JHI JOURNAL OF HEALTH INFORMATICS

InDeCS: Método automatizado de classificação de páginas Web de Saúde usando mineração de texto e Descritores em Ciências da Saúde (DeCS)

InDeCS: Automated method for classification of health Web pages using text mining and Health Sciences Descriptors (DeCS)

Alex Esteves Jaccoud Falcão[1]
Felipe Mancini[1]
Thiago Martini da Costa[1]
Anderson Diniz Hummel[1]
Fabio Oliveira Teixeira[1]
Daniel Sigulem[2]
Ivan Torres Pisa[2]

Descritores: Internet; Saúde; Classificação; Inteligência artificial; Sistemas de Recuperação de informação

RESUMO

Introdução: A quantidade de páginas web tem crescido exponencialmente, potencialmente levando conhecimento a mais pessoas, mas com a desvantagem de dificultar a localização de informação relevante e confiável. **Objetivo**: Apresentar resultados sobre a construção de um método automatizado de classificação e indexação de páginas web de saúde. **Métodos**: Foram selecionados endereços de páginas web classificadas manualmente como "saúde" e "não-saúde". Em uma segunda etapa foi calculada a similaridade dos termos do conteúdo destas páginas web com os Descritores de Ciência em Saúde (DeCS). Utilizando os valores de similaridade foram desenvolvidos e ajustados parâmetros de classificadores automatizados. **Resultados:** Para os experimentos foram coletadas 1.132 páginas web, separadas nas bases "saúde", "não-saúde" e "Merck", gerando mais de 3 milhões de composições até 3-gramas. Experimento usando as bases "saúde" e "não-saúde" resultou acerto, sensibilidade, especificidade e área sob a curva ROC de, respectivamente, 85,10%; 0,81; 0,88 e 0,92. Experimento usando as bases "Merck" e "não-saúde" resultou, respectivamente, 97,44%; 0,92; 1,00 e 0,98. **Conclusão**: Os resultados preliminares da investigação sobre o uso de métricas da mineração de texto e vocabulários controlados para aperfeiçoar o resultado de buscadores web especificamente para a área da saúde se mostraram significativos.

Descriptors: Internet; Health; Classification; Artificial intelligence; Information systems

ABSTRACT

Introduction: The amount of webpages has growing strongly, potentially leading knowledge to more people, but with the disadvantage of hindering relevant and reliable information. **Objective**: To present results of an automated method to classify and indexing health webpages. **Methods:** It was selected and classified webpages manually as health (saúde) and non-health (não-saúde). On a second step it was calculated the similarity between the webpages terms and the Health Science Descriptors (DECS). Automated classifiers parameters were developed using these similarities values. **Results:** For this experiment were collected 1,132 webpages, separate in "saúde", "não-saúde" and "Merck" databases, generating more than 3 million of 3 grams compositions. The experiment using the "saúde" and "não-saúde" databases resulted hit, sensitivity, specificity and area under ROC curve, respectively, 85.10%, 0.81, 0.88 and 0.92. The other experiment using the "Merck" and "não-saúde" databases resulted respectively, 97.44%, 0.92, 1.00 and 0.98. **Conclusion:** The preliminary results of this text mining metric using controlled vocabularies to improve the result of web search engines specifically for health were significant.

Autor Correspondente:
Alex Esteves Jaccoud Falcão
e-mail: **falcao-pg@dis.epm.br**

[1] *Programa de Pós-graduação em Informática em Saúde, Universidade Federal de São Paulo - UNIFESP - São Paulo (SP), Brasil.*
[2] *Departamento de Informática em Saúde, Universidade Federal de São Paulo - UNIFESP - São Paulo (SP), Brasil.*

Artigo recebido: **10/10/2008**
Aprovado: **10/03/2009**

INTRODUÇÃO

A quantidade de páginas web tem crescido vertiginosamente. Atualmente estima-se que exista mais de 182 milhões de servidores web[1], o que representa bilhões de páginas web com conteúdos bastante diversificados. Se por um lado este universo de informação em expansão potencialmente leva conhecimento a mais pessoas, por outro apresenta desvantagens[2], em especial quanto à dificuldade do usuário avaliar se a informação encontrada é relevante e confiável.

As ferramentas de busca têm tido um papel primordial na recuperação de informações na web. O Google, por exemplo, que se tornou o maior e mais utilizado buscador nos EUA[3], apresenta-se como uma excelente ferramenta para encontrar informação. No entanto, mesmo com os melhores buscadores da atualidade, encontrar informação relevante em um domínio específico de conhecimento diante dessa grande quantidade de páginas web permanece uma tarefa árdua.

A área de saúde, especificamente, merece distinção. Além dos seus profissionais, toda a comunidade tem utilizado a web cada vez com maior freqüência para encontrar informação sobre saúde. De acordo com o Centro de Estudos Sobre as Tecnologias da Informação e da Comunicação[4] calcula-se que no ano de 2007 em torno de 32% das atividades de usuários da web no Brasil estavam relacionadas à procura de informação nesta área. Mas ainda há dificuldade em recuperar informação qualificada na área de saúde, como exemplificado por Keselman, Browne e Kaufman[5]. Por meio de um estudo foi exposto a 20 usuários leigos um caso clínico cujo diagnóstico (angina) não havia sido revelado. Foi solicitado aos usuários que identificassem um diagnóstico utilizando como apoio a ferramenta de busca Medline Plus (*http://www.nlm.nih.gov/medlineplus*). Os usuários identificaram erroneamente a doença como infarto agudo do miocárdio, sendo apontado pelos autores como causa desse equívoco o nível de conhecimento dos usuários na área de doenças cardíacas e o fato de que a ferramenta de busca não retornava informações relevantes.

Tang e Ng[6] mostram ainda que a estratégia de utilizar buscador web de propósito geral, como o Google, para recuperação de páginas da área de saúde com o propósito de auxiliar o entendimento de doenças e identificação de diagnóstico não é efetiva, devido à vasta quantidade de informação que é recuperada e à baixa relevância das páginas web para o contexto desejado. Corroborando com esses resultados, Abraham e Reddy[7] criticam a acurácia tanto de buscadores web de propósito geral quanto de buscadores web específicos na recuperação de páginas para a área da saúde. A falta de especificidade das páginas web retornadas, que incluem páginas comerciais com propaganda de produto, é a principal característica que incomoda aos usuários[8].

Assim, o crescimento do volume de informação disponível na web atrelado à demanda dos usuários de informação em saúde potencializa a necessidade de aprimorar os buscadores web populares com foco na precisão das páginas web de saúde retornadas. O objetivo deste artigo é apresentar os primeiros resultados sobre a construção de um método automatizado – aqui denominado InDeCS – para classificação de conteúdos provenientes de páginas web de saúde, inicialmente identificando-os como "saúde" ou "não-saúde", a partir do uso de técnicas de mineração de texto aliadas a uma medida de similaridade de termos aos Descritores em Ciência da Saúde (DeCS)[9].

MATERIAS E MÉTODOS

O presente estudo é parte de trabalho aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de São Paulo sub número 0851/08 e foi conduzido em três etapas conforme apresentado na Figura 1. Na primeira etapa foram selecionadas páginas web e seus conteúdos foram classificados por voluntários como "saúde" ou "não-saúde". Na segunda etapa foi calculada a similaridade dos termos do conteúdo das páginas web selecionadas com os Descritores em Ciência da Saúde (DeCS), do Centro Latino Americano e do Caribe de Informação em Ciências da Saúde (BIREME). O DeCS é um vocabulário estruturado trilingüe (português, espanhol e inglês) com foco na área da saúde e baseado em coleções de termos organizados para facilitar o acesso à informação. A BIREME utiliza o DeCS na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos e outros tipos de materiais[10].

Na etapa final, utilizando os conjuntos de dados de similaridade para cada página web classificada, foram desenvolvidos e ajustados parâmetros dos classificadores automatizados com a ferramenta gratuita, de código aberto, para mineração de dados da Universidade de Waikato chamada Weka[11].

Chamamos de InDeCS a união das 3 etapas descritas anteriormente, incluindo as características de indexação utilizando a similaridade ao DeCS e a classificação automatizada dos conteúdos web.

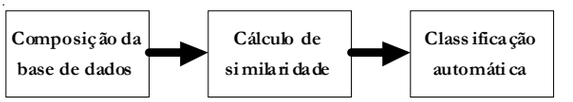

Figura 1 - Fluxo para cálculo InDeCS.

Composição da Base de Dados

Utilizando uma interface web desenvolvida em linguagem de programação PHP (*http://www.php.net*) para seleção dos endereços web, inicialmente foram selecionados por cinco voluntários páginas web escolhidas arbitrariamente de diversos temas. Durante esta seleção, o conteúdo e o corpo do texto de cada página web, já sem os marcadores HTMLs[12], foram armazenados em dois arquivos separadamente e o endereço web foi cadastrado em um banco de dados

MySQL (*http://www.mysql.com*). Após este cadastro, foi realizada a classificação das páginas web por 4 avaliadores como "saúde" ou "não-saúde". A página somente foi rotulada definitivamente como "saúde" se 3 ou 4 avaliadores a tivessem classificado como tal (75% de concordância). Analogamente, a página foi classificada como "não-saúde" somente se houve um mínimo de 75% de concordância.

Uma terceira base de dados composta por páginas web do Manual Merck de Informação Médica – Saúde para a Família[13] foi construída utilizando um robô desenvolvido em linguagem Perl (*http://www.perl.org*). Este robô armazenou o conteúdo de cada página web do Manual Merck em um arquivo e também armazenou em outro arquivo o texto sem os marcadores HTMLs. Esta base de dados foi classificada integralmente como "saúde" e denominada nos experimentos como "Merck".

Cálculo de Similaridade

A similaridade dos termos das páginas web armazenadas em nossa base de dados com os termos do DeCS é um dos elementos principais do motor do classificador entre "saúde" e "não-saúde" desenvolvido. Antes do cálculo de similaridade fez-se necessário um pré-processamento dos dados armazenados nas bases de dados construídas. Foi desenvolvido um algoritmo em Perl para separar o texto das páginas em sentenças e de cada sentença foram removidos termos que ocorrem freqüentemente como conjunções, preposições e artigos, denominados *stopwords*[14]. Os termos de cada sentença foram separados em grupos de um, dois ou três termos vizinhos adjacentes. A Tabela 1 mostra um exemplo da distribuição da sentença "otite é um termo médico utilizado para indicar uma infecção de ouvido" nestes agrupamentos.

Tabela 1 - Distribuição da sentença "otite é um termo médico utilizado para indicar uma infecção de ouvido" em agrupamentos com até 3 termos vizinhos adjacentes.

Um termo	Dois termos	Três termos
Otite	Otite termo	Otite termo médico
Termo	Termo médico	Termo médico utilizado
Médico	Médico utilizado	Médico utilizado indicar
Utilizado	Utilizado indicar	Utilizado indicar infecção
Indicar	Indicar infecção	Indicar infecção ouvido
Infecção	Infecção ouvido	
Ouvido		

Este processamento resultou em 3.596.746 composições de termos distribuídos nos agrupamentos para todas as páginas deste experimento. Então, cada composição de termos dos grupos foi submetida a um serviço web do BIREME[15] para uma consulta de similaridade ao DeCS. Esta similaridade é calculada com base no método Term Frequency–Inverse Document Frequency (TF-IDF) com trigramas, utilizado em recuperação de informações e mineração de textos, resultando um valor entre 0 e 1. Nota-se que similaridade 1 significa que existe algum termo cadastrado no DeCS que é exatamente igual ao termo submetido para avaliação. Analogamente, similaridade 0 significa que não existe termo com qualquer semelhança. Valores intermediários entre 0 e 1 indicam seu nível de similaridade com algum termo do DeCS. O resultado desta consulta é retornado em formato XML[16] e deste resultado é extraído somente o maior valor da similaridade.

Para cada página web foi construído um histograma de todas as similaridades dos termos contidos. Os valores de similaridade foram divididos em dez intervalos e para cada intervalo foi contabilizada a freqüência de termos. O Gráfico 1 mostra a representação de um histograma de uma página "saúde" escolhida arbitrariamente.

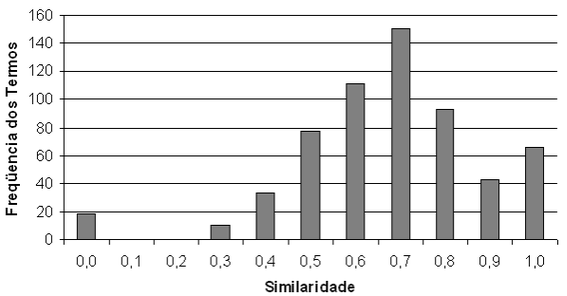

Gráfico 1 - Histograma da similaridade de termos ao vocabulário DeCS de uma página web em "saúde".

Classificadores Automáticos

Para realizar a classificação automática das páginas web dos grupos "saúde", "não-saúde" e "Merck" foram utilizadas técnicas de mineração de dados que realizam inferências baseadas na classificação manual e nos histogramas de similaridade. No entanto, é necessário um pré-processamento do histograma para que essa tarefa seja passível de ser executada.

Após a obtenção destes histogramas foi calculada a freqüência relativa de cada intervalo, que é a razão entre a freqüência de termos em cada intervalo e a soma das freqüências absolutas dos termos em todos os intervalos do histograma. Adicionaram-se ainda duas outras informações relativas ao histograma: a soma quadrática e a média ponderada das freqüências relativas dos intervalos. Estes dados foram armazenados no formato Attribute Relation File Format (ARFF), um arquivo padrão para entrada de dados no software Weka.

O cálculo da freqüência relativa dos intervalos do histograma foi realizado da seguinte maneira:

$$Nij = \frac{Dij}{\sum_{0}^{10} Dij}, \quad (1)$$

sendo D_{ij} a freqüência absoluta dos termos para o intervalo *j* da página *i*, e N_{ij} a freqüência relativa dos termos para o intervalo *j* da página *i*.

A soma quadrática dos intervalos e a média ponderada dos intervalos de uma página foram determinadas da seguinte maneira, respectivamente:

$$Si = \sum_{0}^{10} (Nij)^2 , \quad (2)$$

sendo S_i a soma quadrática dos intervalos dos histogramas de cada página *i*;

$$Mi = \sum_{0}^{10} (Nij * Vj) , \quad (3)$$

sendo M_i a média ponderada dos intervalos da página *i*, e V_j o valor médio de cada intervalo.

Foi utilizado o software Weka para treinar os classificadores automáticos. Essa ferramenta possui um amplo espectro de algoritmos de inteligência artificial utilizados em mineração de dados, como por exemplo, indução de árvores de decisão, vizinhos mais próximos (VMP) e redes neurais artificiais (RNA)[17].

Definiu-se como metodologia de treinamento e teste uma validação cruzada com 10 subgrupos (*10 fold cross validation*)[18]. Para avaliar e comparar os algoritmos quanto à acurácia na classificação utilizamos como métrica para seleção do algoritmo a porcentagem de acertos, a sensibilidade, a especificidade e a área sob a curva ROC[19]. O cálculo da especificidade foi realizado com o objetivo de determinarmos a proporção de páginas web com conteúdos "não-saúde" classificadas corretamente. Inversamente, a sensibilidade foi utilizada para determinarmos a proporção de páginas web com conteúdos de "saúde", ou "Merck", classificadas corretamente. A partir da especificidade e sensibilidade foi calculada a curva ROC, cuja área abaixo da curva foi utilizada para determinar o melhor classificador de padrões para uma determinada tarefa[20].

RESULTADOS

Para a composição da base de dados foi coletado um total de 1.132 páginas web. Destas, 608 (53,7%) páginas web são provenientes do Manual Merck, 256 (22,6%) páginas web foram selecionadas manualmente pelos voluntários e classificadas como "saúde", 268 (23,7%) páginas web foram também foram selecionadas manualmente pelos voluntários, porém classificadas como "não-saúde".

Nas 1.132 páginas web coletadas foram examinadas 740.644 composições para um termo, 1.296.780 composição de dois termos e 1.426.760 composição para três termos, totalizando 3.464.184 composições, para todas as páginas web analisadas, conforme dados da Tabela 2. Os valores de 0,1 e 0,2 não apresentam dados devido ao fato de que o mecanismo de similaridade ao DeCS utilizado neste experimento não disponibiliza tais valores.

Tabela 2 - Distribuição dos grupos de termos nas faixas do histograma da similaridade.

Similaridade	Quantidade		
	Um termo	Dois termos	Três termos
0,0	14.854	6.771	6.360
0,1	0	0	0
0,2	0	0	0
0,3	9.574	7.129	24.159
0,4	22.381	120.281	347.478
0,5	72.023	424.921	537.670
0,6	138.331	372.681	345.865
0,7	208.634	251.331	110.579
1,8	138.230	83.539	44.733
0,9	53.967	20.321	8.738
1,0	82.650	9.806	1.178
Total	740.644	1.296.780	1.426.760

A partir do armazenamento das páginas web o cálculo do histograma para uma página web consumiu em média 7 minutos de processamento no servidor utilizado. Para as 1.132 páginas web processadas em 3 lotes paralelos, o processamento totalizou em torno de 130 horas. Vale ressaltar que para o cálculo de similaridade ao DeCS foi realizada uma chamada a um serviço web externo, em outro servidor, o que tornou esta etapa mais demorada

Tabela 3 - Porcentagem de acerto (%), sensibilidade (sen), especificidade (esp) e área sob a curva ROC (ROC) calculados para o algoritmo de vizinhos mais próximos (VMP), redes neurais artificiais (RNA) e regressão logística (Logística) para o experimento usando base de dados "saúde" e "não-saúde".

	Um termo				Dois termos				Três termos			
	%	sen	esp	ROC	%	sen	esp	ROC	%	sen	esp	ROC
VMP	81,13	0,77	0,86	0,88	85,10	0,81	0,88	0,92	84,42	0,84	0,91	1,00
RNA	80,19	0,79	0,81	0,88	84,55	0,83	0,86	0,91	82,74	0,83	0,82	0,89
Logística	80,37	0,79	0,82	0,87	84,91	0,80	0,88	0,91	82,59	0,79	0,86	0,90

Tabela 4 - Porcentagem de acerto (%), sensibilidade (sen), especificadade (esp) e área sob a curva ROC (ROC) calculados para o algoritmo de vizinhos mais próximos (VMP), redes neurais artificiais (RNA) e regressão logística (Logística) para o experimento usando base de dados "Merck" e "não-saúde".

	Um termo				Dois termos				Três termos			
	%	sen	esp	ROC	%	sen	esp	ROC	%	sen	Esp	ROC
VMP	97,44	0,92	1,00	0,98	95,86	0,97	1,00	1,00	97,11	0,91	1,00	1,00
RNA	96,71	0,80	0,88	0,91	95,32	0,88	0,99	0,96	96,45	0,91	0,99	0,98
Logística	97,25	0,92	1,00	0,98	94,56	0,87	0,98	0,96	95,59	0,89	0,98	0,98

em termos de processamento.

Foi realizada uma análise exploratória com algoritmos da ferramenta Weka por meio de uma programação no padrão do software. A Tabela 3 e a Tabela 4 apresentam os valores de porcentagem de acerto, sensibilidade, especificidade e área sob a curva ROC para os algoritmos que apresentaram a melhor acurácia na classificação de páginas em "saúde" para um, dois e três termos analisados.

Considerando a Tabela 3, o algoritmo VMP usando dois termos para análise apresentou melhor acurácia na classificação de páginas web de saúde usando as bases "saúde" e "não-saúde", com valores para porcentagem de acerto, sensibilidade, especificidade e área sob a curva ROC de, respectivamente, 85,10%; 0,81; 0,88 e 0,92. Considerando a Tabela 4, o algoritmo VMP usando um termo para análise apresentou melhor acurácia na classificação de páginas web de saúde usando as bases "Merck" e "não-saúde", apresentando valores de porcentagem de acerto, sensibilidade, especificidade e área sob a curva ROC de, respectivamente, 97,44%; 0,92; 1,00 e 0,98.

DISCUSSÃO

Iniciado este estudo, os autores acreditavam que as páginas web com conteúdos em saúde poderiam ser identificadas simbolicamente a partir da similaridade com termos do DeCS. Foi desenhado um histograma de similaridade com estes termos para cada página web com o objetivo de encontrar estas distribuições visualmente distintas entre conteúdos da área da saúde e de outras áreas. Porém, no DeCS existem termos que podem pertencer a áreas que não são apenas da saúde como, por exemplo, "telefone celular", que pertence à subclasse "telefone" da categoria "ciências da informação".

Optou-se em utilizar técnicas de mineração de dados para analisar os histogramas, o que mostrou ser uma abordagem válida, tendo em vista a acurácia obtida pelos algoritmos usados. Foi observado que o algoritmo de vizinhos mais próximos (VMP) é mais eficiente, se comparado aos algoritmos de RNA e regressão logística, apresentando as melhores taxas de acerto, sensibilidade, especificidade e área sob curva ROC para praticamente todas as bases de dados. Provavelmente isto ocorreu devido à influência da heurística do algoritmo VMP – relacionar uma página web com as páginas mais próximas – ser mais relevante para a base de dados estudada, se comparado a RNA e regressão logística que, genericamente, traçam planos, hiperplanos ou detectam relações estatísticas não-lineares entre os atributos analisados.

Apesar de, em princípio, considerarmos o agrupamento de três termos (3-gramas) como mais representativo para o cálculo de similaridades[21], neste estudo o agrupamento de dois termos para o experimento "Merck x não-saúde" e um termo para o experimento "saúde x não-saúde" apresentaram melhor acurácia. Isto ocorreu devido ao fato de que uma grande quantidade dos termos pertencentes ao DeCS é de termos únicos, determinando assim maior grau de similaridade para um e dois termos.

Como resultado dos classificadores temos que o experimento "Merck e não-saude" apresentou melhor acurácia na classificação de conteúdos que o experimento "saúde x não-saúde". Na opinião dos autores, isso se deve pela característica do Manual Merck de Saúde ter um público alvo definido e uma única linha editorial, o que facilita a correlação entre seu conteúdo e o torna mais distinto do conteúdo "não-saúde", porém, para confirmar a veracidade deste raciocínio é necessário um estudo mais aprofundado.

É importante ressaltar que não se uniu as bases de dados "saúde" e "Merck" porque os resultados dos classificadores não refletiriam os objetivos deste estudo. O foco deste estudo foi determinar a acurácia na classificação de páginas web de saúde a partir de um experimento com páginas retornadas pelo Google e também uma análise com base em um manual disponível na web com informações de saúde voltadas para a família. A união das bases de dados está sendo foco das análises atuais dos autores.

A fim de determinar um conjunto de dados relevante pesquisadores[22] utilizam milhões de páginas web em estudos exploratórios para análise e classificação automática de conteúdos na web selecionados por robôs. Contrário a esta abordagem, nesse trabalho foi utilizada uma abordagem de percepção humana para a coleta da base de dados. Desta maneira, apesar dos conjuntos de páginas web serem ínfimos se comparados à quantidade de páginas disponíveis na web, o processo manual de coleta e classificação da base de dados foi a única estratégia que possibilitou determinar se uma página é de saúde ou não com precisão.

Os resultados apresentados são relevantes, porém, o desempenho do classificador pode ser aprimorado com a investigação, por exemplo, de outro conjunto de descritores da saúde, por meio do processamento dos histogramas em cachê e também pela otimização das rotinas de obtenção dos histogramas. Em especial, a obtenção de outra base de páginas web representativas com maior número de páginas pode melhorar o desempenho do classificador. Uma das estratégias é utilizar portais web populares sobre saúde, como o próprio Manual Merck. Além disto, é possível que o uso de outros descritores ou outras métricas de obtenção do histograma permita obter uma descrição matemática mais adequada das páginas em relação ao seu contexto. Os autores têm realizado novos experimentos com uma lista de palavras com 10 mil termos, populares, relacionados à saúde, e também pelo mapeamento dos termos do DeCS com conceitos da Unified Medical Language System (UMLS) (*http://umlsinfo.nlm.nih.gov*).

Cabe ressaltar que embora essa metodologia tenha sido aplicada em páginas web que estavam em língua

portuguesa brasileira, a abordagem proposta pode ser aplicada em outros idiomas, utilizando-se outros descritores de saúde independente do idioma.

CONCLUSÃO

O método apresentado mostrou-se significativo para classificar conteúdos de páginas web entre "saúde" e "não-saúde". Os resultados da porcentagem de acerto, sensibilidade, especificidade e área sob a curva ROC são, respectivamente, 85,10%; 0,81; 0,88 e 0,92, ao classificar páginas web utilizando conteúdos selecionados manualmente. Ao utilizar uma base específica como o Manual Merck, os valores obtidos para porcentagem de acerto, sensibilidade, especificidade e área sob a curva ROC foram, respectivamente, 97,44%; 0,92; 1,00 e 0,98. Estes são os resultados preliminares da investigação sobre o uso de métricas da cientometria e correlações entre termos de vocabulários controlados para potencialmente aperfeiçoar o resultado de buscadores web especificamente para a área da saúde.

REFERÊNCIAS

1. Netcraft [homepage on the Internet]. Web Server Survey; September 2008 [updated 2008 Sep 30; cited 2008 Oct 10]. Available from: http://news.netcraft.com/archives/2008/09/30/september_2008_web_server_survey. html.
2. Fogg BJ, Soohoo C, Danielson DR, Marable L, Stanford J, Tauber ER. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In: DUX '03: Proceedings of the 2003 conference on Designing for user experiences. New York, NY, USA: ACM Press; 2003. p. 1-15.
3. Search Engine Watch [homepage on the Internet]. Burns E. c2008-09 [updated 2008 Sep 02; cited 2008 Oct 10]. Available from: http://searchenginewatch.com / showPage.html?page=3630718
4. Pesquisa sobre o uso das Tecnologias da Informação e da Comunicação no Brasil: TIC Domicílios e TIC Empresas 2007 [homepage on the Internet]. Comitê Gestor da Internet no Brasil. c2008-03 [atualizada 2008 Mar 14; citado 10 Out 2008]. Disponível em: http://www.cetic.br/usuarios/tic/2007/index.htm.
5. Keselman A, Browne A, Kaufman D. Consumer health information seeking as hypothesis testing. J Am Med Inform Assoc. 2008;15(4): 484-95.
6. Tang H, Ng JH. Googling for a diagnosis-use of Google as a diagnostic aid: internet based study. BMJ. 2006; 333(7579):1143-5.
7. Abraham J, Reddy M. Quality of healthcare websites: a comparison of a general-purpose vs. Domain-Specific Search Engine. In AMIA Symposium Proceedings; 2007 Oct 11; Chicago, ILL. p. 858.
8. Toms E, Latter C. How consumers search for health information. Health Informatics J. 2007;13(3): 223-35.
9. BIREME. [homepage on the Internet]. DeCS - Descritores em Ciências da Saúde. c1999-03 [atualizada 2008 Fev; citado 2008 Out 10]. Disponível em: http://decs.bvs.br.
10. Pellizzon RF. Pesquisa na área da saúde: Base de dados DeCS (Descritores em Ciências da Saúde). Acta Cir. Bras. 2004;19(2): 153-63.
11. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann; 1999.
12. W3C [homepage on the Internet]. HTML Tutorials. c1999 [cited 2008 Oct 10]. Available from: http://www.w3schools.com/html/default.asp.
13. Berkow R, Beers M, Bogin R, Fletcher A. Manual Merck de informação médica: saúde para a família [Internet]. Disponível em: http://www.msd-brazil.com/msdbrazil/patients/manual_Merck/prefacio.html
14. Hersh W. Information retrieval: A health and biomedical perspective (Health Informatics). 3rd ed. Springer; 2008.
15. Tardelli AO, Anção MS, Packer AL, Sigulem D. An implementation of the trigram phrase matching method for text similarity problems. Stud Health Technol Inform 2004;103:43-9.
16. W3C [homepage on the Internet]. XML Tutorial. c1999 [cited 2008 Oct 10]. Available from: http://www.w3schools.com/xml/default.asp.
17. Duda RO, Hart PE, Stork DG. Pattern classification (2nd Edition). Wiley-Interscience; 2000.
18. Burnham KP, Anderson D. Model selection and multi-model inference. Springer; 2002.
19. Massad E, Menezes RX, Silveira PSP, Ortega NRS. Métodos quantitativos em medicina. São Paulo: Manole Ltda; 2004.
20. Metz C. Basic principles of ROC analysis. Semin Nucl Med.1978;8(4):283-98.
21. Adams ES, Meltzer AC. Trigrams as index element in full text retrieval: observations and experimental results. In: ACM Conference on Computer Science; 1993 Feb 16-18; Indianapolis. Proceedings. New York: ACM; 1993 p.433-9.
22. Chakrabarti S, Puniyani K, Das S. Optimizing scoring functions and indexes for proximity search in typeannotated corpora. In: International World Wide Web Conference; 2006 May 23-26; Edinburgh, Scotland. Proceedings. New York: ACM; 1993 p.717-26.