	1009	1,5%
658.3	185	1,1%	728	1,1%
658.4	484	2,8%	2362	3,4%
658.5	176	1,0%	788	1,1%
658.7	51	0,3%	406	0,6%
658.8	405	2,3%	1935	2,8%
659.1	93	0,5%	202	0,3%
660	93	0,5%	323	0,5%
660.2	53	0,3%	232	0,3%
700	591	3,4%	1661	2,4%
710	31	0,2%	100	0,1%
711.4	61	0,4%	321	0,5%
712	24	0,1%	201	0,3%
790	132	0,8%	403	0,6%
791.4	68	0,4%	238	0,3%
796	247	1,4%	707	1,0%
800	371	2,1%	1297	1,9%
813	144	0,8%	363	0,5%
860	40	0,2%	109	0,2%
869	37	0,2%	149	0,2%
869.9	274	1,6%	688	1,0%
900	403	2,3%	1162	1,7%
981	106	0,6%	396	0,6%
Total	17421	100,0%	68543	100,0%

APÊNDICE B – TABELA DE GRANDES ÁREAS - *CLUSTER*

Grupo	Sub grupo	Descrição do Grupo	AS	Descrição do AS
1		Direito	341.5	Direito penal
1		Direito	342.1	Direito civil
1		Direito	341.4	Direito processual
1		Direito	341.3	Direto administrativo
1		Direito	341.2	Direto constitucional
1		Direito	342.2	Direito comercial
1		Direito	341.6	Direito do trabalho
1		Direito	340.1	Filosofia do Direito
1		Direito	340	Direito
2		Marketing	658.8	Marketing
2		Marketing	658.4	Planejamento Estratégico
3		Contabilidade	658.1	Organização e finanças
3		Contabilidade	657.4	Contabilidade auditoria
3		Contabilidade	657	Contabilidade
4		Administração	658	Administração
4		Administração	658.5	Administração da produção
4		Administração	338	Produção
4		Administração	658.3	Administração de pessoal
4		Administração	158	Psicologia aplicada
5		Literatura	869.9	Literatura portuguesa
5		Literatura	800	Literatura
6		Artes	700	Artes
7		História	900	Geografia, história e biografia
8		Economia /Política	330	Economia
8		Economia /Política	320	Ciência política
8		Economia /Política	306	Antropologia
8		Economia /Política	300	Ciências sociais
9		Filosofia da educação	370.1	Filosofia da educação
10		Filosofia	100	Filosofia
11		Metodologia Cientifica	001.4	Metodologia da pesquisa
12		Saúde	610	Medicina
12		Saúde	616	Doenças
12		Saúde	617.6	Odontologia
12		Saúde	617	Cirurgia e tópicos relativos
12		Saúde	618	Obstetrícia e ginecologia
12		Saúde	615.8	Fisioterapia
12		Saúde	612	Fisiologia humana
12		Saúde	614	Saúde pública
12		Saúde	362.1	Assistência médica-social
13		Pedagogia	378.1	Administração universitária
13		Pedagogia	378	Ensino superior
13		Pedagogia	371.2	Administração educacional

13		Pedagogia	370.9	História da educação
13		Pedagogia	371.3	Métodos de instrução e estudo
13		Pedagogia	372	Educação elementar
13		Pedagogia	371	Generalidades da educação
13		Pedagogia	370.7	Estudo e ensino da educação
13		Pedagogia	370	Educação
14		Psicologia	616.8	Psiquiatria
14		Psicologia	153	Inteligência
14		Psicologia	155.4	Psicologia infantil
14		Psicologia	155	Psicologia genética
14		Psicologia	150	Psicologia
15		Ciência e Tecnologia	628.5	Saneamento
15		Ciência e Tecnologia	628.3	Esgotos
15		Ciência e Tecnologia	628	Engenharia sanitária e municipal
15		Ciência e Tecnologia	630	Agricultura
15		Ciência e Tecnologia	574.5	Ecologia
15		Ciência e Tecnologia	551	Geologia
15		Ciência e Tecnologia	620	Engenharias
15		Ciência e Tecnologia	600	Ciências aplicadas
15		Ciência e Tecnologia	500	Ciências puras
16		Letras	420	Língua Inglesa
16		Letras	469	Língua Portuguesa
16		Letras	400	Línguas e linguagem
16		Letras	420.7	Estudo e ensino da língua inglesa
16		Letras	410	Lingüística
16		Letras	372.4	Leitura
17	1	Química	547	Química orgânica
17	1	Química	541.3	Físico química
17	1	Química	540	Química
17	2	Farmácia	615.1	Farmacologia
17	2	Farmácia	615	Farmacologia e terapêutica
17	3	Biologia	574	Biologia
17	3	Biologia	570	Ciências biologicas
17	4	Engenharias	660.2	Engenharia química
17	4	Engenharias	660	Tecnologia química
17	4	Engenharias	621.3	Engenharia elétrica
17	4	Engenharias	515	Cálculo
17	4	Engenharias	510.7	Estudo e ensino da matemática
17	4	Engenharias	658.7	Administração de materiais
17	4	Engenharias	869	Literatura Portuguesa
17	4	Engenharias	860	Literatura de língua Espanhola
17	4	Engenharias	610.7	Estudo da medicina
17	4	Engenharias	380	Comércio, comunicação e transporte
17	4	Engenharias	338.6	Organização da produção
17	4	Engenharias	382	Comércio internacional
17	4	Engenharias	342.3	Direito internacional privado
17	4	Engenharias	341.1	Direito internacional público
17	4	Engenharias	320.1	Estado
17	4	Engenharias	712	Urbanismo
17	4	Engenharias	710	Arquitetura

17	4	Engenharias	460	Línguas Espanhola
17	4	Engenharias	152.3	Psicomotrocidade
17	5	Filosofia	501	Filosofia da ciência
17	5	Filosofia	331	Relações industriais
17	5	Filosofia	361	Serviço social
17	5	Filosofia	193	Filosofia alemã
17	5	Filosofia	303	Processos sociais
17	5	Filosofia	152	Psicologia experimental
17	6	Ficção	813	Ficção americana
17	7	Outros	133	Parapsicologia
17	7	Outros	028.5	Literatura infanto - juvenil
18		TI / Matemática	519.5	Estatística
18		TI / Matemática	510	Matemática
18		TI / Matemática	005.1	Programação
18		TI / Matemática	006	Métodos especiais de computação
18		TI / Matemática	005.7	Bases de dados
18		TI / Matemática	005	Software
18		TI / Matemática	004	Ciência da computação
19	1	Ciências Sociais	338.9	Desenvolvimento econômico
19	1	Ciências Sociais	333.7	Recursos naturais
19	1	Ciências Sociais	711.4	Planejamento urbano
19	1	Ciências Sociais	307.7	Sociologia Urbana
19	1	Ciências Sociais	304.2	Ecologia humana
19	2	Educação Física / Turismo e Lazer	796	Educação Física
19	2	Educação Física / Turismo e Lazer	790	Lazer
19	2	Educação Física / Turismo e Lazer	380.1	Comércio (Turismo)
19	3	Serviço Social	981	Brasil história
19	3	Serviço Social	305	Estrutura social
19	3	Serviço Social	362	Assistência Social
19	3	Serviço Social	360	Serviço social
19	3	Serviço Social	302	Interação social
19	4	Sociologia	303.4	Mudança social
19	4	Sociologia	301	Sociologia
19	5	Religião	200	Religião
19	6	Comunicação Social	650	Administração
19	6	Comunicação Social	791.4	Cinema, rádio e televisão
19	6	Comunicação Social	659.1	Marketing e propaganda
19	6	Comunicação Social	302.2	Comunicação de massa
19	7	Generalidades	000	Generalidades

ANEXO A – TABELAS DE GRANDES ÁREAS - BIBLIOTECA

Código CDD	Descrição grande área	Código grande área
000	Generalidades	000
001.5	Generalidades	000
007	Generalidades	000
040	Generalidades	000
001.4	Metodologia Cientifica	001.4
004	Processamento de Dados	004
020	Biblioteconomia	020
028.6	Biblioteconomia	020
028.5	Literatura Infanto-Juvenil	028.5
030	Dicionários e Enciclopédias	030
100	Filosofia	100
160	Filosofia	100
150	Psicologia	150
200	Religião	200
300	Sociologia	300
302.24	Sociologia	300
304.3	Sociologia	300
304.2	Ecologia e Meio Ambiente	304.2
333.7	Recursos Naturais	304.2
363.7	Problemas e Serviços Ambientais	304.2
574.5	Ecologia e Meio Ambiente	304.2
581.5	Ecologia das Plantas	304.2
591.5	Ecologia dos Animais	304.2
627	Engenharia Hidráulica	304.2
628	Engenharia Sanitária	304.2
310	Estatística Demográfica	310
320	Ciência Política	320
330	Economia	330
334	Economia - Cooperativas	330
340	Direito	340
350	Administração Publica	350
360	Serviço Social	360
363.79	Poluição	360
370	Educação	370
380	Comercio, Comunicação e Transportes	380
380.146	Comércio e Serviços	380
380.145	Turismo	380.145
390	Costumes	390
392	Folclore	390
395	Etiqueta	390
391	Moda	391
400	Linguagem	400

500	Ciências Puras	500
510	Matemática	510
520	Astronomia	520
530	Física	530
540	Química	540
550	Geologia	550
560	Paleontologia	560
570	Biologia	570
574.6	Biologia econômica	570
580	Botânica	580
581.6	Economia Botânica	580
590	Zoologia	590
591.6	Zoologia Econômica	590
600	Tecnologia	600
610	Medicina	610
610.74	Medicina	610
614	Saúde Publica	610
615.9	Medicina - Toxicologia	610
617.7	Medicina - Oftalmologia	610
619	Medicina Experimental	610
610.73	Enfermagem	610.73
615	Farmácia	615
615.8	Fisioterapia	615.8
617.6	Odontologia	617.6
620	Engenharia	620
621.4	Engenharia do Calor	620
625	Engenharia Rodoviária	620
629	Engenharia	620
621.3	Engenharia Elétrica	621.3
621.382	Telecomunicações	621.3
624	Engenharia Civil	624
630	Agricultura	630
635	Agricultura - Horticultura	630
634.9	Engenharia Florestal	634.9
640	Economia Domestica	640
650	Administração e Serviços Auxiliares	650
651	Secretariado	650
658	Administração	650
658.9	Administração	650
659.2	Relações Publicas	650
657	Contabilidade	657
302.23	Comunicação de Massa	658.8
658.8	Marketing e Propaganda	658.8
659	Propaganda	658.8
660	Engenharia Química	660
665	Engenharia Química - Tecnologia Industrial	660
666	Cerâmica	660
668	Engenharia Química - Engenharia Produtos Orgânicos	660
678	Engenharia Química - Elastomeros	660

664	Tecnologia de Alimentos	664
667	Engenharia Têxtil	667
677	Engenharia Têxtil	667
690	Construção Civil	690
700	Artes	700
730	Artes	700
780	Musica	700
710	Arquitetura e Urbanismo	710
790	Educação Física e Recreação	790
800	Literatura	800
900	Historia	900
930	Historia	900
910	Geografia	910
920	Biografia	920