XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

GT8 - **Informação e Tecnologia**
Modalidade de apresentação: Comunicação Oral

UM MÉTODO DE INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS: APLICAÇÃO EM LAUDOS DE EXAMES RADIOLÓGICOS

EDBERTO FERNEDA
Universidade Estadual Paulista "Júlio de Mesquita Filho"
MARIA CRISTIANE BARBOSA GALVÃO
Universidade de São Paulo
JOELI ESPÍRITO SANTO ROCHA

Resumo: Apresenta um método estatístico de indexação automática de documentos baseado em diversas pesquisas realizadas nessa área. Esse método foi utilizado no desenvolvimento de um sistema de indexação automática de laudos de exames radiológicos. Testes utilizando-se um *corpus* com aproximadamente cinco mil documentos apontam resultados positivos e promissores, principalmente no campo da saúde, no qual a disponibilização rápida e precisa informação tem um papel fundamental para a melhoria na qualidade de vida da população do país.

1 Introdução

Indexar um documento é um processo que visa a representação de seu conteúdo temático através da utilização de um conjunto de palavras ou termos de indexação. Os termos de indexação servem também como pontos de acesso mediante os quais um documento é localizado e recuperado em um sistema de informação.

A primeira tentativa de automatizar o processo de indexação foi realizada no final da década de 1950. Se nessa época tratava-se de uma teoria progressista, atualmente a indexação automática torna-se uma necessidade, diante do acelerado aumento na disponibilização da informação eletrônica e a proliferação de documentos de textos completos.

Anderson e Perez-Carballo (2001) citam o baixo custo da indexação automática e sua facilidade de aplicação a grandes conjuntos de documentos como o encontrado na Internet, um importante fator de incentivo ao desenvolvimento de métodos de indexação automática. Outro argumento em favor da indexação automatizada está na homogeneidade do processo quando realizados por algoritmos computacionais. O resultado da indexação realizada por seres humanos pode variar de um indexador para outro, bem como de um mesmo indexador em momentos diferentes. Um sistema de computador irá realizar a indexação de maneira uniforme, utilizando-se sempre dos mesmos critérios para o qual foi programado, independentemente da quantidade de documentos existente no *corpus* ou de qualquer outro fator externo.

O objetivo principal deste trabalho é apresentar um método estatístico de indexação automática de documentos. Este método, implementado em um sistema computacional, foi aplicado na indexação de laudos de exames radiológicos e apresentou resultados expressivos.

Na seção 2, é apresentado brevemente algumas iniciativas relacionadas à indexação automática de fontes de informação na área médica Na seção 3 é descrito um método de indexação automática de documentos baseado em diversos outros métodos desenvolvidos ao longo de décadas de pesquisas nessa área. A partir desse método foi implementado um sistema, denominado SintagMed, utilizado na indexação de um *corpus* documental composto por aproximadamente cinco mil laudos de exames radiológicos. Esse sistema é apresentado detalhadamente na seção 4. Os resultados experimentais do processo de indexação do *corpus* são apresentados e avaliados na seção 5. Por fim, na seção 6, são apresentadas as consideração finais sobre o projeto aqui descrito.

2 Indexação automática de laudos de exames radiológicos

É consenso entre alguns autores que apesar dos avanços tecnológicos e dos progressos ocorridos nos sistemas de informação, as instituições ainda se deparam com grandes dificuldades para produzir, gerenciar, disponibilizar e acessar informações relevantes. Entende-se que mesmo a informação estando em meio digital, ainda é pouco organizada, sendo difícil acessá-la, pesquisá-la e filtrá-la (ALVARENGA, 2006; NORUZI, 2007; SVENONIUS, 2000).

Estas afirmações são também válidas para o campo da saúde, especificamente, para a informação clínica, ou seja, informação sobre pacientes, disponível em ambientes digitais ou em outros suportes, já que muitos sistemas de informação clínica têm sido desenvolvidos sem levar em conta os processos necessários para a posterior recuperação da informação – fato que gera certas restrições a estes sistemas no momento de busca por informação precisa.

No campo da saúde a indexação automática pode ser aplicada a muitos sistemas, dentre os quais: a análise de resumos de alta hospitalar de pacientes; laudo de exames; prontuários de pacientes.

A National Library of Medicine (NLM) é uma das instituições de saúde que mais investe no desenvolvimento métodos de processamento automático de textos para a indexação automática. Seu programa *Indexing Initiative* (II) tem como objetivo investigar métodos de indexação automática,

agregando projetos como: atribuir automaticamente descritores do *Medical Subject Headings* (Mesh) aos artigos de periódicos indexados em sua base de dados, processos automáticos para identificação de nomes químicos de substâncias, processos automáticos para lidar com terminologia de ligação molecular, terminologia de medicamentos e genes, e termos anatômicos. Em muitos processos de indexação automática desenvolvidos pela NLM, os termos identificados nos textos são comparados com o UMLS (*Unified Medical Language System*). Todavia, tais iniciativas não abarcam a língua portuguesa, e priorizam a informação de caráter bibliográfico e não a informação clínica.

Outra instituição que lida com o tratamento da informação bibliográfica relacionado à área da saúde e que está começando a implantar métodos de indexação automática é a base de dados da Literatura Latino-Americana e do Caribe em Ciências da Saúde (LILACS), base de dados cooperativa da Rede de Biblioteca Virtual em Saúde (BVS). A LILACS já usa a indexação automática ao transferir os artigos da Scielo Brasil para sua base de dados. Tal iniciativa abarca a língua portuguesa, mas restringe-se igualmente à informação bibliográfica.

Além do uso de processos de indexação automática pelas instituições que lidam diretamente com o tratamento da produção científica no campo da saúde, outro ramo que a indexação automática tem ganhado espaço no campo da saúde é a recuperação de informações de textos em formatos eletrônicos sem estruturação, principalmente em hospitais. Informações desestruturadas e em texto livre, geradas há muitos anos e armazenadas em banco de dados começam a ser utilizadas como: ferramenta de investigação epidemiológica; para aperfeiçoamento dos cuidados aos pacientes; para a identificação de pacientes com sintomas e doenças raros ou com doenças e sintomas similares; para o estudo e comparação entre as diferentes formas de tratamento e diagnósticos adotados (CASTILLA, 2007; ALEXANDRINI, 2005)

Embora o Brasil já utilize documentos como fontes de informação em saúde (Atestado de Óbito, Certidão de Nascimento, Autorização de Internações Hospitalares) para o levantamento de dados a serem usados na formulação de indicadores de saúde, disponíveis em sistemas nacionais (Sistemas de Informação do Programa Nacional de Imunização (SI–PNI), Sistema de Informações de Mortalidade (SIM), Sistema de Informações de Nascidos Vivos (SINASC), Sistema de Autorização de Internação Hospitalar (AIH), e muitos outros), outras fontes de informações poderiam ser melhor aproveitadas, como: atas, relatórios, exames etc., pois, normalmente somente são aproveitadas em instituições de saúde quando faltam dados estatísticos apropriados para analisar uma hipótese ou problema local (PEREIRA, 2003).

Cada vez mais exames de pacientes são produzidos e a busca de informações pelos profissionais da saúde (médicos, enfermeiros, assistentes sociais, terapeutas, radiologistas etc.) aumenta, sendo necessários sistemas automatizados que façam a codificação de exames de pacientes (ALEXANDRINI, 2005).

O campo da saúde é uma área que demanda o desenvolvimento de métodos automáticos de organização e recuperação de informações. Porém, para que sistemas computacionais de indexação obtenham melhores resultados, os documentos no meio digital devem apresentar alguma estruturação e padronização. Para tanto, ao se criar sistemas de informação é importante saber *a priori* qual o tipo de necessidade informacional dos usuários o sistema precisa atender, e a partir disso elaborar documentos padronizados e normalizados para que se possa implantar métodos automatizados e/ou até mesmo métodos semi-automatizados de organização e recuperação de informações (SVENONIUS, 2000)

3 Método de indexação automática de documentos

Os primeiros métodos de indexação automática, desenvolvidos por Luhn (1957) durante suas atividades na IBM, baseavam-se nos estudos de Zipf, que décadas antes observara a utilização

repetida de certas palavras na comunicação escrita e oral. Luhn considerava que o vocabulário existente em um documento deveria ser a base para a análise de seu conteúdo. Para ele, após a retirada palavras vazias tais como artigos, preposições etc, as melhores palavras para indexação seriam as de freqüência média.

Seguindo essa linha estatística da indexação automática, no início da década de 1970 Spärk Jones (1972) propôs um método de ponderação de termos, o IDF (*Inverse Document Frequency*), "que mede a escassez de aparição de um termo em uma coleção". Essa forma de ponderação é ainda muito usada juntamente com a freqüência da palavra/termo em um documento, medida conhecida pela sigla TF-IDF (*Term Frequency - Inverse Document Frequency*).

Segundo Robredo (1982), a indexação automática baseia-se "na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas". De acordo com o mesmo autor, esse processo pode identificar termos, pares de termos ou até frases significativas que expressem o conteúdo do documento, e pode-se dizer que é semelhante ao processo de leitura-memorização.

O método de indexação automática apresentado neste trabalho baseia-se em trabalhos de Marie-Françoise Bruandet do *Laboratoire Génie Informatique de Grenoble,* França (BRUANDET, 1989).

Seja **T** o conjunto de palavras/termos existentes em um *corpus* documental. O primeiro elemento a ser definido é uma medida para avaliar a ligação contextual entre dois termos. A i-ésima ocorrência de um termo ***x*** do vocabulário **T**, simbolizado por $\mathbf{w_x(i)}$, é definido por suas coordenadas:

$$w_x(i) =) ND_x(i), NF_x(i), NP_x(i) *$$

$ND_x(i)$ identifica um determinado documento; $NF_x(i)$ representa o número da frase no documento e $NP_x(i)$ a posição da palavra na frase do texto do documento. Tem-se assim um "endereço" que identifica cada palavra em cada um dos documentos do *corpus*.

Para cada par de termos **(*x*, *y*),** define-se uma distância **d** entre a i-ésima ocorrência de ***x*** e a j-ésima ocorrência de ***y***. Essa distância é calculada pela diferença entre a posição relativa de cada uma dessas palavras no texto do documento. Porém essa medida só será considerada se o par de palavras estiver em uma mesma frase de um mesmo documento. Assim, a distância **d** pode ser representada matematicamente utilizando a seguinte fórmula:

$$d(w_x(i), w_y(j)) = \begin{cases} NP_x(i) \square NP_y(j) & se \begin{cases} ND_x(i) = ND_y(i) \\ e \\ NF_x(i) = NF_y(i) \end{cases} \\ 0, \textit{caso contrário} \end{cases}$$

Seja **F** uma função que expressa a força de ligação entre duas palavras. Ela é definida como sendo o inverso da distância **d** definida acima:

$$F(w_x(i), w_y(j)) = \frac{1}{d(w_x(i), w_y(j))}$$

Portanto, a força de ligação entre um par de palavras é inversamente proporcional à distância entre essas palavras. Quando menor a distância entre duas palavras, maior a força de ligação (F) entre elas. Quando maior a distância, menor será o valor de F.

Seja **b** a função do somatório de **F** para todas as ocorrências das palavras ***x*** e ***y*** no conjunto de documentos do *corpus*:

$$\mathbf{b}(x, y) = \sum_i \sum_j F(w_x(i), w_y(j))$$

Uma definição preliminar da medida de associação entre ***x*** e ***y*** é dada por:

$$M_1(x, y) = \frac{b(x, y)}{f(x, y)}$$

onde *f(x, y)* é o número de ocorrências (freqüência) do par *(x, y)* no conjunto de documentos.

$$0 \leq M_1(x, y) \leq 1$$

Utilizando apenas $M_1(x, y)$, a força de ligação poderá valer 1 quando dois termos *x* e *y* forem adjacentes em uma mesma frase e só aparecerem uma única vez no *corpus*. Tal situação representa uma exceção significativa na utilização da medida M_1. Para eliminar tal problema é introduzido um fator de correção **k**, que é função da freqüência f(x, y) do par (x, y). O fator **k** recebe o valor zero quando f(x, y) vale 1 e tende a 1 conforme f(x, y) aumenta. Assim, uma nova medida M_2 é definida como:

$$M_2(x, y) = k(x, y) - M_1(x, y)$$

$$com \quad k(x, y) = \frac{(f(x, y) - 1)^n}{f(x, y)^n}$$

A variável ***n*** é um parâmetro inteiro definido por experimentação. Um aumento de ***n*** tende a reforçar as ligações muito freqüentes e atenuar as ligações pouco freqüentes. O parâmetro ***n*** tem seu valor definido de forma empírica, podendo ser utilizado para ajustar a aplicação de acordo com finalidades específicas.

A introdução do parâmetro ***k*** permite uma melhor adaptação da medida M_2 às necessidades da aplicação. Com este parâmetro é possível atuar sobre a freqüência com a qual os pares de termos devem se relacionar para serem considerados como relacionados a um mesmo conceito.

A medida $M_2(x, y)$ pode ser também representada através da seguinte fórmula geral:

$$M_2(x, y) = \frac{\sum_i \sum_j F(w_x(i), w_y(j))}{f(x, y)} - \frac{(f(x, y) - 1)^n}{f(x, y)^n}$$

3.1 Construção da Matriz Termo-Termo

Os valores da medida M_2 são armazenadas em uma **matriz termo-termo**, como mostrado no exemplo da Figura 1, onde as letras p, t, u, w, x, y, z representam palavras ou termos extraídos dos documentos.

Termo	p	t	u	w	x	y	z
p	-	0,12	0	0	0	0,18	0
t	-	-	0,21	0,19	0,23	0,18	0,17
u	-	-	-	0,18	0,13	0	0
w	-	-	-	-	0	0	0
x	-	-	-	-	-	0,27	0,26
y	-	-	-	-	-	-	0,23
z	-	-	-	-	-	-	-

Figura 1: Matriz termo-termo contendo os valore da medida M_2

É importante observar a simetria da matriz gerada, além do fato da diagonal ser desconsiderada. Essas características permitem uma redução significativa do espaço de armazenamento das informações contidas na matriz, refletindo em um aumento de velocidade de execução das rotinas de cálculo realizadas com essa matriz.

A partir da matriz contendo os valores de M_2 (Figura 1) constrói-se uma **matriz binária termo-termo**. Neste processo é possível eliminar as ligações mais fracas, de acordo com um limite mínimo preestabelecido. A matriz binária apresentada na Figura 2, por exemplo, é construída a partir da matriz da Figura 1, utilizando um limite mínimo igual a 0,14. Dessa forma são eliminadas as ligações mais fracas (t, p), (x, u), reduzindo ainda mais a quantidade de informações a serem armazenadas.

Termo	p	t	u	w	x	y	z
p	-	0	0	0	0	1	0
t	-	-	1	1	1	1	1
u	-	-	-	1	0	0	0
w	-	-	-	-	0	0	0
x	-	-	-	-	-	1	1
y	-	-	-	-	-	-	1
z	-	-	-	-	-	-	-

Figura 2: Matriz binária termo-termo

3.2 Cliques

A matriz binária da Figura 2 pode ser apresentada na forma de um grafo, como apresentado na Figura 3.

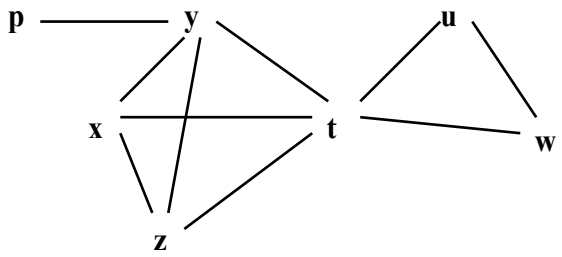

Figura 3: Grafo representando a matriz binária

A partir do grafo da Figura 3 é possível extrair os seus subgrafos máximos completos, chamados **cliques** (subgrafos cujos nós estão todos conectados entre si), representado como na Figura 4.

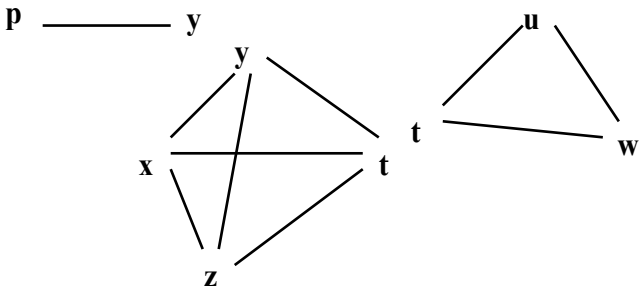

Figura 4: Cliques extraídos da matriz binária termo-termo

Diversos algoritmos para a extração de cliques de um grafo estão disponíveis no domínio da teoria dos grafos. Um deles é apresentado por Reingold (1977). Através dos cliques obtém-se uma

representação da matriz sem perda de informações. Cada clique fornece um conjunto de termos completamente conectados entre si. Pode-se dizer que cada clique representa um conceito ou uma idéia contida no conjunto de textos do *corpus*. Pode-se ainda considerar que cada clique representa um termo de indexação ou sintagma.
A partir do método de indexação automática aqui descrito, foi desenvolvido um sistema computacional cuja finalidade principal e imediata era a indexação de um *corpus* documental composto por cerca de 5000 laudos médicos de exames radiológicos. Este software foi denominado SintagMed.

4 O Sistema SintagMed

O sistema SintagMed foi desenvolvido utilizando a linguagem Delphi 7.0 em ambiente Windows. Como meio de armazenamento foi utilizado a versão gratuita do sistema gerenciador de banco de dados Interbase. Na Figura 5 é apresentada a janela de apresentação do sistema.

Figura 5: Janela de apresentação do SintagMed

Nas primeiras observações realizadas no *corpus* documental verificou-se que os laudos de exames radiológicos, mesmo em suporte eletrônico, trazem em sua redação erros de digitação, excesso de abreviaturas, siglas e variações terminológicas que dificultam a indexação seja a realizada por humano, seja a realizada automaticamente através de *software*. Significa dizer que a cultura da escrita médica em suporte papel, de alguma forma, permanece no suporte digital, dificultando igualmente sua inteligibilidade.
Assim sendo, foi necessário desenvolver alguma forma de normalizar palavras, termos, abreviações e siglas utilizados pelos médicos na elaboração de seus laudos. Optou-se por utilizar uma simples substituição de uma determinada palavra, sigla ou abreviação por uma palavra normalizada. Esta tarefa é realizada utilizando uma lista pré-definida de palavras seguida da palavra normalizada correspondente. A Figura 6 apresenta a janela do sistema onde esta lista é cadastrada.

Figura 6: Janela para o cadastramento de palavras e suas respectivas formas normalizadas

Algumas palavras possuem valor semântico irrelevante na composição dos termos de indexação. Essas palavras, conhecidas como "*stop words*", são desconsideradas durante o processo de indexação. Além de artigos, preposições, conjunções etc, são também desconsiderados os números, as unidades de medidas e algumas abreviações. A Figura 7 apresenta a janela para o cadastramento de *stop words*.

Figura 7: Janela para o cadastramento de palavras vazias (*stop words*)

O processo de extração de termos de indexação é executado em duas fases. Na primeira fase são extraídas as palavras dos textos, excetuando-se as *stop words* e normalizando as demais utilizando-se a tabela previamente cadastrada. O sistema permite definir a abrangência da indexação. É possível limitar a indexação a um determinado conjunto de laudos com determinadas características comuns, tais como o tipo de exame e a região (parte do corpo) onde foi realizado o exame. A Figura 8 apresenta a janela onde são executada a primeira fase de indexação, a extração e normalização de palavras.

Figura 8: Primeira fase da indexação: extração de palavras

Em uma segunda fase são calculadas as forças de ligação entre as palavras conforme metodologia apresentada na Seção 3. O resultado dessa operação é um conjunto de termos compostos por duas ou mais palavras (sintagmas). Antes de se realizar este processo é possível definir o valor de parâmetros que afetarão diretamente nos resultados. Uma definição detalhada desses dois parâmetros foi apresentada na Seção 3. A Figura 9 mostra a janela onde é feita a extração dos termos de indexação.

Figura 9: Segunda fase da indexação: extração de termos de indexação

O sistema SintagMed possui diversos relatórios e telas de consulta utilizados para aferir a precisão dos resultados obtidos no processo de indexação.

5 Resultados experimentais

A comparação entre os dois tipos de indexação, automática e manual, é realizada para se verificar as diferenças e semelhanças entre os termos selecionados por programas de um computador e pelo homem. De acordo com os resultados obtidos, avalia-se a aplicabilidade de uma ou outra técnica. Segundo Salton (1969; 1972), a grande maioria de testes comparativos entre descritores atribuídos manual e automaticamente chega a um resultado aproximado de 60% de compatibilidade entre uma linguagem e outra.

Carroll e Roeloffs (1969) realizaram estudos comparativos entre indexação manual e automática aplicando a análise de correlação estatística. Verificaram que os termos obtidos pelos indexadores humanos foram semelhantes aos da indexação automática, mas levando-se em conta os custos de

contratação e treinamento de mão-de-obra especializada, e a inerente inconsistência na indexação humana, a indexação automática apresenta grandes vantagens .

Salton (1972), Boyce; e Lockard (1975) realizaram suas experiências na área médica. O primeiro comparou os mesmos documentos indexados manualmente, utilizando vocabulário controlado, e indexados automaticamente pelo sistema SMART *(System for the Mechanical Analysis and Retrieval of Text*) utilizando termos livres do resumo. Salton verificou que numa indexação automática somente com truncagem de palavras, a indexação manual torna-se mais efetiva cerca de 15% a 20%. Quando se utiliza um controle através de tesauros e dicionários, a eficiência da indexação automática é semelhante à da manual.

Durante estudo empírico realizado no contexto deste projeto, foi utilizado um *corpus* documental delimitado a partir dos 660.000 laudos de exames radiológicos produzidos pelo Centro de Ciências das Imagens e Física Médica do Hospital das Clínicas de Ribeirão Preto até o ano de 2005. Dos 660.000 laudos foram selecionados aleatoriamente 5.000 laudos de exames radiológicos para análise de suas estruturas informacionais. Para viabilizar uma comparação entre um indexador humano e o software SintagMed, foram selecionados 344 laudos, sendo 23 flebografias, 273 radiografias contrastadas, 20 mamografias e 28 planigrafias. O processo de indexação empregado pelo indexador humano foi dividido em três etapas: leitura e análise do laudo; identificação do conteúdo informacional do laudo e seleção das palavras-chave presentes no laudo que melhor pudessem representar o seu conteúdo. Empregou, desta forma, uma indexação por linguagem natural pautada na identificação do conteúdo informacional do texto completo.

Após os processos de indexação realizados pelo indexador humano e pelo software Sintagmed, os produtos da indexação obtidos foram comparados quanto à quantidade e qualidade dos termos obtidos.

Dessa forma, sobretudo para o processo de indexação automática, foi necessário cadastrar uma lista de palavras a serem desconsideradas no processo de indexação (*stopwords*); uma lista de equivalência entre siglas e abreviações e a sua respectiva forma extensa. Somente após a constituição destas listas o software Sintagmed realizou uma indexação adequada, identificando os sintagmas ou termos compatíveis ao conteúdo informacional de cada laudo de exame radiológico.

O resultado da indexação realizada pelo indexador humano e pelo software Sintagmed apresentou grande semelhança quanto à qualidade semântica. Todavia, o ser humano em todos os tipos de exame radiológicos identificou e selecionou um número maior de termos para representar o conteúdo semântico analisado, conforme apresentado na Tabela 1 e Tabela 2.

Tabela 1 – Comparação quantitativa de termos resultantes da indexação humana e da indexação automática

Exame radiológico	Quantidade de termos resultantes da indexação humana	Quantidade de termos resultantes da indexação automática	Termos equivalentes
Flebografia	170	84	60
Mamografia	56	59	34
Planigrafia	99	34	19
Radiografia com contraste	558	466	206
Total	883	643	319

Tabela 2 – Relevância dos resultantes da indexação automática, quando comparadas ao resultado obtido pelo indexador humano

Exame radiológico	Relevância dos termos resultantes da indexação automática
Flebografia	71,43%
Mamografia	57,63%
Planigrafia	55,88%
Radiografia com contraste	44,21%
Total	49,61%

Em relação ao tempo gasto, buscou-se testar o Sintagmed em computadores de fácil acesso em países subdesenvolvidos ou em desenvolvimento, a fim de verificar sua aplicabilidade em contextos economicamente menos favorecidos. Esta decisão considerou o conhecimento que os autores da pesquisa possuem acerca da realidade de muitas unidades de saúde existentes, por exemplo, nos países lusófonos.

Dessa forma, o Sintagmed instalado em um computador com processador AMD atlhon, 900 MHz e 512Mb de memória RAM, realizou a indexação de 344 laudos radiológicos em 4 horas. Já o indexador humano demorou 20 horas para ler os laudos, compreendê-los e indexá-los sem o auxílio de dicionários, tesauros ou terminologias, ou seja, elaborando igualmente, a indexação por linguagem natural.

6 Conclusão

Em que pese o caráter experimental do estudo, seus achados são relevantes, quais sejam: identificou a carência de estudos sobre a indexação automática que considerem a língua portuguesa, observou a dificuldade de sensibilização dos desenvolveres de softwares para as problemáticas enfrentadas por países lusófonos; observou a necessidade de estudos voltados para a indexação automática retrospectiva de massas de informações clinicas já existentes em suporte digital; verificou as dificuldades de indexação por humano e automatizada no que tange a inteligibilidade dos textos médicos em suporte digital, dada a existência de siglas, abreviaturas, erros de digitação e variação terminológica.

Quanto a proposta de comparação a que se propôs o estudo, houve grande similaridade entre a indexação realizada por indexador humano e via software Sintagmed. No entanto, o software teria melhor desempenho caso os textos dos laudos fossem de melhor qualidade: utilizassem abreviaturas e siglas de forma padronizada; tivessem menor quantidade de erros de digitação, variações terminológicas e neologismos.

Entende-se que o método de indexação automática e o software desenvolvido (SintagMed) apresentaram-se como promissores e adequados às instituições de saúde. No caso do Brasil, ferramentas como estas são essenciais para se tentar recuperar uma imensa quantidade de informação em suporte digital que foi sendo acumulada nas últimas décadas, sem que houvesse uma paralela preocupação com ferramentas eficientes de organização e recuperação de informação.

Não seria exagero dizer que a imensa massa de informação em saúde em suporte papel foi e está sendo paulatinamente substituída, no Brasil, por uma grande massa de informação caótica em bases de dados e sistemas de informação, motivos pelos quais as ferramentas de indexação automática com aplicabilidade em língua portuguesa são e serão tão valiosas neste país, que se constitui como um grande produtor de informações em saúde.

Outro dado importante que reitera a importância deste estudo é o fato de serem ausentes em língua portuguesa terminologias em saúde. O Brasil, ao longo de sua história, não se preocupou em estabelecer um planejamento lingüístico e terminológico nacional expressivo. Dessa forma, há de um lado uma grande carência de terminologias, vocabulários e tesauros especializados em língua portuguesa, e de outro uma demanda por estudos aprofundados indexação automática, com aplicabilidade nesta língua.
A história da indexação automática possui um grande déficit no que tange aos estudos envolvendo documentos em língua portuguesa. Tal situação, precisa ser revista por meio de esforços e pesquisas de cooperação internacional, sobretudo, entre os países lusófonos.

Abstract

This work presents a statistical method of automatic indexing based on several researches carried out in the area. This method was used in the development of an automatic indexing system and applied on reports of radiological examinations. Some tests using a corpus with approximately five thousand documents show promising results, mainly in the field of health, in which the providing of quick and accurate information has a critical role to improving the quality of life of country's population

Referências

ALVARENGA L. Organização da informação nas bibliotecas digitais. In: **Organização da informação: princípios e tendências. Brasília**, DF: Briquet de Lemos, 2006. pp.76-98.

ALEXANDRINI F. **Desenvolvimento de uma metodologia de interpretação, recuperação e codificação inteligente de laudos médicos independente de idioma**. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, 2005.

ANDERSON, J. D.; PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: machine indexing, and the allocation of human versus machine effort. **Information Processing and Management**, v.37, pp.255–277, 2001.

BOYCE, Bert; LOCKARD, Marta. Automatic and manual indexing performance in a small file of medical literatura. **Bulletin of Medicai Library Association**, 63 (4):378-85, Oct. 1975.

BRUANDET, Marie-France. **Construction Automatique d'une Base de Connaissances du Domaine dans un Systeme de Recherche d'Informations**. Document fourni pour la soutenance du Diplôme d'Habilitation à Diriger des Recherches de L'Université Joseph Fourier de Grenoble. Março, 1989.

CARROLL, John M.; ROELOFFS, Robert. Computer selection of keywords using word-frequency analysis. **American Documentation**, 20 (3):227-33, July 1969.

CASTILLA A.C. Instrumento e investigação clínico-epidemiológica em cardiologia fundamentado no Processamento de Linguagem Natural. São Paulo. Tese (Doutorado em Cardiologia) - Faculdade de Medicina da Universidade de São Paulo, 2007

LUHN, H.P. A statistical approach to mechanized encoding and searching of literary information. **IBM Journal of Research and Development**, v.1, n.4, pp. 309-317, 1957.

NORUZI A. Folksonomies: (un)controlled vocabulary? **Knowledge Organization**. 2007: 33, (4): p.199-203, 2007.

PEREIRA, M.G. **Epidemiologia: Teoria e Pratica**. 7 ed. Rio de Janeiro: Editora Guanabara Koogan, 2003.

REINGOLD, M.E. NIEVERGELT, J. DEO, N. **Combinatorial Algorithms – Theory and Practice**. Englewood Cliffs: Prentice Hall, 1977.

ROBREDO, Jaime A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. O., ed. **Estudos Avançados em Biblioteconomia e Ciência da Informação**. Brasília, ABDF, 1982. v. 1, p. 236-74.

SALTON, G. A comparison between manual and automatic indexing systems. **Computing Reviews**, 10 ,(6):274, June. 1969.

SALTON, G. A new comparison between conventional indexing and automatic text processing. **Journal of the American Society for Information Science**, 23 (2):75-84, Mar./Apr. 1972.

SPÄRCK-JONES, K. A statistical interpretation of term specificity and its application in retrieval, **Journal of Documentation**, v.28, pp. 11-21, 1972.

SVENONIUS E. Information organization. In: **The intellectual foundation of information organization**. Cambridge: MIT Press, 2000. pp. 1-14.