Amilton Souza Martha

RECUPERAÇÃO DE INFORMAÇÃO EM CAMPOS DE TEXTO LIVRE DE PRONTUÁRIOS ELETRÔNICOS DO PACIENTE BASEADA EM SEMELHANÇA SEMÂNTICA E ORTOGRÁFICA

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina para obtenção do Título de Mestre em Ciências

São Paulo

2005

Amilton Souza Martha

RECUPERAÇÃO DE INFORMAÇÃO EM CAMPOS DE TEXTO LIVRE DE PRONTUÁRIOS ELETRÔNICOS DO PACIENTE BASEADA EM SEMELHANÇA SEMÂNTICA E ORTOGRÁFICA

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina para obtenção do Título de Mestre em Ciências

Orientador: Carlos José Reis de Campos

São Paulo

2005

Martha, Amilton Souza
Recuperação de Informação em campos de texto livre de Prontuários Eletrônicos do Paciente baseada em semelhança semântica e ortográfica /Amilton Souza Martha. - São Paulo, 2005.
xii, 91f.

Tese (Mestrado) – Universidade Federal de São Paulo. Programa de Pós-graduação em Informática em Saúde.

Título em inglês: Information Retrieval from free text of Electronic Patient Records based on semantic similarity and approximate string matching

1.Recuperação de Informação. 2. Semelhança Semântica. 3. Semelhança Ortográfica. 4. Textos Livres 5. Prontuário Eletrônico do Paciente

UNIVERSIDADE FEDERAL DE SÃO PAULO
ESCOLA PAULISTA DE MEDICINA
DEPARTAMENTO DE INFORMÁTICA EM SAÚDE

Programa de Pós-Graduação em Informática em Saúde

Coordenador: Daniel Sigulem
Vice-Coordernador: Carlos José Reis de Campos

DEDICATÓRIA

Dedico esta obra aos meus pais José e Fátima, à minha irmã Nanci e à minha esposa Valéria que sempre me apoiaram e incentivaram.

AGRADECIMENTOS

À família, pelo amor e dedicação.

Aos pós-graduandos de Informática em Saúde que tanto contribuíram para minha formação.

Aos docentes do programa de pós-graduação em Informática em Saúde.

Ao meu orientador, prof. Carlos José Reis de Campos, pelo estímulo, dedicação e oportunidades que muito agregaram em minha formação.

Ao coordenador do Departamento de Informática em Saúde, prof. Daniel Sigulem, que acreditou na minha capacidade e sempre me apoiou nas horas de necessidade.

SUMÁRIO

Lista de Anexos

Índice de Figuras

Índice de Tabelas

Índice de Gráficos

1 - Resumo

A Recuperação de Informações é a ciência que estuda a criação de algoritmos para recuperar informações, principalmente provenientes de textos livres, que constituem a maior parte da informação em forma digital disponível nos dias atuais, sobretudo após a Internet.

É evidente a necessidade de técnicas para recuperar informações dessa grande massa. Mecanismos de busca como Google®, Altavista®, Yahoo® e outros são indispensáveis para encontrar informações espalhadas na Internet em páginas da Web (arquivos PDF, TXT, HTML e outros) nos dias atuais.

Na área da saúde, muitas informações também se encontram na forma de textos livres como os artigos científicos em bases de dados específicas da saúde como o Medline que possuem ferramentas de busca como Pubmed.

Prontuários Eletrônicos do Paciente (PEP) também possuem informações em textos livres como o histórico ou evolução do paciente. Os profissionais da saúde que inserem informações podem utilizar termos sinônimos, jargões médicos, abreviaturas ou mesmo terem erros de ortografia. Para esses casos, a recuperação de informações com essas variações pode ser algo não trivial.

Foram utilizadas duas bases de dados de PEP´s de clínicas distintas, sendo a primeira com 6732 histórias clínicas e a segunda com 26072 histórias. Foi desenvolvido um software chamado SIRIMED (Sistema de Indexação e Recuperação de Informações Médicas) que permitiu mostrar que a recuperação de informações baseada em semelhança semântica com um thesaurus médico (DeCS – Descritores em Ciências da Saúde) e semelhança ortográfica, baseada em um algoritmo de *stemming*, juntamente com *edit distance*, pode melhorar a quantidade de termos recuperados numa busca, em média de 30% comparada com a busca tradicional direta, que faz somente a busca do termo exato.

A média de falsos positivos encontrados é menor que 0,5% nas duas bases de dados, o que não compromete o resultado do aumento de recuperação conseguido.

2 – Histórico

2.1 – História do armazenamento de dados médicos

A história da medicina é tão antiga quanto à própria história do homem que, como ser racional e pelo próprio instinto de sobrevivência, teve que aprender a curar os males que ocorriam em seus semelhantes.

As primeiras formas de cura e de medicina primitiva experimentadas pelo homem assemelham-se ao comportamento animal, isto é, instintivo: o uso da saliva e da imersão do machucado ou da ferida em água, a variação de temperatura (como uma compressa primitiva), a aplicação de lama e de vegetais na região infectada (Lopes, 1999).

A cura, em suas primeiras manifestações, esteve ligada à religião e curandeirismo (Lopes,1999). Durante esse período, o conhecimento médico foi transmitido de pai para filho, como um dom divino, não havendo registro detalhado desse conhecimento.

2.1.1 – Os registros antigos

Com o surgimento da escrita por volta de 3000 a.C., iniciou-se a fase do registro histórico do conhecimento humano. A cidade da Babilônia, durante a dinastia Hamurábi (1728-1686 a.C.), deixou inúmeras tábuas de argila com conhecimento de botânica e zoologia que são preservadas até os dias atuais (Lopes, 1999).

Os egípcios também deixaram uma série de papiros datados dos séculos XIV, XV e XVI a.C. que são os documentos mais antigos relacionados à medicina egípcia (Lopes, 1999).

Porém, foi após Hipócrates (século IV a.C.) que a medicina teve sua transição do caráter mitológico para o uso do pensamento lógico científico, quando se estabeleceu uma abordagem racional para encontrar as explicações para os fenômenos naturais.

Os estudos de Hipócrates foram reunidos na grande biblioteca de Alexandria no século IV a.C., que se compõe de 72 livros e 59 tratados, o que

leva a crer que a compilação colheu textos de diversos professores e praticantes da medicina. A hipótese de que a obra de Hipócrates tenha sido escrita por mais de uma pessoa é aceita hoje em dia (Lopes, 1999).

Após a racionalização da prática da medicina, o registro médico tornou-se prática entre os adeptos ao pensamento pós-hipocrático.

2.1.2 – O registro em Papel

> "...avalia-se em 20 mil volumes, incluindo-se os panfletos, a quantidade de publicações cujo conteúdo é acrescentado ao saber humano a cada ano; e, a não ser que essa massa seja armazenada com ordem e que se especifiquem bem quais os meios que nos irão expor os respectivos conteúdos, tanto a literatura como a ciência ficarão esmagadas sob o próprio peso ..."
>
> (Henry, 1851, *apud* Kent, 1972).

A cada ano, cerca de 300.000 novas referências são adicionadas ao banco de dados MEDLINE que indexa artigos científicos na área da saúde (Hersh *et al.*, 2001). À medida que a humanidade evolui, a complexidade da tarefa de armazenar os registros cresce. A ciência da Recuperação de Informações não é nova e nem começou em meios eletrônicos. Criar condições de armazenamento de informações para posterior recuperação é uma preocupação muito antiga, pois mesmo em papel, a quantidade de informações sempre foi grande e era necessário criar mecanismos para facilitar sua recuperação.

Porém, a área teve grande impulso com o surgimento da www (*World Wide Web*) nos anos 90 com milhares de páginas espalhadas pelo mundo representando o conhecimento humano mundial de fácil acesso, mas que precisava de mecanismos que facilitassem encontrar as informações requeridas.

A Web está se tornando um repositório universal de conhecimento humano e cultura na qual disponibiliza um compartilhamento sem precedentes de idéias e informações numa escala nunca vista anteriormente (Baeza-Yates & Ribeiro-Neto, 1999).

2.1.2.1 – Vantagens e Desvantagens

Com o crescente aumento do volume de informações sobre o paciente, o registro médico tradicional em papel não tem sido mais suficiente para suprir todas as necessidades dos usuários da saúde. Os médicos de hoje não cuidam de uma pequena quantidade de pacientes, como faziam os médicos da família de antigamente; hoje podem ter milhares de pacientes, o que torna impossível lembrar dos detalhes clínicos de cada um.

Além disso, existem alguns problemas críticos no prontuário em papel, como a falta de sistemática na inclusão de dados, o extravio e a redundância de informações e a dificuldade de recuperação seletiva das mesmas.

Na tentativa de permitir que alguns dados sejam encontrados mais facilmente, esses podem aparecer mais de uma vez no prontuário médico em papel (redundância) e com isso pode o prontuário tornar-se maior do que deveria e fazer a busca de informações mais ineficiente (Shortliffe & Barnett, 2001).

Fazer pesquisas com a leitura de vários prontuários em papel na busca de informações para a pesquisa clínica pode ser uma aventura tediosa, pois a informação está espalhada no prontuário. Imagine procurar todos os pacientes que mencionaram determinado sintoma, de uma cerca faixa etária e foram tratados com uma determinada droga, num período de tempo. Encontrar esses casos em centenas ou milhares de prontuários em papel de um hospital ou clínica pode ser uma tarefa dispendiosa e não totalmente eficaz.

2.2 – Prontuário Eletrônico

A necessidade de armazenamento de dados nos dias atuais não é um luxo e sim uma necessidade. Décadas atrás, o custo do armazenamento era muito alto, mas com a evolução da computação, os sistemas de hoje são capazes de armazenar cada vez mais informações e a um custo cada vez menor.

Devido às dificuldades relatadas anteriormente sobre o prontuário em papel surgiu, então, a necessidade de criação do Prontuário Eletrônico do

Paciente (PEP), visando uma melhoria no controle dos dados do paciente para melhor recuperação posterior.

O PEP surgiu, inicialmente, com o objetivo principal de controle de dados administrativo-financeiros de um hospital ou clínica visando o planejamento estratégico da instituição; porém, atualmente, os dados clínicos também assumem um papel fundamental nesse processo (Shortliffe & Blois, 2001).

Porém, transformar um registro médico em papel para registro eletrônico, não é tarefa fácil. A maioria das informações médicas está em texto manuscrito e converter essa informação e introduzi-la em um banco de dados pode ser tarefa árdua e não completamente eficiente.

2.2.1 – Dado x Informação

Antes de podermos falar de recuperação de informações, precisamos definir o que vem a ser informação e, por conseqüência, a diferença entre dados e informações.

Nota-se em muitas publicações que dados e informações são tratados como termos sinônimos. Inclusive livros específicos sobre Bancos de Dados não possuem diferenciação clara entre os conceitos: “Um sistema de gerenciamento de banco de dados (SGBD) consiste em uma coleção de dados inter-relacionados...” e logo no parágrafo seguinte temos: “Os sistemas de bancos de dados são projetados para gerenciar grandes grupos de informações.” (Korth & Silberschatz, 1995).

De acordo com Pereira, em termos computacionais, dados representam uma abstração de parte da realidade, ou seja, representam algumas características selecionadas das entidades do mundo real, necessárias para a solução de um determinado problema (Pereira, 1996).

De acordo com Silva Filho, “Dado é a sentença descritiva resultante de um processo de mensuração”, isto é, a mensuração de características observadas por um humano ou não (Silva Filho, 2003). Como exemplo, podemos citar a mensuração do peso, da data de nascimento, do histórico familiar, do sexo, da taxa de glóbulos brancos etc.

Van Bemmel conceitua dado como a representação de observações ou conceitos apropriados para comunicação, interpretação e processamento por humanos ou máquinas. Dados interpretados formam as informações (Van Bemmel, 1999).

Os dados são processados a partir de um conhecimento e podem gerar informação. Por exemplo, sendo um dado de entrada o raio de uma circunferência e o conhecimento que o perímetro tem como fórmula $P = 2.\pi.raio$, podemos chegar à informação do valor do perímetro da circunferência (Pereira, 1996).

Em teoria, a sentença descritiva de um processo de mensuração resulta nos "dados brutos" (Silva Filho, 2003). Entende-se por dados brutos aqueles usados para obtenção de informações por meio de conhecimentos anteriores. Mensurar que o peso de uma pessoa é 165 quilos e sua altura é 1,62 m são dados, ao passo que saber que a mesma possui obesidade mórbida é uma informação extraída desses dados.

Podemos perceber que dados estão vinculados à obtenção, registro e armazenamento de características medidas ou observadas enquanto que a informação está vinculada à recuperação, análise e uso dos dados registrados.

Pereira resume que dado é aquilo que entra em um processo informatizado e informação é aquilo que sai (Pereira, 1996).

Portanto, conclui-se que ***dado*** é a sentença descritiva resultante do processo de mensuração ou observação de uma determinada característica por um ser humano ou máquina, enquanto que ***informação*** é o resultado do tratamento e interpretação desses dados por uma inteligência humana ou computacional.

Sendo mais específico, podemos considerar um dado médico como o resultado de uma observação isolada de um paciente, por exemplo, a leitura da temperatura, a contagem de glóbulos vermelhos do sangue, o passado histórico de rubéola ou a leitura da pressão sanguínea. Uma pressão de 120x80 mmHg, por exemplo, pode ser inserida no banco de dados como "pressão normal", porém, se a informação separada de pressão sistólica e

pressão diastólica for importante, os dados devem ser registrados separadamente (Shortliffe & Barnett, 2001).

Segundo Shortliffe & Barnett, os dados médicos podem ser divididos em três grandes grupos, sendo o primeiro os dados numéricos, aqueles mensuráveis numericamente como alguns testes laboratoriais, temperatura, pulso, pressão e outros e podem ser armazenados como dados estruturados, que serão melhor detalhados adiante. O segundo grupo compreende dados narrativos, incluindo a descrição dos sintomas pelo paciente, respostas de questões apresentadas pelo médico, histórico familiar e social do paciente e outras observações que o médico ache relevante para posterior consulta que são armazenados na forma de textos livres e melhor detalhados também adiante e são o foco desse trabalho. Por último, as imagens e os gráficos, que geralmente são adquiridas por máquinas ou desenhadas pelo médico (Shortliffe & Barnett, 2001).

2.2.2 – Armazenamento em Banco de Dados

Dados são o centro de todo o processo de decisão na saúde, portanto, eles devem ser confiáveis, completos e bem estruturados. Um programa que disponibilize operações de armazenamento e recuperação de dados, controle de acesso e de transações de dados é chamado Sistema de Gerenciamento de Bancos de Dados (SGBD) (Van Bemmel, 1999).

Os bancos de dados (BD) são estruturas organizadas de dados que proporcionam formas ágeis de recuperação de informações. O objetivo fundamental de um banco de dados é a posterior recuperação, portanto, uma informação armazenada não tem nenhum valor se não puder ser recuperada.

Mais especificamente, um Banco de Dados Médico é um conjunto pluridimensional de informações de saúde ou doença dos indivíduos (pacientes), associando elementos importantes para a ação médica, tendo como fim último a maior eficiência dos serviços prestados.

Existem hoje muitos modelos de bancos de dados, que podem ser divididos em quatro categorias: BD em arquivos seqüenciais, BD Hierárquico, BD em Rede e BD Relacionais (Silva, 2001). Nos últimos anos está surgindo

uma nova modalidade que une as características de BD Relacionais com o paradigma de orientação a objetos, chamados de BD Objeto-Relacionais.

Sem dúvida, o modelo predominante no mercado é o Banco de Dados Relacional, ou melhor, Sistema de Gerenciamento de Banco de Dados Relacionais (SGBDR) baseado nos conceitos de campos, registros e tabelas. Um campo é a unidade básica de um banco de dados, sendo responsável por armazenar um determinado dado, como nome do paciente, diagnóstico, peso ou pressão. Um conjunto de campos relacionados forma um registro, como o conjunto de dados de uma ficha clínica de um paciente específico. Ao conjunto de registros, denominamos tabela que, analogamente, pode ser comparado ao arquivo de fichas de pacientes.

Quando convertemos o prontuário em papel para o prontuário eletrônico, os dados do paciente são representados por um modelo de algumas centenas de campos. Em outras palavras, somente uma parte da realidade médica é representada pelos sistemas atuais de informação (Lovis *et al*., 2000). É uma visão simplista demais enxergar os dados do paciente como apenas colunas de números (Shortliffe & Barnett, 2001).

Uma grande vantagem nos bancos de dados relacionais é o uso de uma linguagem comum de acesso às informações, denominada SQL (*Structured Query Language*), que permite manipular as informações nas tabelas (inserir, excluir, alterar ou pesquisar). Atualmente, esta linguagem garante seu sucesso no mercado de banco de dados porque está integrada em praticamente todos os produtos de SGBDR´s e Objeto-Relacionais (Silva, 2001). Dentre eles, Oracle, Microsoft SQL Server, MySQL, MS-Access e PostGreSQL.

2.2.3 – Dados Estruturados

Os diferentes tipos de dados médicos que um sistema de prontuário eletrônico contem podem ser armazenados basicamente de duas formas: textos livres ou dados estruturados.

Dado estruturado é aquele que possui uma faixa de valores pré-definidos. Esta situação tem muitas vantagens, pois podemos validar os conteúdos dos campos e acionar determinadas funções, quando o conteúdo se

encontra fora dessa faixa especificada, que poderão gerar alertas para situações incomuns ou de risco (Lovis *et al.*, 2000; Shortliffe & Barnett, 2001).

2.2.3.1 – Vantagens e Desvantagens de Dados Estruturados

Dados estruturados limitam o médico por pré-definir um conjunto de termos médicos para capturar a informação do paciente. Muitos autores citam o ato da anamnese como uma arte e que depende muito da experiência do médico que examina o paciente, tornando a padronização das informações algo discutível.

Os bancos de dados tradicionais são ótimos para armazenar informações que possuam estruturas e relações fáceis de ser identificadas e extraídas, como escolher o tipo de informação (texto, número), tamanho ou conteúdo. Porém, nem todas as informações são tão facilmente estruturadas como textos, imagens e gráficos.

Dados capturados de uma forma estruturada são mais fáceis para o uso em pesquisas clínicas, para troca com outros serviços de saúde, sistemas de apoio à decisão e acesso à literatura biomédica on-line (Mulligen *et al.*, 1998).

2.2.4 – Texto Livre

A necessidade de padronização nos impulsiona ao uso de categorias pré-definidas e vocabulários controlados, enquanto a necessidade de expressar livremente, sem distorcer um dado do paciente, nos remete ao uso de textos livres (Sager *et al.*, 1994).

Transformar um dado formatado em texto é uma tarefa relativamente fácil, porém a recíproca não é verdadeira.

Devido à semelhança do processo de inserção de informações com o prontuário em papel, o texto livre é o modo favorito dos usuários dos sistemas de prontuário eletrônico.

O uso da formatação de um texto livre em dados estruturados provoca uma significante perda de informações e erros de classificação (Lovis *et al.*, 2000). A observação de alguma reação estranha do paciente durante uma

consulta, informações sobre a família do paciente ou a situação econômica do mesmo, são informações que podem ser importantes e dependem da experiência do médico que conversa com o paciente, mas fica difícil registrar essas informações para que outro médico possa ter a mesma informação, apenas acessando os registros computadorizados estruturados (Shortliffe & Barnett, 2001).

2.2.4.1 – Vantagens e Desvantagens de Textos Livres

Em textos livres, a capacidade de narrativa de textos para representar a realidade das condições do paciente é limitada apenas pelo autor ou pelos limites técnicos do sistema. Essa grande vantagem associa-se com uma grande desvantagem, pois o controle do conteúdo de documentos é deixado a cargo do usuário e a incoerência entre um documento e outro é quase inevitável (Lovis *et al*., 2000).

As publicações biomédicas e o considerável volume de dados clínicos são criados e armazenados em documentos de texto livre. No entanto, computadores não foram criados para processar textos livres eficientemente e os métodos de pesquisa, como o SQL, não são facilmente empregados para manipular textos livres (Chu, 2002).

Enquanto textos livres são convenientes para tarefas como revisão de prontuários por médicos, eles apresentam graves obstáculos para a criação de gráficos, busca, sumarização e análise estatística (Johnson,1999).

Apesar da informação em texto livre ser difícil de indexar e, conseqüentemente, de recuperar, é largamente utilizada (Wives, 1997). Cada vez mais, as instituições médicas têm acesso aos registros de pacientes através de computadores. Muitos dos dados disponíveis já estão em forma textual como resultado da transcrição de relatórios ditados, uso de tecnologias de reconhecimento de voz e diretamente inseridos por profissionais da saúde (Johnson,1999).

3 – Introdução

3.1 - Recuperação de Informações

A Recuperação de Informações – do inglês *Information Retrieval* – é uma ciência que estuda a criação de algoritmos para recuperar informações, principalmente provenientes de textos livres, que constituem a maior parte da informação em forma digital disponível nos dias atuais, sobretudo após a internet e a WWW (*World Wide Web*). É evidente a necessidade de técnicas e algoritmos de busca específicos para recuperar informações seletivas dessa grande massa. Mecanismos de busca como Google®, Altavista®, Yahoo® e outros são indispensáveis para encontrar informações espalhadas na internet nos dias atuais.

Para muitos, recuperação de informação implica na recuperação de qualquer tipo de informação do computador, no entanto, para aqueles que trabalham na área, a idéia é mais específica, consistindo na recuperação de informações de bancos de dados que predominam informações na forma textual (Hersh, 2003).

Linguagens de Bancos de Dados, como o SQL, possuem cláusulas de pesquisa em campos textos como o 'LIKE' que encontra ocorrências exatas de strings ou de substrings em textos, o que podemos chamar de busca direta, porém a Recuperação de Informações trabalha com algoritmos mais elaborados, incluindo processos de indexação, *stemming*, remoção de *stop words* e outros que serão vistos a seguir.

De modo a diferenciar os termos "recuperação de dados" e "recuperação de informações" e, ao mesmo tempo, justificar o título do trabalho, temos que a Recuperação de Dados (*Data Retrieval*) consiste principalmente em determinar quais os documentos de uma coleção que contém as palavras-chave contidas na pergunta do usuário. A principal diferença é que na Recuperação de Informações os textos estão em linguagem natural e, na maioria das vezes, não estão bem estruturados com possibilidade de ser semanticamente ambíguos (Baeza-Yates & Ribeiro-Neto, 1999).

Portanto, a escolha do título de Recuperação de Informações está baseado no conceito de informações visto na seção 2.2.1 e também por estarmos trabalhando com textos livres não estruturados.

A recuperação de informações é a ciência e a prática da identificação e uso eficiente dos dados armazenados. Como já foi dito, uma informação armazenada, que não pode ser recuperada, não tem valor.

Por exemplo, para se buscar o nome do paciente em um BD estruturado, basta percorrer a tabela que possui o campo “nome” e localizar o registro, porém, se os nomes não estiverem distribuídos de uma forma tabular, ou seja, estiverem na forma de texto livre, a tarefa será muito mais árdua (Wives, 1997).

Dados estruturados são mais fáceis de serem tratados por meios computacionais, porque existem linguagens formais como o SQL que permitem sua manipulação e consulta de forma mais concisa e precisa (Loh, 1997).

Porém, os textos livres são estruturas mais complexas para a recuperação, sendo necessário a aplicação de técnicas avançadas de computação como indexação de documentos, tratamento de termos, uso de sinônimos e outros.

3.1.1 - Indexação de Documentos

Um dos métodos bastante utilizados para o registro e posterior recuperação de informações em textos livres é a indexação. A idéia de indexar é produzir um índice menor, porém mais eficiente, para representar o conteúdo original que facilite a recuperação de informações (Hersh *et al.*, 2001).

Desenhos ou outras espécies de registros gráficos, que não se compõem de palavras, são freqüentemente descritos em termos de palavras e linguagem natural. As mais elevadas formas de comunicação entre seres humanos baseiam-se em palavras e na linguagem (Kent, 1972).

A informação, que é o objeto de todas as pesquisas, está contida simbolicamente em registros expressos por palavras. Tais palavras são organizadas de tal forma que produzem uma linguagem natural (Kent, 1972).

Podemos então dizer que as palavras (ou termos) podem descrever o conteúdo de um texto. Nessa dissertação usaram-se as expressões ‘palavras’

e ‘termos’ de forma indistinta uma vez que o processo de indexação proposto indexa termos que são simples palavras.

Desde o início do registro do conhecimento, o homem tem organizado a informação para mais tarde recuperar e usar. Como o volume de informação cresceu rapidamente nos últimos anos, foi necessário construir estruturas de dados especializadas para facilitar o acesso à informação armazenada. Uma estrutura de dados bastante antiga e usada é uma coleção de palavras ou termos selecionados associado a ponteiros que se relacionam com a informação, ou documentos, chamado índice. Índices são a base de todo sistema de recuperação de informações (Baeza-Yates & Ribeiro-Neto, 1999).

A estrutura básica de um índice é uma lista de itens e seus atributos. Os itens do índice são unidades de informação adequadas para a comparação com os termos da pergunta. Em alguns índices, os itens são simplesmente as palavras ou frases encontradas na coleção de documentos, em outros índices são termos escolhidos por homens ou máquinas para representar o conteúdo. Por outro lado, os atributos dos itens descrevem detalhes do item, como o número de documentos que ele ocorreu, sua freqüência no documento, a posição onde se encontra e outros que podem ser necessários para a recuperação (Hersh *et al*., 2001).

Um sistema de indexação pode ter apenas um índice simples, onde o usuário não necessita escolher qual o índice a ser pesquisado, porém sistemas mais complexos possuem índices múltiplos que permitem o acesso ao conteúdo estruturado, onde existem regiões semânticas distintas, por exemplo, nome de autor, data de publicação ou palavras-chave. O benefício desse tipo de índice é que o usuário pode pesquisar em uma região semântica distinta, melhorando o processo de recuperação e evitando que o sistema tenha que pesquisar em todo o índice.

O processo de indexação para a recuperação de informações médicas não é recente e nem começou em meios computacionais. Em 1879, John Shaw Billings criou o *Index Medicus* para ajudar profissionais médicos a encontrar artigos relevantes em jornais. Artigos de jornais eram indexados por nome de autor e assunto do título e, então, eram colocados em volumes separados para

um pesquisador encontrar com maior facilidade um assunto específico (Hersh *et al*., 2001).

Esse índice foi usado por muitas décadas para a recuperação de literatura médica até ser criada em 1966 pela **NLM** (*National Library of Medicine*) uma versão digital chamada **MEDLARS** (*Medical Literature Analysis and Retrieval System*) e posteriormente uma versão on-line chamada **MedLine**.

Até o momento, os artigos do banco de dados Medline são indexados manualmente, onde um grupo de indexadores humanos atribui alguns termos **MeSH** (*Medical Subject Heading*), que também é um vocabulário controlado criado pela NLM, para a recuperação de artigos com o mesmo assunto.

O MeSH é um vocabulário que possui mais de 18000 assuntos organizados hierarquicamente em 15 árvores para representar conceitos em biomedicina. Os assuntos possuem relacionamento explícito entre eles, além de poder existir formas sinônimas.

O DeCS (Descritores em Ciências da Saúde) é um vocabulário estruturado, trilingüe (português, espanhol e inglês), baseados em coleções de termos, organizados para facilitar o acesso à informação (Pellizzon, 2004).

A BIREME, que faz parte do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde, desenvolveu DeCS em 1986 traduzindo e adaptando o MeSH da NLM, para uso na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos e outros tipos de materiais, assim como para ser usado na pesquisa e recuperação de assuntos da literatura científica nas bases de dados LILACS (Literatura Latino-Americana e do Caribe em Ciências da Saúde), MEDLINE, SciELO, e outras (Pellizzon, 2004).

Nele foram também incorporadas as áreas específicas de saúde pública e homeopatia, totalizando 26851 descritores, sendo destes 3656 de Saúde Pública e 1950 de Homeopatia. Por ser dinâmico, registra processo constante de crescimento e mutação registrando a cada ano um mínimo de 1000 interações na base de dados, dentre alterações, substituições e criações de novos termos ou áreas (Bireme, 2005).

O DeCS possui sempre um termo autorizado, que consiste na forma canônica do termo e uma lista de sinônimos para o mesmo.

A NLM possui vários projetos de grande impacto nessa área, onde podemos destacar a UMLS (*Unified Medical Language System*) que tem a intenção de ser uma ponte entre os diversos vocabulários controlados existentes, como o MeSH para literatura médica, o CID (Codificação Internacional de Doenças) da OMS (Organização Mundial da Saúde) para codificação de diagnósticos, o SNOMED (*Systematized Nomenclature of Medicine*) para codificação de informações clínicas, o DeCS e outros. Atualmente, ela está com um projeto para a indexação automática de todos os artigos do Medline para substituir a indexação manual feita atualmente e que é muito custosa (Aronson *et al*., 2004)

3.1.2 - Indexação de Texto Completo

Basicamente, existem duas formas de indexação: a manual e a automática. Na primeira, um indexador humano lê o artigo ou o objeto a ser indexado e escolhe um conjunto de termos que descreverão o seu conteúdo. Nessa indexação podemos recuperar objetos que possuem um mesmo conteúdo semântico, mesmo que escritos de forma diferentes, pois possuem o mesmo termo ou conjunto de termos escolhidos. Porém, isso também pode causar problemas, pois a subjetividade da escolha dos termos pode acarretar a falta de algum aspecto importante para a posterior consulta.

A indexação manual implica numa seleção cuidadosa da terminologia empregada (Kent, 1972). Para tanto, existem atualmente muitos vocabulários controlados médicos como o MeSH, CID, SNOMED e outros (Hersh *et al*., 2001).

Por outro lado, a indexação automática não possui o lado subjetivo da escolha dos termos. O uso mais comum da indexação automática é aquela aplicada a textos completos.

Na indexação automática de textos completos todas as palavras e/ou frases do mesmo podem fazer parte do seu índice. Nesse caso, a freqüência

em que o termo ocorre no texto e no conjunto de todos os textos podem dar uma idéia da capacidade de descrição do termo.

Um dos métodos pioneiros na área foi desenvolvido por Salton em 1960, mas não teve grande sucesso até a década de 90. Por vezes chamado de Modelo de Vetor Espacial, isso porque os documentos podem ser conceituados como vetores de termos com recuperação baseada na similaridade de ângulos entre os vetores da pergunta e dos documentos (Hersh *et al*., 2001).

No processo de indexação descrito por Salton, a primeira etapa consiste em escolher quais as palavras que farão parte do índice. Palavras com alta freqüência na coleção de documentos não são capazes de diferenciar um documento do outro. Essas palavras são chamadas de s*top words* e são normalmente filtradas para a indexação por termos, que incluem artigos, preposições, conjunções e outras, dependendo do tipo de documento (Baeza-Yates & Ribeiro-Neto, 1999; Hersh *et al*., 2001; Wives, 1997; Wives & Loh, 1998).

A eliminação de *stop words* tem importância fundamental na redução do tamanho ao índice, diminuindo 40% ou mais o tamanho do índice criado (Baeza-Yates & Ribeiro-Neto, 1999).

A obtenção da lista de s*top words* (s*top list*) pode ser manual, onde o projetista avalia quais as palavras que não devem ser indexadas, o que pode variar de sistema para sistema (Wives, 1997). O projeto Snowball, criado por Martin Porter, possui uma s*top list* sugerida para o português com 220 termos, incluindo artigos, pronomes, alguns verbos e respectivas variações sintáticas (Snowball, 2002).

A *stop-list* pode variar de aplicação para aplicação. Por exemplo, num banco de dados médico a palavra ‘diagnóstico’ pode ser tão comum que não tem capacidade de representar o conteúdo, portanto pode ser inserida na *stop-list*. Em outros casos, um termo que consta originalmente como *stop-word* pode ser importante num determinado contexto, como por exemplo, o termo ‘não’ que pode ser de extrema importância quando usado como modificador de idéia ou conceito, como a inexistência de um sintoma ou sinal.

Antes de indexar essas palavras, elas devem passar por um processo chamado de *stemming*, que consiste no processo de reduzir as palavras ao seu radical, evitando as variações das mesmas, como por exemplo, as palavras 'desmaiado', 'desmaiar', 'desmaiando', 'desmaios' serão indexadas como o seu radical 'desmaio'.

Martin Porter escreveu algoritmos para *stemming* em várias línguas, incluindo o português. O algoritmo, originalmente em Inglês, foi descrito por Porter em 1980 e consistia na remoção de sufixos comuns. Esse processo, além de permitir uma comparação de termos pela sua raiz comum, diminui o número total de termos no índice do sistema de indexação (Porter, 1980).

Após isso, o método de Salton sugere o cálculo dos pesos das palavras. Esse peso discriminará a capacidade da palavra descrever o texto e será utilizado para ordenar os documentos mais relevantes à pesquisa solicitada. Tipicamente, palavras que são largamente distribuídas entre os documentos não são bons discriminadores e, analogamente, palavras que ocorrem somente em um pequeno número de textos são melhores discriminadores (Wives & Loh, 1998).

Uma maneira bastante utilizada para medir os pesos das palavras é o padrão TF-IDF(*Term Frequency-Inverse Document Frequency*). O IDF de um termo é dado como:

IDF_i = log(número de documentos / número de documentos com o termo i) + 1

E para o cálculo do TF de um termo em determinando texto temos:

TF_{ij} = log(freqüência do termo i no documento j) +1

Desse modo, o peso do termo para determinar sua relevância é dado como:

$W_{ij} = Tf_{ij} * IDF_i$

3.1.2.1 – Correção Ortográfica

Uma das propostas de aprimoramento para a indexação é a correção ortográfica dos termos antes da inserção dos mesmos no índice (Wives, 1997). Para tanto, se faz necessário o uso de um dicionário do português falado no Brasil e usar um processo semelhante dos corretores ortográficos para verificar erros de ortografia.

Porém, um dos problemas desse processo é que a linguagem médica possui muitos termos que não estão em um dicionário da língua portuguesa, dentre eles jargões, abreviaturas e outros termos, tornando difícil a utilização desse método de indexação com correção ortográfica prévia.

3.1.2.2 – Formação de Frases-Termo

Um dos problemas na busca por palavras é a contextualização das mesmas. Muitos sistemas possuem a capacidade de procurar palavras num determinado contexto, ou seja, próximo a outras palavras. Palavras que aparecem próximas a outras podem sinalizar maior relacionamento do que aquelas que estão distantes (Baeza-Yates & Ribeiro-Neto, 1999).

Em geral, as frases-termo não são armazenadas de forma composta, mas sim a posição de cada termo no texto, dando a idéia de distância entre termos. Podemos considerar que a distância de uma palavra da outra no texto cria uma Distância de Contexto.

Alguns métodos são propostos para criação de frases-termo. Um n-grama é definido como uma seqüência ordenada de n palavras retiradas de um documento. Por exemplo, “alguns métodos” e “métodos são” podem ser considerados os dois primeiros bi-gramas da última frase. Um dos problemas desse tipo de indexação é que é dependente da ordem das palavras e muito dependente da proximidade das palavras (Johnson *et al.*, 1998).

Outra proposta é a criação de n-palavras, onde uma coleção de n palavras são retiradas de um texto, ou seja, numa frase de 4 palavras, teremos seis diferentes combinações de duas palavras. Dessa maneira, retiramos a dependência das palavras estarem muito próximas, mas aumentamos a quantidade de combinações possíveis (Johnson *et al.*, 1998).

A idéia pode ser também utilizada não com palavras, mas com caracteres. Algoritmos baseados em trigramas (seqüência de três letras) - *Trigram Matching* - foram propostos para a recuperação por semelhança e implementados com resultados significantes (Tardelli *et al.*, 2004).

3.1.3 – Formulação da Pergunta

O segundo passo no processo de recuperação da informação é a formulação da pergunta (*query*), que é a expressão formal da necessidade do usuário (Baeza-Yates & Ribeiro-Neto, 1999). A pergunta do usuário deve ser convertida em uma forma que o sistema seja capaz de identificar suas necessidades para recuperar.

A forma mais antiga e ainda mais usada de combinação de palavras-chave em perguntas é o uso de operadores booleanos (Baeza-Yates & Ribeiro-Neto, 1999).

Esses operadores, baseados na álgebra de Boole, são operadores lógicos que conectam os termos da pergunta (AND, OR ou NOT). Por exemplo, uma pergunta com os termos “cefaléia AND hipertensão” deve trazer todos os documentos em que os dois termos ocorrem simultaneamente. Em outro caso, uma pergunta “cefaléia OR hipertensão” deve trazer os documentos em que ocorre qualquer um dos dois termos (ou ambos). E por último, o operador NOT exclui documentos com um certo termo, como “cefaléia NOT hipertensão” deve trazer documentos com o termo cefaléia mas que não contenham o termo hipertensão.

Sistemas mais modernos estão procurando utilizar linguagem natural para a formulação da pergunta do usuário, porém há a necessidade de um pré-processamento da pergunta para que se torne algo compreensível ao sistema de recuperação.

O problema da busca usando operadores booleanos é que exige do usuário um conhecimento inicial de como funciona o sistema e como utilizar os operadores.

Estudos de avaliação tem verificado que muitos usuários principiantes se confundem ou usam inapropriadamente os operadores booleanos (Hersh *et al.*, 2001).

Por essa razão, existem várias pesquisas para permitir que o usuário entre com perguntas em linguagem natural, sem formatos pré-definidos e sem a necessidade de treinamento do usuário. Porém, como as perguntas não especificam qual o índice a procurar e nem a relação entre os termos, há a

necessidade de um pré-processamento da pergunta para um formato mais adequado ao processamento da recuperação.

Esse pré-processamento inclui uma parte de análise sintática, onde devemos avaliar a estrutura das palavras envolvidas, remoção de *stop-words* e realização de *stemming*. A segunda parte diz respeito à análise semântica que inclui a expansão para a busca de termos sinônimos e a identificação dos índices onde as palavras devem ser procuradas.

3.1.4 – Recuperação

O processo de recuperação consiste em comparar os termos da pesquisa com os termos do índice e retornar os documentos relevantes à pesquisa ordenados por um critério especificado. Esse critério pode ser em ordem alfabética, ordem cronológica ou por ordem de peso dos termos nos documentos (Hersh *et al.*, 2001; Wives, 1997).

A seguir veremos algumas etapas do processo de recuperação de informações.

3.1.4.1 – Semelhança Ortográfica e Semântica

Existem algumas dificuldades na recuperação de informações e dentre elas podemos citar o problema ortográfico e semântico. Muitos dos sistemas de prontuário eletrônico não possuem corretores ortográficos embutidos, e mesmo os que possuem não são capazes de englobar todos os termos médicos usados. Além disso, pela diversidade de profissionais da saúde que usam o prontuário, há vários estilos pessoais, incluindo abreviações e jargões de cada pessoa ou cada especialidade.

Quanto à ortografia, variações de escrita de uma palavra podem fazer com que a mesma não seja identificada numa busca. Por exemplo, uma busca pela palavra “dores” não achará a palavra “dor”, ou mesmo por erros de ortografia como uma busca por “sefaléia” não trará a palavra “cefaléia”. Alguns dos erros são minimizados pelas técnicas de *stemming* citadas acima, porém

para erros ortográficos necessitamos de outros tipos de tratamento, tanto ortográfico como fonético.

Devido à possibilidade de erros de ortografia, uma pergunta com o termo “dispnéia” não encontrará o termo “dispinéia” pois há um erro de ortografia na inserção. Existem basicamente quatro tipos básicos de erros de ortografia:

1. Inserção: Quando são inseridas na palavra letras ou caracteres a mais, como “cardíaco” e “cardíanco”.
2. Remoção: Quando alguns caracteres são omitidos na transcrição, como “anamnese” e “ananese”.
3. Troca: quando alguns caracteres são substituídos por outros, como “coração” e “corasão”.
4. Inversão: muitas vezes considerado como troca, mas ocorre quando as letras são trocadas de posição em uma palavra, como “epilético” e “epiléitco”.

Em sistemas informatizados, erros de digitação e ortografia constituem uma fonte muito comum de variação entre palavras. Além disso, sistemas de reconhecimento óptico de caracteres (OCR - *Optical Character Recognition*) produzem erros similares. Podemos considerar que quanto menos operações de inserção, remoção, troca e reversão de caracteres for feita para uma palavra se transformar em outra, mais similares elas serão (Hall & Dowling, 1980).

O número mínimo de inserções, remoções ou substituições para uma palavra virar outra é conhecido como *edit distance* (Baeza-Yates & Navarro, 1998).

Esse problema de aproximação de palavras, do inglês *Approximate String Matching*, pode ser definido como: dada uma palavra P de tamanho m, um texto longo T, de tamanho n e uma quantidade máxima de erros permitidos k, encontrar todas as ocorrências onde a palavra P ocorra no texto T com no máximo k erros. Esse enunciado corresponde à Distância de Levenshtein (*Levenshtein Distance*) (Baeza-Yates & Ribeiro-Neto, 1999).

Em relação ao problema semântico, o significado de uma palavra escrita ou falada é determinado pelo contexto, seja em parte, seja em seu todo. Palavras idênticas (homógrafas) diferem no significado se forem usadas em

contextos diferentes. Por outro lado, significados funcionalmente idênticos podem ser transmitidos por palavras diferentes ou sinônimas (Kent, 1972).

O assunto de termos sinônimos é especialmente problemático na medicina, pois a linguagem biomédica possui muitas trocas de termos (Chu, 2002).

De forma simples, um *thesaurus* consiste em uma lista pré-compilada de palavras importantes num dado domínio do conhecimento, onde cada palavra nessa lista possui um conjunto de palavras relacionadas (Baeza-Yates & Ribeiro-Neto, 1999). Um *thesaurus* é mais do que um dicionário de sinônimos, pois na maioria das vezes, está construído de forma hierárquica, possuindo termos mais abrangentes e outros menos abrangentes.

Cada assunto médico ou sub-assunto expressa seu conteúdo quase sempre com tipos de sentenças estereotipadas utilizando palavras específicas (Sager *et al*., 1994).

Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto para busca de documentos (Wives & Loh, 1998).

Uma possibilidade de expansão semântica para recuperação de informações médicas é o uso do DeCS. Procurar pelo termo que está sendo pesquisado e incluir na busca termos sinônimos, todos os termos abaixo e o termo acima na hierarquia, tornando a busca mais abrangente, mas tentando manter a pergunta num mesmo contexto.

3.1.4.2 – Ordenação dos Resultados

Depois de comparar os termos da pergunta com os termos indexados, uma coleção de documentos é recuperada. É necessário mostrar ao usuário em uma determinada ordem.

O método de ordenação mais simples é a alfabética ou cronológica. Nesse método de ordenação, os resultados são ordenados de forma crescente ou decrescente, mas não informam o grau de pertinência do texto com a pergunta realizada.

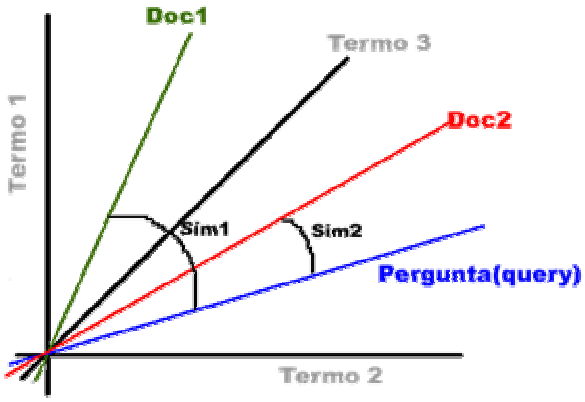
No modelo vetorial de recuperação, tanto a pergunta quanto os documentos são transformados em vetor e então é feita uma ordenação por grau de similaridade, geralmente calculado pelo ângulo de inclinação dos vetores.

Podemos representar um documento como um vetor espacial. Nesse caso, cada documento é transformado em um documento espacial, consistindo em uma coleção de um ou mais termos indexados, onde cada termo recebe um peso que indicará o grau de relevância do termo no documento. Dadas a forma vetorial de dois documentos, é possível medir o grau de similaridade entre eles através do ângulo entre os vetores, isto é, quanto menor o ângulo entre eles, mais similares serão os documentos (Salton *et al.*, 1975).

Para o modelo vetorial, o peso $w_{i,j}$ associado ao par(k_i, d_j) é positivo e não binário. A cada termo, tanto dos documentos quanto da pergunta, é atribuído um peso. Então, o vetor da pergunta q é definido como $q=(w_{1,q},w_{2,q}, ..., w_{t,q})$ onde t é o número total de termos indexados pelo sistema assim como o vetor de cada documento é representado como $d_j=(w_{1,j}, w_{2,j}, ..., w_{t,j})$ (Baeza-Yates & Ribeiro-Neto, 1999).

Podemos simplificar indicando que cada um dos t termos existentes representará uma dimensão no modelo e os pesos desses termos, para cada um dos documentos, representam as coordenadas t dimensionais dos mesmos. Veja um exemplo do uso do *Space Vector* (**Figura 1**) para três termos e o grau de similaridade com a pergunta(*query*).

Figura 1 – Space Vector de 2 documentos e 1 pergunta com 3 termos



Sim1 – Similaridade entre o Documento 1 (Doc1) e a pergunta (query)
Sim2 – Similaridade entre o Documento 2 (Doc2) e a pergunta (query)

3.1.4.3 – Avaliação

A maneira mais comum de medir o desempenho de um sistema é tempo e espaço. Quanto menor o tempo de resposta e menor o espaço usado, melhor será considerado o sistema (Baeza-Yates & Ribeiro-Neto, 1999).

Porém, em sistemas de recuperação de informação, outras métricas são também bastante importantes, denominadas *recall* (abrangência) e *precision* (precisão). Usando a tendência da literatura especializada, optou-se pelo uso dos termos em inglês.

Recall é a fração entre documentos recuperados e relevantes sobre a quantidade de documentos relevantes, enquanto que *Precision* é dada como a fração entre os documentos recuperados e relevantes sobre a quantidade de documentos recuperados (Baeza-Yates & Ribeiro-Neto, 1999; Hersh *et al.*, 2001; Hersh, 2003).

Tabela 1 – Tabela 2x2 para cálculo de recall e retrieval

	Relevantes	Não-Relevantes	Total
Recuperados	a	b	a+b
Não-Recuperados	c	d	c+d
Total	a+c	b+d	a+b+c+d

Podemos utilizar a **Tabela 1** para cálculo das medidas de desempenho do teste diagnóstico. Podemos notar que *recall* equivale à sensibilidade e *precision* equivale ao valor preditivo positivo (VPP) (Hersh, 2003).

$$\text{recall} = \text{sensibilidade} = \frac{a}{a + c} \qquad \text{precision} = \text{VPP} = \frac{a}{a + b}$$

Um dos grandes problemas encontrados para o cálculo de *Recall* e *Precision* é saber qual o total de documentos relevantes existentes na base de dados para poder comparar com os documentos relevantes recuperados pelo sistema. Obviamente, esses precisam ser descobertos por outros mecanismos, que na maioria das vezes, é feito por análise manual de todos os documentos da base, o que pode ser inviável em certos casos.

Pesquisas em recuperação de informações possuem críticas quanto à falta de bancos de dados robustos e consistentes para testes dos algoritmos. No início da década de 90 foi criada a TREC (*Text Retrieval Conference*) com o objetivo de criar bases de dados de testes e testar os algoritmos nas bases para medir o desempenho dos sistemas (Baeza-Yates & Ribeiro-Neto, 1999; Hersh, 2003).

Umas das iniciativas para a área da saúde foi a coleção OHSUMED, apresentada na TREC´94, que consiste em uma sub-coleção dos artigos indexados no Medline de 1987 a 1991. A coleção possui quase 350 mil referências e 101 perguntas que foram criadas por médicos, onde cada uma possui um resumo do paciente com a informação solicitada contendo julgamento de relevância de modo a testar os resultados produzidos por sistemas de recuperação de informações (Hersh *et al*., 1994).

O OHSUMED Corpus, cujo conteúdo é apenas em inglês, está disponível on-line no endereço ftp://medir.ohsu.edu/pub/OHSUMED/ e está dividido em 5 arquivos separados por ano.

4 - Justificativa

A quantidade de informações médicas textuais na forma digital supera qualquer outro tipo de mídia. Essa quantidade tem tido um crescimento exponencial, principalmente devido à internet. Além disso, os prontuários eletrônicos do paciente (PEP´s) tem sido adotados com freqüência cada vez maior devido à dificuldade de gerenciar o volume cada vez maior de informações clínicas.

Esses textos nos PEP´s não possuem formatos específicos e nem regras de formatação, onde o limite é apenas a imaginação do profissional que insere a informação. Nesse tipo de mídia, não existe revisão por pares como na elaboração de artigos científicos da literatura médica e, portanto, podem ocorrer erros de ortografia, uso de sinonímia e jargões específicos de uma determinada área da saúde.

Necessitamos de programas específicos voltados para a área da saúde para a recuperação de informação textual no meio desse caos, pois uma informação que não pode ser recuperada não tem razão da existência.

5 – Objetivos

Esse trabalho visou a análise de algoritmos relacionados à recuperação de informações textuais com dois objetivos principais:

1) Desenvolver um algoritmo de recuperação de informações em campos de textos livres de prontuários eletrônicos do paciente, baseado em semelhança semântica e ortográfica.
2) Comparar a recuperação de informações em textos livres utilizando ferramentas tradicionais de busca direta incorporado ao software Clinic Manager® (Sigulem *et al.*, 1994) com um software desenvolvido pelo autor (SIRIMED), utilizando o algoritmo proposto pelo trabalho.

6 - Materiais e Métodos

6.1 – Materiais

6.1.1 – Bancos de Dados

Para esse trabalho foram utilizados dois bancos de dados, sendo o primeiro banco cedido por uma clínica especializada em neurologia e psiquiatria (Instituto Campos & Cardeal) com 6732 registros de histórias clínicas em textos livres colhidas entre 17/05/2001 e 08/06/2004 e o segundo banco, uma clínica médica especializada em nefrologia e clínica médica (Clínica Médica Sigulem & Mattei S/C Ltda) com 26072 registros de histórias no mesmo formato colhidas entre 14/11/1991 e 16/08/2004. Nesse trabalho, os bancos de dados serão citados como Base 1 e Base 2 respectivamente. Ambos os bancos utilizam a plataforma MS-Access® e fazem parte de um sistema de prontuário eletrônico desenvolvido pelo Departamento de Informática em Saúde (DIS) da UNIFESP chamado Clinic Manager® (Sigulem *et al*., 1994).

Para não comprometer o sigilo das informações, todos os nomes dos pacientes foram trocados por suas iniciais para evitar a identificação e para que seja mantida a confidencialidade dos seus dados. Nesses termos, o projeto foi analisado e aprovado pelo Comitê de Ética em Pesquisa sobre o protocolo CEP 1500/03 datado de 19/12/2003 (**Anexo V**).

6.1.2 – Vocabulário Controlado Médico DeCS

Foi utilizada a versão DeCS 2004 que possui 26851 termos em Português e 31936 formas sinônimas para os termos autorizados. Durante a dissertação, essa base de dados será mencionada como vocabulário DeCS.

O uso da versão DeCS em português para fins acadêmicos foi autorizado pelo diretor da BIREME/OPAS/OMS, Abel L. Packer, conforme **Anexo III**.

6.1.3 – Dicionário Português Ispell

Para executar uma análise estatística de termos que pertencem à língua portuguesa, foi utilizado o dicionário br.ispell do português falado no Brasil, versão 2.4 de outubro de 1999. Esse dicionário está disponível nos termos da licença GNU GPL e pode ser utilizado livremente possuindo 225.502 termos (Karpischek, 1999).

Nesse trabalho, esse dicionário será mencionado como dicionário Ispell. Já existe uma versão br.ispell 3.0 beta4 de 25 de março de 2003, porém suas alterações não incluem novos termos, apenas correção de falhas e novas características que não foram utilizadas nesse trabalho.

Para se ter idéia da abrangência do dicionário utilizado, comparamos a quantidade aproximada de entradas de três grandes dicionários da língua portuguesa:

a) Michaelis[1] com 200 mil termos
b) Aurélio[2], 3ª edição com 435 mil termos
c) Houaiss[3] com 228 mil termos

6.1.4 - Lista de s*top words* do projeto Snow Ball

Para a remoção de *stop words* foi utilizada a lista proposta pelo projeto Snow Ball (http://snowball.tartarus.org/) composta por 220 termos incluindo proposições, artigos, pronomes, verbo estar, haver, ser, ter e suas variações. Essa lista será mencionada durante o trabalho como s*top list* (**Anexo II**).

6.1.5 – Softwares e Hardware

Para a implementação e teste do algoritmo foi utilizada a linguagem de programação Visual Basic 6.0® (VB) devido à ampla experiência pessoal do autor nessa linguagem. A conexão com o banco de dados foi feita com ADO (*ActiveX Data Objects*), com manipulação utilizando SQL.

---

[1] Michaelis (Moderno Dicionário da Língua Portuguesa) - http://www2.uol.com.br/michaelis/
[2] Aurélio, 3ª ed. - http://www.aureliopositivo.com.br/aurelio/prod_aur/default.asp
[3] Houaiss - Dicionário da Língua Portuguesa - http://www.dicionariohouaiss.com.br/index2.asp

Como padrão de comparação de eficiência na busca dos termos, foi utilizado o Clinic Manager® versão 7.0.7.85 – Versão de Avaliação, Registro no INPI – Protocolo nº 00040706, desenvolvido pelo Departamento de Informática em Saúde da Unifesp. Esse programa faz uma busca direta nos textos sem nenhum tratamento, podendo assim comparar os resultados com o uso do algoritmo e sem o uso do mesmo.

A codificação e testes do programa foram executados num PC (*Personal Computer*) com as seguintes configurações: Processador ATHLON XP 2.6 MHz, 512 MB de memória RAM, 80GB de disco rígido.

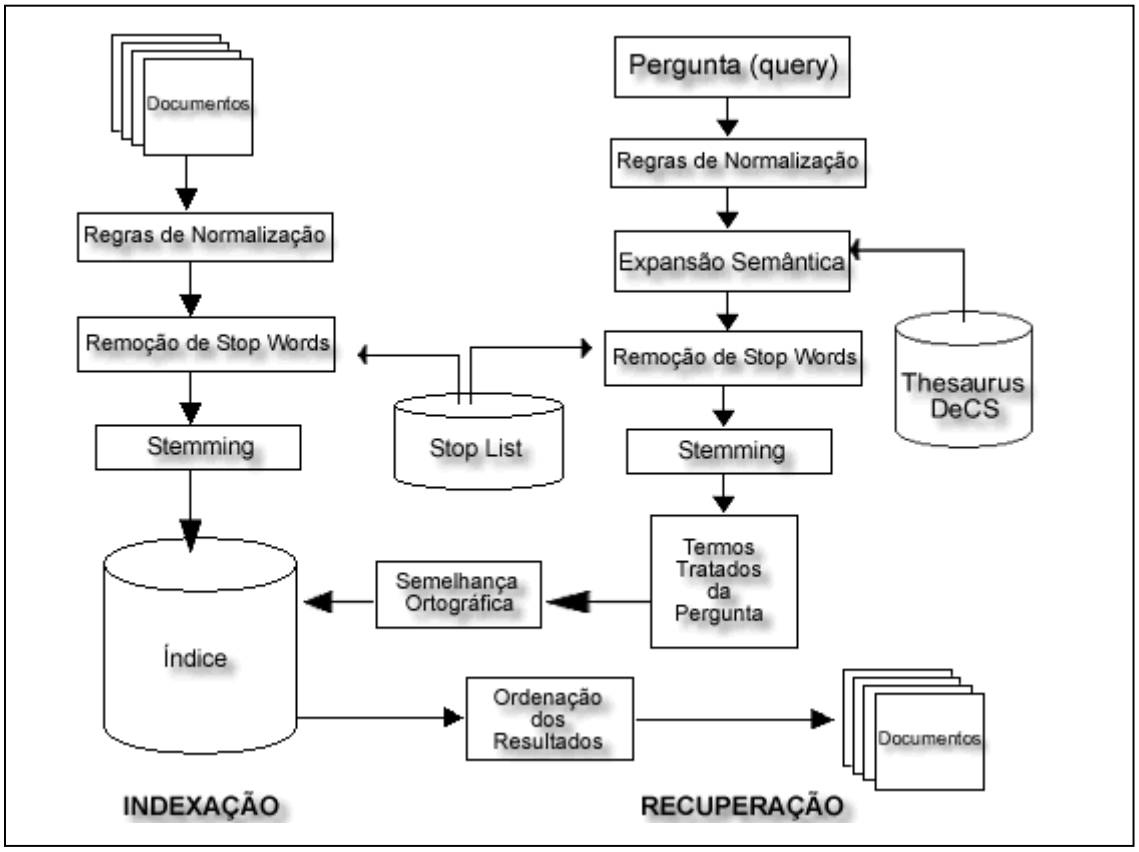
6.2 – Métodos

O método foi dividido em duas etapas. A primeira etapa, chamada de Indexação Automática, consiste na criação do índice de pesquisa que será comparado posteriormente com os termos da pergunta do usuário. A segunda, chamada Recuperação, consiste na recuperação dos textos originais de acordo com uma pergunta do usuário baseado em alguns critérios selecionados pelo usuário como uso de sinonímia, distância de contexto e semelhança ortográfica que serão comentados com mais detalhes adiante.

O algoritmo proposto foi implementado num sistema batizado SIRIMED (Sistema de Indexação e Recuperação de Informações Médicas) com o propósito de testar o algoritmo.

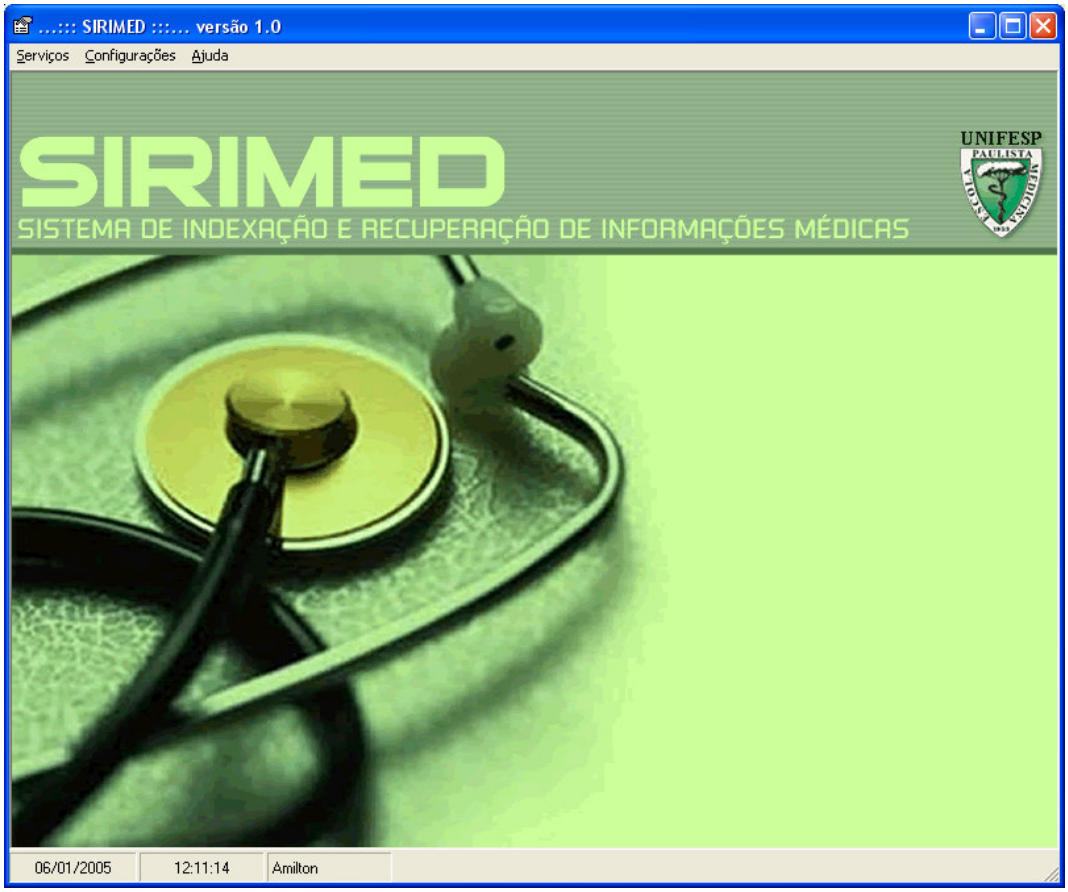
A estrutura do algoritmo pode ser vista na **Figura 2**. O SIRIMED não cria uma cópia dos textos originais e sim cria ponteiros no seu índice invertido que fazem referência ao texto original que é mantido no banco de dados do Clinic Manager®. A tela inicial do sistema é apresentada na **Figura 3**.

Figura 2 - Esquema das etapas de indexação e recuperação do SIRIMED



*Regras de Normalização: remoção de formatação, transformação em minúsculas, remoção de caracteres especiais.

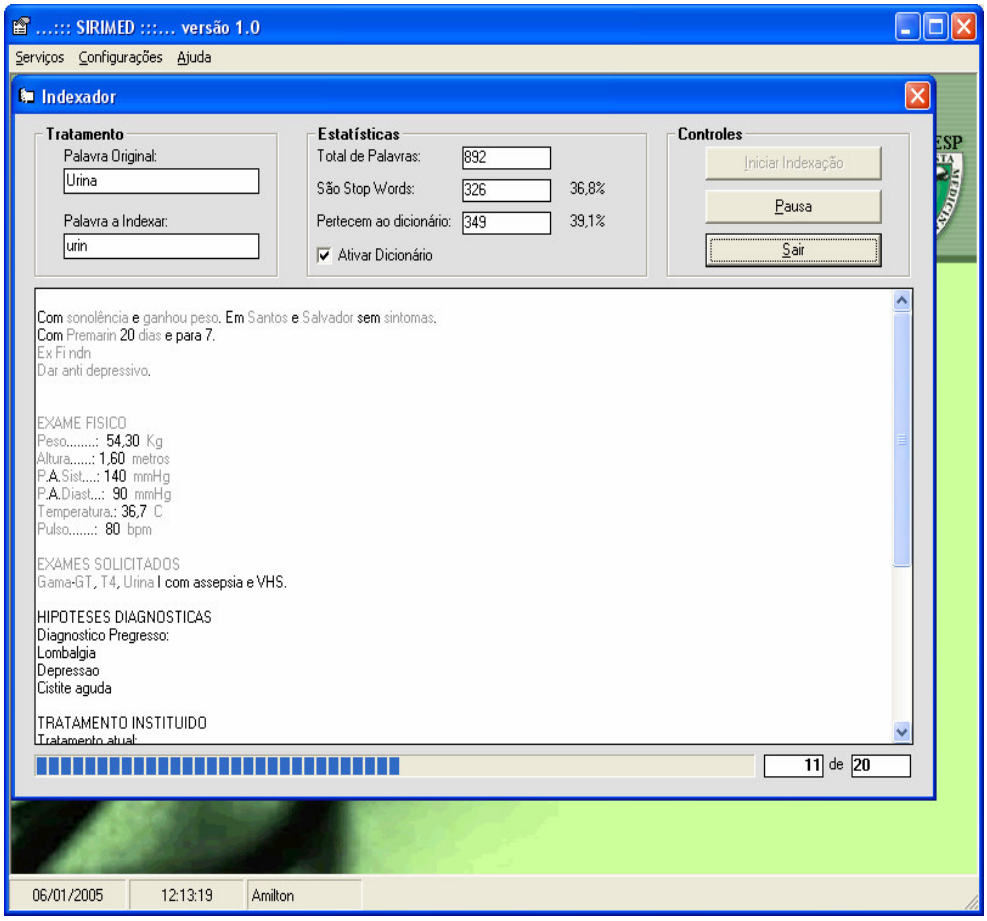
Figura 3 – Tela Inicial do SIRIMED



6.2.1 – Indexação Automática

A primeira etapa do método consiste na Indexação Automática dos textos, ou seja, a criação do índice invertido com os termos e seus atributos para a posterior recuperação. Nessa etapa os termos são tratados e inseridos no índice de forma a facilitar a recuperação posterior. Algumas regras de normalização dos textos foram definidas, como a conversão dos formatos originais para textos puros, transformação de todo o texto em minúsculas e remoção de caracteres especiais.

Figura 4 – Módulo Indexador



6.2.1.1 – Conversão dos diversos formatos para texto

Os textos estavam armazenados no banco de dados no formato RTF (*Rich Text Format*). Primeiramente, os textos foram convertidos para o formato de texto puro utilizando o objeto RichTextBox do Visual Basic.

6.2.1.2 – Normalização dos termos

Para normalizar os termos foram removidos todos os caracteres especiais e substituídos por um espaço. Os caracteres removidos se encontram na **Tabela 2**.

Tabela 2 – Caracteres removidos

! # $ * & ( ) - _ = + | \ [{] } : ; , . ? / < '

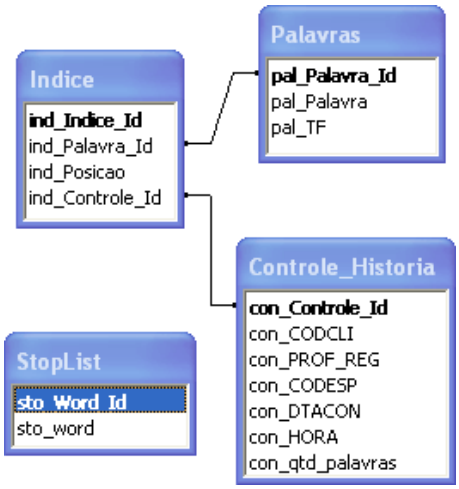
Além disso, as palavras foram convertidas para letras minúsculas e foram removidos caracteres especiais e trocados por caracteres normais, conforme a **Tabela 3**:

Tabela 3 – Caracteres Trocados

de	para
á, à, ã, â, ä	a
é, ê, è, ë	e
í, î, ì, ï	i
ó, ò, ô, õ, ö	o
ú, ù, û, ü	u
ç	c
ñ	n

6.2.1.3 – Índice Invertido

Figura 5 - Relacionamento entre as tabelas do sistema



A tabela “Palavras” possui uma chave primária auto-numérica (pal_Palavra_Id), um campo para o registro da palavra/termo (pal_Palavra) e um campo para o registro da freqüência do termo (*term frequency*) no conjunto de textos (pal_TF).

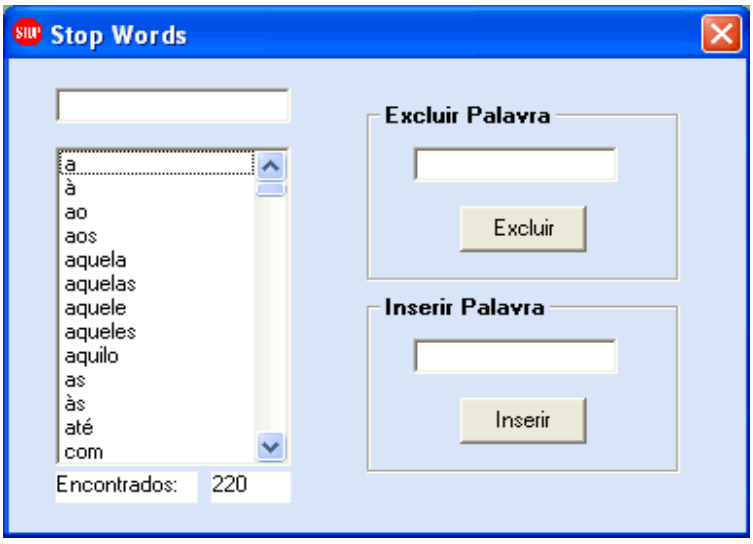
A tabela de histórias do banco de dados original possui uma chave primária composta pelos campos CODCLI (código do cliente/paciente), PROF_REG (registro do profissional), CODESP (código da especialidade do profissional), DTACON (data da consulta) e HORA (hora da consulta). Para facilitar o controle de histórias, foi criada uma tabela auxiliar chamada “Controle_Historia” com uma chave primária auto-numérica que representará a história no banco de indexação, além de possuir um campo “con_qtd_palavras” para o registro da quantidade de palavras indexadas por texto.

A tabela “Indice" representa o índice invertido do sistema, tendo também uma chave primária auto-numérica “ind_Indice_Id”, uma referência para o identificador da palavra/termo encontrada “ind_Palavra_Id”, um controle de posicionamento no texto em que foi encontrada “ind_Posicao" e uma referência para a história que ocorreu o termo “ind_Controle_Id”.

6.2.1.4 – Remoção de s*top words*

Após o tratamento inicial dos textos, eles foram varridos e todas as palavras encontradas, excetuando-se s*top words* e números, passavam para a próxima etapa de tratamento. Para a remoção de s*top words*, foi criada uma tabela de nome “StopList” com a lista sugerida pelo projeto Snow Ball (**Figura 5**).

Figura 6 – Tela de Inserção e Remoção de stop words



O sistema permite inserir ou remover palavras da lista de *stop words* (**Figura 6**). Palavras como ‘diagnóstico’, ‘retorno’, ‘dias’, ‘exame’ são tão comuns que podem ser inseridas na stop list. Nesse sistema foi mantida a lista original de 220 palavras.

6.2.1.5 – Tratamento de *Stemming*

Após a remoção de *stop words*, os termos restantes foram processados pelo algoritmo de *Stemming* para o Português de Martin Porter (SnowBall, 2005) e inseridos no índice invertido. Para cada termo foi contabilizada a freqüência do termo (*Term Frequency*-TF), em qual texto o termo aparece e qual a posição em que ele ocorre.

O princípio desse algoritmo é a remoção de sufixos. A remoção é dividida em 5 passos, sendo o primeiro a remoção de sufixos comuns. No **Anexo IIA** estão listados os sufixos comuns que são removidos ou substituídos pelo sufixo na raiz. Por exemplo, os sufixos “logia” e “logias” são substituídos por “log” como nas palavras “tecnologias” e “biologia”, “encia’ e “encias” são substituídos por ente, como nas palavras “paciência” e “ecidências”. Para os sufixos “ista” ou “oso”, são apenas removidos como nas palavras “especialista” ou “medicamentoso”.

O segundo passo é a remoção de sufixos de verbos regulares, listados no **Anexo IIB**.

Por fim, os passos 3, 4 e 5 se referem a remoção de sufixos residuais, sendo os seguintes: "os", "a", "i", "o", "e", "s".

6.2.1.6 – Análise de pertinência ao vocabulário médico e dicionário

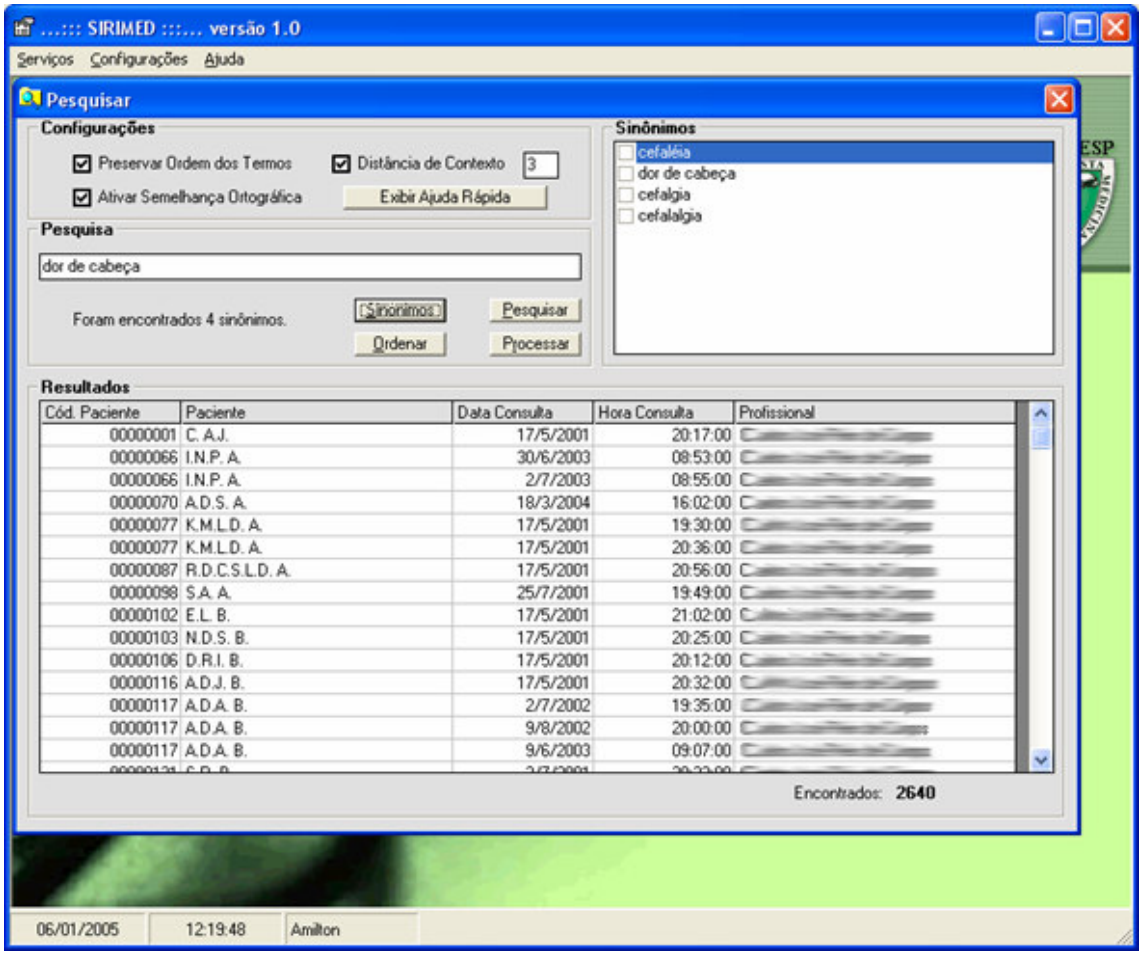
Para cada palavra nos textos, foi verificado se pertence ao dicionário da língua portuguesa Ispell (Karpischek, 1999), se pertence à lista de s*top words* (Snowball, 2002) e se pertence ao vocabulário médico controlado DeCS (Bireme, 2005). Esta estatística mostra a porcentagem de palavras que

pertencem aos vocabulários e, principalmente, a quantidade de palavras que não pertencem e devem ser tratadas para posterior recuperação.

6.2.2 – Recuperação

A segunda etapa da pesquisa consiste na recuperação dos textos baseado numa pergunta do usuário. Nessa tela do sistema (**Figura 7**), o usuário entra com a sua pergunta (*query*) podendo configurar alguns detalhes da busca como Preservação da Ordem dos Termos, Semelhança Ortográfica, Distância de Contexto e Inserção de Sinônimos.

Figura 7 - Módulo de Recuperação do SIRIMED



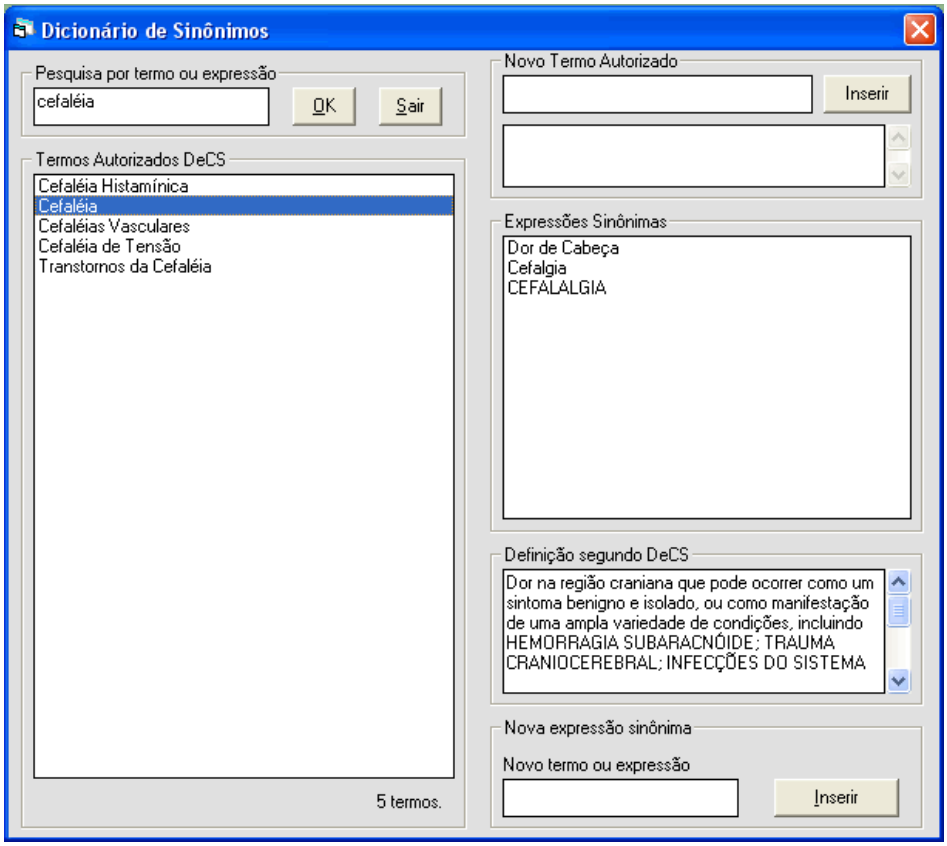
A pergunta deve passar por um pré-processamento semelhante ao da indexação, incluindo a transformação em letras minúsculas, remoção de caracteres especiais e de *stop-words* e por fim *stemming*.

Para termos compostos como ‘dor de cabeça’ ou ‘nega meningite’ foi definida uma distância de contexto, sendo considerado um padrão de três, ou seja, as palavras devem estar numa distância máxima de três termos para que sejam encontradas. Com isso, na busca de ‘dor de cabeça’ também serão encontrados termos com ‘dor forte de cabeça’ ou ‘dor na cabeça’ visto que a palavra ‘de’ é *stop-word*. Além disso, o usuário tem a opção de ‘Preservar a ordem dos termos’ ou não, para recuperar “dor de cabeça forte” para a pergunta “dor forte de cabeça”.

Para essa a funcionalidade de recuperação de termos compostos foi desenvolvida uma rotina recursiva que procura, inicialmente, todas as ocorrências do primeiro termo. Para cada ocorrência, procura se existe o próximo termo numa distância máxima estipulada pelo usuário, mantendo ou não a ordem e assim por diante até o último termo da pergunta.

6.2.2.1 – Semelhança Semântica

Figura 8 – Tela de Controle de Dicionário de Sinônimos



O sistema possui um Thesaurus interno que está carregado com o vocabulário DeCS e pode ser editado pelo usuário (**Figura 8**). Do lado esquerdo da tela o usuário pode fazer uma pesquisa nos termos autorizados do DeCS. Ao clicar em um dos termos autorizados resultantes da pesquisa, é listada do lado direito a lista de sinônimos para a expressão. Além disso, no lado superior direito o usuário ainda pode inserir novos termos autorizados ou inserir novos sinônimos para um termo no canto inferior direito.

Na parte de consulta (**Figura 7**), o usuário pode selecionar a opção de 'Sinônimos' e o sistema lista todos os sinônimos para o termo. Caso seja um termo autorizado, o sistema lista todos os sinônimos, caso seja um sinônimo, o sistema busca qual o termo autorizado equivalente e em seguida lista os sinônimos do mesmo.

Para inserir os termos na pesquisa, o usuário pode selecionar cada sinônimo que achar relevante.

6.2.2.2 – Semelhança Ortográfica

Ao clicar na opção de 'Semelhança Ortográfica' (**Figura 7**), o sistema procura os termos cujas raízes tenham um *edit distance* de, no máximo, um. Esse processo só será realizado se a raiz tiver mais de quatro caracteres. Isso foi definido para evitar que palavras curtas com uma letra de diferença possam ser recuperadas como a mesma, como 'dor' e 'cor' ou 'sono' e 'sino'.

A função que calcula o *edit distance* por ser vista no **Anexo IV**.

6.2.3 – Ordenação dos Resultados

Inicialmente os resultados são exibidos na ordem de inserção das histórias no banco de dados, porém, o usuário pode ordenar os resultados utilizando o algoritmo do vetor espacial para ordenar pela freqüência dos termos da pesquisa. Para isso, basta o usuário clicar no botão 'Ordenar' (**Figura 7**).

6.2.4 – Critérios de Inclusão e Exclusão

Foram selecionadas as 200 raízes dos termos mais freqüentes na coleção de histórias das Bases 1 e 2. Dentre esses termos, foram excluídos aqueles que não representavam sintomas, sinais, diagnóstico ou medicamentos.

Esses critérios visam contextualizar a busca na área médica. Dentre os elementos excluídos, podemos citar nomes pessoais (médicos e pacientes), palavras muito freqüentes, porém que não tinham poder semântico isoladamente como "dias", "exame", "retorno", "data", "diagnóstico" e outras.

Os termos selecionados foram:

Base 1 – Neurologia / Psiquiatria			
1. cefaléia	6. meningite	11. gardenal	16. cbz
2. enxaqueca	7. ronco	12. latejante	17. diabetes
‘3. convulsão	8. tegretol	13. desmaio	18. ansiedade
4. tontura	9. tremor	14. depressão	
5. nervoso	10. vômitos	15. insônia	

Base 2 – Nefrologia / Clínica Médica			
1. dor	5. plaquetas	9. edema	13. febre
2. triglicérides	6. hemoglobina	10. diabetes	14. vioxx
3. creatinina	7. hematocrito	11. moduretic	15. diprospan
4. leucócitos	8. obesidade	12. renitec	16. antak

6.2.5 – Análise Estatística

Neste trabalho, todas as palavras das duas bases de dados foram indexadas, ou seja, toda a população foi trabalhada e não somente uma amostra. Como os resultados obtidos consistem na realidade completa das bases, não há necessidade de aplicar métodos estatísticos para extrapolar os resultados, visto não termos trabalhado com uma amostra.

Podemos considerar que as duas bases de dados são amostras de Prontuários Eletrônicos do Paciente (PEP), porém, nesse caso, os resultados não podem ser extrapolados para a população de PEP´s pois a amostra (n=2) é muito pequena.

Portanto, o trabalho foi analisado como estudo de caso e optou-se por uma análise descritiva dos resultados obtidos ao invés de uma análise inferencial.

A análise estatística do trabalho foi realizada com apoio e orientação da disciplina de Bioestatística do Departamento de Medicina Preventiva da UNIFESP.

7 – Resultados

Tabela 4 – Freqüência de palavras que pertencem ao dicionário e vocabulário médicos.

Base de Dados	Quantidade de Histórias	Quantidade de Palavras	Pertencem ao dicionário Ispell	Pertencem ao vocabulário DeCS
Base 1	6.732	830.471	481.947 (58%)	72.689 (8,8%)
Base 2	26.072	3.990.900	1.716.902 (43%)	289.902 (7,3%)

Podemos notar na **Tabela 4** que, mesmo existindo um corretor ortográfico embutido na inserção das informações médicas em forma de textos livres, somente cerca da metade das palavras pertencem ao dicionário da língua portuguesa Ispell e poderão ter a ortografia corrigida.

Nesse caso, ainda serão necessárias técnicas para recuperação por semelhança ortográfica, visto que não será possível corrigir automaticamente palavras não pertencentes ao dicionário utilizado.

Além disso, num contexto médico, muitos termos usados são jargões locais ou da especialidade, abreviações e termos sinônimos que mesmo um vocabulário especializado não possui, como se pode notar, menos de 9% das palavras utilizadas pertencem ao vocabulário DeCS.

Tabela 5 – Quantidade de *stop words* e tamanho do índice criado

Base de Dados	Base 1	Base 2
Quantidade de palavras	830.471	3.990.900
Quantidade de s*top words*	270.913 **(32,6%)**	1.511.763 **(37,9%)**
Palavras que foram indexadas	559.558	2.479.137
Tamanho do índice sem *Stemming*	26.977 **(4,8%)**	38.179 **(1,5%)**
Tamanho do índice com *Stemming*	19.543 **(3,5%)**	26.781 **(1%)**

A remoção de s*top words* reduz cerca de 40% da quantidade de palavras a serem inseridas no índice, incluindo as palavras constantes na s*top list* e os algarismos conforme **Tabela 5**.

Devido ao uso de *stemming*, o índice também foi reduzido pela união de palavras com o mesmo radical comum na mesma entrada do índice. Na Base 1, de 26.977 palavras reduziu para 19.543 e na Base 2, de 38.179 palavras

reduziu para 26.781, representando uma redução de cerca de 30% no tamanho do índice em ambos (**Tabela 5**).

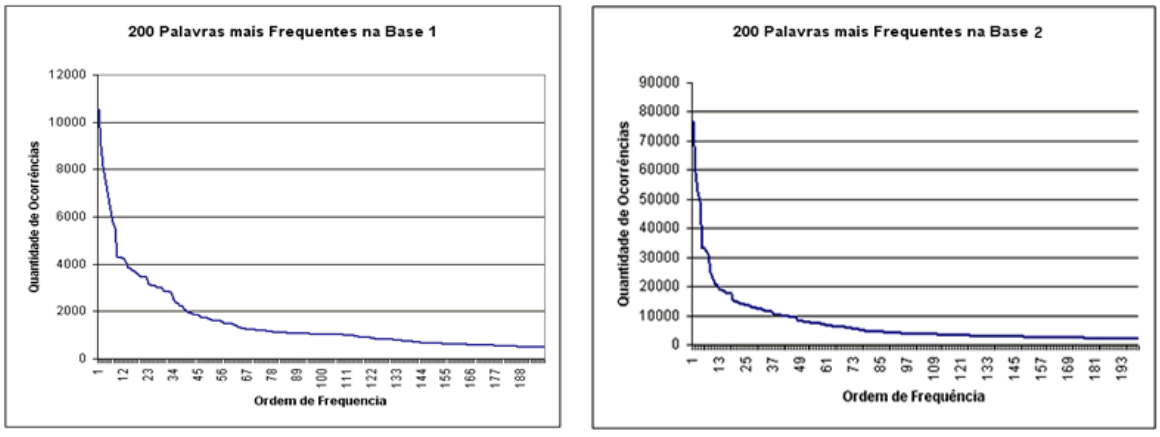
Tabela 6 – Tempo de indexação das Bases de Dados

Base de Dados	Base 1	Base 2
Quantidade de Histórias	6.732	26.072
Quantidade de Palavras	830.471	3.990.900
Média de Palavras por História	123	153
Tempo de indexação com dicionário Ispell	11:36:39 *	50:26:16 *
Tempo de indexação sem dicionário Ispell	05:25:41 *	30:10:17 *
Média de indexação por história com dicionário Ispell	6,2 segundos	7,0 segundos
Média de indexação por história sem dicionário Ispell	2,9 segundos	4,2 segundos

* Tempo expresso em hh:mm:ss

Um dos fatores que deixa a indexação mais lenta é o fato de verificar se cada uma das palavras dos textos pertence ao dicionário Ispell da Língua Portuguesa, como notado na **Tabela 6**. A utilização do dicionário serviu principalmente para fins experimentais e estatísticos, e o software SIRIMED permite não utilizá-lo na indexação.

Gráfico 1 – Distribuição de Freqüência das 200 palavras mais freqüentes nas Bases



No **Gráfico 1** notamos que a distribuição da freqüência das palavras possui tendência exponencial. Isso quer dizer que muitas palavras se repetem

várias vezes, e o índice com menos de 5% da quantidade de palavras indexadas, representa o vocabulário utilizado nas histórias conforme **Tabela 5**.

Na **Base 1**, 5% das palavras, ou seja, cerca de 957 palavras correspondem a 83% de todo o conteúdo indexado. Isso se repete na **Base 2**, onde 5% das palavras, ou seja, cerca de 1340 palavras correspondem a 89% do conteúdo indexado. Isso ratifica a idéia de que conseguimos expressar a maioria das idéias inseridas em textos livres de prontuários eletrônicos com um subconjunto reduzido de palavras.

Tabela 7 – Quantidade de profissionais que inseriram textos

Base de Dados	**Base 1**	**Base 2**
Quantidade de profissionais que inseriram textos	10	6

Com 16 profissionais inserindo textos, podemos remover o viés de vício de escrita de uma determinada pessoa (**Tabela 7**). Claro que vícios de linguagem referente a cada especialidade não foram removidas e nem deveriam, pois o motivo desse trabalho é criar métodos para a recuperação de informações com as características da linguagem médica.

Tabela 8 – Comparação da Recuperação de Histórias com e sem o algoritmo da Base 1

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**%**	**ST+SO**	**%**	**ST+SO+SS**	**%**
desmaio	**258**	357	*38,4*	701	*171,7*	**760**	*194,6*
ronco	**341**	614	*80,1*	614	*80,1*	**614**	*80,1*
tontura	**441**	766	*73,7*	776	*76,0*	**776**	*76,0*
nervoso	**569**	912	*60,3*	912	*60,3*	**912**	*60,3*
tremor	**206**	324	*57,3*	327	*58,7*	**327**	*58,7*
vômitos	**475**	636	*33,9*	651	*37,1*	**651**	*37,1*
insônia	**293**	375	*28,0*	394	*34,5*	**394**	*34,5*
cefaléia	**1988**	2265	*13,9*	2359	*18,7*	**2640**	*32,8*
diabetes	**322**	409	*27,0*	415	*28,9*	**415**	*28,9*
convulsão	**1295**	1312	*1,3*	1581	*22,1*	**1581**	*22,1*
depressão	**333**	386	*15,9*	387	*16,2*	**389**	*16,8*
cbz	**336**	344	*2,4*	344	*2,4*	**390**	*16,1*
enxaqueca	**940**	992	*5,5*	1009	*7,3*	**1059**	*12,7*
ansiedade	**366**	380	*3,8*	383	*4,6*	**383**	*4,6*
latejante	**495**	503	*1,6*	517	*4,4*	**517**	*4,4*
gardenal	**249**	254	*2,0*	256	*2,8*	**256**	*2,8*
tegretol	**299**	303	*1,3*	305	*2,0*	**305**	*2,0*
meningite	**1063**	1066	*0,3*	1080	*1,6*	**1080**	*1,6*
Média			24,8		35,0		38,1

- CM – Clinic Manager,
- ST – Uso de Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Notamos na **Tabela 8** que somente o processo de *stemming* aumenta a recuperação (chegando a 80,1% em ronco), mas que combinado com a semelhança ortográfica, o aumento na recuperação alcança 171% (como em desmaio) e, de acordo com a quantidade de sinônimos encontrados, a recuperação semântica pode até triplicar (194% em desmaio).

Alguns termos como ‘ronco’ e ‘tontura’ tiveram aumento somente com o uso de *stemming* (80,1% e 73,7%) ao passo que outras esse processo aumentou bem menos a recuperação como ‘meningite’ e ‘tegretol’ (0,3% e 1,3%).

Em alguns casos, como nos termos ‘ronco’, ‘nervoso’ e ‘cbz’, a inserção da semelhança ortográfica não aumentou em nada na recuperação.

Tabela 9 – Comparação da Recuperação de Freqüência de Palavras com e sem o algoritmo da Base 1

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**%**	**ST+SO**	**%**	**ST+SO+ SS**	**%**
desmaio	**344**	485	*41,0*	1011	*193,9*	**1249**	*263,1*
ronco	**680**	1224	*80,0*	1244	*82,9*	**1244**	*82,9*
tontura	**511**	921	*80,2*	921	*80,2*	**921**	*80,2*
nervoso	**666**	1111	*66,8*	1111	*66,8*	**1111**	*66,8*
tremor	**377**	490	*30,0*	558	*48,0*	**619**	*64,2*
vômitos	**3577**	4274	*19,5*	4828	*35,0*	**5760**	*61,0*
insônia	**465**	709	*52,5*	719	*54,6*	**719**	*54,6*
cefaléia	**511**	690	*35,0*	705	*38,0*	**705**	*38,0*
diabetes	**480**	530	*10,4*	530	*10,4*	**656**	*36,7*
convulsão	**357**	471	*31,9*	481	*34,7*	**481**	*34,7*
depressão	**1589**	1606	*1,1*	2117	*33,2*	**2117**	*33,2*
cbz	**1519**	1609	*5,9*	1658	*9,2*	**1778**	*17,1*
enxaqueca	**427**	492	*15,2*	495	*15,9*	**500**	*17,1*
ansiedade	**833**	881	*5,8*	891	*7,0*	**891**	*7,0*
latejante	**597**	616	*3,2*	629	*5,4*	**629**	*5,4*
gardenal	**504**	512	*1,6*	527	*4,6*	**527**	*4,6*
tegretol	**455**	469	*3,1*	474	*4,2*	**474**	*4,2*
meningite	**1079**	1082	*0,3*	1097	*1,7*	**1097**	*1,7*
Média			26,9		40,3		48,5

- CM – Clinic Manager,
- ST – Uso de Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

A **Tabela 8** mostra a quantidade de histórias recuperadas e a **Tabela 9** a freqüência de palavras recuperadas, visto que dentro de cada história pode ocorrer a palavra mais de uma vez.

Todavia, as observações comentadas na **Tabela 8** são semelhantes às da **Tabela 9**. Os termos ‘ronco’ e ‘tontura’ ainda possuem o maior aumento na recuperação (80% e 80,2%). Nos termos ‘ronco’, ‘nervoso’ e ‘cbz’ a inserção da semelhança ortográfica também não aumentou em nada na recuperação.

Tabela 10 – Quantidade de variações dos termos encontrados com os algoritmos na Base 1

Termos	CM	ST	Termo Exato	Variações Recuperadas	ST+SO	Variações Recuperadas	ST+SO+SS	Variações Recuperadas
desmaio	344	485	352	9	1011	22	1249	32
tontura	680	1224	690	3	1244	19	1244	19
ronco	511	921	513	9	921	9	921	9
nervoso	666	1111	679	6	1111	6	1111	6
insônia	377	490	466	5	558	10	619	11
cefaléia	3577	4274	4157	9	4828	47	5760	48
tremor	465	709	466	3	719	11	719	11
vômitos	511	690	581	13	705	22	705	22
cbz	480	530	487	5	530	5	656	10
diabetes	357	471	360	8	481	14	481	14
convulsão	1589	1606	1601	3	2117	26	2117	26
enxaqueca	1519	1609	1556	7	1658	23	1778	38
depressão	427	492	490	2	495	4	500	5
tegretol	833	881	836	7	891	14	891	14
gardenal	597	616	602	3	629	10	629	10
latejante	504	512	509	1	527	7	527	7
ansiedade	455	469	469	0	474	3	474	3
meningite	1079	1082	1080	1	1097	9	1097	9

- CM – Termos recuperados no Clinic Manager;
- ST – Termos recuperados com o uso de Stemming
- Termo Exato – Termos recuperados com a string exata
- SO – Termos recuperados com o uso de Semelhança Ortográfica
- SS – Termos recuperados com o uso de Semelhança Semântica
- Variações Recuperadas – quantidade de variações do termo original

Na **Tabela 10** podemos notar que foram recuperadas mais ocorrências do termo original no SIRIMED do que no Clinic Manager®. Note que o termo ‘desmaio’ encontrou 32 variações do termo, incluindo variações ortográficas e sinônimos.

Mesmo para o termo ‘ansiedade’, ainda assim foram encontrados 3 variações do termo na incorporação da semelhança ortográfica.

Gráfico 2a – Evolução da porcentagem de recuperação em cada algoritmos por termo da Base 1

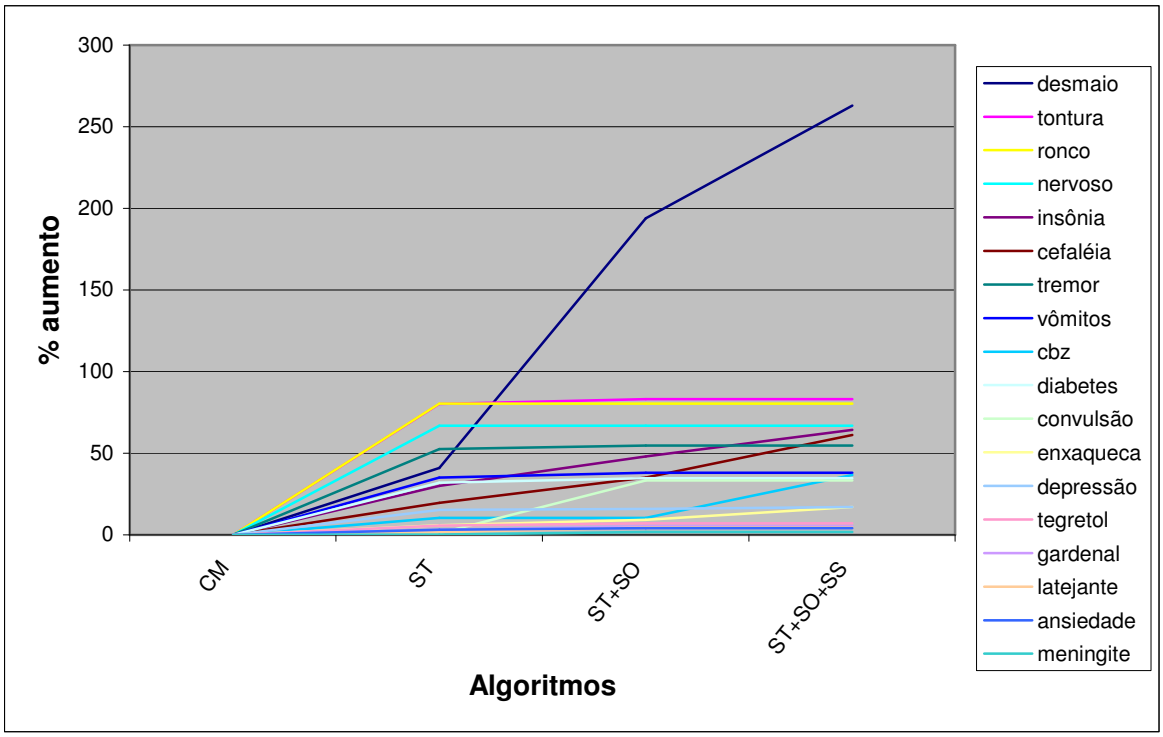
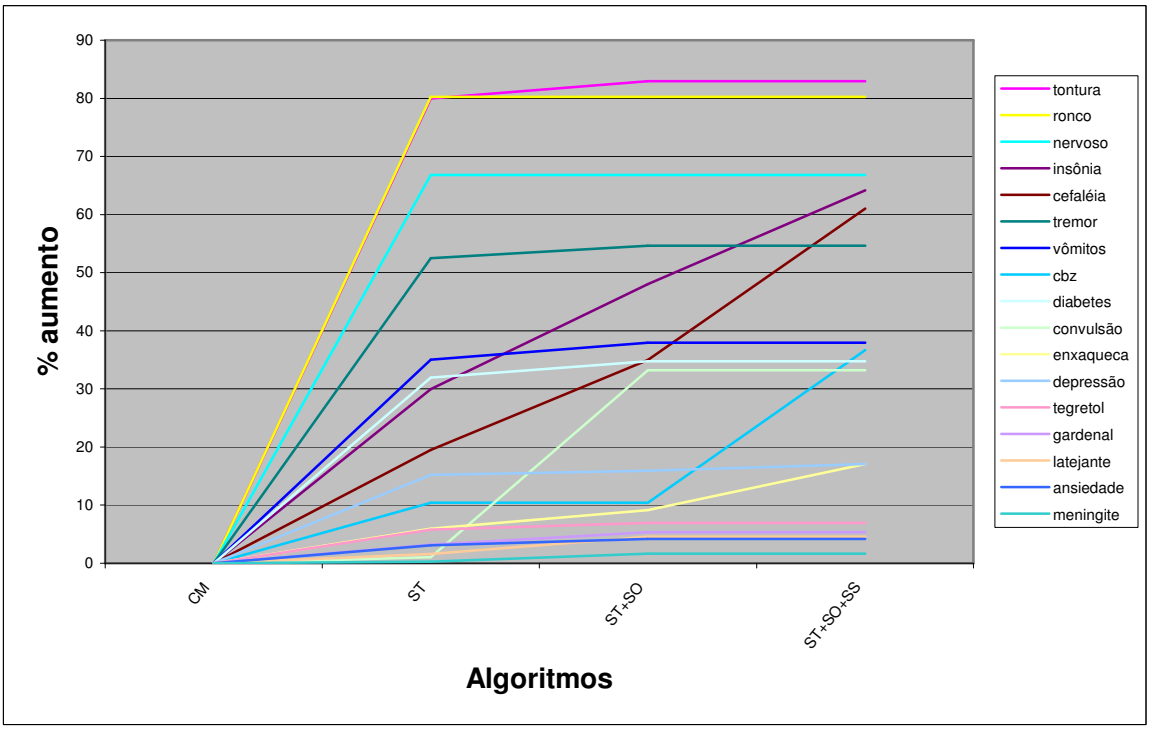


Gráfico 2b – Reprodução do Gráfico 2a sem o termo 'desmaio'



- ST – Uso de Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Os **gráficos 2a e 2b** mostram a evolução da porcentagem de recuperação dos termos com a incorporação de cada algoritmo ao processo. No **gráfico 2a** incluíram-se todos os termos e no **gráfico 2b**, o termo 'desmaio' foi removido por se comportar de forma anormal (outlier).

Nota-se claramente no **gráfico 2b** que existe um grande aumento até o uso do *stemming*, mas a partir dele, há uma tendência à nivelação do aumento na maioria dos termos analisados.

Tabela 11 – Variações Incorporadas na busca dos termos da Base 1

Termos	Termos Adicionados pelo ST	Termos Adicionados pelo SO	Sinônimos Incorporados	Termos Adicionados pela SS
desmaio	desmaiou; desmaiso; desmaiado; desmaiar; desmaiada; Desmaiava; desmaiei; desmaisos; desmaiara	Desmaios; desmaia; desmio; desmais; desamaio; desmame; Desmaou; desmiou; Desmanio; demaio; desmai; demaiou; demaisos	(4) Síncope, ataque por queda, pré-síncope, síncope postural	Siíncope; sínbcope; síncocpe; sincopais; Sincope; pre-síncope; Sincopes; pré-síncopes; sínncope; syncope
tontura	tonturas; Tonturs; tontures	tonntura; toturas; tonrturas; Tontuas; tontua; tointura; Tonturras; tuntura; ontura; tontuura; totura; tonbtura; tunturas; tonura; contura; tntura	(2) sensação de cabeça leve, ortoestase	-
ronco	ronca; roncos; Ronco; Roncador; roncava; roncado; roncando; roncar; roncam; roncadora	-	-	-
nervoso	Nervos; nervosa; nervo; nervosos; nervosas; nervso	-	-	-
insônia	insonia; Insônica; insõnia; insônida; insôni	insônio; insônias; sinsônia; inônia; insolação	(3) distúrbios do início e da manutenção do sono	distúrbios do início e da manutenção do sono
cefaléia	cefaléa; cefalé; cefaleia; cefaléas; Cefáleia; cefaléi; cefaléoa; cefaléis; cefaléios	cafaléia; cealéia; cefaaléia; Cefaéia; cefal; Cefalaéia; cefaleái; cefaléai; Cefaleáis; cefaléais; cefaleéia; cefaleía; cefaleias; cefaleías; cefaléias; cefaléie; Cefaléioa; cefaléios; cefaléua; cefaléui; cefalía; cefálico; cefálicos; cefalie; cefalíe; cefaliea; cefalíea; cefaliéas; cefaliéia; Cefaliu; cefalkéia; ceffaléia; Ceflaléia; cefléia; cegfaléia; ceraléia; cerfaléia; sefaléia	(3) dor de cabeça, cafalgia, cefalalgia	dor de cabeça; dores de cabeça; dor de cabeç; dore de cabeça
tremor	tremores; Tremore; tremors	remor; Tremorres; temores; temor; tremeor; rtremor; tremortes; Trmor	(3) tremor de ação, tremor de intenção, tremor de repouso	-
vômitos	vômitar; vômitou; vomitos; vômito; vomito; vomitado; vomitar; vomitou; vomitava; vomita; võmitos; vómitos; vomitando	vcômitos; vôitos; vômiots; v6omitos; omitindo; vômnitos; vômigos; v^mito; vômtios	(1) Emese	-
cbz	CBZ200mg; CBZ400; CBZ400mg; CBZ1200; CBZ200	-	(1) Carbamazepina	cabamazepina; carbamazapina; carbamazepiana; carbamazepina; Carbanazepina

- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Termos	Termos Adicionados pelo ST	Termos Adicionados pelo SO	Sinônimos Incorporados	Termos Adicionados pela SS
diabetes	diabete; Diabética; Diabético; diabetica; Diabets; diabetico; diabéticos; diabeticos	diabestes; dabetes; deabetes; doiabetes; dibetes; diabetese	-	-
convulsão	convulsao; convulsào; convulsã	convulsão; convulsões; Convulsõa; convulsãoes; convulão; convusão; conjvulsão; convulcão; connvulsão; conuvulsão; convuulsão; Concvulsão; consvulsão; consulsão; Convulsõ; cnvulsão; convuslsão; onvulsão; convlsão; convuklsão; convulção; convulslão; cojnvulsão; CONVULSAN	-	-
enxaqueca	enxaquecoos; enxaquecas; enxaquecosas; enxaquecosos; enxaquec; Enxaquecoso; enxaquecosa	enxqueca; exaqueca; enxaquca; Ensaqueca; enxaqudca; enxaqudeca; enxaqucosos; enxaueca; enxcaqueca; enxaquecaa; ennxaqueca; aenxaqueca; enaqueca; Enxaqqueca; enxaaqueca; Enxqqueca	(5) hemicrania, enxaqueca confusional aguda, enxaqueca complicada, cefaléia hemicrânia, estado migrainosus	hemicrania; cefaléia hemicrânia; cefaléia hemicranica; hemicrania; hemicrânia; hemicranias; hemicranica; hemicrânica; hemicrânicas; hemicranio; hemicrânio; hemicrãnio; himicrania; himicranica; hmicrânica
depressão	depressã; depressao	Depresssão; depresão	(1) Sintomas depressivos	sintomas depressivos
tegretol	Tegretol200mg; Tegretol200cr; Tegretol400mgcr; Tegretol200; Tegretol400; Tegretol400cr; Tegretol400mg	Tegretoil; Tetretol; tegretal; Tegrettol; Teghretol; tegrotol; Tgretol	-	-
gardenal	Gardenal10; Gardenal100; Gardenal50	Gardenol; Grdenal; Gardenmal; Gardeal; Gardenaal; Gardeanal; Garenal	-	-
latejante	latejantes	latenjante; latejente; letejante; tatejante; latejannte; latjante	-	-
ansiedade	-	Ansiedde; ansieade; ansieadade	-	-
meningite	meningites	meninigite; menigite; menngite; meingite; mengingite; miningite; meiningite; memingite	(1) paquimeningite	-

- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Tabela 12 – Comparação na Recuperação de Histórias com e sem o algoritmo da base 2

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**%**	**ST+SO**	**%**	**ST+SO+SS**	**%**
edema	**1121**	2692	*140,1*	2692	*140,1*	**2748**	*145,1*
obesidade	**1906**	3021	*58,5*	3022	*58,6*	**3022**	*58,6*
diprospan	**1357**	1904	*40,3*	1906	*40,5*	**1906**	*40,5*
hematocrito	**2844**	3586	*26,1*	3586	*26,1*	**3586**	*26,1*
dor	**5425**	6785	*25,1*	6785	*25,1*	**6785**	*25,1*
antak	**1344**	1607	*19,6*	1613	*20,0*	**1613**	*20,0*
renitec	**1747**	1939	*11,0*	1941	*11,1*	**1941**	*11,1*
febre	**1757**	1916	*9,0*	1916	*9,0*	**1929**	*9,8*
moduretic	**1818**	1981	*9,0*	1982	*9,0*	**1982**	*9,0*
diabetes	**2212**	2342	*5,9*	2357	*6,6*	**2357**	*6,6*
creatinina	**6138**	6261	*2,0*	6262	*2,0*	**6262**	*2,0*
leucócitos	**3956**	3981	*0,6*	3989	*0,8*	**3989**	*0,8*
vioxx	**1762**	1767	*0,3*	1772	*0,6*	**1772**	*0,6*
hemoglobina	**3609**	3628	*0,5*	3628	*0,5*	**3628**	*0,5*
triglicérides	**6499**	6515	*0,2*	6517	*0,3*	**6517**	*0,3*
plaquetas	**3723**	3728	*0,1*	3730	*0,2*	**3730**	*0,2*
média			21,8		21,9		22,3

- CM – Clinic Manager,
- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

A **Tabela 12** mostra características bastante semelhantes às encontradas na **Tabela 8**, onde o número de histórias possui um grande aumento percentual somente com o *stemming* (edema – 140%). Comparando as **Tabelas 5 e 9**, notamos que a primeira teve um aumento percentual maior com o uso de semelhança ortográfica e semântica, porém na segunda o aumento, na média, é pequeno.

Tabela 13 – Comparação na Recuperação de Freqüência de Palavras com e sem o algoritmo da Base 2

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**%**	**ST+SO**	**%**	**ST+SO+SS**	**%**
edema	**1321**	3146	*138,2*	3146	*138,2*	**3208**	*142,8*
obesidade	**2193**	3327	*51,7*	3334	*52,0*	**3334**	*52,0*
diprospan	**1403**	1950	*39,0*	1955	*39,3*	**1955**	*39,3*
dor	**7700**	10216	*32,7*	10216	*32,7*	**10216**	*32,7*
hematocrito	**2922**	3697	*26,5*	3697	*26,5*	**3697**	*26,5*
antak	**1502**	1765	*17,5*	1773	*18,0*	**1773**	*18,0*
renitec	**2065**	2264	*9,6*	2270	*9,9*	**2270**	*9,9*
febre	**1916**	2084	*8,8*	2084	*8,8*	**2098**	*9,5*
diabetes	**2515**	2730	*8,5*	2752	*9,4*	**2752**	*9,4*
moduretic	**2139**	2304	*7,7*	2311	*8,0*	**2311**	*8,0*
creatinina	**6838**	6961	*1,8*	6964	*1,8*	**6964**	*1,8*
leucócitos	**6507**	6554	*0,7*	6566	*0,9*	**6566**	*0,9*
vioxx	**1980**	1987	*0,4*	1994	*0,7*	**1994**	*0,7*
triglicérides	**7385**	7414	*0,4*	7424	*0,5*	**7424**	*0,5*
hemoglobina	**3728**	3747	*0,5*	3747	*0,5*	**3747**	*0,5*
plaquetas	**3858**	3865	*0,2*	3868	*0,3*	**3868**	*0,3*
média			21,5		21,7		22,1

- CM – Clinic Manager,
- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

A **Tabela 13** confirma os resultados obtidos na **Tabelas 9**, onde a inserção dos algoritmos de semelhança ortográfica e semelhança semântica contribuem muito pouco para o aumento na recuperação dos termos.

Tabela 14 – Quantidade de variações dos termos encontrados com os algoritmos na Base 2

Termos	CM	ST	Termo Exato	Variações Recuperadas	ST+SO	Variações Recuperadas	ST+SO+SS	Variações Recuperadas
edema	**1321**	3146	1331	2	3146	2	3208	10
obesidade	**2193**	3327	3326	1	3334	6	3334	6
diprospan	**1403**	1950	1949	1	1955	5	1955	5
dor	**7700**	10216	7714	13	10216	13	10216	13
hematocrito	**2922**	3697	2922	1	3697	1	3697	1
antak	**1502**	1765	1765	0	1773	0	1773	0
renitec	**2065**	2264	2263	1	2270	6	2270	6
febre	**1916**	2084	1970	4	2084	4	2098	10
diabetes	**2515**	2730	2524	11	2752	18	2752	18
moduretic	**2139**	2304	2302	2	2311	8	2311	8
creatinina	**6838**	6961	6961	0	6964	3	6964	3
leucócitos	**6507**	6554	6511	3	6566	10	6566	10
vioxx	**1980**	1987	1984	3	1994	8	1994	8
triglicérides	**7385**	7414	7389	2	7424	9	7424	9
hemoglobina	**3728**	3747	3747	0	3747	0	3747	0
plaquetas	**3858**	3865	3862	2	3868	4	3868	4

- CM – Termos recuperados no Clinic Manager;
- ST – Termos recuperados com o uso de Stemming
- Termo Exato – Termos recuperados com a string exata
- SO – Termos recuperados com o uso de Semelhança Ortográfica
- SS – Termos recuperados com o uso de Semelhança Semântica
- Variações Recuperadas – quantidade de variações do termo original

Destaca-se na **Tabela 14** o mesmo resultado visto na **Tabela 10**, onde a quantidade de termos recuperados com o termo exato supera os resultados obtidos com o Clinic Manager®.

Alguns termos tiveram uma grande incorporação de variações como ‘diabetes’ (18) ao passo que para outros termos não foi encontrada nenhuma variação como em ‘antak’ e ‘hemoglobina’.

Gráfico 3a– Evolução da porcentagem de recuperação em cada algoritmos por termo da Base 2

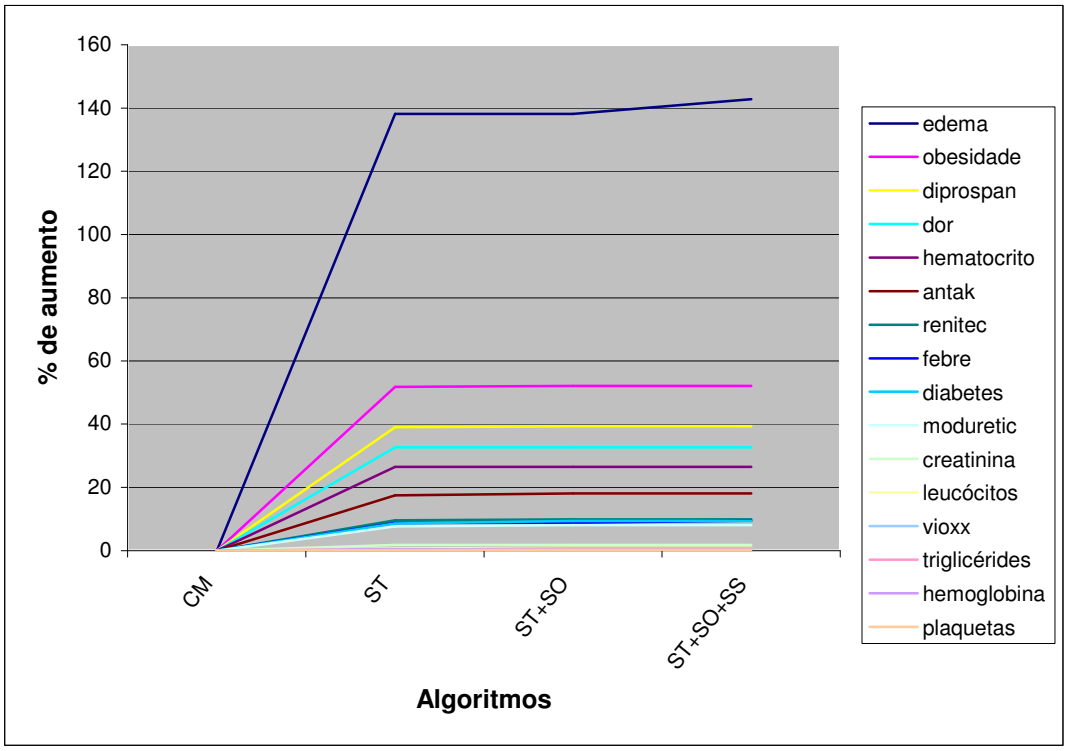
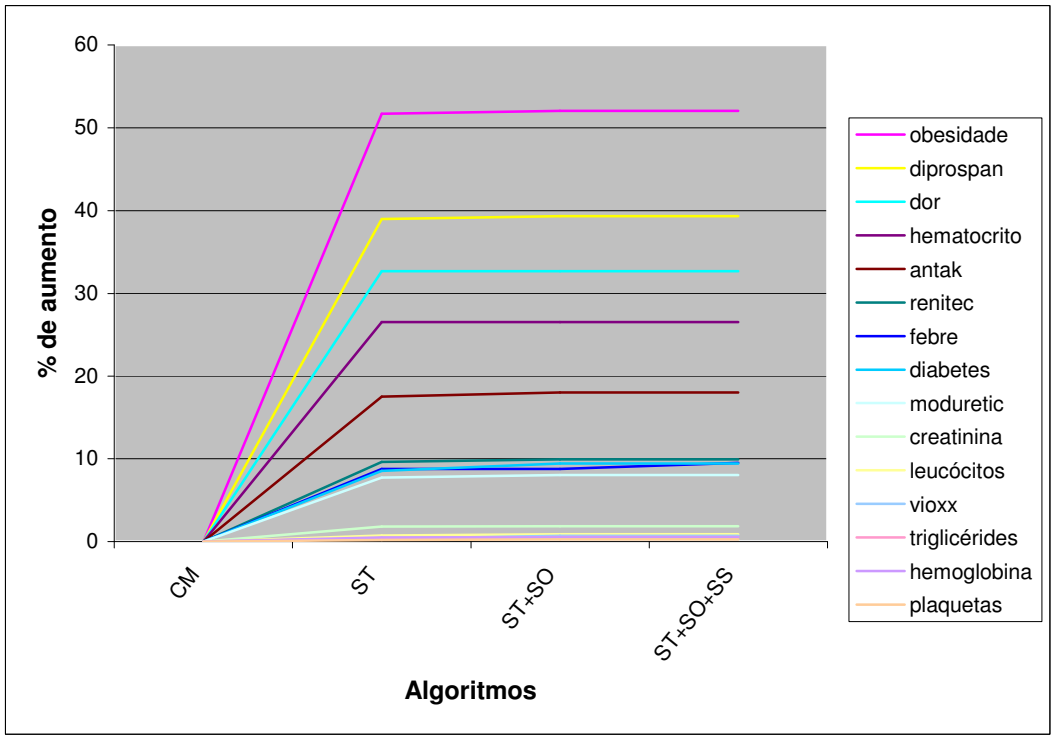


Gráfico 3b– Reprodução do Gráfico 3a sem o termo 'edema'



- CM – Clinic Manager,
- ST – Uso de Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Os **gráficos 3a e 3b** possuem características semelhantes às encontradas nos **gráficos 2a e 2b**, ou seja, o aumento é notável com o uso de *stemming*, porém a inclusão da semelhança ortográfica e semântica possui, na maioria dos casos, um patamar no gráfico com um aumento quase imperceptível visualmente.

De maneira semelhante ao **gráfico 2b**, o termo 'edema' foi removido no **gráfico 3b** por se comportar como um outlier.

Tabela 15 – Variações Incorporadas na busca dos termos da Base 2

Termos	Termos Adicionados pelo ST	Termos Adicionados pelo SO	Sinônimos Incorporados	Termos Adicionados pela SS
edema	edemas; edems	-	Hidropsia, hidropisia	Hidroxi; Hidróxido; Hidrox; Hidroneo; hidrog; Hidróxido; hidrotera; hidrol
obesidade	obesidades	obeidade; obsesidade; obesidae; obesisdade; obsidade	-	-
diprospan	Diprospane	Doprospan; Dipeospan; dprospan; dipospan	-	-
dor	Dores; dorso; Dórico; dors; dorsa; Dôr; Dora; Dore; dorsos; Dórico; dór; dorseios; Doris	-	Sofrimento físico	-
hematocrito	Hematócrito	-	-	-
antak	-	Aantak; Anatak; Antal; AntakI; Antax	-	-
renitec	renitece	rentec; enitec; Renitex; Renitc; renitrec	-	-
febre	febres; febr; febris; febri	-	Doenças febris, enfermidades febris, hipertermia, pirexia	Pirox; Piretamida; pirena; porex; doente febri; hiperermia
diabetes	diabetica; diabética; diabetico; diabético; diabetis; Diabete; diabets; diabeticos; diabeticas; diabetse; diabéticos	dibetes; dabética; Dabetes; Ddiabetes; diabtes; dioabetes; 0Diabetes	-	-
moduretic	moduretice; moduretici	Modurretic; Molduretic; 2Moduretic; Mosduretic; omoduretic; Modutretic	-	-
creatinina	-	cratinina; cretatinina; Cretinina	-	-
leucócitos	leucócitos; leucocit; leucocito	Leuccocitos; lecócitos; leucocsitos; leucocitose; leucocotos; lucocitos; lecocitos	-	-

- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Termos	Termos Adicionados pelo ST	Termos Adicionados pelo SO	Sinônimos Incorporados	Termos Adicionados pela SS
vioxx	vioxxe; Vioxx25; Vioxx`25	Voxx; vioxxe; Vioox; Viozx; oVioxx; voioxx	-	-
triglicérides	Trigliceride; triglicérides	trigliceridese; Trigliceride; Triiglicerides; triglicrides; triglicerdes; trigliceruide; triglicérides; 0Triglicerides; triclicerides;	-	-
hemoglobina	-	-	-	-
plaquetas	plaquetaria; plaqueta	Plaqetas; plaquetose	Trombócitos	-

- ST – Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Tabela 16 –Falsos positivos recuperados na Base 1

Palavra	Falsos positivos encontrados no SIRIMED – Base 1							
	ST	% falsas	ST+SO	% falsas	ST+SO +SS	Sinônimos adicionados	Falsos Positivos Encontrados	% falsos
tremor	709	0	719	0,56	719	(3) tremor de ação, tremor de intenção, tremor de repouso	Temor, temores	0,56
desmaio	485	0	1011	0,2	1249	(4) Síncope, ataque por queda, pré-síncope, síncope postural	desmame	0,2
insônia	490	0	558	0,18	619	(3) distúrbios do início e da manutenção do sono	insolação	0,18
convulsão	1606	0	2117	0,14	2117	-	convulsan	0,14
vômitos	690	0	705	0,14	705	(1) Emese	omitindo	0,14
cefaléia	4274	0	4828	0	5760	(3) dor de cabeça, cafalgia, cefalalgia	-	0
enxaqueca	1609	0	1658	0	1778	(5) hemicrania, enxaqueca confusional aguda, enxaqueca complicada, cefaléia hemicrânia, estado migrainosus	-	0
tontura	1224	0	1244	0	1244	(2) sensação de cabeça leve, ortoestase	-	0
nervoso	1111	0	1111	0	1111	-	-	0
meningite	1082	0	1097	0	1097	(1) paquimeningite	-	0
ronco	921	0	921	0	921	-	-	0
tegretol	881	0	891	0	891	-	-	0
gardenal	616	0	629	0	629	-	-	0
latejante	512	0	527	0	527	-	-	0
depressão	492	0	495	0	500	(1) Sintomas depressivos	-	0
cbz	530	0	530	0	656	(1) Carbamazepina	-	0
diabetes	471	0	481	0	481	-	-	0
ansiedade	469	0	474	0	474	-	-	0
Média		0		0,07				0,07

- CM – Clinic Manager,
- ST – Uso de Stemming
- SO – Uso de Semelhança Ortográfica
- SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Podemos notar na **Tabela 16** que a quantidade de falsos positivos, ou seja, palavras que o sistema recuperou por semelhança, mas que possuem significado diferente do termo de busca, tem um valor muito pequeno.

Na **Tabela 16** pode ser notado ainda que alguns sinônimos incorporados não são de uso trivial, como ‘hemicrania’ para ‘enxaqueca’ ou ‘ortoestase’ para ‘tontura’.

Aqui vale a pena comentar que não é possível definir a quantidade de falsos negativos, ou seja, é possível que alguns termos devam ser recuperados por semelhança semântica ou ortográfica, mas o algoritmo não recuperou.

Tabela 17 – Falsos positivos recuperados na Base 2

Palavra	Falsos positivos encontrados no SIRIMED – Base 2							
	ST	% falsas	ST+SO	% falsas	ST+SO+SS	Sinônimos adicionados	Falsos Positivos Encontrados	% falsas
edema	3146	0	3146	0	3208	Hidropsia, hidropisia	Hidróxido, hidrox,hidroneo, hidrotera,hidrol	1,93
dor	10216	0,68	10216	0,68	10216	Sofrimento físico	Dorso, dórico, Dora, dorseios, Dora, Doris	0,68
febre	2084	0	2084	0	2098	Doenças febris, enfermidades febris, hipertermia, pirexia	Pirox, Piretamida, pirena, porex	0,67
triglicérides	7414	0	7424	0	7424	-	-	0
creatinina	6961	0	6964	0	6964	-	-	0
leucócitos	6554	0	6566	0	6566	-	-	0
plaquetas	3865	0	3868	0	3868	Trombócitos	-	0
hemoglobina	3747	0	3747	0	3747	-	-	0
hematocrito	3697	0	3697	0	3697	-	-	0
obesidade	3327	0	3334	0	3334	-	-	0
diabetes	2730	0	2752	0	2752	-	-	0
moduretic	2304	0	2311	0	2311	-	-	0
renitec	2264	0	2270	0	2270	-	-	0
vioxx	1987	0	1994	0	1994	-	-	0
diprospan	1950	0	1955	0	1955	-	-	0
antak	1765	0	1773	0	1773	-	-	0
Média		0,04		0,04				0,21

Nota-se claramente nas **Tabelas 13 e 14** que poucos termos apresentaram falsos positivos, e mesmo aqueles que apresentaram, possuem uma porcentagem baixa, sendo o maior valor para edema (1,93%).

A palavra ‘edema’, que apresenta como sinônimos ‘hidropsia’ e ‘hidropisia’ trouxe a maior quantidade de falsos positivos pela semelhança com os sinônimos e não com a palavra original de busca.

8 – Discussão

8.1 – Críticas Metodológicas

Na seleção dos termos a serem testados, foram selecionados os termos mais freqüentes em ambas as bases de dados. Com isso, sendo os termos mais freqüentes, ou seja, mais utilizados, a quantidade de erros de ortografia é mais elevada, o que mostra que há muitas variações ortográficas que precisam ser recuperadas por semelhança ortográfica conforme **Tabelas 11 e 15**, por outro lado, por serem termos muitos comuns, os sinônimos para eles não surtem grande contribuição.

Provavelmente, na busca de termos mais incomuns, a incorporação de sinônimos terá uma contribuição mais expressiva na recuperação por semelhança semântica.

Não houve priorização na otimização dos algoritmos quanto à velocidade. O tempo de indexação, principalmente para a primeira indexação, ainda está muito grande e é necessário aprimorar os métodos e melhorar a eficiência do algoritmo e diminuir o tempo. Para as indexações posteriores, a cada inserção, remoção ou modificação de história, o tempo não é tão crucial, pois, na média, leva menos de 4 segundos para indexar uma história (**Tabela 6**).

O fato de ter usado um algoritmo de *stemming* que remove somente sufixos (SnowBall, 2005) pode levar a perda de termos que possuem também prefixos como acéfalo em relação à cefaléia. Como a maioria dos prefixos possui a idéia de negação, modificando o significado, como acéfalo (sem cérebro), inconsciente (sem consciência) e outros, esse problema é minimizado.

Porém, a indexação utilizando *stemming* aumentou a quantidade de termos recuperados, pois conseguiu recuperar termos com mesmo radical. Por outro lado, perdeu-se na precisão, pois não foi mais possível diferenciar ‘dor’ de ‘dores’ pois estão no índice com o mesmo radical.

Para esse trabalho não foi analisado o conteúdo semântico da palavra, ou seja, a busca pelo termo ‘mente’ irá trazer todos os termos homógrafos,

incluindo o verbo mentir ou o substantivo mente, sendo privilegiado a recuperação dos termos por semelhança ortográfica.

Nesse trabalho, não foram utilizados algoritmos de recuperação fonética. A Recuperação Fonética é usada em aplicações como a recuperação de nomes próprios, onde o nome soletrado é usado para identificar outras palavras parecidas ou com pronúncia similar (Zobel & Dart, 1996). Uma das grandes aplicações desse tipo de algoritmo é na busca de nomes em grandes bases de dados, como páginas amarelas, onde o nome ditado pelo solicitante e escrito na forma que ele entende e o sistema procura nomes com a mesma pronúncia como “Michael”, “Mykon” ou “Maicow”.

A maioria dos algoritmos fonéticos simplesmente trocam as letras por códigos numéricos que representam o seu som, entre eles podemos citar o Soundex, Phonix e Methafone. Todos esses algoritmos foram desenvolvidos para o idioma inglês, utilizando o som dos fonemas dessa língua.

Soundex usa códigos para o som das letras e transforma uma String em uma forma canônica de até quatro caracteres, preservando a primeira letra (Zobel & Dart, 1996). A seguir podemos ver na **Tabela 18,** a conversão das letras por códigos.

Tabela 18 – Tabela de Conversão do Soundex

Código	Representa as letras
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Note que o algoritmo ignora as vogais, o Y, H e W. O problema desse algoritmo e de suas variações como o Phonix, é que pode gerar códigos iguais para sons diferentes, como é o caso de ‘Catherine’ e ‘Cotroneo’ (código C365)

e palavra como ‘Geronimo’ (Soundex G655) e ‘Jeronimo’ (Soundex J655) serão consideradas diferentes, além se ser criado para fonemas da língua inglesa.

Além disso, variações fonéticas de ‘G’ por ‘J’, ‘X’ por ‘S’, ‘Ç’ por ‘S’ são supridas pelo algoritmo de semelhança ortográfica de *edit distance* utilizado, portanto, justificando o não uso da semelhança fonética.

Não usamos nenhuma base de dados de testes para calcular o *recall* e *precision* com documentos e perguntas previamente estabelecidas, pois, a maioria dessas, como a coleção OHSUMED apresentada na TREC (Hersh, 2004), porque, além de todos os documentos da coleção estarem em inglês, o material é uma sub-coleção do MEDLINE, onde os artigos passaram por revisão por pares, praticamente não possuindo erros de digitação, o que não é o mundo real de prontuários eletrônicos do paciente. A maioria dos algoritmos de recuperação de informações não levam em conta erros ortográficos, que é um fator importante a ser considerado em PEP´s.

8.2 – Discussão dos Resultados

Nesse trabalho preferimos criar um novo software (SIRIMED), ao invés de usar um programa de indexação e recuperação de documentos textuais existente no mercado, para aplicar os algoritmos propostos contextualizados para a área de medicina e saúde.

Existem alguns softwares genéricos de indexação e recuperação de documentos textuais, alguns livres e outros comerciais, com características semelhantes, porém sem foco específico na área da saúde. Dentre eles podemos citar o Swish-e[4] (*Simple Web Indexing System for Humans – Enhanced*) que possui algoritmos de *stemming*, recuperação fonética (*Soundex e Metaphone*), uso de coringas e expressões regulares somente para o inglês, cujo idioma já possui esses algoritmos bem definidos. Além desse, outro software é o ht://Dig[5], possuindo as mesmas características do anterior, além da inclusão de sinônimos, também em inglês. Ambos os mencionados são softwares livres.

[4] Swish-e - http://swish-e.org/
[5] ht://Dig - http://www.htdig.org/

A Universidade de Glasgow disponibilizou sua ferramenta de indexação e recuperação chamada Terrier[6] (TERabyte RetrIEveR), licenciada como software livre, que utiliza algoritmos de *stemming* e remoção de *stop words* (para o inglês) e não possui algoritmos de busca fonética, semântica ou semelhança ortográfica.

Além dos programas mencionados, existe o dtSearch[7], que é um software comercial, adaptável para o português, mas com algumas características que só podem ser aplicadas em inglês, como a pesquisa fonética. Possui um Thesaurus em inglês incorporado (WordNet), embora também permita usar sinônimos em português.

Da mesma maneira, sites de busca como o Google®, Altavista®, Yahoo® e outros não provêm uma ferramenta que inclua sinônimos, pois indexam conteúdo de várias línguas. Além disso, não possuem suporte a termos semelhantes ortograficamente, onde os erros de sintaxe são perdidos nos resultados obtidos. Alguns sites de busca, como o Google, sugerem termos se o mecanismo acha que busca foi escrita errada, porém, não incorpora à sua busca e simplesmente substitui o termo caso a sugestão seja aceita.

Apesar de não serem ferramentas *Open Source*, os algoritmos de ordenação dos resultados são conhecidos, como por exemplo, o algoritmo de *Page Rank* do Google[8], porém o critério de ordenação usado não se aplicada a prontuários médicos por não terem título, palavras chave (META TAGS) ou links apontando para o seu conteúdo como existem nas páginas Web.

As diferenças de recuperação da palavra exata entre o Clinic Manager® e o Sirimed podem ser atribuídas ao processo de recuperação utilizado (**Tabelas 7 e 11**). Todos os textos do Clinic Manager® são gravados no formato RTF e, enquanto o mesmo faz uma busca direta nos textos (converte a pergunta para RTF), o Sirimed usa um processo de indexação, onde os textos são convertidos para o modo texto. Nota-se que a maioria dos sistemas de busca atuais (Google, Terrier, Altavista, Yahoo etc.) usam o processo de indexação do conteúdo para agilizar e melhorar a busca dos documentos.

---

[6] Terrier - http://ir.dcs.gla.ac.uk/terrier/

[7] dtSearch - http://www.multidoc.com.br/prod/dts/apres.htm

[8] Google's PageRank Explained (http://www.webworkshop.net/pagerank.html)

Palavras longas com radical incomum como na palavra meningite (meningit), quase não possuem variações morfológicas, sendo que o aumento pode estar ligado a variações de plural, por outro lado, palavras que possuem muitas variações morfológicas como a palavra ronco (ronc), tem um aumento muito maior, principalmente por ser substantivo primitivo do verbo roncar, conforme visto na **Tabela 9**.

Além disso, palavras muito curtas podem ter o seu radical muito comum. Isso acontece, por exemplo, com a palavra 'dor', cuja recuperação por radical comum trouxe palavras como 'dorso', 'dorico', 'Dora' ou 'Doris' conforme visto na **Tabela 13**. Para evitar esse tipo de problema, o usuário pode inserir outra palavra que especificasse o conteúdo semântico, como 'dor de cabeça' ou mesmo 'dor forte' que elimina grande quantidade desses falsos positivos.

Outro caso é a inserção de sinônimos que possuem também radical comum, como edema que possui como sinônimo o termo hidropisia, que recuperou palavras como 'hidróxido' ou 'hidroneo' (**Tabela 13**).

Muitas vezes a inclusão de sinônimos não ajuda no aumento da recuperação por serem sinônimos não usuais. Por exemplo, como sinônimo de febre temos 'pirexia', porém esse termo não é usual. Por outro lado, um sinônimo de cefaléia é 'dor de cabeça', o que faz a recuperação aumentar muito mais, conforme visto nas **Tabelas 6 e 10**.

A inserção de algoritmos de semelhança na recuperação pode trazer resultados que não são esperados. Nas **Tabela 16 e 14** mostramos quantos falsos positivos o sistema trouxe. A maioria dos erros se encontra na recuperação de palavras que possuem o radical ortograficamente semelhante. Como exemplo, a palavra 'vômito' trouxe na aproximação semântica a palavra 'omitindo', isso porque a raiz de 'vomitar' é semelhante à de 'omitir'. Outro exemplo é a palavra 'tremor' que trouxe na semelhança ortográfica a palavra 'temor (es)'.

O thesaurus incorporado no SIRIMED (DeCS) não é capaz de incorporar jargões médicos das diversas áreas da saúde, nem mesmo as siglas utilizadas. Para tanto, o sistema permite ampliar o thesaurus, inserindo novos sinônimos a

ele, conforme ilustra a **Figura 8**. O termo “cbz” foi uma incorporação manual feita para o medicamento “carbamazepina”.

O Departamento de Informática Médica da Universidade de Freiburg (Alemanha) juntamente com o programa de pós-graduação em informática aplicada da PUC do Paraná vem desenvolvendo uma pesquisa nos últimos anos sobre indexação e recuperação de textos médicos baseado em morfemas. Morfemas são sub-palavras retiradas na formação das palavras, por exemplo, o termo miocardite é composto pelos morfemas mio (músculo), card (coração) e ite (inflamação). A idéia é substituir os morfemas por um identificador comum e indexá-lo, mesmo entre língua diferentes. Para tanto, estão criando um sistema denominado MORPHOSAURUS, que consiste de bases terminológicas e as rotinas de normalização de textos (Schulz *et al*., 2002).

Mesmo com esses problemas na indexação com o método de *stemming* puro, o algoritmo utilizado no SIRIMED se mostrou bastante eficaz no aumento da quantidade de termos recuperados. Nessa pesquisa foram estudados vários algoritmos, que juntos, tem um objetivo mais prático do que isoladamente.

8.3 – Comentários Finais

A maior parte da informação médica na forma digital está em formato de textos livres, tanto em banco de dados de artigos científicos, páginas na Web, livros virtuais ou mesmo em PEP. Em uma empresa, cerca de 80% das informações está na forma textual (Tan, 1999). Toda essa informação precisa de técnicas capazes de extrair o que se necessita delas.

No PEP, o médico prefere usar textos livres para a inserção de dados do paciente do que dados estruturados, pela liberdade de expressão e semelhança do prontuário em papel. Por isso, criar técnicas mais elaboradas para recuperar informações em textos livres é mais atraente para usuários de PEP do que fazer com que sejam preenchidos formulários com vários campos estruturados para que o sistema seja capaz de “entender” o que está sendo inserido.

O processamento da informação, por parte dos sistemas, inserida de modo estruturado é muito mais fácil que àquela inserida no formato de textos livres, isso porque já existem ferramentas bastante aprimoradas para a manipulação desse tipo de informação como *Store Procedures*, *Triggers* e linguagens como SQL embutidas em sistemas gerenciadores de banco de dados como Oracle®, MS-SQLServer® ou IBM-DB2®. Por outro lado, para o usuário, preencher uma seqüência de campos estruturados em formulários, navegar por árvores de decisão para escolher o melhor item não é nada amigável e demanda muito tempo.

Processos de recuperação de informações em textos livres podem ter muitas aplicações práticas. Um exemplo é o uso de programas para analisar informações de determinados usuários suspeitos como o Carnivore.

Carnivore é um sistema de monitoramento da Internet desenvolvido pelo FBI em julho de 2000 e é instalado nos provedores de serviço de internet (Bennett *et al*., 1999). Além disso, seu irmão mais novo, o Altivore veio como uma resposta open source, porém com o mesmo poder do Carnivore.

Outras aplicações que podem ser mencionadas são a Extração de Palavras-chave de textos em Português, cujo processo desempenha um importante papel na indexação de documentos. Pesquisas para tratar o inglês são muitas, porém ainda é necessário avançar os estudos nesse tópico para o português como foi feito em Pereira, Souza & Nunes, 2002.

A entrada de dados também pode ser feita com o auxílio de sistemas de reconhecimento de voz que traduzem a fala do médico para textos livres e posteriormente, podem ser recuperadas com os algoritmos utilizados nesse trabalho, visto que atender o paciente e preencher formulários eletrônicos simultaneamente pode dificultar a relação médico-paciente. Além disso, existem casos que o profissional da saúde está com as mãos ocupadas para registrar no prontuário, como o profissional que está executando um ultra-som ou um médico que está realizando uma necrópsia e que precisa ditar o que está vendo.

Há uma necessidade de técnicas que não fiquem presas somente aos termos, mas ao seu significado, tanto individual como no contexto. Um dos

exemplos disso são os modificadores de significado, como ‘nega’, ‘não’, ‘anti’ e outros. Por exemplo, em um dos bancos de dados usado, a busca por meningite trouxe 1066 histórias onde o termo aparece, porém destes, 1024 possuem o termo ‘nega’ que alterava o contexto, porém nesse trabalho optamos pela busca de termos sem contexto.

A classificação de documentos também é uma área da recuperação de informações que visa classificar documentos de acordo com o seu conteúdo, ou seja, os termos presentes nele. Na área da saúde, pesquisas estão sendo feitas para encontrar relacionamentos entre documentos, até então, sem relacionamento direto, usando conceitos de *data mining* (mineração de dados) para encontrar conhecimento implícito em textos médicos.

Também na área médica, a tentativa de extração automática de diagnósticos é um grande desafio pelos muitos motivos vistos durante a leitura desse trabalho, dentre eles, podemos citar as formas sinônimas de inserção de termos, erros de ortografia, variações morfológicas, usos de jargões e abreviaturas.

Em inteligência artificial, trabalhos têm sido direcionados em programação para o computador entender a linguagem natural. Procedimentos automáticos para processar e entender linguagem natural estão sendo desenvolvidos. Em pseudo-linguística estão investigando o mecanismo pelo qual o cérebro humano entende a linguagem (Rijsbergen, 1979).

Desse modo, pudemos notar que a quantidade de informações não recuperadas em uma busca direta é muito grande. O mundo real dos prontuários eletrônicos do paciente possuem muitos erros de ortografia e uso de sinonímia e a necessidade de melhorar os algoritmos de busca ainda é muito evidente para que os sistemas de recuperação de informações sejam capazes de reconhecer coisas semelhantes assim como os seres humanos o fazem.

9 – Conclusão

1) Foi desenvolvido um software chamado SIRIMED para indexação e recuperação de informações em campos de textos livres de prontuários eletrônicos do paciente baseada em semelhança semântica utilizando um thesaurus médico (DeCS) e semelhança ortográfica utilizando algoritmos de *stemming* e *edit distance;*

2) Foi realizada uma comparação na recuperação de termos utilizando 34 termos selecionados (18 na base 1 e 16 na base 2), utilizando a busca tradicional com comparação de string exata embutida no software Clinic Manager® e o novo software SIRIMED com o novo algoritmo. Os resultados mostraram que a recuperação aumentou, em média, 30%, sendo que a maior contribuição foi pelo processo de *stemming*, uma pequena contribuição pelo *edit distance* e uma contribuição menor ainda por semelhança semântica. A quantidade de falsos positivos foi menor de 0,5% em ambas as bases, não comprometendo os resultados obtidos.

10 - Referências

ARONSON, A. R.; MORK, J. G.; GAY, C.W.; HUMPHREY, S.M.; ROGERS, W.J. **The NLM Indexing Initiative's Medical Text Indexer**. MEDINFO 2004 - Proceedings of the Eleventh World Congress on Medical Informatics, San Francisco, CA. IOS Press., p. 268-72, 2004.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York, ACM Press, 1999.

BAEZA-YATES, R.; NAVARRO, G. **Fast approximate string matching in a dictionary**. In Proc. 5th South American Symposium on String Processing and Information Retrieval (SPIRE'98), IEEE CS Press, p. 14-22, 1998.

BENNETT, K.; CHILES, P.; JACOBS, M. (1999). **An educator's guide to privacy**. [On-line]. Disponível em : http://lrs.ed.uiuc.edu/wp/privacy-2002/encryption.html (25/09/2005)

BIREME, **DeCS - Descritores em Ciências da Saúde**, disponível em http://decs.bvs.br/P/decswebp.htm (25/09/2005)

CHU, S.S. **Information Retrieval and health / clinical management**, Yearbook of Medical Informatics, 2002.

HALL, P.A.V.; DOWLING, G.R. **Approximate String Matching**, Computing Surveys, v. 12, p. 381-402, 1980.

HERSH, W.R**. Information Retrieval : a Health and Biomedical Perspective**. 2nd ed. New York : Springer, 2003.

HERSH, W.R.; DETMER, W.M.; FRISSE, M.E. **Information Retrieval Systems** In: SHORTLIFFE EH, PERREAULT LE, WIEDERHOLD G, FAGAN LM. Medical Informatics: Computer Applications in Health Care and Biomedicine. 2nd ed. New York: Springer, p. 539-72, 2001.

HERSH, W.; BUCKLEY, C.; LEONE, T. J.; HICKAM, D. **OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research,** In Proc. ACM SIGIR '94, p. 192-201, 1994.

JOHNSON, S.B. **A Semantic Lexicon for Medical Language Processing**. Journal of the American Informatics Association (JAMIA), v. 6, n. 3, p. 205-18, 1999.

KARPISCHEK, R.U. **Dicionário br.ispell**- versão 2.4 (10/1999), Disponível em http://www.ime.usp.br/~ueda/br.ispell/ (25/09/2005)

KENT, A. **Manual da Recuperação Mecânica da informação,** São Paulo, Polígono, 1972

KORTH, H.F.; SILBERSCHATZ, A. **Sistema de Bancos de Dados**, 2ª ed. São Paulo: Makron Books, p.1, 1995.

LOH, S. **Descoberta de Conhecimento em Bases de Dados Textuais**. Maio de 1997 – Disponível em http://mozart.ulbra.tche.br/~loh/apostilas/dc-texto.htm (25/09/2005)

LOPES, A.C. **Primórdios da Medicina. Clínica Médica – Passado, Presente, Futuro**. São Paulo: Lemos Editorial e Gráficos Ltda, 1999. Disponível em http://www.alziravelano.com.br/clinica_medica.htm (25/09/2005)

LOVIS, C.; BAUD, R.H.; PLANCHE, P. **Power of expression in the electronic patient record: structured data or narrative text?**. International Journal of Medical Informatics, v. 58, n. 1, p. 101-10, 2000.

MULLIGEN, E.M.V.; STAM, H.; GINNEKEN, A.M.V. **Clinical Data Entry**. Proceedings of AMIA (American Medical Informatics Association) Annual Fall Symposium, p. 81-5, 1998, Disponível em http://www.amia.org/pubs/symposia/D004709.PDF (25/09/2005).

PELLIZZON, R. F. **Pesquisa na área da saúde: 1. Base de dados DeCS (Descritores em Ciências da Saúde).** Acta Cir. Bras., v. 19, n. 2, p. 153-163, 2004.

PEREIRA, M.B.; SOUZA, C.F.R.; NUNES, M.G.V. **Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português,** Revista Eletrônica de Iniciação Científica da SBC (Sociedade Brasileira de Computação), Ano II, v. 2, n. 1, 2002.

PEREIRA, S.L. **Estruturas de Dados Fundamentais: Conceitos e Aplicações**, São Paulo: Érica, p. 2-3, 1996.

PORTER, M.F. **An algorithm for suffix stripping**, v. 14, n. 3, p. 130-7, 1980, Disponível em http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html (25/09/2005)

RIJSBERGEN, C.J. Van **Information Retrieval**, London: Butterworths, 1979, Disponível em http://www.dcs.gla.ac.uk/Keith/Preface.html (25/09/2005)

SAGER, N.; LYMAN, M.; NHAN, N.; TICK, L.J. **Natural Language Processing and the Representation of Clinical Data**, Journal of the American Medical Informatics Association (JAMIA), v. 1, n. 2, p. 142-60, 1994.

SALTON, G.; WONG, A.; YANG, C.S. **A vector space model for automatic indexing**. Communications of the ACM, v. 18, n.11, 1975.

SCHULZ, S.; NOHAMA, P.; BORSATO, E.P.; MATIAS, L.J.D. **Indexação e Recuperação Automática de Textos Médicos**, VIII CBIS (Congresso Brasileiro de Informática em Saúde), Sessão Pôster, 2002, Disponível em www.avesta.com.br/anais/dados/trabalhos/287.pdf (25/09/2005)

SHORTLIFFE, E.H.; BARNETT, G.O. **Medical Data: Their Acquisition, Storage, and Use** In: SHORTLIFFE EH, PERREAULT LE, WIEDERHOLD G, FAGAN LM. Medical Informatics: Computer Applications in Health Care and Biomedicine. 2nd ed. New York: Springer; p. 41-75, 2001.

SHORTLIFFE, E.H.; BLOIS, M.S. **The Computer Meets Medicine and Biology: Emergence of a Discipline** In: SHORTLIFFE EH, PERREAULT LE, WIEDERHOLD G, FAGAN LM. Medical Informatics: Computer Applications in Health Care and Biomedicine. 2nd ed. New York: Springer; p. 3-40, 2001.

SIGULEM, D. ; CARDOSO, O.L. ; GIMENEZ, S.S.F.X. ; CEBUKIN, A. ; ANÇÃO, M. S. **Clinic Manager System**. In: World Congress on Medical Physics and Biomedical Engineering, Rio de Janeiro. Physics in Medicine and Biology, 1994. v. 1. p. 556-556, 1994

SILVA, E.K.O. **Um Estudo sobre Sistemas de Banco de Dados Cliente/Servidor**. João Pessoa/PB, 2001 - Monografia apresentada ao Curso de Processamento de Dados da Faculdade Paraibana de Processamento de Dados – Disponível em http://www.biblioteca.sebrae.com.br/bte/bte.nsf/subarea2?OpenForm&AutoFramed&jmm=_n9574cjqi9mql8ia384_ (25/09/2005)

SILVA FILHO, A.A. - **O que é "dado"?,** Programa de Pós-graduação em Informática em Saúde, Unifesp (não publicado), 2003.

SNOWBALL, **Portuguese stemming algorithm**, última atualização em 10/09/2002 - disponível em http://snowball.tartarus.org (25/09/2005)

TAN, Ah-Hwee, **Text mining: the state of the art and the challenges**, Proceedings of the Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases, Beijing, p.65-70, 1999.

TARDELLI, A.O.; ANCAO, M.S.; PACKER, A.L.; SIGULEM, D. **An Implementation of the Trigram Phrase Matching Method for Text Similarity Problems**. Stud Health Technol Inform, p. 103:43-9, 2004.

VAN BEMMEL, J.H. , **Handbook of Medical Informatics** , WEBSITE v3.3 [on-line], Atualização em 25/03/1999, Disponível em http://www.mihandbook.stanford.edu/handbook/home.htm (25/09/2005)

WIVES, L.K.; LOH, S. **Recuperação de Informações usando a Expansão Semântica e a Lógica Difusa**. In: CONGRESO INTERNACIONAL EN INGENIERIA INFORMATICA, ICIE, 1998.

WIVES, L.K. **Indexação de Documentos Textuais**. Trabalho desenvolvido para a disciplina de Sistemas Bancos de Dados, ministrada no curso de mestrado em Ciência da Computação da UFRGS. Porto Alegre: CPGCC da UFRGS, Junho de 1997 – Disponível em http://www.inf.ufrgs.br/~wives/ publicacoes/IDT.pdf (25/09/2005).

ZOBEL, J.; DART, P. **Phonetic string matching: lessons from information retrieval**, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.166-172, 1996, Disponível em http://www.seg.rmit.edu.au/research/ download.php?manuscript=105 (25/09/2005)

11 – Abstract

Information retrieval is a science that investigates models and techniques to recover information, mainly from free texts, that are the majority digital information after the internet advent.

The necessity of techniques to recover information from this great mass of data is evident. Search engines like Google®, Altavista®, Yahoo® and others are indispensable to find information at Internet in PDF, TXT or HTML files.

In the health context, a lot of information is registered as free texts like scientific articles into specific health databases like Medline which has specific search engines (Pubmed).

Electronic Record Patient (ERP) has also free text information to describe patient's history or evolution. The health professional who inserts information can use synonymous or medical terms, abbreviations or even make orthography mistake. In these cases, the recovery of the information with these variations could be not trivial.

Two ERP databases from distinct clinics had been used. The first one had 6732 clinical histories and second had 26072 histories. A software called SIRIMED (Sistema de Indexação e Recuperação de Informações Médicas) was developed to show that recovery of the information based in both similarity semantics with a medical thesaurus (DeCS – Descritores em Ciências da Saúde) and approximate string matching (based on stemming and edit distance algorithm) can improve approximately 30% the amount of terms recovered if compared to traditional method, which searches only the exact string matching.

The false positives average is less than 0.5% for both databases and, therefore, it doesn't prejudice the obtained results.

Anexo I – Stop List sugerida por Porter

de
a
o
que
e
do
da
em
um
para
é
com
não
uma
os
no
se
na
por
mais
as
dos
como
mas
foi
ao
ele
das
tem
à
seu
sua
ou
ser
quando
muito
há
nos
já
está
eu
também
só
pelo
pela
até
isso
ela
entre
era
depois
sem
mesmo
aos
ter
seus
quem
nas
me
esse
eles
estão
você
tinha
foram
essa
num
nem
suas
meu
às
minha
têm
numa
pelos
elas
havia
seja
qual
será
nós
tenho
lhe
deles
essas
esses
pelas
este
fosse
dele
tu
te
vocês
vos
lhes
meus
minhas
teu
tua
teus
tuas
nosso
nossa
nossos
nossas
dela
delas
esta
estes
estas
aquele
aquela
aqueles
aquelas
isto
aquilo
estou
está
estamos
estão
estive
esteve
estivemos
estiveram
estava
estávamos
estavam
estivera
estivéramos
esteja
estejamos
estejam
estivesse
estivéssemos
estivessem
estiver
estivermos
estiverem
hei
há
havemos
hão
houve
houvemos
houveram
houvera
houvéramos
haja
hajamos
hajam
houvesse
houvéssemos
houvessem
houver
houvermos
houverem
houverei
houverá
houveremos
houverão
houveria
houveríamos
houveriam
sou
somos
são
era
éramos
eram
fui
foi
fomos
foram
fora
fôramos
seja
sejamos
sejam
fosse
fôssemos
fossem
for
formos
forem
serei
será
seremos
serão
seria
seríamos
seriam
tenho
tem
temos
tém
tinha
tínhamos
tinham
tive
teve
tivemos
tiveram
tivera
tivéramos
tenha
tenhamos
tenham
tivesse
tivéssemos
tivessem
tiver
tivermos
tiverem
terei
terá
teremos
terão
teria
teríamos
teriam

Anexo IIA – Lista de Sufixos Comuns

"eza", "ezas", "ico", "ica", "icos", "icas", "ismo", "ismos", "avel", "ivel", "ista", "istas", "oso", "osa", "osos", "osas", "amento", "amentos", "imento", "imentos", "adora", "ador", "acao", "adoras", "adores", "acoes", "logia", "logias", "encia", "encias", "iva", "ivo", "ivas", "ivos", "ira", "iras", "amente", "mente", "idade", "idades"

Anexo IIB – Listas de Sufixos de Verbos Regulares

"ada", "ida", "ia", "aria", "eria", "iria", "ara", "ara", "era", "era", "ira", "ava", "asse", "esse", "isse", "aste", "este", "iste", "ei", "arei", "erei", "irei", "am", "iam", "ariam", "eriam", "iriam", "aram", "eram", "iram", "avam", "em", "arem", "erem", "irem", "assem", "essem", "issem", "ado", "ido", "ando", "endo", "indo", "ara~o", "erao", "irao", "ar", "er", "ir", "as", "adas", "idas", "ias", "arias", "erias", "irias", "aras", "aras", "eras", "eras", "iras", "avas", "es", "ardes", "erdes", "irdes", "ares", "eres", "ires", "asses", "esses", "isses", "astes", "estes", "istes", "is", "ais", "eis", "ieis", "arieis", "erieis", "irieis", "areis", "areis", "ereis", "ereis", "ireis", "ireis", "asseis", "esseis", "isseis", "aveis", "ados", "idos", "amos", "amos", "iamos", "ariamos", "eriamos", "iriamos", "aramos", "eramos", "iramos", "avamos", "emos", "aremos", "eremos", "iremos", "assemos", "essemos", "issemos", "imos", "armos", "ermos", "irmos", "eu", "iu", "ou", "ira", "iras"

Anexo III - Aprovação do uso do DeCS Português pela BIREME

Amilton,

Autorizado uso do DeCS versão no idioma Português para fins acadêmicos na tese "Recuperação de Informações em Campos de Textos Livres de Bancos de Dados médicos baseado em semelhança semântica e ortográfica" do aluno Amilton Souza Martha, RG: 22.268.742-3, aluno regularmente matriculado no mestrado de Informática em Saúde do Departamento de Informática em Saúde da Unifesp/SP sob o número de matrícula 0410016.

Abel L. Packer
BIREME/OPAS/OMS, Diretor

-----Original Message-----
From: amilton-pg@dis.epm.br [mailto:amilton-pg@dis.epm.br]
Sent: Tue 2/1/2005 11:04 AM
To: Packer, Abel Laerte (BIR)
Cc: Tardelli, Adalberto Otranto (BIR)
Subject: Cessão de direitos de uso do DeCS

Ao Diretor BIREME/OPAS/OMS
Abel Packer

Venho por meio deste solicitar autorização para o uso da versão em português do vocabulário controlado DeCS (Descritores de Ciências da Saúde) para uso como thesaurus no protótipo SIRIMED (Sistema de Indexação e Recuperação de Informações Médicas) como parte da dissertação de mestrado com o título de "Recuperação de Informações em Campos de Textos Livres de Bancos de Dados médicos baseado em semelhança semântica e ortográfica" do aluno Amilton Souza Martha, RG: 22.268.742-3, aluno regularmente matriculado no mestrado de Informática em Saúde do Departamento de

Informática em Saúde da Unifesp/SP sob o número de matrícula 0410016.

Informo que seu uso será exclusivamente para fins acadêmicos.

Atenciosamente,

Amilton Souza Martha

Anexo IV – Edit Distance ou Levenshtein Distance

```
'*********************************
'*** Get minimum of three values
'*********************************

Private Function Minimum(ByVal a As Integer, _
                         ByVal b As Integer, _
                         ByVal c As Integer) As Integer
Dim mi As Integer

  mi = a
  If b < mi Then
    mi = b
  End If
  If c < mi Then
    mi = c
  End If

  Minimum = mi

End Function

'*********************************
'*** Compute Levenshtein Distance
'*********************************

Public Function LD(ByVal s As String, ByVal t As String) As Integer
Dim d() As Integer ' matrix
Dim m As Integer ' length of t
Dim N As Integer ' length of s
Dim i As Integer ' iterates through s
Dim j As Integer ' iterates through t
Dim s_i As String ' ith character of s
Dim t_j As String ' jth character of t
Dim cost As Integer ' cost

  ' Step 1

  N = Len(s)
  m = Len(t)
  If N = 0 Then
    LD = m
    Exit Function
  End If
  If m = 0 Then
    LD = N
    Exit Function
```

```
  End If
  ReDim d(0 To N, 0 To m) As Integer

  ' Step 2

  For i = 0 To N
    d(i, 0) = i
  Next i

  For j = 0 To m
    d(0, j) = j
  Next j

  ' Step 3

  For i = 1 To N

    s_i = Mid$(s, i, 1)

    ' Step 4

    For j = 1 To m

      t_j = Mid$(t, j, 1)

      ' Step 5

      If s_i = t_j Then
        cost = 0
      Else
        cost = 1
      End If

      ' Step 6

      d(i, j) = Minimum(d(i - 1, j) + 1, d(i, j - 1) + 1, d(i - 1, j - 1) + cost)

    Next j

  Next i

  ' Step 7

  LD = d(N, m)
  Erase d

End Function
```

Anexo V – Aprovação do Comitê de Ética em Pesquisa

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo