Models for automatic retrieval of health information on the Web

Ana Marilza Pernas[1.2]
[1]Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, RS, Brasil
ana.pernas@inf.ufrgs.br

Jonas Bulegon Gassen[2]
[2]Instituto de Informática Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil
jbgassen@inf.ufrgs.br

José Palazzo M. de Oliveira[2]
[2]Instituto de Informática Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil
palazzo@inf.ufrgs.br

ABSTRACT
Evaluate the data available in Web pages is necessary to allow the suggestion of adequate content to a specific public, for such, an idea is to develop an automatic detection mechanism of the quality present in Web pages content. Due to its quickly and not standardized growth, the content freely available in the Web has reached very large proportions, becoming extremely difficult to automatic manage this large mass of data. One of the main reason for this complexity is the not appliance of standards for its construction, which should lead to a standardized access. This article describes the development of models and techniques to locate, standardize and extract the content in web pages associated with health issues. After that, the objective is to provide an appropriate content to evaluate the quality of a web page according to specific metrics.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**: Information filtering*;* Retrieval models*;* Selection process.

J.3 **[Life and Medical Sciences]**: Health; Medical information systems.

General Terms
Management, Design, Experimentation, Standardization, Verification.

Keywords
Location and extraction models, standardization, quality metrics.

1. INTRODUCTION
One of the biggest challenges in Web systems is to automatically deal with the large amount of data available in the Web as a large distributed system. As a consequence of its rapid not standardized growing, the Web reached a huge proportion that became extremely difficult to be efficiently controlled, being the automatic management of data very complex.

The automatic web-data management is especially desired when the subject is the quality evaluation of a given web page for its subsequent recommendation to a specific public. In the case of health information content it is necessary a careful check on the validity of the information before an effective recommendation. We know that check all health information on the Web to determine if is reliable it's very difficult, but we can evaluate a number of indicators present in a web page to try to guarantee a minimum of quality. Examples of indicators, or metrics, could be: sponsors, if exists and who are; the subject of the page; the ways of contact with the responsible by the web page; the web page readability; if the content is continuously updated (freshness).

To achieve this objective there are organizations that grant certificates of quality in form of quality-stamps to ensure the quality of the information presented on a website. That is the case of Web Mèdica Acreditada[1] (WMA), an organization of the Medical Association of Barcelona that offers "a quality programme addressed to medical websites. Through a voluntarily certification process, websites that follow that programme meet a set of quality criteria, making possible a trustworthy virtual community on the Internet for general public, patients and health professionals" [1]. Other examples of organizations which grant quality seals to websites are the Internet Content Rating Association (ICRA) [2] and the Internet Quality Agency (IQUA) [3].

Even with this kind of organization to ensure the quality of health information on the Web by a formal request, the best alternative would be to orchestrate this practice performing an automated analysis of the content presented in health web pages. However, given the size of the Web nowadays this task would be really complex and hard, as an example the indexing of the Web content by the Yahoo! indicates that "the production of the Web Map, an index of the World Wide Web that is a critical component of search {takes} (75 hours elapsed time, 500 terabytes of MapReduce intermediate data, 300 terabytes total output)" using larger clusters of 3500 nodes [19]. In a post published in July 2008 in the Google Blog the Web size was evaluated as already exceeding 1 trillion of URLs (Uniform Resource Locator) [4].

To enable automatic retrieval of data from web pages about health, some models and techniques are presented here to support: (i) location of health web-data (ii) standardization and (iii) (ii) extraction of this data, based on pre-established criteria. Conscious of the complexity of dealing to the entire Web, the location, standardization and extraction models presented here are applied to specific search engines.

The remainder of this article is structured as follows. Section 2 explores some related works. Section 3 describes the general vision of this work, showing our starting point and the relevant data in estimating the quality of a web page. Section 4 is associated with the task of automatic location of web page about health, explaining its general model. Section 5 present the task of automatic extraction data from the localized web pages pre-

[1] http://wma.comb.es/

senting problems related to the lack of standardization. Section 6 shows a case study developed on pages related with Alzheimer's disease. Finally section 7 presents the conclusions and future works.

2. RELATED WORKS

An example of project related to the analysis of quality content on health web pages is the QUATRO Project (The Quality Assurance and Content Description Project) and its successor, QUATRO+ [5]. The objective of this project is to offer a common machine readable vocabulary to certify the quality of health information on the Web. One of the results of the QUATRO project consists on a vocabulary that presents a list of descriptors and their definitions to be used as a base for creation of quality seals. Among the participants of the QUATRO project is the aforementioned Web Mèdica Acreditada – WMA.

An old work, but that can still be referenced by its vision is the so called "*Oh, yeah?*" button, proposed by Tim Berners-Lee in [6]. In this proposition, the author mention how important is the Web say to the users something about the information being presented. This "*Oh, yeah?*" button would be responsible when the user is not so confident in a web page content to show a number of reasons to trust in that information. In a practical view, the button would access some meta information about the content and show that to the user.

Related to information extraction, the Project AQUA (Assisting Quality Assessment) [7] proposes to automate parts of the work manually done by organizations that offers quality labels to websites, making this process easily and, consequently, increasing the number of sites with quality seals. The objective is to crawl the Web to locate unlabelled health web resources, suggest machine readable labels for them, according to predefined labeling criteria, and monitor them. To do that, the crawling mechanism uses Google[2] and Yahoo![3] search engines to do a meta-search engine on the Web, collecting and filtering (to ignore sub-paths of URLs already in list and removes URLs having already a content label) the resulting URLs. This work is very similar to our work, but data related with authorship information and inLinks and outLinks are not mentioned.

Another related work is described in [8] which aim to extract a number of quality indicators defined by organizations like *HONCode* to establish the quality content of a health website. The work looks to detect measurable indicators in the evaluated website, searching this information in HTML tags and meta-tags. However, in our view, only this simple analysis could not cover a great set of websites.

3. GENERAL MODEL

The general model presented in this work intends to cover critical points considered to achieve the final goal: evaluate the quality of a Web page about health. In a general view it is necessary to accurately automate the following steps:

- **Localization** – answer to the question: given an item X stored in some dynamic set of nodes in the system, how to find it? [9].
- **Standardization** – relates to existing standards in presentation of data in web pages which may indicate ways to achieve automatic access (automatic extraction) of this data.
- **Extraction** – once the web page is located, how to obtain and properly manage its data to be evaluated?
- **Quality** – is the analysis of collected data where metrics are applied to define the level of quality present in the evaluated web page, according to a specific user profile.

As we mentioned in section 1, in this work only the phases of Location**,** Standardization and Extraction are treated. The starting point in developing models was the research described in [10]. In [10], the main points to be collected in a Web page were analyzed, in order to determine its quality. This analysis resulted in an ontology of quality, which was reduced to the model depicted in the Figure 1 [11]. This reduction was made to simplify the task of automatic retrieval because obtaining this main data about a health web page is already possible to define a first quality estimative.

As we can see in Figure 1, information about the web page is obviously necessary to define its quality. Information about author is very important because if the content was written or revised by a specialist in the subject, possibly the quality will be enhanced. Is either important define if the web page is sponsored by some organization and if is a recognized organization, for example, a governmental organization, university or industry.

The E-R (Entity-Relationship) model of the Figure 1 describes the main classes for determining the page quality: (i) the page itself, with data related with its title, language, authorship, references and dates of creation and update; (ii) the web page author, with data related with contact and expertise in the subject covered by the web page; (iii) the organization that sponsor the Web page (if exists); (iv) the links of another pages or sites in the Web pointing to this (inLinks); and (v) which are the web links from this page to other web pages (outLinks).

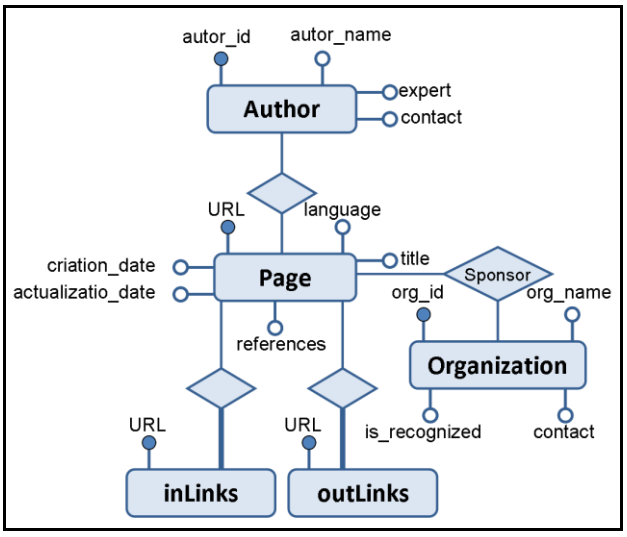

Figure 1. E-R model proposed to evaluate the content of health web pages [11].

[2] http://www.google.com

[3] http://yahoo.com

In the development of the techniques to collect data on the model of Figure 1, was noted the difficulty in obtain information about the date of creation and update of web pages. This is very important in determining the quality of a web page, because expresses the content freshness. We observed that this happen in general or by the absence of this kind of information or because is presented in a non machine readable format, like in an image. Thus, the set of data in which was possible to automatic extract information in this work is related to the site language and its inLinks and outLinks. About the authorship, it was possible to determine the author expertise and contact (e-mail address).

The following sections describe the models and techniques developed to perform the automatic location and extraction of web pages to obtain the set of data described on the model presented in the Figure 1.

4. AUTOMATIC LOCATION

One of the main points in the system's operation was the automatic location of health web pages to be evaluated and extracted. The results achieved must be returned to the evaluation process in order of relevance, i.e. according to how they satisfy the search criteria [11]. For more details regarding the criteria of quality, we recommend the reading of the work developed in [10].

Given the estimated size of the Web, find the more appropriate page to the user's intention is not a trivial task. Its necessary a sophisticate algorithm, combined with massive computational power, to accomplish this task. For that reasons, in this work was chose to use search engines specifically oriented for health-pages retrieval. In this category of search engines the search is performed on a database of previously indexed and constantly updated pages. Some of the specialized search engines are SearchMedica[4] and MedStory[5], which are specific to the medical field. In the next sections is present the model and prototype developed to locate health web pages, but not focusing on specific search engines – as this choice is application-dependent. The specific configuration of this model to a practical application is presented in section 6, with a Case Study about Alzheimer's disease.

4.1 Location Model

The model presented in the Figure 2 is based on the interaction among the application, a database and the Internet. It is application dependent because it needs the existing pattern for presenting the results of the search engine used in the application, i.e., the internal standard to display the results to the user.

Initially, the data about the search engine (pattern related) is required by the system. Then, the criterion to be used by the search engine is recovered from the database and for each criteria is stored the first URLs returned as answer by this search engine. For more details about this model see [11]. The database stores the identification of the search engine used, as well as their model. This is necessary because different engines offer specifics return forms. Thus, it is possible to extend the system, adding new patterns of search engines or removing them from the database. A prototype was developed for retrieval and indexing of returned results as well as for its storage in the database [11].

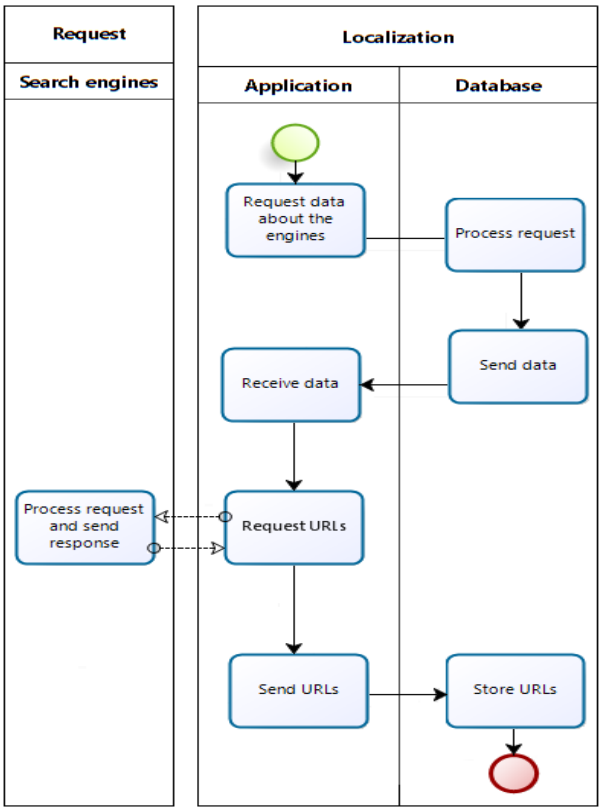

Figure 2. Model to locate data in the Web
(modified from [11])

5. DATA EXTRACTION

After recovering the URLs from the search engines, obtained at the location step, the system starts the next task, which extracts the content that will be used to evaluate the quality of the web page. As this task was idealized to occur in an automatic way a search is performed for specific attributes that are relevant to determine the page quality, those attributes can be found in the Figure 1. The initial supposition was that it should be possible to find a standardization way that allows encompassing the maximum portion of data existent in web pages in order to make the data extraction easier and faster. However, as presented in a more detailed way in the 5.1 subsection, it was not possible to use this approach because until today there are no standardization models for naming the tags or to place data in pages. Therefore, specific strategies were proposed, trying to attend the necessary requirements to the search of each data existent in the ER model (Figure 1), as: site language; authorship data; inLinks and outLinks.

The next sections present each one of those strategies, starting with troubles and adopted strategies for standardization of web pages content and structure.

5.1 Standardization

The considered standardization has the objective to define the ways in which the content of a web page could be structured. That research is applied later as an information source to the

[4] http://www.searchmedica.com

[5] http://www.medstory.com

extraction phase. The structure and content of search engines, as well as web pages, is extremely dynamic. Consequently, to overcome this constantly changes the standardization model adopted here apply terms defined in the WorldNet [12] trying to found synonyms for the keywords found in web pages. Examples of synonyms that could be obtained for the term "author", based in the WorldNet, are: "writer", "generator", "source".

There are others interesting approaches that can be applied as a standardization strategy. An example is the application of ontologies [13], where they can be used to support decisions about which fields must to be analyzed; these ontologies could, either, be enriched by the WorldNet terms. The use of terms from WorldNet is interesting to find synonyms of the fields that could appear in the pages and be used in the extraction. Unfortunately, there are no guarantees that the website developer has used some synonymous term to define fields (as author or writer, for the author´s name field). So, is important to think in alternative strategies.

In this work, the recognition of HTML tags in web pages had generated special attention. Techniques for that are applied in areas as Information Storage and Retrieval to identify the relevant topic of a page, since that definition is very important to the information filtering phase [14]. Some of the HTML tags by it self can provide good information for the definition/search of patterns, for example, the tags: *<h1>*, *<h2>*, *<h3>*, *<h4>* could indicate field patterns, as: title, author´s name and contact.

5.2 Extracting the Site Language

These step objectives find out the language in which the text of the web page was written. In this project the system informs the URL of the health web page that must to be analyzed. The first step refers to search for meta tags that indicate the language of the web page as defined in the Dublin Core pattern: *meta tag language* (1)[11].

<meta name="language" content="english"/> (1)

If the system didn't found the meta tag, is necessary to make a deeper analysis on the content of the web page to verify that. In this case, some programs that analyses the text in order to detect the language can be used. Examples of analyzers used in the application are: Fuzzums[6], Applied Language[7] e Google Language Detection[8]. After obtaining the page content subsets of text, as well as the keyword used in this search, are extracted from the content. In the most part of the cases tested the analysis made by two of those programs was enough to detect the web page language. When the analysis made by two programs disagreed, one extra program was used to reach a conclusion [11].

5.3 Extraction of InLinks and OutLinks

Supposing the website X, inLinks are all the links existing in the Web that point to X, outLinks are the links of the website X that point to sites that are hosted in other domains. The inLinks and outLinks of each website should be stored for posterior analysis. InLinks are used for some search engines in order to build a ranking of the results. Websites that have many inLinks receive a higher score, and consequently, appear in a better position in the ranking. In this work, inLinks are recovered as follows [11]:

- An request is submitted to the Browser Google Web: http://www.google.com.br/search?q=link%3A+<url_target_page>;
- After recover the results page, for each one of the listed pages, the XPath expression: //cite is executed;
- All lists of URLs obtained in each page are stored.

For the recovery of outLinks the steps are [11]:

- Recover the page which intends to get the outLinks;
- Execution of the expression XPath: //a[not(contains(@href = 'DOMAIN'))]/@href, where DOMAIN represents the domain of the desired page;
- Through the expression cited in the previous step, the duplicated links that are hosted in the same domain of the page are removed;
- When counting the number of outLinks, the duplicates are removed and the links that point to the same domain are counted as only one entry.

5.4 Extraction of Author's Information

Data about the website author are important for determine the quality of the presented content. If the author is considered a specialist on the subject, the web page will have more credibility then the others, written by people not recognized as such. However, this kind of information is not easily determined in a manual way and in case of automatic treatment of data it became even more complex. There are some techniques that could be applied to help in analyzing data about the author, as is the case of h-index [15], which applies the number of author's publications and citations. In case of a complex task, where an individual technique does not solve the problem, in this work was applied a solution described in [16] where techniques for data extraction about pages authorship are employed. The authorship model is defined by the combination of vocabularies defined in the Dublin Core pattern [17] and in the FOAF (Friend Of A Friend) ontology [18], for descriptions related to the authors expertise. Another tool for automatic extraction obtains the author's organization; his or her electronic mail address; the website address; the author's number of publications and h-index [16].

6. CASE STUDY

The models described in the previous sections were developed as prototypes to evaluate the proposed approach. During the tests the topic used for search was "Alzheimer Disease". The first step was focused in the automatic location of data, the considered tasks were:

- Study of the existent search engines, as well as analysis of each one of them in order to identify which ones were appropriate to search data about health;
- Definition of the criteria that will be used for the searches related to the topic "Alzheimer Disease".

The section 6.1 below presents details of the search engines. About the criteria, without a real population of users to develop

[6] http://www.fuzzums.nl/~joost/talenknobbel/index.php

[7] http://www.appliedlanguage.com

[8] http://www.google.com/uds/samples/language/detect.html

this case study, was chosen to use common criteria from the topic. A more detailed description about the applied criteria is presented in the subsection 6.2.

6.1 Health and General Search Engines

The queries have two kinds of results: texts that are appropriated to general public; technical texts with more detailed information of the topic. During the search engine analysis, three of them have been choose. That was necessary for automating the location of web pages, because the developed tool should be prepared for a specific pattern of the engines. Among the analyzed ones, the following were adopted [11]:

- **Medstory:** Search in the Web and sorts the results based on several keywords, indicating which ones occur more frequently. These keywords, about healthcare, belong to a predefined static list, grouped by categories, as: drugs, symptoms, procedures and so on. This list predefined by the website allows refining the results presented to the user. After the search, several complementary categories of data related to the topic are presented to the user. These categories consist of pre-processed information, stored in a specific database.
- **SearchMedica:** Allows setting up the search scope in which the search engine will search: the whole Web or sites defined by the user. Also, the search can be made over a determined health area (e.g. cardiology, geriatrics, dermatology, etc.). The results page has an area in which several related information are presented.
- **Google:** Allows search in the whole Web, so, can find since information for laymen until news or scientific work. However, there are no guarantees about the reliability of the presented data in the resulting websites, unlike what happens in the Medstory, which search in a recognized source internally indexed. At the same time, searches that encompass a bigger number of pages can find more relevant results for the user. In these cases, algorithms as PageRank™ [20], used by this search engine, try to rank the results based on reliability.

6.2 Search Criteria

As described in the introduction of this case study, the search criteria applied were chosen according to the topic "Alzheimer Disease", so, they are application dependent. Thus, an analysis should be done in order to choose good criteria for more coherent results. This issue refers to business analysis [21] and not to a development issue, i.e., an analysis preferably held by an expert in the area, in order to provide the closer criteria that could be used by laymen and experts in its researches. Also, keywords were separated into categories; they are listed below [11]:

- "General information" - introductory information. Keywords: alzheimer and alzheimer introduction.
- "Treatment methods" - has two goals: for practitioners, interested in new methods for treatment; and for laymen, interested in know about its treatment. Keyword: alzheimer's treatment.
- "Diagnosis" - looks to provide information about the first symptoms of the disease. Keywords: alzheimer's diagnostic and alzheimer's symptoms.
- "Drugs" - intend to provide information about drug interactions, allergic reaction and efficiency. Keywords: alzheimer drugs interactions and alzheimer drug treatment.
- "Case study" - has focus on expert users, which may be searching for information to base their researches about a disease, for example. Keyword: alzheimer case study.
- "Tips" – has the objective to answer questions as: ways to deal with a patient with Alzheimer; how is possible do help; what are the risk factors of the disease. Keyword: alzheimer's practical tips.

These search criteria allows the automatic location of websites by the execution of searches over the used search engines (subsection 6.1), as well as store the 10 first URLs, associated with each search engine. After the location and storing the URLs, the system can go further - data extraction.

6.3 Extraction

This step intends to extract the following attributes: page language, authorship data, inLinks and outLinks. Each one of those attributes was described during the previous sections. In order to develop the proposed solution as a prototype some tools were used to support retrieve data in the websites, they are: Web Harvest[9] e XPather [11]:

- The Web-Harvest provide an API (Application Program Interface) that allows: consult Web servers; recover the HTML page; convert it in XHTML(Extensible Hypertext Markup Language) file and apply some manipulation techniques of XML (Extensible Markup Language) documents, as XPath, XQuery and XSLT (Extensible Stylesheet Language Transformations) to extract the desired information;
- The XPather allows search for patterns in websites, being used in a semi-automatic manner.

7. CONCLUSIONS AND FUTURE WORK

This article described the development of models, techniques and prototypes aiming the automatic retrieval of content presented in web pages with health subject. After the automatic location, standardization and extraction of page's data the objective is to deliver normalized data to perform evaluations on the quality of the health associated web page. The process of developing the model for automatic retrieval of data anticipated a series of challenges and steps to overcome and the process was directed to work with these problems. For location were applied well known search engines as they use quite efficient algorithms for Web search. The search for patterns for presenting data in web pages has led to the conclusion that despite the already existence of standards such as Dublin Core, in practice, these patterns are not strictly applied by the vast majority of pages, or is partially applied. Thus, the need of some kind of standardization for the development of these pages is clearly needed. As a future work, it would be necessary to end the process of evaluating the quality of a health web page making the task of quality evaluation itself automatic.

[9] http://web-harvest.sourceforge.net/

8. ACKNOWLEDGMENTS

This work was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, Brazil, Edital Universal - MCT/CNPQ - 14/2010. We would also like to thank the group of students of the discipline CMP112, PPGC/UFRGS, that developed the initial work on this subject.

9. REFERENCES

[1] Web Medica Acreditada – WMA. Retrieved March 21, 2011, from: <http://wma.comb.es/>.

[2] Internet Content Rating Association - ICRA. Retrieved March 21, 2011, from: <http://www.fosi.org/icra/>.

[3] Internet Quality Agency – IQUA. Retrieved March 21, 2011, from: <http://www.iqua.net/>.

[4] Alpert, J. and Hajaj, N. We knew the Web was big... The Oficial Google Blog. Retrieved March 20, 2011, from: <http://googleblog.blogspot.com/2008/07/we-knew-Web-was-big.html>.

[5] The Quality Assurance and Content Description Project – QUATRO+. Retrieved March 28, 2011, from: <http://legacy.quatro-project.org/>.

[6] Berners-Lee, T. 1997. Cleaning up the User Interface, Section—The "Oh, yeah?"-Button, Retrieved March 28, 2011, from: http://www.w3.org/DesignIssues/UI.html.

[7] Stamatakis, K., Chandrinos, K., Karkaletsis, V., Mayer, M.A., Gonzales. D.V., Labsky, M., Amigo, E. and Pöllä, M. 2007. AQUA, a system assisting labeling experts assess health Web resources. In Proceeding of Symposium on Health Information Management Research – ISHIMR 2007.

[8] Wang, Y., Liu Z. 2007. Automatic detecting indicators for quality of health information on the Web, International Journal of Medical Informatics, 76(8), 575-582.

[9] Balakrishnan, H., Kaashoek, M. F., Karger, D., Morris, R., Stoica, I. 2003. Looking up data in P2P systems. In *Communications of the ACM*. DOI=http://doi.acm.org/10.1145/606272.606299.

[10] Lichtnow, D., Jouris, A., Bordignon, A., Pernas, A. M., Levin, F. H., Nascimento, G. S., Silva, I. C. S., Gasparini, I., Teixeira, J. M., Rossi, L. H. L., Oliveira, O. E. D., Schreiner, P., Gomes, S. R. V., Oliveira, J. P. M. de. 2009. Relato e Considerações sobre o Desenvolvimento de uma Ontologia para Avaliação de Sites da Área de Saúde. *Cadernos de Informática* (UFRGS), v. 4, p. 7-46.

[11] Pernas, A. M., Palazzo, J. M. de O., Santos, A.H., Donassolo, B.M., Bezerra, C.B., Manica, E., Kalil, F., Soares, L.S., Svoboda, L.H., Mesquita, M.P., Torres, P.R., Petry, R.L., Santos, R.L., Leithardt, V.R.Q. 2009. Relato sobre o Desenvolvimento de Modelos para Obtenção Automática do Conteúdo de Sites sobre Saúde. *Cadernos de Informática* (UFRGS), v. 4, p. 47-91.

[12] Wordnet: a lexical database for the English. Princeton University. Retrieved March 20, 2011, from: <http://wordnet.princeton.edu/wordnet/download/>.

[13] Tiun, S., Abdullah, R. and Kong, T.E. 2001. Automatic Topic Identification Using Ontology Hierarchy. In *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing - CICLing '01*. Springer-Verlag, London, UK.

[14] Liu, B., Chin, C. and Ng, H. 2003. Mining Topic-Specific Concepts and Definitions on the Web. In *The 12th International World Wide Web Conference - WWW 2003*, Budapest, Hungary, May 20-24.

[15] Hirsch, J. E. 2005. An index to quantify an individual's scientific research output. *PNAS* 102 (46), 16569–16572.

[16] Lichtnow, D., Pernas, A. M., Manica, E., Kalil, F., Oliveira, J. P. M. de, Leithardt, V. R. Q. 2010. Automatic Collection of Authorship Information for Web Publications. In: *Proceedings of 6th International Conference on Web Information Systems and Technologies – WEBIST*. v. 1. p. 339-344. Lisboa: INSTICC. Valencia.

[17] Dublin Core Metadata Initiative. Retrieved March 22, 2011, from: <http://dublincore.org/>.

[18] Brickley, D. and Miller, L. FOAF Vocabulary Specification 0.98. Namespace Document 9 August 2010. Marco Polo Edition. Retrieved March 20, 2011, from: <http://xmlns.com/foaf/spec/>.

[19] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The Hadoop Distributed File System, IEEE 26th Symposium on Mass Storage Systems and Technologies, 2010, ISBN: 978-1-4244-7152-2, p.1-10.

[20] Brin, S., Page. L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*. Elsevier Science Publishers, Amsterdam, The Netherlands. 1998.

[21] Witten, I.H. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. 2a ed. Morgan Kaufmann. 629p. 2005. ISBN-13:978-0-12-088407-0.