BRAZILIAN HEALTH-RELATED CONTENT WEB SEARCH PORTAL

Presentation on a Method for its Development and Preliminary Results

Felipe Mancini, Alex Esteves Jaccoud Falcão, Anderson Diniz Hummel, Thiago Martini Costa
Cristina Lucia Feijo Ortolani, Fabio Teixeira and Ivan Torres Pisa
Department of Health Informatics, Federal University of São Paulo, Rua Botucatu 862, São Paulo, SP, Brazil

Keywords: Internet, health, Information storage and retrieval, Pattern recognition system.

Abstract: The increase in the amount of available information on the world wide web is inexorable, which, on one hand, provides the web user with more information. On the other hand, however, web searches become increasingly more difficult to handle due to the increasing number of retrieved documents. The present study is a proposal of development for a Brazilian search portal specific for health-related content. The aim of such development is to provide web users, mainly the non-specialist ones, with the largest number possible of web pages relevant to their search terms and inferred search intentions. The proposed search portal integrates web mining-based filters and a decision-making support tool. The preliminary study results show that among the algorithms tested to incorporate a filter module specific for health-related content - artificial neural networks, logistic regression and nearest neighbor clustering (NNC) -, the application of NNC resulted in the automated web health-related content classifier with the best performance for sensitivity and specificity 0.92 and 1.00 respectively.

1 INTRODUCTION

The world wide web content increases inexorably each day pushing and being pushed by the technological and economical development. Health-related activities are one contributor to this increase. In Brazil, 30% of web searches concern to information on health conditions related topics (CETIC, 2007).

However, adequacy and quality of available information are two relevant issues regarding health-related web content. A number of criteria for assessing adequacy and quality have been devised since 1997 (Lopes, 2004) such as the Health On Net Foundation code (HON, 2008) and the criteria for quality assessment by Health Information Technology Institute (HITI, 2000).

These accrediting institutions have as aim ensuring the suitability of heath information made public on the web. However, the design of major web search engines does not prioritize certified websites. Chang et al. (2007) reported that in a search concerning the health field, the first HON certified reference was the 763th on the list of sites retrieved by Google™.

This particular feature of major search engines makes the system of quality certification for health information on web unavailable for internet users. A Brazilian study showed the usefulness and potential appreciation of search engines incorporating a feature that skims high quality health information - 83% of accesses to health-related sites didn't use popular web search engine (Silva, 2006).

There is an implied credibility in health sites. The Harris Poll (2008) reported that 60% of American web users believe entirely in the content of the available health pages. This figure rises to 90% among Brazilian users (Silva, 2006). This makes it even more important to provide Brazilians accessing the worldwide web with assuredly accurate health web pages.

This paper shows a method that aims at implementing a Brazilian web portal for search and retrieval of health focusing on the needs and interests of the general public.

Mancini F., Esteves Jaccoud Falcão A., Diniz Hummel A., Martini Costa T., Lucia Feijo Ortolani C., Teixeira F. and Torres Pisa I. (2009).
BRAZILIAN HEALTH-RELATED CONTENT WEB SEARCH PORTAL - Presentation on a Method for its Development and Preliminary Results.
In *Proceedings of the International Conference on Health Informatics*, pages 306-310
DOI: 10.5220/0001552303060310

2 METHODS

The architecture of this search portal, called In Health Search Model (InHealth), has a services-oriented architecture (SOA) (Erl, 2007), therefore, all modules to be incorporated to it should follow interoperation standards based on web-services (Papazoglou, 2007).

InHealth comprises fundamentally 3 modules – indexing module (IndHealth), inference module (InfHealth) and a graphical user interface module (IntHealth). The interaction of these 3 modules within InHealth architecture is outlined in Figure 1.

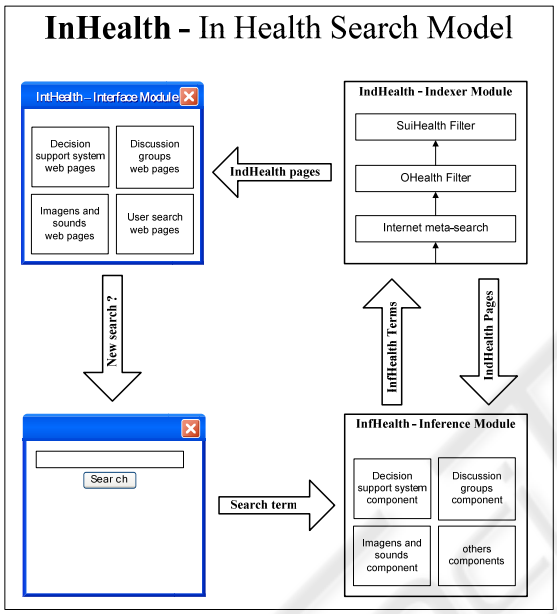


Figure 1: In Health Search Model modules interaction.

As indicated in Figure 1, after a search term is entered, the web service Inf Health performs a pre-processing and selects related search terms supplied to IndHealth, which retrieves sites relevant sites supplied to IntHealth. IndHealth, InfHealth and IntHealth are described below.

2.1 IndHealth

The portal is provided with a meta-search engine (Dunford II, 2008). incorporated in IndHealth module which works with Google and Yahoo© search engines.

IndHealth performs an index-based retrieval using as indexing system the web content mining (Kosala & Blockeel, 2000), which is a process to find target information in documents subject to a search. A number of techniques, such as information retrieval (Hersh, 2003) and artificial neural networks, can be used to implement this process (Haykin, 1999).

The indexing module uses 2 filters: only-health filter (OHealth) and suitability health filter (SuiHealth) have been developed in Perl (Perl, 2008), PHP (PHP, 2008), Java (Java, 2008) and MySQL (MySQL, 2008) database.

2.1.1 OHealth Filter

OHealth filter was incorporated into IndHealth aiming at retrieving only web pages with health information. Excluding commercial sites and product advertisements, would save users searching health issues the unwanted task of doing it themselves (Tom & Latter, 2007).

Basically, this filter analyzes the similarity between the web page content and the content of a dictionary or descriptor for the health field, such as DeCS (DeCS, 2008) or MeSH (MeSH, 2008). The OHealth filter works with the assumption that the greater the similarity between the web page under analysis and health content descriptors, the higher the probability of it being a web page focusing on health issues.

2.1.2 SuiHealth Filter

Tang and Ng (Tang and Ng, 2006) reported that using general-purpose searching engines, such as Google, may render ineffective when the internet search aims at reaching a diagnosis or understanding better a health condition due to the web user coming across with a large amount of irrelevant retrieved information. Abraham and Reddy (Abraham and Reddy, 2007) also showed some criticism regarding the adequacy of both general-purpose and specialized search engines for web user retrieving information in the health field. Therefore, SuiHealth filter focus on retrieving the web content which is the most relevant and suitability possible to the user's search term.

The development of this filter ensues an investigation carried out by Falcão (http://telemedicina6.unifesp.br/healthrank) using the social media (O'Reilly, 2005) to determine sets of terms for automatic ranking and indexing relevant health web content.

2.2 InfHealth

Using a search term, this module makes inferences concerning the health content that the user would like to be retrieving.

Decision-making support system (SAD) can be used for this goal (Musen et al., 2006). SAD reads the user's search term as a health condition sign or symptom and retrieves from and internal database all possible health conditions including such sign of symptom, which are fed into IndHealth as search terms.

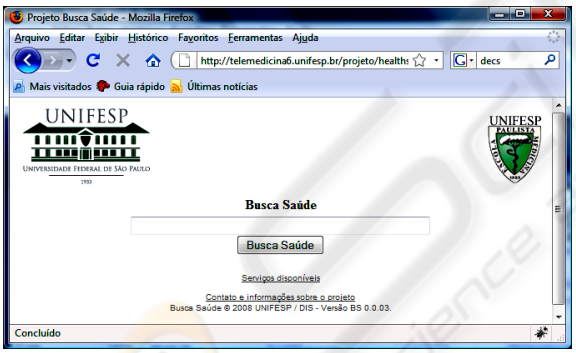
Lepidus (Silva & Roque, 2000) is an SAD focusing on general medical practice that has shown to be highly effective in providing its user with the correct diagnosis in 84% of the cases where signs and symptoms of a health condition is presented. Its incorporation to IntHealth may result in improved performance.

Other inference component can also integrate IntHealth, such as those working with images, audio-visual features, forum groups, and scientific reports, particularly available on Virtual Health Library (Bireme, 2008).

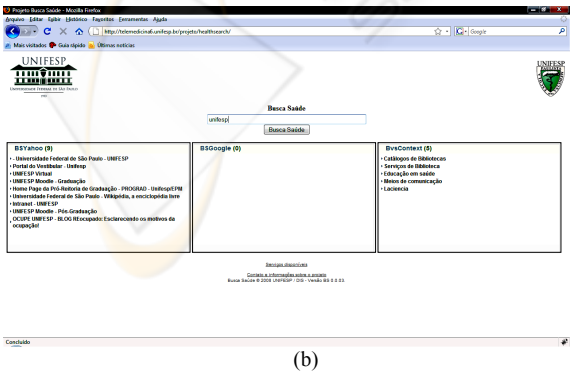
This component is development in PHP and Java languages in conjunction with MySQL database.

2.3 IntHealth

IntHealth, the GUI module of the portal system, has been developed using Ajax (Ajax, 2008) and PHP languages, and the Mini Ajax framework (Mini Ajax, 2008). Its GUI should follow the heuristics concepts of Nielsen (2005).



(a)



(b)

Figure 2: GUI proposed for InHealth portal for health-related content web search.

Figure 2, also available at http://telemedicina6.unifesp.br/healthsearch, show the GUI of the InHealth web pages. Figure 2 (a) shows the home page exhibiting just one box to enter the search term. Figure 2 (b) shows the page with the outcome of InHealth processing in a three-box layout proposed by Mini Ajax. Each InfHealth inference is processed on an asynchronous, independent thread due this framework.

3 PRELIMINARY RESULTS

The present study focus on the preliminary results of OHealth filter. This assessment involved 3 steps as indicated in Figure 3. Firstly, 608 web pages from the Merck Manual for Medical Information on Health for the Family (Berkow et al., 2003) and 268 non-health web pages were selected.

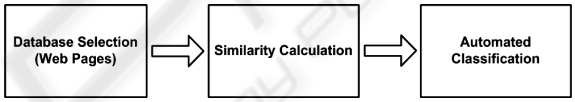


Figure 3: Automated content classifier flow chart.

Next, the Term Frequency–Inverse Document Frequency (TF-IDF) statistical model (Tardelli et al., 2004) and stemming (Hersh, 2003) were used to determine the similarity between the web pages selected with the DeCS. The stopwords (Hersh, 2003) was unappreciated for the set of words analyzed.

Finally, using similarity data sets for each web page analyzed, classifier was development with the use of free tools and open code for data mining, Weka, developed by the University of Waikato (Witten & Frank 2005). The goal of this classifier was automatically recognition health-related content.

ANN, logistic regression (LR) (Bishop, 2007), and nearest-neighbor clustering (NNC) (Duda, et al., 2000) were used as aid algorithms for similarity measurement, since they showed the best rates of specificity and sensitivity as determined by the area under the ROC curve (AUC) (Metz, 1978) when applying the 10-fold cross validation test (Burnham & Anderson, 2004). Others algorithms were tested as bayes net, linear regression, and normalized Gaussian radial basis function network (Bishop, 2007).

Specificity and sensitivity rates and AUC for ANN, LR and NNC are presented in Table 1.

Table 1: Sensitivity (SE) and specificity (SP) rates, and area under ROC curve (ROC) for nearest-neighbor clustering (NNC), artificial neural networks (ANN), and logistic regression (LR) classification algorithms, tested in OHealth filter.

	SE	SP	ROC
NNC	0.92	1.00	0.98
ANN	0.80	0.88	0.91
LR	0.92	1.00	0.98

At the current stage of the study, NNC showed the best performance as aid algorithm for automated classification of web health content with the best rates of specificity and sensitivity, and the largest area under the ROC curve – 0.92, 1.00, 0.98 respectively.

4 DISCUSSION

In the present study, nearest-neighbor clustering (NNC) was able to recognize all Merck health web pages – specificity 1 – and was able to filter 92% of web pages with content not related to health issues – sensitivity 0.92.

Despite OHealth showed a satisfactory performance in classifying health web pages, its represented a single editor (Merck). Further tests should be run including health web pages from different sources and different editorial characteristics.

Another point requiring further investigation is classification time performance. With the present design, OHealth took approximately 7 minutes to carry out the classifying calculations for each web page. Other approaches to assess similarity have been proposed such as white lists and memory hash calculation.

Other software components integrating the InHealth portal system are still in stages of design and test, as the Lepidus adaptation and the GUI improvement.

The authors also plan to implement a questionnaire-based assessment tool to be applied on the general and specialist web users in order to compare the performance of InHeath with other general-purpose search engines such as Google.

5 CONCLUSIONS

The usefulness of a search portal for web pages with health-related content is potentially enormous, and the challenge of its implementation is motivating.

The preliminary results of the present study show that web mining techniques can improve the specificity of search for health information on the world wide web.

REFERENCES

Abraham, J., & Reddy, M. (2007). Quality of Healthcare Websites: A Comparison of a General-Purpose vs. Domain-Specific Search Engine. *AMIA Symposium Proceedings*, (p. 858).

Ajax. (01 of 01 of 2008). *Ajax*. Access in 11 of 07 of 2008, available in Ajax: http://www.w3schools.com/Ajax/Default.Asp

Berkow, R., Beers, M., Bogin, R., & Fletcher, A. (01 of 01 of 2003). *Manual Merck de Informação Médica: Saúde para a Família*. Access in 08 of 07 of 2008, available in Merck: http://www.msd-brazil.com/msdbrazil/patients/manual_Merck/prefacio.html

Bireme. (01 of 01 of 2008). *VHL*. Access in 11 of 07 of 2008, available in VHL: http://www.bireme.br/php/index.php?lang=en

Bishop, C. (2007). *Pattern Recognition and Machine Learning.* Springer: New Jersey.

Burnham, K., & Anderson, D. (2004). *Model Selection and Multi-Model Inference.* Berlim: Springer.

CETIC. (01 of 11 of 2007). *TIC Domicílios e usuários 2007*. Access in 07 of 07 of 2008, available in CETIC: http://www.cetic.br/usuarios/tic/2007/rel-int-10.htm

Chang, P., Hou, I., Hsu, C., & HF, L. (2006). Are Google or Yahoo a good portal for getting quality healthcare web information. *AMIA Annu Symp Proc*, (p. 878).

DeCS. (01 of 01 of 2008). *DeCS - Health Sciences Descriptors*. Access in 07 of 07 of 2008, available in http://decs.bvs.br/I/homepagei.htm

Duda, R., Hart, P., & Stork, D. (2000). *Pattern Classification.* New York: Wiley-Interscience.

Dunford II, T. (2008). *Advanced Search Engine Optimization: A Logical Approach.* Maui: American Creations of Maui.

Erl, T. (2007). *SOA Principles of Service Design.* Prentice Hall: New York.

Falcão AEJ. *HealthRank: Construção e Avaliação de um Software para Medir Adequação à Códigos de Ética e Relevância de Websites em Saúde Utilizando Métodos de Mídia Social e Indicadores Automatizados*. Master Thesys –Federal University of São Paulo, 2008.

Haykin, S. (1999). *Neural Networks: a Comprehensive Foundation.* New Jersey: Prentice-Hall.

Hersh, W. (2003). *Information Retrieval : a Health and Biomedical Perspective.* New York: Springer.

HITI. (12 of 06 of 2000;). *HITI*. Access in 02 of 22 of 2008, available in HITI: http://hitiweb.mitretek.org/docs/policy.html

HON. (07 of 01 of 2008). *HON*. Access in 10 of 07 of 2008, available in HON: http://www.hon.ch/

Java. (11 of 07 of 2008). *Java*. Access in 11 of 07 of 2008, available in Java: http://java.sun.com/

Kosala, R., & Blockeel, H. (2000). Web Mining Research: a Survey. *ACM SIGKDD Exploration* , pp. 1-15.

Lopes, I. (2004). New paradigms for evaluation of the information quality health retrieved on the web. *Ciência da Informação* , pp. 81-90.

MeSH. (04 of 01 of 2008). *MeSH*. Access in 14 of 07 of 2008, available in MeSH: http://www.nlm.nih.gov/mesh/

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nucl Med* , pp. 283-298.

MiniAjax. (01 of 01 of 2008). *MiniAjax*. Access in 11 of 07 of 2008, available in MiniAjax: http://miniajax.com/

Musen, M., Shahar, Y., & Shortliffe, E. (2006). *Clinical Decision-Support Systems.* New York: Springer-Verlag.

MySQL. (01 of 01 of 2008). *MySQL*. Access in 09 of 07 of 2008, available in MySQL: http://www.mysql.com/

Nilsen, J. (01 of 01 of 2005). *Ten Usability Heuristics*. Access in 11 of 07 of 2008, available in Ten Usability Heuristics: http://www.useit.com/papers/heuristic/heuristic_list.html

O'Reilly, T. (30 of 09 of 2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. Access in 10 of 07 of 2008, available in http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/ what-is-web-20.html

Papazoglou, M. (2007). *Web Services: Principles and Technology*. Prentice Hall: New York.

Perl. (01 of 01 of 2008). *Perl*. Access in 08 of 07 of 2008, available in Perl: http://www.perl.org/about.html

PHP. (2008 of 01 of 01). *PHP*. Access in 2008 of 07 of 17, available in PHP: www.php.net

Poll, T. H. (31 of 07 of 2008). *Harris Poll Shows Number of "Cyberchondriacs" – Adults Who Have Ever Gone Online for Health Information– Increases to an Estimated 160 Million Nationwide*. Access in 11 of 07 of 2008, available in The Harris Poll: http://www.harrisinteractive.com/harris_poll/index.asp?PID=792.

Silva, R., & Roque, A. (2000). Clinical medical diagnosis using a signal-processing approach. *Conference on Mathematics and Engineering Techniques in Medicine and Biological* (pp. 13-18). Las Vegas: CSREA Press.

Silva, WM. *Navegar é preciso: Avaliação of impactos do uso da internet na relação médico-paciente*. Master Thesis – University of São Paulo, 2006.

Tang, H., & NG, J. (10 of 11 of 2006). Googling for a diagnosis—use of Google as a diagnostic aid: internet based study. *British Medical Journal* , pp. 1143-1145.

Tardelli, A., Anção, M., Packer, A., & Sigulem, D. (2004). An implementation of the trigram phrase matching method for text similarity problems. *Stud Health Technol Inform* , pp. 43-49.

Toms, E., & Latter, C. (2007). How consumers search for health information. *Health Informatics Journal* , pp. 213-223.

Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* San Francisco: Morgan Kaufmann.