ADRIANA ZANELLA MARTINHAGO

DESENVOLVIMENTO DE UM SISTEMA DE RECUPERAÇÃO DE INFORMAÇÃO PARA A ÁREA HOSPITLAR

Monografia de final de curso apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação, para a obtenção do título de Bacharel.

APROVADA em __ de _____ de _____.

Prof. _____

Prof. _____

Prof. _____
UFLA
(Orientador)

LAVRAS
MINAS GERAIS - BRASIL

DEDICATÓRIA

Dedico este trabalho primeiramente aos meus pais, Eide Zanella Martinhago e José Carlos Martinhago, pois sem a força e o amor deles eu não teria conseguido chegar até aqui. Aos meus irmãos, em especial à Dariana que sempre me ajudou e apoiou em todos os momentos.

Á minha família que mesmo de longe sempre me apoiaram e confiaram em mim. Aos meus amigos pelo companheirismo e ajuda na hora dos apertos.

Ao meu amor, Douglas, pela compreensão, paciência e pelo amor que teve comigo durante esses anos que estamos juntos.

Por fim a Deus, por me dar forças para vencer todos os obstáculos impostos até aqui e por ter colocado no meu caminho todas essas pessoas maravilhosas.

AGRADECIMENTO

Agradeço a minha orientadora, Olinda, por me ajudar na concretização deste trabalho.

Ao pessoal do Hospital Vaz Monteiro, principalmente o Willian, que me auxiliou muito neste trabalho, me mostrando que sou capaz de fazer muita coisa.

Aos amigos do curso, em especial, a Carmen, Ricardo, Érika, Muzamba, Flávia, Chambinho, Samuel, Paulo, aprendi muito com vocês, não sei o que seria de mim sem vocês no meu caminho, vou sentir muita falta.

Aos meus amigos de infância, Soraia, Milla, Fabiane, Aline, Júnior Macuco, Bira, Eliane, valeu por tudo.

Às minhas amigas de Palmas Kelly, Gi e Kharina pelo apoio.

Ao pessoal da TreinaSoft, Elvis, Luciana, Maikel, Leandro, Fernanda, Diego, por tudo que me ensinaram, cresci muito no tempo que trabalhei com vocês.

À todos que de uma maneira ou outra passaram pelo meu caminho e me ajudaram de alguma forma.

Sei que posso pecar em não citar o nome de alguém que tanto ajudou na concretização deste trabalho, mas desde já peço desculpas e agradeço.

DESENVOLVIMENTO DE UM SISTEMA DE RECUPERAÇÃO DE INFORMAÇÃO PARA A ÁREA HOSPITALAR

RESUMO

Todas as organizações em saúde geram um grande volume de documentos contendo informações sobre pacientes. Esses documentos são fontes de dados preciosos, que precisam ser trabalhados adequadamente para que possam gerar informações relevantes quando buscados.

É por esse motivo que este trabalho foi feito, cuja proposta é de desenvolver um Sistema de Recuperação de Informação junto com um Banco de Dados Relacional para área hospitalar, com uma interface fácil e amigável, para ajudar a diminuir o tempo de busca por informações necessárias – tanto em documentos quanto em dados de pacientes – que sirvam de apoio aos médicos nas suas consultas e outros profissionais da área.

DEVELOPMENT OF A SYSTEM INFORMATION RETRIEVAL FOR THE HOSPITAL AREA

ABSTRACT

All organizations in health generate a great volume of documents contend information on patients. These documents are sources of precious data, that need to be worked adequately that they can generate excellent information when searched.

For this reason that this work was fact, whose proposal is to develop a together System of Information Retrieval with a Relationary Database for hospital area, with an easy and friendly interface, to help to diminish the time of search for necessary information - as much in documents how much in data of patients - that they serve of support to the doctors in its consultations and other professionals of the area.

SUMÁRIO

LISTA DE FIGURAS **1**

1. INTRODUÇÃO **2**

2. REFERENCIAL TEÓRICO **4**

2.1. BANCO DE DADOS 4
2.1.1. Banco de Dados Relacional *4*
2.1.2. Modelagem de dados usando o Modelo Entidade/ Relacionamento (MER) *5*
2.1.3. Sistemas Gerenciadores de Banco de Dados *5*
2.1.4. Banco de Dados e a Web *6*
2.2. A METODOLOGIA OOHDM *7*
2.3. SISTEMA DE RECUPERAÇÃO DE INFORMAÇÃO 8
2.3.1. Paradigma do Sistema de Recuperação de informação *9*
2.3.2. Modelos de Recuperação de Informações *12*
2.3.3. Recuperação de informações textuais *14*
2.3.4. Busca e recuperação *21*
2.3.5. Bibliometria *23*
2.4. RECUPERAÇÃO DE INFORMAÇÃO NAS BIBLIOTECAS DIGITAIS 25

3. METODOLOGIA **29**

3.1. TIPOS DE PESQUISA 29
3.1.1. Observação participante *29*
3.1.2. Pesquisa Documental *30*
3.1.3. Pesquisa Bibliográfica *30*
3.2. FERRAMENTAS UTILIZADAS NO DESENVOLVIMENTO 31
3.3. AMBIENTE DE TRABALHO 32

4. RESULTADOS E DISCUSSÕES **33**

4.1. O HOSPITAL 33
4.2. MODELAGEM 36
4.3. ESTRUTURA E FUNCIONAMENTO DO SISTEMA 38

5. CONCLUSÃO **47**

5.1. TRABALHOS FUTUROS 47

6. BIBLIOGRAFIA **49**

7. ANEXOS **52**

LISTA DE FIGURAS

Figura 1: Esboço da metodologia OOHDM 7
Figura 2 – Paradigma da recuperação de informações 10
Figura 3 – Processo de abstração 11
Figura 4 – Problema no processo de descrição de uma consulta 12
Figura 5 – Função Similaridade 15
Figura 6: Exemplo *Thesaurus* para o termo "computador" 19
Figura 7. Arquivo Invertido utilizando *array* ordenado 20
Figura 8: Diagrama simplificado do processo de busca a informação 21
Figura 9: Base de dados disponíveis na BIREME 27
Figura 10: Pagina principal do Akwanmed 28
Figua 11: Estrutura Administrativa do Hospital Vaz Monteiro 34
Figura 12: Estrutura de Hardware 35
Figura 13: Modelagem Entidade Relacionamento 37
Figura 14: Modelagem OOHDM da página Arquivos do sistema 37
Figura 15: Estrutura Básica do Sistema 38
Figura 16: Tela1 – Página Principal do Sistema 39
Figura 17: Tela2 – Pesquisar (Consultar) arquivos 39
Figura 18: Tela 3 – Resultados da pesquisa em Arquivos 40
Figura 19: Tela 4- Consultar Pacientes 40
Figura 20: Tela 5 – Visualização dos Dados do Paciente selecionado 41
Figura 21: Tela 6 – Orientações de Uso (Ajuda) 41
Figura 22: Tela 7 – Cadastro de Pacientes 42
Figura 23: Tela 8 – Fazer *upload* do arquivo 43
Figura 24: Estrutura interna do Sistema 44
Figura 25: Estrutura do arquivo invertido depois da ordenação 46

1. INTRODUÇÃO

A necessidade de informações vem crescendo a cada dia e o mundo de forma geral está trabalhando com mais informações e dados de todos os tipos. Em parte, essa quantidade de dados e informação é sustentada pelo fenômeno da globalização. Devido a isso, a busca por informações relevantes está cada vez mais em alta nas empresas de modo geral.

Na área hospitalar não é diferente, pois o uso da informática na área médica está aumentando a cada dia. A informatização na área hospitalar começou de forma sutil e hoje é considerada indispensável. Hoje é possível realizar inúmeras tarefas com o computador dentro de um hospital. Por exemplo: Sistemas de agendamentos de pacientes, cadastro de prontuários, sistemas especialistas de auxílio ao diagnóstico, exames que usam imagens como fonte de diagnóstico e pesquisa, entre outros.

Segundo [RUIZ 01] com a popularização da Internet, o acesso ao conhecimento ficou mais fácil em todas as áreas. Atualmente, os profissionais da área da saúde consultam base de dados estatísticos por intermédio de um *browser*, buscam artigos no mundo e tomam rápido conhecimento das últimas descobertas e estudos nas suas respectivas especialidades.

Num ambiente médico é produzido um grande volume de documentos diariamente, que retratam as características, experiências e resultados obtidos. Esses documentos são fontes de dados preciosos, que precisam ser tratados adequadamente para que possam gerar informações úteis quando buscados.

O objetivo deste trabalho é auxiliar os profissionais da área de saúde na busca por essas informações úteis no dia a dia, para isso, foi desenvolvido um Sistema de Recuperação de Informação com um Banco de Dados Relacional para o Hospital Vaz Monteiro de Assistência à Infância e a maternidade, com uma interface fácil e amigável. Diminuindo assim o tempo de busca por

informações relevantes e de pacientes (que estarão armazenadas no Banco de Dados Relacional) que servirão de apoio aos médicos nas suas consultas e à outros profissionais na área de saúde no seu trabalho.

O conteúdo deste projeto está divido da seguinte maneira: no capítulo 2 serão apresentados conceitos de Banco de Dados e Recuperação de Informação, mostrando o funcionamento e a importância de ambos para a realização do projeto.

No Capitulo 3, tem-se a metodologia de desenvolvimento deste projeto, ou seja, o que, quando, como e onde o trabalho foi desenvolvido.

No Capítulo 4 são apresentados os resultados e discussões, com a explicação do funcionamento do sistema, e apresentando algumas telas.

Por fim, serão apresentados a conclusão, Capítulo 5 e a bibliografia, Capítulo 6.

2. REFERENCIAL TEÓRICO

2.1. Banco de Dados

Os Bancos de Dados (BD) e a tecnologia de bancos de dados vêm evoluindo ao longo do tempo, fazendo com que os sistemas de BD se tornem componentes essenciais no cotidiano da sociedade moderna e impulsionando o crescimento do uso de computadores. Ao longo de um dia, quase todos nós encontramos diversas atividades que envolvem alguma interação com um banco de dados. Por exemplo, se vamos a um banco para depositar ou retirar dinheiro, ao supermercado comprar mercadorias, provável que essas atividades envolvam o acesso de alguém a um banco de dados.

Uma definição completa para o termo Banco de Dados é um conjunto de dados armazenados, cujo conteúdo informativo representa, a cada instante, o estado atual de uma determinada aplicação [ELM 02].

Como todas áreas da computação, os sistemas de bancos dados também evoluíram e continuam evoluindo, dia após dia. Antigamente a principal preocupação era em isolar o banco de dados da aplicação, facilitando a vida do programador. Atualmente, a principal preocupação gira em torno da manipulação dos dados complexos.

O tipo de Banco de dados utilizado neste projeto foi o Banco de dados Relacional.

2.1.1. Banco de Dados Relacional

Banco de dados relacionais são o que nós podemos chamar de meio mais eficiente para administração de bases de dados. Muitos programadores que estão migrando para este novo conceito de administração de bases de dados ficam

perplexos com as possibilidades apresentadas pelos meios mais modernos, com novos recursos como, por exemplo, real integridade dos dados e dispositivos de segurança.

Num BD relacional os dados relacionados têm que possuir interesses comuns e têm que ser ligados à realidade. Os dados são matéria-prima de forma crua, fatos que podem ser gravados e que possuem algum significado implícito.

2.1.2. Modelagem de dados usando o Modelo Entidade/ Relacionamento (MER)

O modelo Entidade/Relacionamento (MER) é um modelo de dados conceitual de alto nível gráfico e muito popular, freqüentemente utilizado para o projeto conceitual dos dados, que servirá de base para o projeto físico [ELM 02].

Os conceitos do MER foram projetados para serem compreensíveis a usuários, descartando detalhes de como os dados são armazenados. Atualmente, o MER é usado principalmente durante o processo de projeto da base de dados.

2.1.3. Sistemas Gerenciadores de Banco de Dados

Sistemas Gerenciadores de Bancos de Dados (SGBD ou DBMS – *Database Management System*) são sistemas que gerenciam Banco de Dados, ou são linguagens utilizadas para manter os Banco de Dados [ELM 02].

Em outras palavras, um SGBD é um pacote de programas que permite a definição, construção e manipulação de um Banco de Dados, onde a definição é a especificação e descrição detalhada dos tipos de dados a serem armazenados, a construção é o processo de carga inicial dos dados em um meio de armazenamento controlado pelo SGBD e a manipulação é a execução de operações de consulta, recuperação e atualização de dados específicos.

O SGBD utilizado neste projeto foi MYSQL que está descrito no capítulo 3.

2.1.4. Banco de Dados e a Web

A *World Wide Web* (*WWW*) – popularmente conhecida como "*Web*" – foi originalmente desenvolvida na Suiça CERN (Conselho Europeu para pesquisa nuclear), no início dos anos 1990, como um sistema de serviços de informações hipermídias de grande escala, para que cientistas biólogos compartilhassem informações. Hoje em dia essa tecnologia permite que qualquer pessoa tenha acesso à *internet*, acesso universal a essas informações compartilhadas e a *Web* contém centenas de páginas *Web* ao alcance de milhões de usuários [ELM 02].

A *Web* é um importante fator no planejamento de ambientes de computação, tanto para fornecer acesso externo aos sistemas e informações da empresa para fornecedores como para fins de *marketing* e propaganda. Ao mesmo tempo, devido a requisitos de segurança, empregados de algumas organizações são confinados a operar dentro de Intranets (sub-redes que não podem ser livremente acessados pelo mundo exterior).

A tecnologia atual vem se alternando rapidamente de páginas da *Web* estáticas para páginas dinâmicas, nas quais o conteúdo pode estar em um constante estado de fluxo [ELM 02].

À medida que a *Web* passava por suas transformações mais recentes, tornou-se necessário permitir que usuários acessem não só sistemas de arquivos, mas também Banco de Dados e SGBDs para dar suporte ao processamento de consultas, geração de relatórios e assim por diante.

2.2. A Metodologia OOHDM

Segundo [MAG 02] a metodologia OOHDM (Object Oriented Hypermidia Design Model) considera o processo de desenvolvimento de um aplicativo como um processo contendo quatro atividades, baseando-se em uma mistura de estilos iterativos e incrementais de desenvolvimento, em cada etapa um modelo é construído ou enriquecido. A figura 1 mostra o esboço da metodologia OOHDM.

Atividades	Produtos	Mecanismos	Interesse do Projetos
Modelagem conceitual	Classes, sub-sistemas, relacionamentos, perspectivas de atributos	Classificação, composição, generalização, especialização	Modelagem da semântica do domínio da aplicação
Projeto de Navegação	Nos, elos, acesso, contextos de navegação, transformações navegacionais	Mapeamento entre objetos conceituais e de navegação para descrição da estrutura geral da navegação	Leva em conta o perfil do usuário e a tarefa, ênfase em aspectos cognitivos e arquiteturas
Projeto de Interface Abstrata	Objetos de interface abstrata, relações e eventos externos, transformação de interface	Mapeamento de objetos de navegação e objetos de interface	Modelagem de objetos, implementa metáforas, descrição de interface para objetos navegacionais
Implementação	Aplicação em execução	Aqueles fornecidos pelo ambiente alvo	Desempenho Completitude

Figura 1: Esboço da metodologia OOHDM
Fonte: [MAG 02]

O modelo do projeto é independente da implementação no sentido em que, embora possa levar em consideração alguma configuração de

implementação, não é condicionada por um ambiente de implementação em particular.

2.3. Sistema de Recuperação de Informação

Recuperação de Informação (RI) ou *Information Retrievel* (IR), é a tarefa de encontrar documentos relevantes a partir de um *corpus* ou conjunto de textos em resposta a uma necessidade de informação de um usuário [LEE 98].

Os Sistemas de Recuperação de Informações (SRI) foram originalmente desenvolvidos para gerenciar a enorme quantidade de literatura científica que vinha sendo produzida desde a década de 40, sendo que atualmente a RI é usada em muitas universidades, corporações e livrarias públicas para prover o acesso a livros, jornais e outros documentos; além de ser usada comercialmente em banco de dados que contêm milhões de informações separadas por área [FRA 92].

Todos os Sistemas de Recuperação de Informações possuem como meta: fazer com que o usuário encontre a informação que está precisando rapidamente de modo que este usuário não necessite analisar ele próprio, todas as informações existentes na base de informações [KRU 97].

As informações encontradas pelo sistema de recuperação de informações devem ser relevantes, ou seja, as informações devem ser importantes para o usuário dentro do contexto que ele está procurando.

O objetivo de um sistema de recuperação de informações é de minimizar o custo (*overhead*) de um uma pessoa localizar uma informação necessária. O custo pode ser entendido como o tempo gasto pelo usuário para completar a tarefa de recuperação de uma informação relevante [KOW 97].

Um conceito muito importante em Sistema de Recuperação de Informações é o de documento. Quando falamos em documento logo imaginamos um texto impresso. Documento é mais que isso, um documento

pode ser uma pintura, uma imagem, um vídeo, um gráfico ou qualquer outro objeto que transmita informação para quem o vê ou lê. Um documento pode ser considerado um conjunto de dados, onde cada dado representa uma informação. Não há como falar em documentos sem falar em informações. As informações estão contidas em um documento em forma de dados, onde um dado contém uma informação quando de alguma forma faz com que uma pessoa modifique seu estado de conhecimento, ou seja, um documento contém informações quando alguém lê ou vê este documento e aprende alguma coisa com ele.

O conceito de informação relevante vai um pouco além, pois uma informação pode ser relevante para uma pessoa e não relevante para outra.

Segundo [BAE 99], relevância é a parte central de um sistema de recuperação de informações. O objetivo maior da recuperação de informação é conseguir recuperar todos os documentos que são relevantes a uma consulta de um usuário e o menor número possível de documentos não relevantes.

Relevância está no fato da informação pertencer ao contexto do que o usuário deseja naquele momento [KRU 00]. Stefano Mizzaro afirma que a informação está diretamente relacionada com o usuário, com a sua necessidade de informação e com o momento que isto ocorre. [MIZ 97].

2.3.1. Paradigma do Sistema de Recuperação de informação

Aqui será mostrada de forma abstrata como funciona o processo de recuperação de informação e todos os agentes envolvidos neste processo. O sistema de recuperação de informações funciona como uma interface entre o usuário e os documentos onde estão contidas as informações. Em um sistema de recuperação de informação automático, o sistema é encarregado de receber a consulta feita pelo usuário e compará-la com todos os documentos contidos em sua base de dados e descobrir quais os documentos relevantes para o usuário. A

figura 2, mostra os três pontos chaves que devem ser trabalhados com atenção no paradigma da recuperação de informação são eles: processo de abstração (modelagem do sistema), descrição da necessidade do usuário (linguagem consulta) e processo de *matching* (casamento que o sistema de recuperação de informações faz entre a consulta do usuário e as informações do sistema).

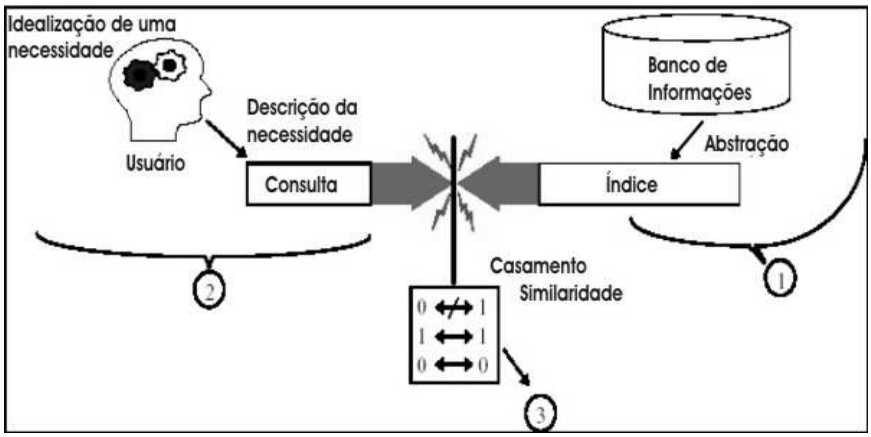


Figura 2 – Paradigma da recuperação de informações
Fonte: [KRU 97]

- Abstração de informações - É através das características de um documento que o SRI é capaz de localizá-lo como relevante ou não para o usuário. O sistema de recuperação de informações de alguma forma deve poder identificar as características de um objeto e descrevê-lo através delas [KRU 00]. Esta descrição é uma modelagem. Em um sistema de recuperação de informações modelar a informação que o sistema irá tratar é uma tarefa importante e difícil, pois vários problemas podem surgir em decorrência de uma modelagem incorreta da informação. A Figura 3 demonstra o processo de abstração, onde as informações são analisadas manualmente ou automaticamente. Após a

análise as características são armazenadas, conforme o modelo, em uma representação interna.

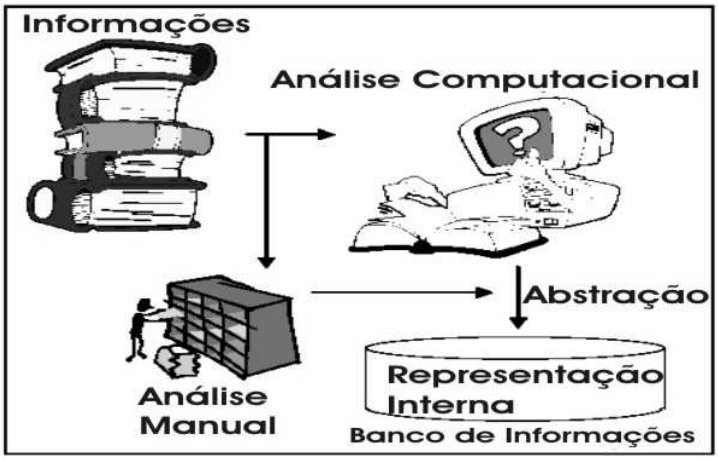


Figura 3 – Processo de abstração
Fonte: [KRU 00]

- Descrição da necessidade do usuário - Dentro de recuperação de informação, o usuário pode ser qualquer um que tenha necessidade de algum tipo de informação e queira buscar estas informações. Muitas vezes o usuário não sabe ao certo o que ele quer e isto dificulta na hora de recuperar as informações relevantes (ver figura 4). A tarefa principal do usuário de um sistema de recuperação de informações é conseguir traduzir a sua necessidade de informação em uma consulta, escrita em uma linguagem fornecida pelo sistema. São através desta consulta feita pelo usuário que o sistema será capaz de determinar quais as informações mais relevantes para este usuário.

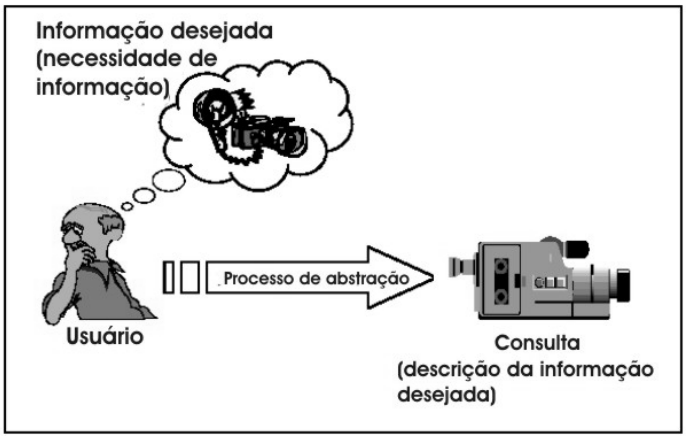


Figura 4 – Problema no processo de descrição de uma consulta
Fonte: [KRU 00]

- O processo casamento (*matching*) - Segundo [KRU 97] é um processo de identificação de quais informações são relevantes para a consulta do usuário. Este processo procura identificar a similaridade entre as informações armazenadas no sistema e a descrição de informação que o usuário deseja. Muitos problemas podem surgir com este processo, pois para fazer a consulta o usuário tem que abstrair a informação, através do processo de casamento o sistema de recuperação de informações também faz uma abstração das informações, o que pode acarretar na perda de características importantes da informação.

2.3.2. Modelos de Recuperação de Informações

Será apresentada aqui uma breve descrição dos modelos clássicos da recuperação de informações, que apresentam estratégias de busca de documentos relevantes para uma consulta. Apesar de terem sido desenvolvidos dentro do

escopo de documentos textuais, os modelos de recuperação de informações podem ser utilizados em qualquer tipo de documento.Os modelos clássicos são:

- **Modelo *booleano*** – baseia-se na teoria de conjuntos, onde cada documento é representado por um conjunto de palavras (termos) [KRU 97]. Quando uma pessoa faz a consulta, ela deve indicar as palavras (elementos) que os documentos (conjuntos) resultantes devem ter para que sejam retornados. Sendo assim, somente os documentos que possuem as mesmas palavras (interseção) que a consulta serão retornados. Os operadores lógicos mais comuns utilizados na consulta são o *and* (união), *or* (interseção) e o *not* (negação). Em teoria o modelo *booleano* é um dos que apresenta melhores resultados, pois permite que o usuário especifique consultas complexas, detalhadas e bem definidas [KRU 00].
- **Modelo vetorial** – cada documento é representado por um vetor associado que indica o grau de importância (denominada peso) desse no documento [KRU 00]. As palavras armazenadas nesse vetor são todas as palavras da coleção e não somente as palavras presentes no documento. Os documentos são representados como vetores no espaço de termos, onde termos são ocorrências únicas nos documentos. As consultas são representadas da mesma forma e a distância entre vetores pode medir a relação do documento com uma consulta.
- **Modelo probabilístico** – neste modelo, busca-se saber a probabilidade de um certo documento ‘A’ ser ou não ser relevante para uma dada consulta ‘B’. Tal informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer da coleção [RIJ 79]. Para saber a relevância de um documento em

relação a uma determinada consulta é feito um cálculo de probabilidade de cada um dos documentos da coleção ser relevante à consulta dada. Para cada termo da consulta seu grau de relevância é identificado no documento. A informação de relevância de um termo é calculada estatisticamente com bases na freqüência desse termo nos documentos da coleção.

2.3.3. Recuperação de informações textuais

Os sistemas de recuperação de informações são classificados de acordo com o tipo de informação que manipulam. Há um sistema de recuperação de informação correspondente para cada tipo de informação (documento). Existem informações do tipo visuais (imagens), textuais (textos) e multimídia (vídeos e sons).

Os sistemas de recuperação de informações em forma de texto (textuais) que segundo [KRU 00] são sistemas desenvolvidos para indexar e recuperar documentos do tipo textual, ou seja, documentos cujas informações estão descritas através da linguagem natural (linguagem normalmente utilizada pelo homem para comunicação. Ex: Português, inglês...).

A busca de informações textuais é diferente da busca de informações realizadas nos Banco de Dados tradicionais. Para Salton, os Bancos de Dados tradicionais preocupam-se com o armazenamento, manutenção e a recuperação de informações disponíveis explicitamente no sistema. O que não é o caso dos Bancos de Dados Textuais onde a informação está implícita na forma de linguagem natural [SAL 83].

Exemplificando, quando queremos buscar dados de uma pessoa em um Banco de Dados Tradicional, percorremos no Banco de Dados a tabela que possui o *Atributo* nome e localizamos o registro (*Tupla*) que possui o nome

desejado. Isso não acontece um Banco de Dados Textual, pois os dados não estariam distribuídos de forma de tabela, e o texto é uma seqüência de caracteres, ou seja, não possui atributos, impossibilitando a pesquisa do nome. Para que a localização fosse feita em uma Base de Dados Textuais seria necessário analisar *caracter-por-caracter*, o que não é conveniente.

Para [KRU 97] uma forma eficiente de acesso aos dados de um documento é através do contexto, ou seja, o assunto que o documento se refere. Para identificar este assunto é necessário analisar as palavras (termos) que este documento contém. Sendo assim os documentos são identificados de acordo com os termos que eles contêm, portanto, a localização de um documento desejado pelo usuário dá-se a partir da identificação da similaridade entre os termos fornecidos pelo usuário e os termos que identificam os documentos, como mostra a Figura 5.

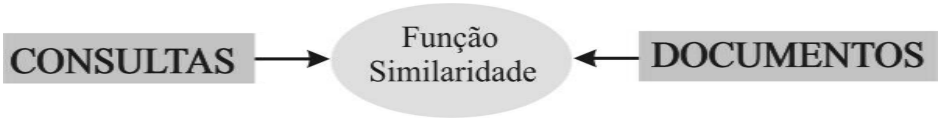


Figura 5 – Função Similaridade
Fonte: Elaborado pelo autor

Teoricamente pode ser feita uma comparação direta entre estes termos, mas na prática é difícil estabelecer esta relação de similaridade entre estes termos devido a alguns problemas, como o Problema do Vocabulário, onde as palavras utilizadas pelo sistema podem ser diferentes das palavras utilizadas pelo usuário, e o Problema da Busca Incerta (*Search Uncertainly*), onde os usuários não sabem quais são as melhores palavras que identificam o assunto que querem localizar [KRU 97].

Segundo [BAE 99] usar o conjunto de todas as palavras numa coleção para indexar seus documentos gera muito ruído para a tarefa de recuperação. Para reduzir o ruído deve-se diminuir o conjunto de palavras usadas para

representar um documento, pois nem todas as palavras em um texto são igualmente importantes. Geralmente substantivos (ou grupo de substantivos) representam mais o conteúdo de um documento do que outros tipos de palavras.

O pré-processamento dos documentos em uma coleção simplesmente diminui o tamanho do vocabulário, melhorando a performance do SRI, mas pode gerar alguns problemas, como: Usuário que usa palavras que não aparecem no vocabulário devido este controle, por exemplo, uma consulta com os termos "Um conto de Fadas", poderiam não ser retornado nenhum documento mesmo se tivesse a frase idêntica neles, pelo fato dos termos "Um" e "de" não aparecerem no vocabulário.

É importante que fique claro para o usuário do SRI que tipos de palavras ele deve usar em suas consultas [BAE 99].

Por este problema, máquinas de busca na *web* costumam ignorar o pré-processamento e fazer a indexação *full text*, ou seja, todas as palavras do texto são indexadas, pois apesar do ruído, é mais eficiente para usuários leigos (da internet).

2.3.3.1. Pré-processamento em Documentos

Pode ser dividido basicamente em cinco operações sobre textos:

- **Análise Léxica** - A Análise Léxica, nada mais é do que a identificação de palavras presentes nos documentos, analisando-se as seqüências de caracteres no texto. Salton [SAL 83] aconselha fazer um *Dicionary lookup*, ou seja, comparar as seqüências de caracteres retiradas do texto com um dicionário a fim de validar sua existência e corrigir possíveis erros ortográficos. Este processo de validação torna-se bastante útil, especialmente quando o documento apresenta muitos caracteres inválidos ou

palavras com erros gramaticais [KRU 97]. Diversas técnicas adicionais de padronização podem ser aplicadas [BAE 92]: a passagem de todos os caracteres para maiúsculo (ou minúsculo); a substituição de múltiplos espaços e tabulações por um único espaço; a padronização de datas e números; a eliminação de hífens. Se uma técnica for adotada, ela também deve ser utilizada em cima da consulta do usuário.

- **Eliminação de Stopwords** - Existem algumas palavras presentes em um documento textual que são utilizadas com o intuito de conectar frases. Essas e outras palavras, pertencentes a classes de palavras cuja finalidade é auxiliar a estruturação da linguagem (tais como conjunções e preposições), não necessitam ser incluídas na estrutura de índice [KRU 00]. Segundo [BAE 99], esta etapa tem como objetivo filtrar palavras com valores discriminatórios baixos para a tarefa de recuperação. Também pode ser considerada uma técnica de compressão de textos. Pode diminuir o tamanho do texto em até 40%. Há estudos que oferecem listas de *stopwords*, contendo todas as palavras que não devem ser indexadas. A esta estrutura foi atribuído o nome de *Stoplist* ou D*icionários Negativos* e podem ser utilizadas livremente na elaboração de ferramentas que realizem o processo de remoção de *stopwords*.

- **Stemming** - O *Stemming* (corte) tem como objetivo remover prefixos e sufixos, permitindo assim a recuperação de variações sintáticas das palavras, ou seja, é uma técnica que consiste em identificar os radicais das palavras e adicioná-las no arquivo de índice desta forma. Assim, todas as palavras que possuem o mesmo radical, e portanto com significados similares são reconhecidos pelo mesmo identificador. Existem conflitos na área sobre o fato de *stemming* trazer melhorias de desempenho [BAE 99]. Uma vantagem

segundo [FRA 92] é que além de eliminar as variações morfológicas das palavras e aumentar a precisão das consultas, o método de *stemming* também é capaz de reduzir o tamanho de um índice em até 50%. A desvantagem deste método é que ele pode tornar as palavras muito abrangentes, pois os termos específicos são eliminados. Neste caso os documentos específicos não são recuperados, Em outras palavras, pode melhorar a abrangência, mais piorar a precisão de resposta a uma consulta. Devido a isso, sugere-se que as palavras sejam indexadas utilizando a forma ortográfica encontrada nos documentos e que o usuário encarregue-se de especificar que variações morfológicas ele deseja durante o processo de consulta [KRU 00].

- **Seleção de termos de indexação -** Os arquivos de índice de um SRI geralmente consomem muito espaço, podendo chegar a 300% do espaço correspondente aos documentos originais. Esse tamanho pode ser diminuído excluindo-se alguns termos de menor importância dos documentos. Assim, há uma redução no espaço de dimensões que modelam os documentos [KRU 00]. A seleção de termos de indexação tem como objetivo determinar que palavras serão utilizadas como elementos de indexação (substantivos são mais representativos que adjetivos, verbos, advérbios, etc.). As técnicas de seleção de termos relevantes podem ser baseadas no peso dos termos ou na sua posição sintática. Podem ser usados todas as palavras ou escolher algumas mais significativas. Esta etapa pode ser feita de forma manual ou automática.

- **Construção de estruturas de categorização de termos -** Estas construções, tais como *Thesaurus*, tem como objetivo extrair estruturas diretamente representadas no texto, para permitir, por exemplo, a expansão

de consultas. A origem do termo *thesaurus* deve-se à obra elaborada por Roget [ROG 58], com primeira versão editada em 1852. O Thesaurus é uma estrutura hierárquica de palavras, permitindo que o usuário descubra os relacionamentos entre estas palavras, as quais são geralmente agrupadas em classes onde cada classe possui um termo chave, que a identifica. Na figura 6 é apresentado um exemplo de um *Thesaurus* para o termo "computador". Segundo [SAL 83] as vantagens de utilizar-se esta estrutura são: Primeiro, ajudar o usuário na formulação de consultas, indicando termos mais adequados que possivelmente recuperem um número de informação maior; E segundo familiarizar o usuário com o vocabulário do sistema. O *Thesaurus* pode ser construído manualmente, por um especialista que identifica os relacionamentos entre as classes de palavras, ou automaticamente, através de técnicas estatísticas de correlação que identificam relações entre palavras e colocam-nas em classes. Essas técnicas podem ser encontradas em [CHE 96] e [SAL 83].

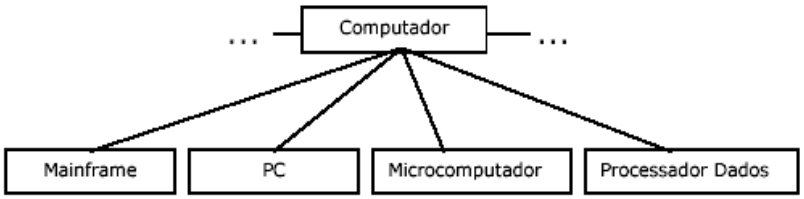


Figura 6: Exemplo *Thesaurus* para o termo "computador"
Fonte: Elaborado pelo autor (adaptada de [BAE 99])

2.3.3.2. Arquivos de indexação

As três principais técnicas de construção de arquivos de indexação são: as árvores desufixos, os arquivos de assinatura e os arquivos invertidos [BAE 99]. A seguir será descrito somente arquivo invertido, o qual foi utilizado na implementação deste projeto.

Arquivos Invertidos

Segundo [BAE 92], arquivos invertidos podem ser definidos como uma lista ordenada (ou indexada) de palavras chaves (ou atributos), onde cada um destes termos contém *links* para o documento que o possui, como pode ser observado na figura 7.

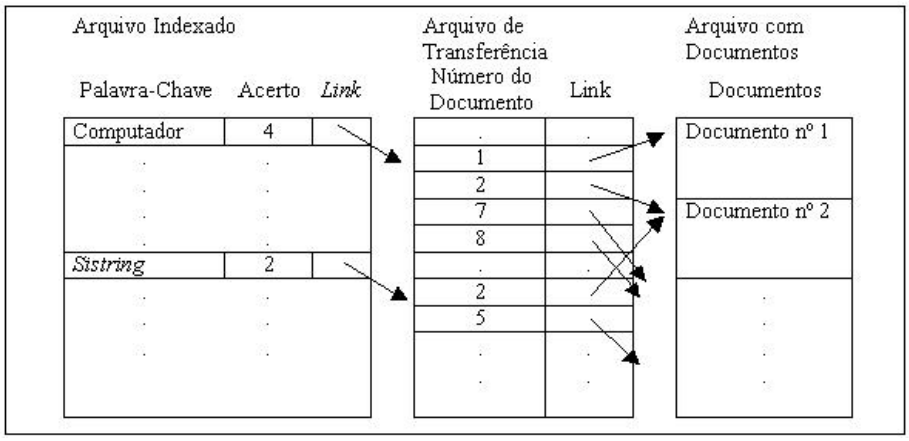


Figura 7. Arquivo Invertido utilizando *array* ordenado
Fonte: [BAE 92]

O uso de arquivos invertidos provê um eficiente método de pesquisa. Porém, [BAE 92] alerta algumas restrições para este método que, além de necessitar de mais espaço para armazenagem (as estruturas de dados dos arquivos invertidos aumentam de 10 % até mais de 100% o tamanho do arquivo em si), necessitam de atualização do índice de acordo com eventuais mudanças no conjunto de dados.

Devido à sua rapidez de acesso e à sua facilidade de identificação de documentos relevantes a um termo, essa estrutura é uma das mais utilizadas em Sistemas de Recuperação de Informações [KOW 97].

2.3.4. Busca e recuperação

Quando o usuário passa a interagir com um SRI diz-se que ele está buscando informações (*information seeking*). Nesta fase o usuário tem que mapear sua necessidade de informação para uma linguagem abstrata, a linguagem utilizada pelo SRI, a fim de descrevê-la. Essa descrição é a única forma de especificação que o SRI tem da necessidade de informação do usuário. É através dela que o SRI vai poder identificar os itens de que o usuário necessita e analisar se esses itens são relevantes para ele [KRU 00]. O processo de busca a informação pode ser representado pelo diagrama apresentado na figura 8.

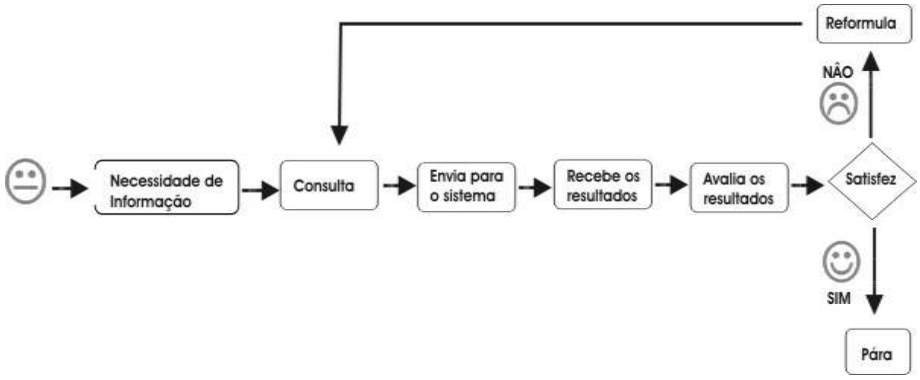


Figura 8: Diagrama simplificado do processo de busca a informação
Fonte: Elaborado pelo autor (adaptado de [BAE 99])

Durante esse processo de especificação diversos problemas podem ocorrer. Esses problemas variam desde o conhecimento do usuário em relação ao sistema, passando por problemas de especificação ou abstração (pois é difícil descrever uma necessidade através de um formalismo) e finalizando com o fato do usuário não ter certeza sobre sua real necessidade de informação (se o usuário necessita informação sobre algum assunto é porque ele não sabe muito sobre ele).

O processo pode ser dividido em três etapas que são a formulação de consultas, a identificação de itens relevantes e a visualização dos resultados.

2.3.4.1. Formulação de consulta

Formulação de consulta é o formalismo como qual o usuário comunica-se com o sistema. É nela que o usuário especifica sua necessidade de informação, definindo a que assuntos os documentos devem pertencer quando retornados.

Segundo [BAE 99], existem diferentes tipos de consultas que podem ser atribuídas a SRI, dependendo do modelo de recuperação que o sistema adote, isto é, um sistema *full text* não responderá ao mesmo tipo de consulta de um sistema baseado em ordenação de palavras chaves (como em máquinas de busca *web*).

Uma consulta é formulada a partir da necessidade de informação de um usuário. Nesta forma mais simples, uma consulta é composta de palavras chaves e documentos que contenham estas palavras são procurados. Consultas baseadas em palavras chaves são populares porque são intuitivas, fáceis de expressar e permitem ordenação rápida. Assim, uma consulta pode ser composta simplesmente por uma palavra (*single-word*), ou pode ser uma combinação complexa de operação envolvendo várias palavras (*multiple-word*). Em ambos os casos este tipo de consulta é chamada de consulta básica [BAE 99].

2.3.4.2. Identificação de itens relevantes (técnicas de "casamento")

O SRI faz a identificação dos documentos relevantes a uma consulta comparando as características da consulta com as características dos documentos presentes na base de documentos.

Essa análise de similaridade de características geralmente depende do modelo conceitual adotado pelo SRI e é feita por uma classe de funções de similaridade (que foi explicada na secção 2.3).

2.3.4.3. Visualização e análise dos resultados

Os algoritmos de busca não conseguem obter informações relevantes com 100% de abrangência e precisão [KOW 97].

Existem diversas formas do SRI apresentar o resultado para o usuário e existem diversas informações que podem ser mostradas para o usuário enquanto ele interage com o SRI [KOW97].

Uma dessas formas é a navegação (*browsing*), que consiste em mostrar para o usuário uma lista de documentos com seus respectivos índices. Tal lista pode ser *navegada* pelo usuário e os documentos relevantes selecionados de modo que o usuário possa ver seu conteúdo. Os documentos podem possuir *links* entre eles, indicando sua similaridade.

Geralmente essa lista é ordenada em uma espécie de ranking onde os documentos mais relevantes aparecem primeiro [HAR 92]. É necessário que o SRI informe para o usuário o porquê de eles terem sido recuperados, uma forma é selecionar trechos do documento que contenham as palavras da consulta e mostrar para o usuário. Essa técnica é conhecida por seleção (*highlight*) [KOW97].

2.3.5. Bibliometria

Bibliometria (sub-área da biblioteconomia) é uma área que contém métricas para avaliar a eficiência de um sistema de recuperação de informação. O objetivo por trás da utilização de medidas de eficiência é indicar se o sistema

realmente consegue recuperar grandes informações relevantes para o usuário, ao mesmo tempo em que consegue excluir os itens irrelevantes.

Pelo fato destas métricas necessitarem de informações como número de documentos relevantes a uma consulta e o sistema não ter como fornecê-las, estes tipos de métricas não usadas na prática, principalmente em sistemas de recuperação de informações comerciais.

Segundo [KRU 00], existem algumas coleções públicas de documentos preparadas especialmente para processo de avaliação de sistemas de recuperação de informações (principalmente acadêmicos). Dentre estas coleções estão as coleções TREC (*Text Retrievel Conference*), que oferecem consultas pré-definidas e alguns conjuntos de documentos relevantes a cada uma das consultas. Assim, o resultado de uma busca realizada em um sistema pode ser comparada com o conjunto de documentos que ela deveria retornar.

Dentro da computação as métricas mais importantes para a avaliação do resultado e do desempenho de um SRI são: abrangência (*recall*) e precisão (*precision*), *fall*-out e *effort* [KOW 97 e SAL 83]. Dessas, *recall* e *precision* são as mais utilizadas e estão apresentadas a seguir.

- Abrangência (*Recall*) – serve para indicar a proporção de itens relevantes, recuperados em uma resposta a uma consulta do usuário. É utilizada para medir a habilidade do sistema recuperar todos os itens relevantes. A fórmula usada é a seguinte:

$$Recall = \frac{n\_recuperados\_relevantes}{n\_possíveis\_recuperados}$$

- Precisão (*Precision*) – mede a proporção de itens recuperados que são realmente relevantes, ou seja, mede a habilidade do sistema em manter

os documentos irrelevantes fora do resultado da consulta. A fórmula que mede a precisão é a seguinte:

$$Precision = \frac{total\_inf\_relevantes\_encontradas}{total\_inf\_encontradas}$$

Apesar destas duas medidas serem muito usadas, há quem critique sua utilidade. Isso porque a abrangência não garante que os documentos recuperados sejam realmente úteis para o usuário. Geralmente a quantidade de documentos relevantes é estimada através de métodos estatísticos, onde somente alguns documentos são analisados e sua categoria identificada. Existem outras formas de medir a eficiência de um SRI, mas geralmente estas medidas são mais difíceis de serem interpretadas e exige um esforço computacional maior.

Alguns fatores devem ser levados em conta quando se for construir um sistema de recuperação de informações. Toda pessoa que fizer uma consulta deve ser avisada se conseguiu ou não o que queria e avisar também porque este resultado foi obtido, pois muitos usuários podem cometer erros de digitação durante a consulta expressando incorretamente sua necessidade. Em [SAL 83] completa-se que é necessário analisar também o esforço intelectual ou físico realizado pelo usuário na elaboração de uma consulta, na condução da busca e na análise dos resultados.

2.4. Recuperação de Informação nas Bibliotecas Digitais

Os gregos se esforçaram para construir uma biblioteca e conseguiram reunir mais de 500.000 volumes na cidade de Alexandria. Segundo a tradição, todos os mercadores que passassem por lá deveriam colocar seus papiros à disposição para cópia. Surgiam assim as bibliotecas, que desde a sua criação até hoje são os maiores centros de informação do mundo guardando tesouros

preciosíssimos, considerados atualmente como símbolo de poder - a informação. Porém, à medida que o tempo vai avançando, novas mudanças são exigidas dos organismos para que se acompanhe o processo evolutivo [CUN 99].

As bibliotecas estão entre as primeiras instituições a adotar sistemas de recuperação de informações. Na primeira geração tais sistemas consistiam basicamente da automação de tecnologias existentes (tais como catálogos) e permitiam pesquisas baseadas no nome do autor e no título. Na segunda geração aumentaram-se as funcionalidades de busca que foram adicionadas com pesquisas permitidas por palavra-chave, assunto e algumas funcionalidades de busca mais complexa. Na terceira geração, que é a que esta se desenvolvendo, o foco esta sendo dado a melhoras nas interfaces gráficas, características de hipertexto, formatos eletrônicos e sistemas de arquitetura aberta [BAE 99].

O conceito de biblioteca digital (BDI) representa um processo gradual e evolutivo como resultante da utilização do computador nas últimas décadas [CUN 00]. Elas surgem como uma resposta ao fenômeno da explosão da quantidade de informação, disponíveis na rede Internet. Sendo este um dos maiores fenômenos em termos de processamento da informação no século XX. Uma biblioteca digital não necessita necessariamente conter o conteúdo das informações, mas sim, prover acesso até elas.

Na área médica existem várias bibliotecas digitais prontas que servem de consulta para todos os interessados na área de saúde. Uma delas é a Biblioteca Virtual em Saúde – BIREME (*www.bireme.br*), nesta biblioteca digital pode-se fazer pesquisas em várias bases de dados diferentes, como pode ser mostrado na Figura 9. Uma delas é a MEDLINE, que é uma base de dados da literatura internacional na área médica e biomédica, produzida pela NLM (*National Library of Medicine*, USA). Esta biblioteca, junto com suas base de são sistemas complexos que permitem o usuário encontrar a informação desejada através de diversas opções de refinamento. Porém para um usuário comum, são sistemas

complexos por possuírem muitas características específicas de sistema de informação bibliográfica. Este fato torna a utilização desses sistemas uma tarefa árdua, principalmente para iniciantes.

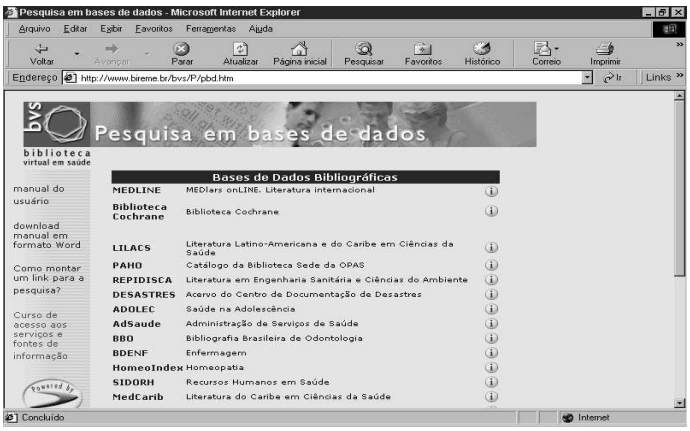


Figura 9: Base de dados disponíveis na BIREME
Fonte: http://www.bireme.br/bvs/P/pbd.htm

Um outro exemplo é o sistema de busca especializado na área de saúde chamado *Akwanmed* (*www.Akwanmed.com.br*). Esse sistema une as características dos sistemas de busca tradicionais da Internet à organização em estruturas hierárquicas dos diretórios, permitindo o acesso integrado a mais de 12 milhões de referências a documentos existentes em portais e bibliotecas médicas digitais. Além disso, o sistema possibilita que a busca seja feita através de conceitos e temas na área de saúde.

A Figura 10, mostra a página principal do Akwanmed.

Figura 10: Pagina principal do Akwanmed
Fonte: http://www.akwanmed.com.br/

3. METODOLOGIA

3.1. Tipos de Pesquisa

Pesquisa é uma indagação minuciosa ou exame crítico e exaustivo na procura de fatos e princípios; uma diligente busca para averiguar algo. A pesquisa não é apenas procurar a verdade; é encontrar respostas para questões propostas, utilizando métodos científicos [MAR 82].

A pesquisa sempre parte de um tipo de problema, de uma interrogação. Dessa maneira, ela vai responder as necessidades de conhecimento de certo problema ou fenômeno.

Os métodos de pesquisas utilizados neste trabalho foram a observação participante, a pesquisa documental e pesquisa bibliográfica (os dois últimos utilizados para coleta de dados).

3.1.1. Observação participante

Este tipo de pesquisa consiste na participação real do pesquisador com a comunidade ou grupo (Hospital Vaz Monteiro). O observador se incorpora ao grupo, confunde-se com ele. Fica próximo quando um membro do grupo esta estudando e participa das atividades [MAR 82].

Em geral, são apontadas duas formas de observação participante:

- Natural: o observador pertence à mesma comunidade ou grupo que investiga;
- Artificial: o observador interage-se ao grupo com a finalidade de obter informações;

A observação participante foi realizada junto aos funcionários do Hospital Vaz Monteiro com o intuito de identificar as necessidades dos mesmos

em relação ao propósito do projeto e colocar em prática no sistema. A externalização do conhecimento foi feita através de diálogos e observações.

3.1.2. Pesquisa Documental

Documentos são todos os materiais escritos que podem servir como fonte de informação para a pesquisa científica e são divididos em:

- arquivos públicos – que podem ser nacionais, estaduais e municipais e as informações encontradas neste tipo de arquivo são muito amplas e de grande utilidade para a pesquisa científica.
- arquivos particulares – que pertencem a instituições de ordem privada ou a domicílios particulares como bancos, igrejas, indústrias entre outras;

Neste trabalho foi utilizada a documentação de arquivos particulares do Hospital Vaz Monteiro. Através da análise desta documentação foi possível verificar quais arquivos eram mais importantes para serem buscados no sistema, saber um pouco mais sobre o hospital para melhor desenvolver o trabalho.

3.1.3. Pesquisa Bibliográfica

O motivo da escolha deste tipo de pesquisa é porque este trata do levantamento da bibliografia publicada e que tenha relação com o que foi proposto. Tendo como finalidade, colocar os pesquisadores em contato direto com maioria daquilo que foi escrito sobre o assunto.

Neste trabalho foi realizado um levantamento bibliográfico tanto em livros quanto em periódicos, trabalhos acadêmicos e *sites* da *Internet* que estivessem relacionados com o tema do trabalho a fim de obter idéias diferentes e inovadoras nas áreas.

Dentro deste levantamento bibliográfico foram incluídas pesquisas que definiram a viabilidade entre usar ou não determinadas tecnologias e ferramentas na construção do sistema, tais como métodos de indexação, linguagens mais viáveis, entre outros.

3.2. Ferramentas Utilizadas no Desenvolvimento

Para modelar a estrutura do Banco de Dados (que contêm informações referentes aos pacientes) utilizado no projeto, optou-se pelo modelo MER (Modelo Entidade/Relacionamento), por ser um modelo conceitual de alto nível gráfico, muito popular e freqüentemente utilizado.

A modelagem do sistema (a interface), foi feita utilizando a modelagem OOHDM (*Object-Oriented Hypermedia Design Method*). Foi utilizada esta pelo fato desta ser uma nova abordagem para representação de aplicação hipermídia.

Para fazer esta modelagem foi utilizado o *ArgoUML*, por ser um software gratuito, de fácil instalação, configuração e utilização. É uma ferramenta Java que possui uma linguagem que permite uma modelagem orientada a objeto eficiente em termos de diagrama.

O Sistema de Recuperação de Informação do Hospital Vaz Monteiro foi desenvolvido na plataforma *Web* e como tal, é acessado através de um *browser*, sendo composta por interface gráfica padrão *web*.

A interface da aplicação foi desenvolvida utilizando o software *Dreamweaver MX da Macromedia Ind.*. Ele foi escolhido devido a sua portabilidade e por ser um dos programas de *desing web* mais fáceis de se manipular. Segundo [LOW 01] O *Dreamweaver* também emite relatórios, em *HTML*, permite a administração remota do *site*, contém geração automática de rotinas em *Javascript*, criação e uso de *Templates*, para a geração com maior produtividade de páginas de seu *site*.

O servidor *Web* utilizado foi o *Apache*, por ser de fácil instalação e utilização

Como linguagem de programação foi usada PHP, por ser uma linguagem de programação *Server-Side scripts* para criar *sites* dinâmicos, os quais retornam para o cliente uma página criada em tempo real [JUN 00]. A escolha por usar PHP se deve ao fato da inexistência de custo e a facilidade de uso.

O Sistema Gerenciador de Banco de Dados (SGBD) utilizado foi o *MySQL*., por ser um banco de dados relacional que utiliza linguagem SQL (*Structured Query Language*) e pelas suas inúmeras vantagens, dentre elas: é uma tecnologia gratuita, instalação simples, administração é fácil, é uma tecnologia multiplataforma, entre outros.

3.3. Ambiente de Trabalho

Este sistema foi desenvolvido no Hospital Vaz Monteiro de Assistência à Infância e à Maternidade, inaugurado em 10 de fevereiro de 1946, situado na cidade de Lavras, sendo considerado de Utilidade Pública Federal, Estadual e Municipal. Regido por Estatuto e dirigido por Diretoria leiga não remunerada eleita bienalmente. Mais informações na seção 4.1.

4. RESULTADOS E DISCUSSÕES

4.1. O Hospital

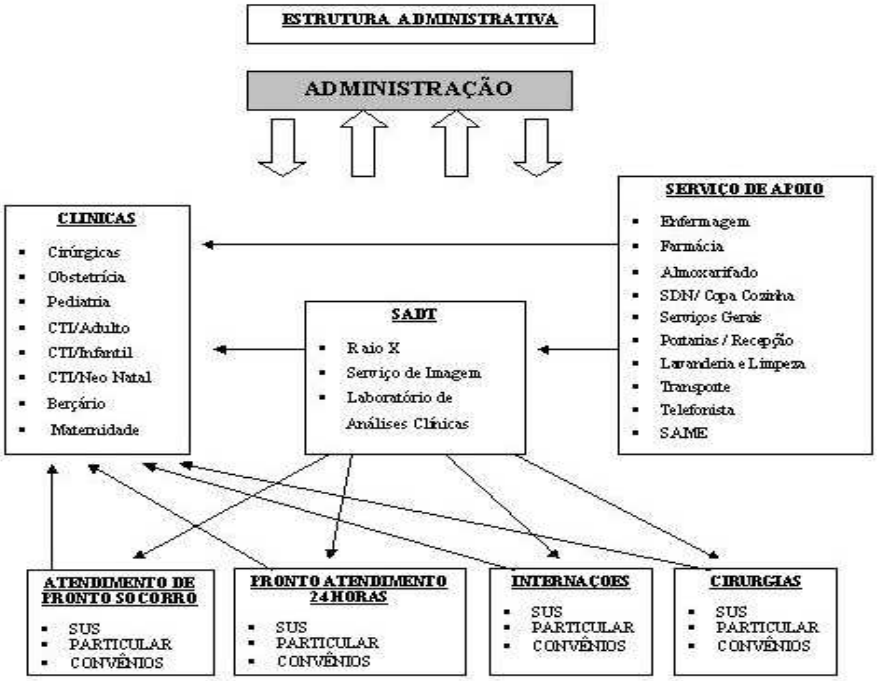
O Hospital Vaz Monteiro de Assistência a Infância e à Maternidade (HVM) foi fundado em 1941, por três homens de formações e origens diferentes. Tinha como objetivos prestar assistência à maternidade e ao recém-nascido, bem como fornecer vacinação às crianças, cujos índices de mortalidade preocupavam os fundadores.

Apesar de sua origem estar ligada a assistência à infância e à maternidade, o hospital hoje atende muitas especialidades, dentre elas: ginecologia, pneumologia, cardiologia, urologia, oftalmologia, entre outras.

Para atender todas as especialidades, O HVM possui 96 leitos assim distribuídos.

- Ala C – 13 leitos distribuídos em 4 quartos com banheiro completo em cada um;
- Ala A – 8 apartamentos com um banheiro em cada unidade, um leito para acompanhante;
- Ala B – 21 leitos distribuídos em 9 quartos com banheiro completo;
- Maternidade – 12 leitos distribuídos em 8 quartos com banheiros, sendo que 4 são apartamentos;
- Berçário – 10 leitos e três incubadoras;
- Pediatria – 21 leitos distribuídos em 9 quartos com banheiro, sendo 2 apartamentos;
- CTI Adulto – 5 leitos equipados;
- CTI Infantil e NeoNatal – 7 leitos, sendo 3 infantis, 1 de isolamento e 3 incubadoras;

O Hospital Vaz Monteiro de Assistência à Infância e à Maternidade tem uma estrutura bastante burocrática, pois tem como ponto chave o seu núcleo operacional formado basicamente pelos médicos e enfermeiros. Seu quadro de funcionários conta com um total de 180 empregados, divididos entre pessoal de escritório, enfermagem, serviços gerais, copa/cozinha, administração e suporte técnico [CAR 03]. A Figura 11 mostra a estrutura administrativa do Hospital.



Figua 11: Estrutura Administrativa do Hospital Vaz Monteiro
Fonte: [CAR 2003]

O seu corpo clínico é formado por 76 médicos de diferentes áreas, que trazem em sua bagagem os recursos e conhecimentos atuais necessários para o melhor desempenho em sua função.

A empresa possui uma boa infra-estrutura tecnológica, quando comparada a outros estabelecimentos da cidade e região, que também optaram por informatizar-se. No total são 16 computadores e 15 impressoras, divididos por setores como mostrado na Figura 12. Os computadores são ligados em rede, conectados a internet de alta velocidade durante 24 horas, para melhor comunicação. Todo o serviço é realizado na plataforma Windows.

SETOR	COMPUTADOR	IMPRESSORA
Servidor	D 1000	
Contabilidade	PENTIUN 1	LX 300
Recursos Humanos	PENTIUN 1	LX 300
Administração	DURON 750	HP 640
Tesouraria	CELERON 300	HP 640
Recepção	CELERON 300	LX 300
Laboratório 1	PENTIUN 1	LX 300
Laboratório 2	PENTIUN 1	LX 300
Laboratório 3	PENTIUN 1	LX 300
Portaria 1	CELERON 300	LX 300
Portaria2	PENTIUN 3	LX 300
Raio X	PENTIUN 1	APOLO
Farmácia	ATLHON 700	LX 300
Faturamento 1	DURON 850	LX 300
Faturamento 2	K6/2	LX 300
Faturamento 3	DURON 850	LX 300

Figura 12: Estrutura de Hardware
Fonte: [CAR 03]

Na busca continuada para melhor servir, o hospital vem cada vez mais se capacitando e especializando no atendimento a pacientes, sendo hoje uma referência na região em diversas especialidades, tanto pela qualidade do serviço que vem prestando, assim como pelos recursos que vem colocando à disposição.

4.2. Modelagem

O SRI (Sistema de Recuperação de Informação) desenvolvido, possui um Banco de Dados que é independente do sistema, ou seja, é simplesmente um Banco de Dados comum para armazenar informações referentes aos pacientes. Neste Banco de Dados não é usada nenhuma técnica de recuperação de informação, o cadastro, armazenamento e consulta é feito utilizando comandos SQL.

Para se fazer a modelagem do SRI (Sistema de Recuperação de Informação) desenvolvido, foi necessário dividir o sistema em duas partes. A primeira é o Banco de Dados que possui as informações referentes aos pacientes, para esta parte a modelagem foi feita através do MER (*Modelagem Entidade/Relacionamento)*. A segunda parte é a interface do sistema e foi usada a modelagem OOHDM.

A Figura 13 mostra a modelagem do Banco de Dados criado para cadastrar, alterar, consultar as informações referentes aos pacientes.

Pelo fato do sistema ter sido desenvolvido para a plataforma *Web* a modelagem do sistema foi feita por páginas. Na Figura 14 é mostrada a modelagem OOHDM da página Arquivos, através desta página o usuário pode acessar as outras páginas, inserir arquivo ou buscar arquivos.

Quando inserimos arquivos através do sistema, estamos simplesmente carregando este arquivo no *servidor* (computador central), não existe um Banco de Dados que armazena estes arquivos. É chamado de base de dados (ou base de informação) o caminho (diretório) onde os arquivos são armazenados.

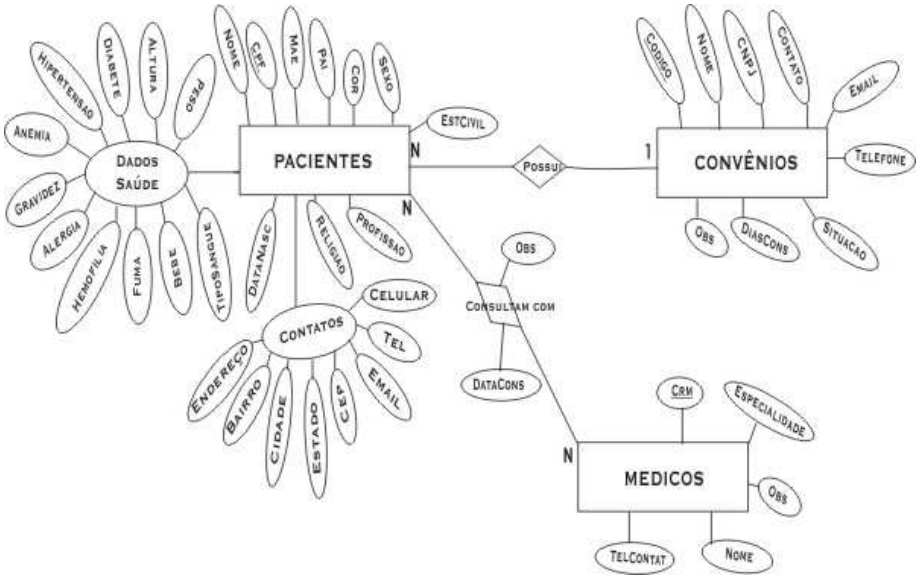


Figura 13: Modelagem Entidade Relacionamento
Fonte: Dados da Pesquisa, 2003

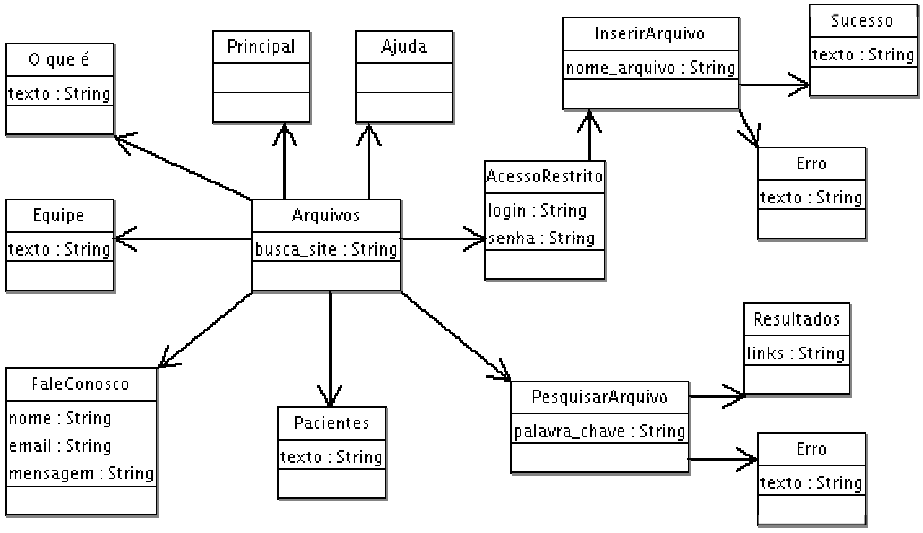


Figura 14: Modelagem OOHDM da página Arquivos do sistema
Fonte: Dados da Pesquisa, 2003

As modelagens feitas em OOHDM das outras páginas do sistema podem ser consultadas em Anexo A.

4.3. Estrutura e Funcionamento do sistema

Para melhor explicação, a estrutura do sistema foi dividida em duas partes: Estrutura Básica e Estrutura Interna.

- **Estrutura Básica**

Através desta estrutura podemos ter uma visão geral da interface do sistema.

Como mostra a Figura 15, o sistema está dividi do em 3 partes principais: a Consulta, a Orientação de Uso e a Área Restrita. Essas partes podem ser acessadas através da página principal. A Figura 16 mostra a tela da página principal do sistema.

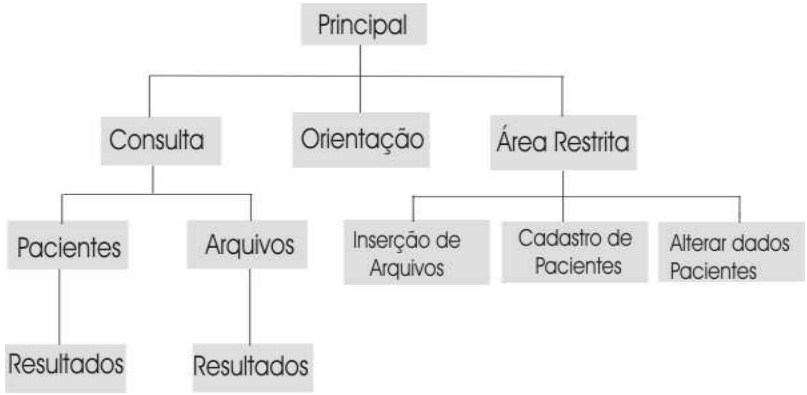


Figura 15: Estrutura Básica do Sistema
Fonte: Dados da Pesquisa, 2003

Figura 16: Tela1 – Página Principal do Sistema
Fonte: Elaborado pelo autor

A parte de **Consulta** permite dois tipos básicos de consulta:

- Consulta em Arquivos, onde o usuário digitará apenas a palavra desejada (Figura 17) e todos os arquivos relacionados com aquela palavra serão retornados com links para os determinados arquivos; como mostra Figura 18.

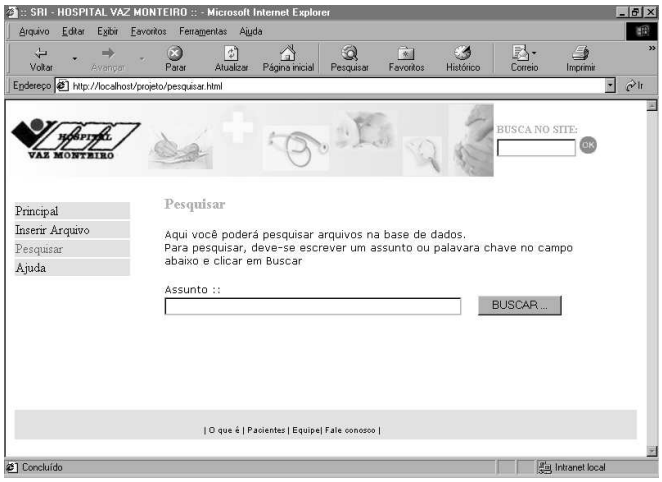


Figura 17: Tela2 – Pesquisar (Consultar) arquivos
Fonte: Elaborado pelo autor

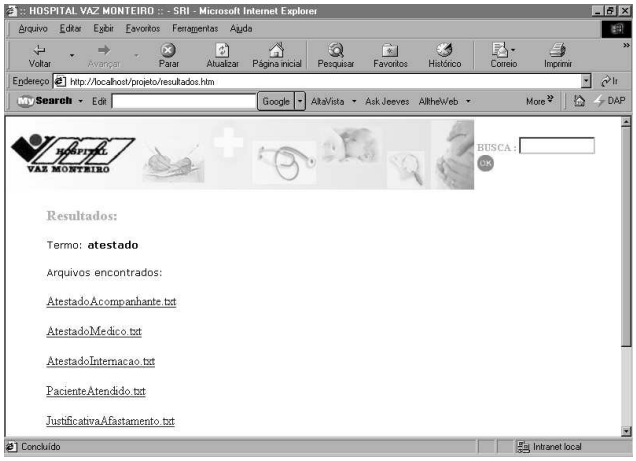


Figura 18: Tela 3 – Resultados da pesquisa em Arquivos
Fonte: Elaborado pelo autor

- Consulta Pacientes, onde o usuário irá escolher o nome do paciente desejado (Figura 19) e o sistema retornará os dados referentes ao mesmo (Figura 20).

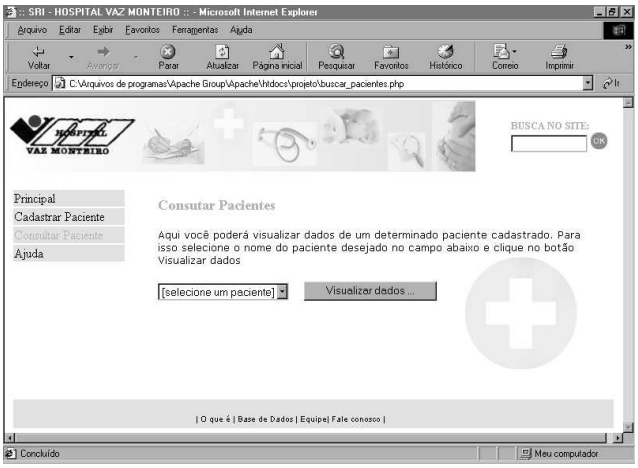


Figura 19: Tela 4- Consultar Pacientes
Fonte: Elaborado pelo autor

Figura 20: Tela 5 – Visualização dos Dados do Paciente selecionado
Fonte: Elaborado pelo autor

A parte de **Orientação** trata-se de uma área explicativa, com o objetivo de orientar o usuário sobre o funcionamento do Sistema, explicando a ele como fazer as consultas. A tela pode ser visualizada na Figura 21.

Figura 21: Tela 6 – Orientações de Uso (Ajuda)
Fonte: Elaborado pelo autor

Na parte de **Área Restrita** o acesso é restrito ao profissional e ao responsável pelo acervo. É um lugar onde os profissionais da saúde podem, através de um formulário, cadastrar (Figura 22) ou alterar informações sobre os pacientes e fazer *upload* (inserir arquivos no servidor) dos arquivos para a base de dados (Figura 23), os quais poderão ser usados nos resultados das consultas.

Foi decidido colocar estes itens como área restrita, devido ao fato de modificarem as condições do Banco de Dados e da Base de Informação (onde estão os arquivos).

Figura 22: Tela 7 – Cadastro de Pacientes
Fonte: Elaborado pelo autor

Figura 23: Tela 8 – Fazer *upload* do arquivo
Fonte: Elaborado pelo autor

- **Estrutura Interna**

A Estrutura interna mostra o funcionamento do sistema que é composto por quatro módulos funcionais como mostra a Figura 24.

O primeiro módulo é o motor de busca e indexação, onde são processadas as consultas, recuperados os documentos e indexados os novos arquivos. O segundo módulo corresponde á interface, camada intermediária entre o usuário e o motor de busca e indexação. O terceiro módulo é o navegador *web*, com o qual o usuário acessa a interface. O quarto módulo é o SQL, onde são processadas as informações relacionadas ao Banco de Dados.

O módulo navegador é responsável por fazer a conexão do usuário com a interface do sistema. A conexão é feita através de uma requisição do usuário, ou seja, o usuário tem necessidade de uma ação no sistema oferecida pelo sistema.

Por exemplo: cadastrar pacientes, procurar arquivos, inserir arquivos, alterar pacientes, entre outros.

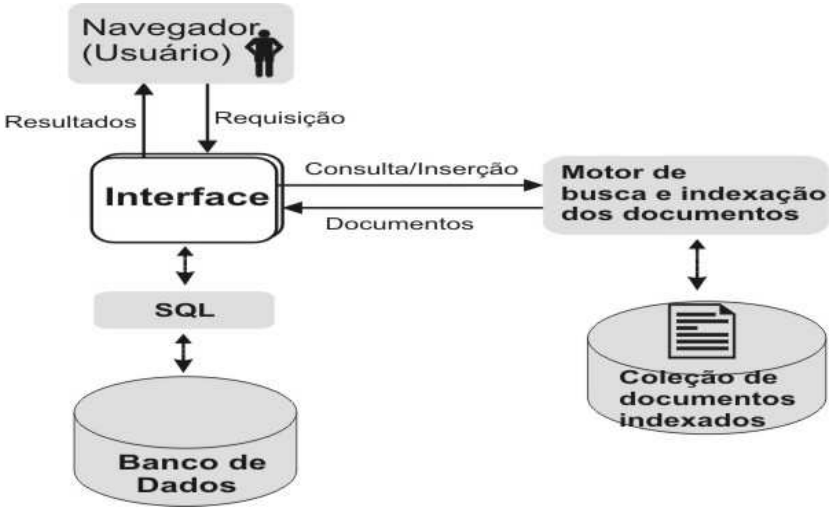


Figura 24: Estrutura interna do Sistema
Fonte: Dados da Pesquisa, 2003

O módulo interface tem a responsabilidade de traduzir a requisição do usuário e transferir aos módulos correspondentes. Por exemplo: Se a requisição do usuário foi *cadastrar paciente*, a interface vai direcionar esta requisição para o módulo SQL. Se a requisição foi *procurar arquivo*, a interface vai direcionar para Módulo de busca indexação.

O módulo SQL é responsável por interagir com o Banco de Dados onde estão armazenadas as informações dos pacientes. Esta interação é feita através de comandos SQL para inserir, alterar ou consultar dados.

No módulo Motor de busca e indexação é feito todo o trabalho de recuperação de informação. Este módulo é dividido em duas partes:

- **Inserção de Arquivos** – Quando o usuário insere um arquivo no sistema é feita a indexação automática deste arquivo. Esta fase preocupa-se com a construção do *vocabulário*, ou seja, um arquivo

contendo todas as palavras chaves do arquivo. Para tanto, um conjunto de regras deve ser pré-estabelecido para a extração de palavras dentro do arquivo. Estas regras deverão tratar espaços, sinais de pontuação, prefixos e outras estruturas gramaticais (identificação de palavras ou análise léxica, explicado na seção 2.3.3.1). Em conjunto com tais regras, pode-se construir uma lista (*stopwords*) de artigos, preposições e palavras em geral que não poderão ser indexadas (remoção de *stopwords* explicado na seção 2.3.3.1), evitando com isso o armazenamento de padrões sem grande importância em buscas futuras. O *vocabulário* deve ser um arquivo (chamado de arquivo invertido, seção 2.3.3.2) armazenado em memória secundária. Este arquivo é incrementado com novos termos toda vez que o usuário insere um novo documento. Depois de montar o *vocabulário* deve-se ordenar seguindo a ordem alfabética dos termos indexados, a ordenação é feita de modo simples, comparando as palavras e verificando qual é maior e trocando as posições. Não é preciso verificar eliminar termos duplicados, pois quando se encere um termo novo no arquivo é feita uma verificação para ver se este termo já existe, ou seja, o termo só será inserido no arquivo quando não existir dentro do mesmo. A estrutura do arquivo de indexação após a ordenação pode ser descrita como mostra a figura 24.

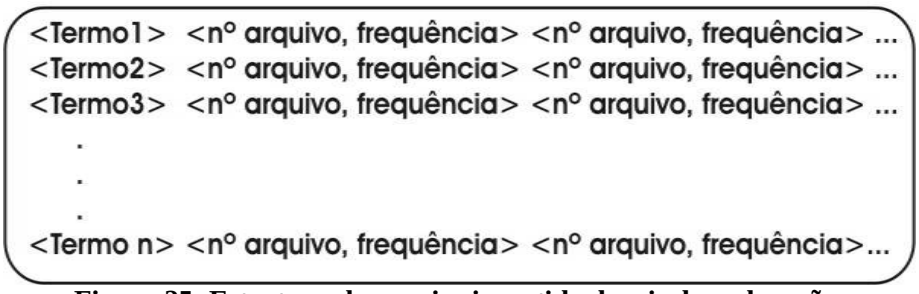


Figura 25: Estrutura do arquivo invertido depois da ordenação
Fonte: Dados da Pesquisa, 2003

- **Consulta de arquivos** – Para fazer uma consulta, o usuário deve digitar a palavra desejada. Depois de clicar em buscar, o sistema faz uma busca no arquivo invertido (arquivo de indexação) criado anteriormente a procura da palavra digitada. Quando o sistema acha a palavra (termo), ele continua lendo o arquivo e verificando o número do arquivo que contém o termo e a freqüência com que este termo aparece. Quanto maior a freqüência mais relevante o arquivo é para o usuário, ou seja, nos resultados os arquivos com maiores freqüências virão primeiro. Os resultados são mostrados com links apontando para os arquivos encontrados.

O Sistema possui uma restrição que atrapalha um pouco o seu funcionamento, ele só consegue fazer a indexação de arquivos com extensão txt.

5. CONCLUSÃO

Durante os últimos anos, um volume crescente de informações na área de saúde tem sido registrado em várias bases de dados, nos mais diversos domínios do conhecimento. Considerando que os recursos informacionais estão cada vez mais acessíveis aos usuários finais, o principal problema é saber como acessar tais recursos de forma fácil e precisa.

É neste sentido que este trabalho foi criado, apresentando o desenvolvimento de um Sistema de Recuperação de Informação para área hospitalar que facilita a busca por informações relevantes dentro de um hospital.

O objetivo principal do trabalho foi alcançado, o sistema foi desenvolvido, mas ainda não foi implantado no hospital.

O uso de um sistema destes traz muitos benefícios para a empresa como um todo, pois médicos e funcionários terão acesso a dados de pacientes e arquivos importantes para o dia a dia do funcionamento do hospital. Além da inclusão de pacientes e arquivos para futuras consultas.

5.1. Trabalhos Futuros

Levando em consideração o seguinte trabalho, que está centrado em informações textuais, seria interessante integrar ao sistema a utilização de *Thesaurus* ou sinônimos (que foi explicado na seção 2.3.3.1), melhorando assim os resultados finais.

Utilizar a tecnologia de recuperação de informação para criar busca em *sites* e base de dados *online* da área de saúde, possibilitando que a busca seja feita através de conceitos e temas desta área, deixando o sistema mais completo de forma que possa ser usado por usuários comuns na busca por informações nesta área.

A questão da apresentação e visualização de documentos encontrados também poderia ser expandida, ou seja, mostrar para o usuário o porquê de eles terem sido recuperados. Uma forma de fazer isso é selecionar os trechos do documento que contenham as palavras da consulta e mostrá-los para o usuário.

Fazer a avaliação do sistema para medir a eficiência do mesmo, utilizando a técnica de Bibliometria, que foi explicado na seção 2.3.5.

Melhorar o método de ordenação do arquivo invertido.

6. BIBLIOGRAFIA

[BAE 92] BAEZA-Yates, Ricardo A. String Searching Algorithms. In: FRAKES,William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992.

[BAE99] BAEZA-YATES, R., RIBEIRO, B. N. **Modern Information Retrieval**. New York, N.Y : Addison-Wesley, 1999.

[CAV 78] CAVALCANTI, C. R. **Indexação e tesauro: metodologia e técnica**, Brasília, ABDF, 1978.

[CAR03] CARNEIRO, A. C. V. M **Processo e Impactos da Informatização: O caso do Hospital Vaz Monteiro em Lavras, MG.**, Lavras, Minas Gerais, 2003

[CHE96] CHEN, H. at alli. **A concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval:**
An Experiment on the Worm Community System, 1996. Disponível em http://ai.bpa.arizona.edu/papers. Último acesso em julho/2003.

[CUN00] CUNHA, M. B. da. **Construindo o futuro: a biblioteca universitária brasileira em 2010**. Ciência da Informação, Brasília, Jan/Abr 2000. Disponível na Internet http://www.ibict.br/cionline/. Último acesso em maio/2003.

[CUN99] CUNHA, M. B. da. **Desafios na construção da biblioteca digital.** Ciência da Informação, Set/Dez 1999. Disponível na Internet http://www.ibict.br/cionline/. Último acesso em 25/05/03.

[ELM02] ELMASRI, R , NAVATHE, S. B. **Sistemas de Banco de Dados Fundamentos e Aplicações**. Rio de Janeiro, RJ, 2002

[FRA92] - FRAKES, W. B.; BAEZA-YATES, R.. **Information Retrieval: Data Structures & Algorithms**. New Jersey: Prentice Hall, 1992.

[HAR92] HARMAN, D. et al. Inverted Files. In: FRAKES, W. B.; BAEZAYATES, R. A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992.

[JUN 00] JUNIOR, F. C., **Programando para a Web com PHP/MYSQL.** POLI – UPE – Engenharia Eletrônica, Agosto/2000. Disponível em: www.nied.unicamp.br/~zeh/flatec/apostilas/PHPManual2.pdf..Último acesso 06/08/2003

[KRU 97] KRUG, L. **Um estudo sobre técnicas de recuperação de informações com ênfase em informações textuais** – Universidade Federal do Rio Grande Do Sul – Porto Alegre, 1997 – Instituto de informática

[KRU 00] KRUG, L. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva –** Universidade Federal do Rio Grande do Sul – Porto Alegre, 2000 – Instituto de informática

[KOW 97] - KOWALSKI, G. **Information Retrieval Systems: Theory and Implementation**. Boston: Kluwer Academic Publishers, 1997.

[LEE 98] - LEE, R. W.. **Pesquisa jurisprudencial inteligente**. Tese (Doutorado) – Engenharia da Produção, Universidade Federal de Santa Catarina, 1998, Florianópolis.

[LOW 01] LOWER,Y.W.. **Dreamweaver 4: A Bíblia**, Editora Campus, 2001.

[MAG 02] MAGALHÃES S. G. **Autoria em hipermedia: O modelo OOHDM aplicado à gestão de eventos.** [Monografia de graduação]. UFLA, DCC, 2002.

[MAR 82] MARCONI M. A. & LAKATOS E. M.; **Tecnicas de Pesquisa**,São Paulo, Editora: Atlas 1982;

[MIZ 97] - MIZZARO, S. **Relevance: The Whole History**. Journal of the American Society for Information Science, New York: John Wiley & Sons. v.48, n.9. 1997.

[ROG 58] ROGET, P. M.; ROGET, J. L.; ROGET, S. R**. Thesaurus of English Words and Phrases**. London: Longmans, Green and Co., 1958.

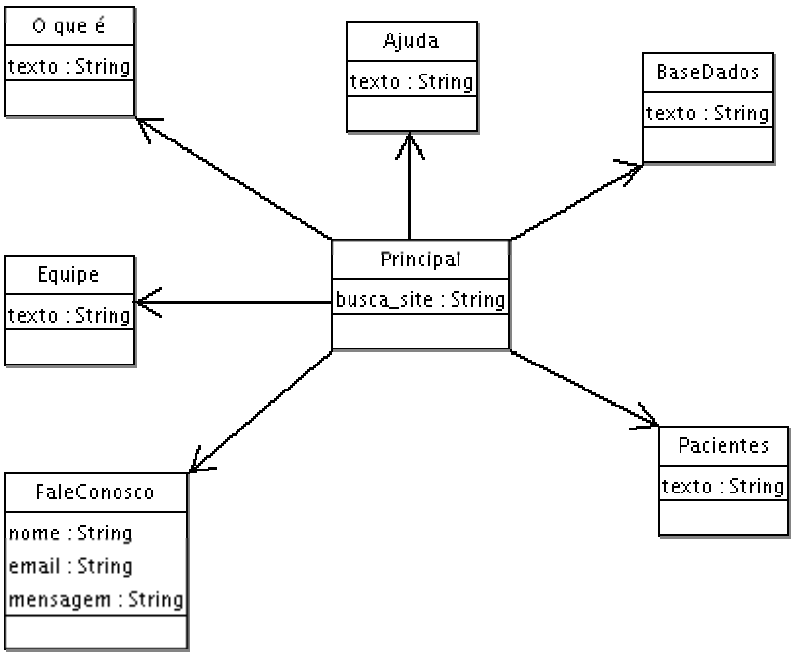
[RUIZ 01] RUIZ, M. A. **Medicina e Internet.** Disponível em http://www.miranet.com.br/medicina/informatica.htm, 2001. Último acesso em 01/06/2003.

[SAL 83] SALTON, G.; MACGILL, M. J. **Introduction to Modern Information Retrieval**. New York: McGRAW-Hill, 1983.

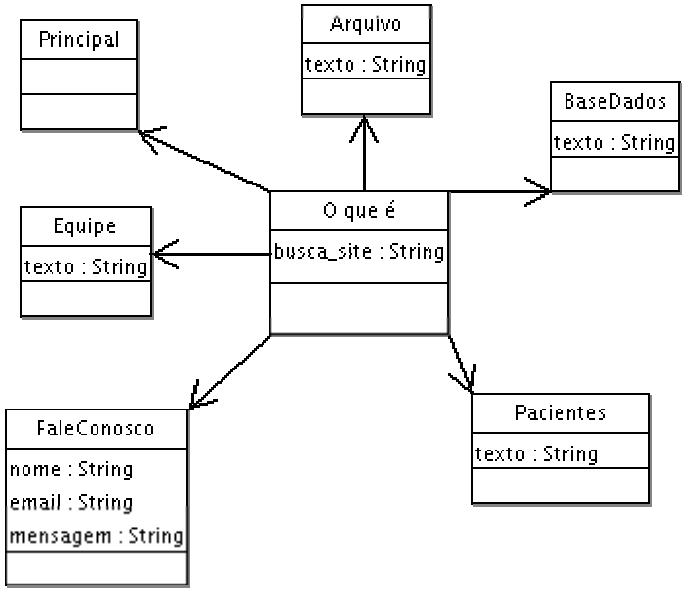
7. ANEXOS

ANEXO A: Figuras da modelagem do sistema feita em OOHDM

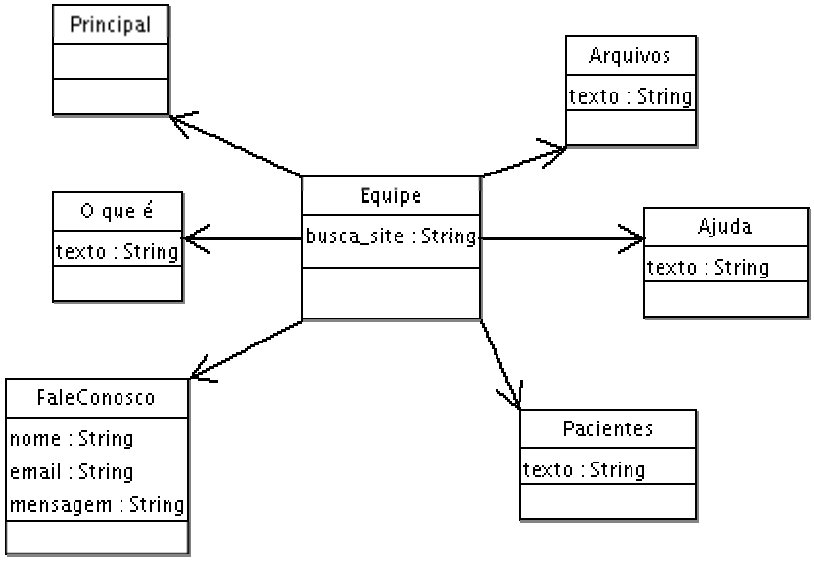
Página Principal,possui links para todas as outras páginas



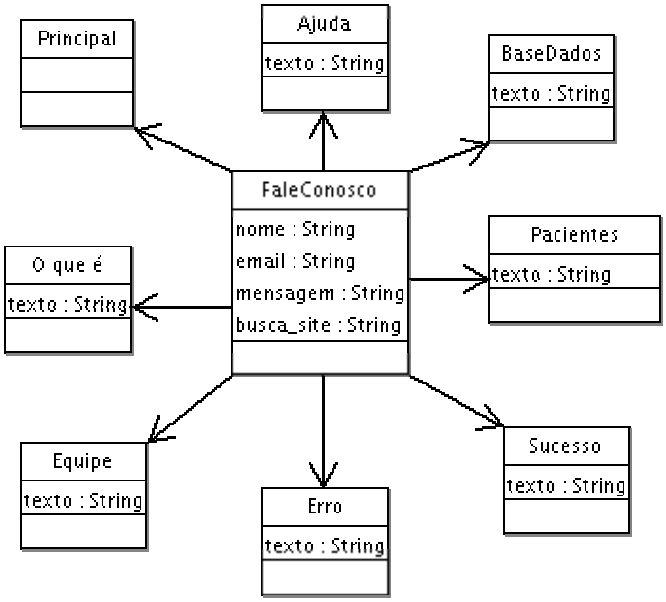
Página O que é, onde descreve o que é o projeto e pra que serve.



Página Equipe. Nesta página estão os nomes das pessoas envolvidas no projeto.



Página Fale conosco. Onde o usuário pode tirar dúvidas, dar sugestões, fazer críticas mandando sua mensagem.



Página Pacientes: Através desta página o usuário pode cadastrar, alterar, ou buscar pacientes. A parte de cadastrar e alterar possui acesso restrito, ou seja, o usuário deverá ter uma senha de acesso para fazer estas funções