JHI JOURNAL OF HEALTH INFORMATICS

Recuperação de Informações em Campos de Texto Livres de Prontuários Eletrônicos do Paciente Baseada em Semelhança Semântica e Ortográfica

Information Retrieval from Free Text Fields of Electronic Patient Records, Based on Semantic and Orthographic Portuguese Similarity

Amilton Souza Martha[1]
Carlos José Reis de Campos[2]
Daniel Sigulem [3]

Descritores: Armazenamento e Recuperação da Informação, Sistemas Computadorizados de Registros Médicos, Serviços de Informação

RESUMO

A maior parte da informação médica em forma digital se encontra na forma de textos livres como nos sites de medicina e saúde, artigos científicos em banco de dados da literatura biomédica e em prontuários eletrônicos do paciente (PEP). Muitos problemas podem ocorrer em sistemas de recuperação de informações médicas como o uso de sinonímia, erros de digitação e variações semânticas na linguagem médica. Para analisar a quantidade de informações que são perdidas em sistemas de busca tradicionais, que fazem um busca do termo exato, foram selecionados 34 termos médicos de duas bases de dados de PEPs e pesquisados com um algoritmo tradicional de busca direta embutido em um PEP chamado Clinic ManagerÒ e um sistema desenvolvido batizado SIRIMED que embutiu algoritmos de semelhança semântica (incorporação de sinônimos) e semelhança ortográfica (*edit distance+stemming*).Os resultados mostram que a recuperação dos termos aumenta em cerca de 30% em relação à busca tradicional, com uma quantidade de falsos positivos baixa (menos de 1%), o que mostra que muitas informações são perdidas normalmente.

Keywords: Information Storage and Retrieval, Medical Records Systems, Computerized, Information Services

ABSTRACT

Most medical information in digital form occurs in internet health sites, biomedical literature databases and electronic patient record (EPR). Many problems can be found in medical information retrieval systems like problems of synonyms, mistakes on typing and semantic variations in medical language. To analyse the amount of lost infomation in traditional information retrieval systems which use exact string matching, 34 medical terms were selected from two databases of EPRs and they were analysed with traditional search found in EPR System called Clinic Manager and a new system called Sirimed developed by autor which was added algorithms to semantic approximate (synonymous) and approximate string matching (edit distance + stemming).

The results show that information retrieval was improved in 30% in compare with the traditional search, with little amopunt of false positives (less 1%), that show a lot of lost informations.

Autor Correspondente:
Amilton Souza Martha
R. Botucatu, 862 - Vila Clementino - São Paulo - SP - Brasil
CEP: 04023-062
e-mail:
amiltonmartha@katusis.com.br

[1] *Meste em Informática em Saúde pela Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[2] *Professor Associado da Disciplina de Informática em Saúde pela Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*
[3] *Professor Titular da Disciplina de Informática em Saúde pela Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil.*

Artigo recebido: **08/09/2008**
Aprovado: **31/08/2010**

INTRODUÇÃO

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

Foi somente com o surgimento da escrita por volta de 3000 a.C., que se iniciou a fase do registro histórico do conhecimento humano. Porém, foi após Hipócrates (século IV a.C.) que a medicina teve sua transição do caráter mitológico para o uso do pensamento lógico-científico. Após a racionalização da prática da medicina, o registro médico tornou-se prática entre os adeptos ao pensamento pós-hipocrático.

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. 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[1].

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.

O Prontuário Eletrônico do Paciente (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[2].

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. 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. Por último, as imagens e os gráficos, que geralmente são adquiridas por máquinas ou desenhadas pelo médico[1].

Dados Estruturados x Textos Livres

Dado estruturado é aquele que possui uma faixa de valores pré-definidos, como sexo, idade, altura ou código da doença. 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[1,3].

Dados estruturados são mais fáceis de serem tratados por meios computacionais, pois existem linguagens formais como o SQL (*Structured Query Language*) que permitem sua manipulação e consulta de forma mais concisa e precisa[4]. Além disso, 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[5].

Por outro lado, eles 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.

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[6].

A conversão de um texto livre em dados estruturados provoca uma significante perda de informações e erros de classificação[3]. 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 estruturados[1].

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[3].

Além disso, 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[7].

Apesar da informação em texto livre ser difícil de indexar e, conseqüentemente, de recuperar, é largamente utilizada[8]. 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[7].

Recuperação de Informações

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 tornou-se necessário a criação de mecanismos para facilitar sua recuperação.

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.

Indexação de Documentos

A idéia de indexar é produzir um índice menor, porém mais eficiente, para representar o conteúdo original que facilite a recuperação de informações[9].

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 reproduzem uma linguagem natural[10].

O índice é uma forma de estruturação de dados bastante antiga e usada, consistindo de uma coleção de palavras ou termos selecionados associado a ponteiros que se relacionam com a informação, ou documentos. Índices são a base de todo sistema de recuperação de informações[11].

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.

A indexação manual implica numa seleção cuidadosa da terminologia empregada[10]. Para tanto, existem atualmente muitos vocabulários controlados médicos como o MeSH, CID, SNOMED e outros[9].

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.

Um dos métodos pioneiros na área foi desenvolvido por Salton em 1960, mas não teve grande sucesso até a década de 90. Chamado de Modelo de Vetor Espacial (*Vector Space Model*)[12], propõe que os documentos podem ser representados como vetores de termos com recuperação baseada na similaridade de ângulos entre os vetores da pergunta e dos documentos[9].

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[8-9,11,13].

Outro processo interessante na indexação é o chamado *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[14].

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[13].

Uma maneira bastante utilizada para medir os pesos das palavras é o padrão TF-IDF(*Term Frequency-Inverse Document Frequency*), que calcula o peso que cada palavra representa para o texto baseado na freqüência que ocorre no texto e no conjunto de documentos, onde:

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

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

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

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 ordem alfabética, ordem cronológica ou ordem de peso dos termos nos documentos[8-9].

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, especialidade ou região.

Quanto à ortografia, variações de escrita de uma palavra podem fazer com que a mesma não seja identificada em uma 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.

Existem basicamente quatro tipos básicos de erros de ortografia:

Inserção: Quando são inseridas na palavra letras ou caracteres a mais, como "cardíaco" e "cardíanco".

Remoção: Quando alguns caracteres são omitidos na transcrição, como "anamnese" e "ananese".

Troca: quando alguns caracteres são substituídos por outros, como "coração" e "corasão".

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[15].

O número mínimo de inserções, remoções ou substituições para uma palavra virar outra é conhecido como *edit distance*[16]. Esse enunciado corresponde à Distância de Levenshtein (*Levenshtein Distance*)[11].

O assunto de termos sinônimos é especialmente problemático na medicina, pois a linguagem biomédica possui muitas trocas de termos[17].

Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas com ela, utilizando então este conjunto para busca de documentos[13].

Uma possibilidade de expansão semântica para recuperação de informações médicas é o uso do DeCS (Descritores em Ciências da Saúde)[18], incorporando à pesquisa todos os termos sinônimos da palavra.

Em sistemas de recuperação de informação, algumas métricas são bastante importantes para a avaliação do sistema, denominadas *recall* (abrangência) e *precision* (precisão).

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[9,11,19].

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 visual de todos os documentos da base, o que pode ser inviável em muitos casos.

O objetivo desse trabalho foi desenvolver um software com algoritmos para recuperação de informações por semelhança semântica e ortográfica e comparar os resultados com um sistema de busca em textos tradicional que somente busca o termo exato.

MÉTODOS

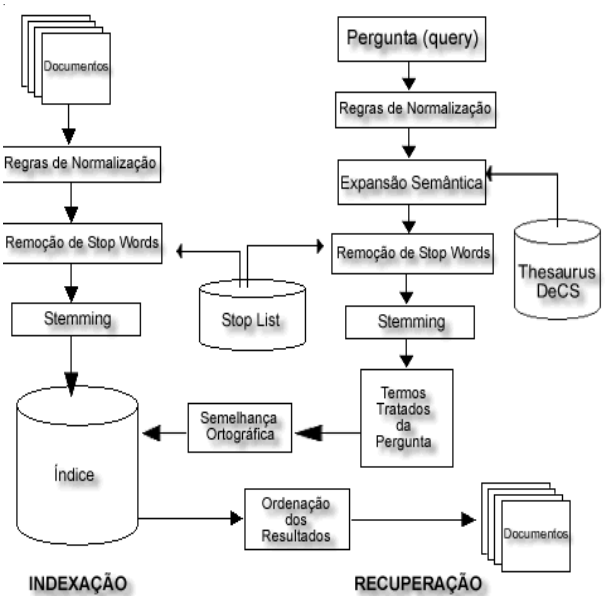
Para esse trabalho foram utilizados dois bancos de dados, sendo o primeiro banco (Base 1) cedido por uma clínica especializada em neurologia e psiquiatria (Instituto Campos & Cardeal) com 6732 registros de histórias clínicas em textos livres e o segundo banco (Base 2), uma clínica médica especializada em nefrologia e clínica médica (Clínica Médica Sigulem & Mattei S/C Ltda) com 26072 registros de histórias no mesmo formato. Ambos fazem parte de um sistema de prontuário eletrônico desenvolvido pelo Departamento de Informática em Saúde (DIS) da UNIFESP chamado Clinic Manager® no formato Access (MDB)[20].

Foi utilizada a versão DeCS 2004 que possui 26851 termos em Português e 31936 formas sinônimas para os termos autorizados.

Para a remoção de *stop words* foi utilizada a lista proposta pelo projeto Snow Ball (http://snowball.tartarus.org/) composta por 220 termos[21].

Para a comparação de eficiência na busca dos termos, foi utilizado o Clinic Manager® versão 7.0.7.85 – Versão de Avaliação. Esse programa faz uma busca exata pelo termo nos textos sem nenhum tratamento, podendo assim comparar os resultados com o uso do algoritmo e sem o uso do mesmo.

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.



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

Figura 1 - Etapas do SIRIMED

O algoritmo proposto foi implementado num sistema batizado SIRIMED (Sistema de Indexação e Recuperação de Informações Médicas), cuja estrutura pode ser vista na **Figura 1**.

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 2**.

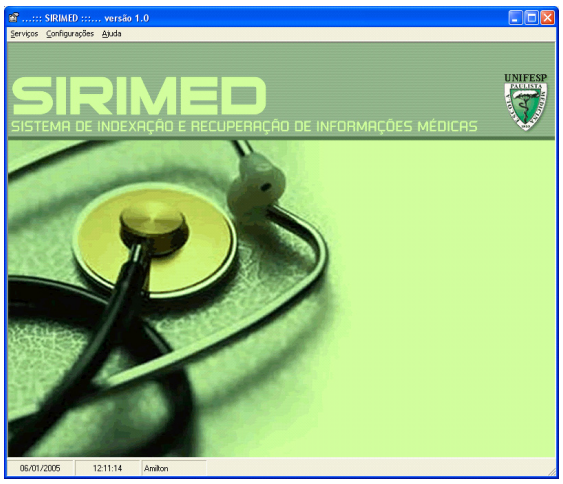


Figura 2 – Tela Inicial do SIRIMED

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 (RTF) para textos sem formatação, transformação do texto em minúsculas e remoção de caracteres especiais, citados a seguir.

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

Além disso, as palavras foram convertidas para letras minúsculas e foram removidos as acentuações.

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. A lista se *stop words* poderia ser editada conforme **Figura 3.**

Após a remoção de *stop words*, os termos restantes foram processados pelo algoritmo de *Stemming* para o Português de Martin Porter[21] 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.

A segunda etapa da pesquisa consiste na recuperação dos textos baseado numa pergunta do usuário. Nessa tela do sistema (**Figura 4**), o usuário

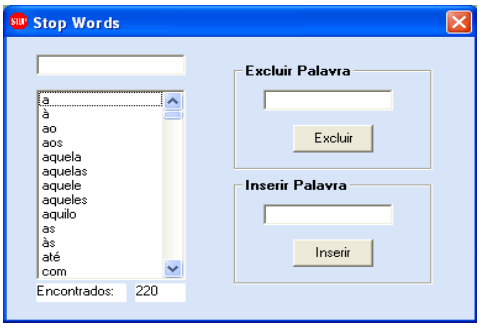


Figura 3 – Tela de edição de *stop words*

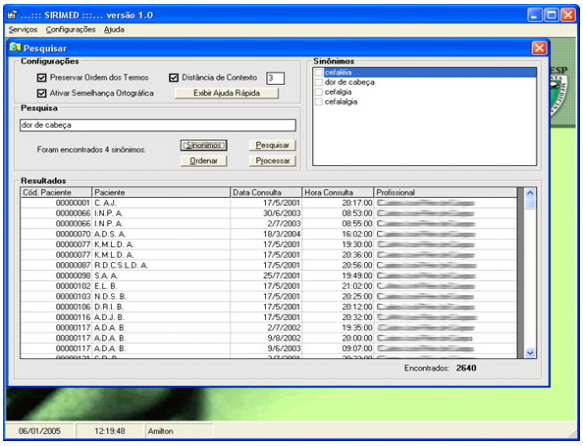


Figura 4 – Tela de pesquisa do SIRIMED

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.

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 a 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*. O sistema possui um Thesaurus interno que é carregado com o vocabulário DeCS e pode ter termos incluídos, excluídos ou alterados pelo usuário.

RESULTADOS

Após a indexação de todas as histórias, 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, diagnósticos ou medicamentos.

Esses critérios visam contextualizar a busca na área médica. Dentre os elementos excluídos, podemos citar nomes pessoais (médicos e pacientes), palavras muito

Tabela 1 – Quantidade de *stop words* e tamanho do índice criado

Base 1 – Neurologia / Psiquiatria	
1. cefaléia	10. vômitos
2. enxaqueca	11. gardenal
3. convulsão	12. latejante
4. tontura	13. desmaio
5. nervoso	14. depressão
6. meningite	15. insônia
7. ronco	16. cbz
8. tegretol	17. diabetes
9. tremor	18. ansiedade
Base 2 – Nefrologia / Clínica Médica	
1. dor	9. edema
2. triglicérides	10. diabetes
3. creatinina	11. moduretic
4. leucócitos	12. renitec
5. plaquetas	13. febre
6. hemoglobina	14. vioxx
7. hematocrito	15. diprospan
8. obesidade	16. antak

Base de Dados	Base 1	Base 2
Qtde de palavras	830.471	3.990.900
Qtde de *stop words*	270.913 **(32,6%)**	1.511.763 **(37,9%)**
Palavras indexadas	559.558	2.479.137
Índice sem *Stemming*	26.977 **(4,8%)**	38.179 **(1,5%)**
Índice com *Stemming*	19.543 **(3,5%)**	26.781 **(1%)**

freqüentes, porém que não tinham poder semântico isoladamente como "dias", "exame", "retorno", "data", "diagnóstico" e outras. Os termos selecionados foram:

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 1**.

Devido ao uso de *stemming*, o índice também foi reduzido pela união de palavras com o mesmo radical comum na mesma entrada do índice.

Notamos na **Tabela 2** e na **Tabela 3** que somente o processo de *stemming* aumenta a recuperação (chegando a 80% em ronco), mas que combinado com a semelhança ortográfica, o aumento na recuperação alcança quase 194% (como em desmaio) e, de acordo com a quantidade de sinônimos encontrados, a recuperação semântica pode até triplicar (263% em desmaio).

Alguns termos como 'ronco' e 'tontura' tiveram aumento somente com o uso de *stemming* (80,1% e 73,7%) ao passo que em 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.

Podemos notar na **Tabela 4** 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 4** 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.

Nota-se claramente nas **Tabelas 4 e 5** 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.

Tabela 2 – Recuperação de Freqüência de Palavras com e sem o algoritmo da Base 1

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**% aumento**	**ST+SO**	**% aumento**	**ST+SO+SS**	**% aumento**
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

Tabela 3 – Recuperação de Freqüência de Palavras com e sem o algoritmo da Base 2

		Recuperação pelo SIRIMED					
Termos	**CM**	**ST**	**% aumento**	**ST+SO**	**% aumento**	**ST+SO+SS**	**% aumento**
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 (busca exata); ST – Uso de Stemming, SO – Uso de Semelhança Ortográfica; SS – Uso de Semelhança Semântica (incorporação de sinônimos)

Tabela 4 –Falsos positivos

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

DISCUSSÃO

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 indica que há muitas variações ortográficas que precisam ser recuperadas por semelhança ortográfica, por outro lado, por serem termos muitos comuns, os sinônimos para eles não surtirão grande contribuição. Provavelmente, na busca de termos mais incomuns, a incorporação de sinônimos terá uma contribuição mais expressiva na recuperação por semelhança semântica.

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.

Nenhum algoritmo de semelhança fonética foi utilizado pois 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

Tabela 5 – 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	-	3146	-	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	-	2084	-	2098	Doenças febris, enfermidades febris, hipertermia, pirexia	Pirox, Piretamida, pirena, porex	0,67
triglicérides	7414	-	7424	-	7424	-	-	-
creatinina	6961	-	6964	-	6964	-	-	-
leucócitos	6554	-	6566	-	6566	-	-	-
plaquetas	3865	-	3868	-	3868	Trombócitos	-	-
hemoglobina	3747	-	3747	-	3747	-	-	-
hematocrito	3697	-	3697	-	3697	-	-	-
obesidade	3327	-	3334	-	3334	-	-	-
diabetes	2730	-	2752	-	2752	-	-	-
moduretic	2304	-	2311	-	2311	-	-	-
renitec	2264	-	2270	-	2270	-	-	-
vioxx	1987	-	1994	-	1994	-	-	-
diprospan	1950	-	1955	-	1955	-	-	-
antak	1765	-	1773	-	1773	-	-	-
Média		0,04		0,04				0,21

algoritmos foram desenvolvidos para o idioma inglês, utilizando o som dos fonemas dessa língua. 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.

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[22], 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 PEPs.

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[1] (*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[2], possuindo as mesmas características do anterior, além da inclusão de sinônimos, também em inglês. Ambos os mencionados são softwares livres.

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 o termo foi escrito incorreto, porém, não incorpora à sua busca e simplesmente substitui o termo caso a sugestão seja aceita.

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 2**.

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 5**. 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.

Outra observação importante 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 5**).

A inserção de algoritmos de semelhança na recuperação pode trazer resultados que não são esperados. Nas **Tabela 4 e 5** mostramos quantos falsos positivos o sistema trouxe. A maioria dos erros se encontra na

[1] Swish-e - http://swish-e.org/
[2] ht://Dig - http://www.htdig.org/

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)'.

CONCLUSÃO

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 maior do que isoladamente.

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[23]. 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.

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.

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.

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.

REFERÊNCIAS

1. Shortliffe EH, Barnett GO. 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. 2. ed. New York: Springer; 2001. p. 41-75.
2. Shortliffe EH, Blois MS. 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. 2. ed. New York: Springer; 2001. p. 3-40.
3. Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? Int J Med Inform. 2000; 58(1): 101-10.
4. Loh S. Descoberta de conhecimento em bases de dados textuais. [Citado 1997 maio 17]. Disponível em: http://mozart.ulbra.tche.br/~loh/apostilas/dc-texto.htm.
5. Mulligen EMV, Stam H, Ginneken AMV. Clinical data entry. Proceedings of AMIA 1998 Symposium. [cited 2006 march 19]. Avaliable from: http://www.amia.org/pubs/symposia/D004709.pdf.
6. Sager N, Lyman M, Nhan N, Tick, LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc (JAMIA). 1994; 1(2):142-60.
7. Johnson SB. A semantic lexicon for medical language. J Am Med Inform Assoc (JAMIA). 1999; 6(3): 205-18.
8. Wives LK. 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. [Citado 2005 set 25]. Disponível em http://www.inf.ufrgs.br/~wives/ publicacoes/IDT.pdf.
9. Hersh WR, Detmer WM, Frisse ME. Information retrieval systems In: Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM. Medical informatics: computer applications in health care and biomedicine. 2. ed. New York: Springer; 2001. p. 539-72.
10. Kent A. Manual da recuperação mecânica da informação, São Paulo: Polígono; 1972.
11. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press; 1999.
12. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. In: Proceeding of the Communications of the ACM. 1975; 18(11): 613-20.
13. Wives LK, Loh S. Recuperação de informações usando a expansão semântica e a lógica difusa. In: Anais do Congreso Internacional en Ingenieria Informatica; 1998 Abril; Buenos Aires Argentina.
14. Porter MF. An algorithm for suffix stripping. 1980;14(3):130-7. Disponível em: http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html. Acesso em: 18/08/2010.
15. Hall PAV, Dowling GR. Approximate string matching. Computing surveys. 1980; 12(4):. 381-402.
16. Baeza-Yates R, Navarro G. Fast approximate string matching in a dictionary. In: Proceeding of the 5th South American Symposium on String Processing and Information Retrieval (SPIRE'98); 1998 september 9-11; Santa Cruz de la Sierra, Bolívia.1998. p. 14-22.
17. Chu SS. Information Retrieval and health / clinical management. Yearbook of Medical Informatics. 2002.
18. Bireme. DeCS - Descritores em Ciências da Saúde. [citado 2010 ago19]. Disponível em http://decs.bvs.br/P/decswebp.htm.
19. Hersh WR. Information retrieval : a health and biomedical perspective. 2. ed. New York: Springer; 2003.
20. Sigulem D, Cardoso OL, Gimenez SSFX, Cebukin A, Anção MS. Clinic manager system. In: Proceeding of the World Congress on Medical Physics and Biomedical Engineering; 1994 August 21-26; Rio de Janeiro, Brasil; 1994 v.1. p.556.
21. Snowball. Portuguese stemming algorithm. [citado 2002 set 10]. Disponível em: http://snowball.tartarus.org.
22. Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceeding of the 26th Annual International ACM SIGIR Conference; 1994 3-6 July; Dublin, Ireland; 1994. p.192-201.
23. Tan Ah-Hwee. Text mining: the state of the art and the challenges. In: Proceedings of the Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases; 1999; Beijing. p.65-70.