ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-688

Automating Accreditation of Medical Web Content

Vangelis Karkaletsis and **Pythagoras Karampiperis** and **Konstantinos Stamatakis**[1] and **Martin Labský** and **Marek Růžička** and **Vojtěch Svátek**[2] and **Enrique Amigó Cabrera**[3] and **Matti Pöllä**[4] and **Miquel Angel Mayer** and **Angela Leis**[5] and **Dagmar Villarroel Gonzales**[6]

Abstract. The increasing amount of freely available health-related web content generates, on one hand, excellent conditions for self-education of patients as well as physicians, but on the other hand entails substantial risks if such information is trusted irrespective of low competence or even bad intentions of its authors. This is why medical web resources accreditation by renowned authorities is of high importance. However, various health web content surveys show that the proportion of accredited web resources is insufficient due to the difficulty of the labeling authorities to cope with the amount and dynamics of the medical web. In this paper, we address the problem of automating the accreditation of medical web content. To this end, we present a system which provides the infrastructure and the means to organize and support various aspects of the daily work of labeling experts, exploiting web content collection and information extraction techniques.

1 INTRODUCTION

The number of health information web sites and online services is increasing day by day. On the other hand, patients continue to find new ways of reaching health information and more than four out of ten health information seekers say the material they find affects their decisions about their health [1, 2]. However, it is difficult for health information consumers, such as the patients and the general public, to assess by themselves the quality of the information because they are not always familiar with the medical domains and vocabularies [3]. Although there are different opinions about the need for accreditation of health web sites and adoption by Internet users [4], different organizations around the world are working on establishing standards of quality in the accreditation of health-related web content. The European Council supported an initiative within "eEurope 2002" to develop a core set of "Quality Criteria for Health Related Websites" [5]. These criteria may be used as a basis in the development of user guides, voluntary codes of conduct, trust marks, accreditation systems, adopted by relevant parties, at European, national, regional or organizational level. There are two major mechanisms in medical quality labeling:

- Filtering portals: the web resources are classified according to predetermined criteria in order to facilitate a quick access to quality reviewed information. Examples of this mechanism are: "Catalog and Index of French-speaking Medical Sites" (CISMEF) [17], "Intute service - its Health and Life Sciences branch" [16] from UK, "Agency for Quality in Medicine" (AQUMED) [19] from Germany.
- Third party accreditation: an organization evaluates the quality of the web site according to a set of criteria. Compliance with those criteria is showed with a logo or trust mark on the homepage. The HONCode of the Health on the Net Foundation [18], the URAC Accreditation Program [20], the Web Mèdica Acreditada [18] trustmark are the most well known quality seals.

The main problem that these mechanisms face is the need for a continuous review and control of the accredited or classified web sites that means a huge amount of human effort. This stress on content quality evaluation contrasts with the fact that most of the current Web is still based on HTML, which only specifies how to layout the content of a web page addressing human readers. This "current web" must evolve in the next years, from a repository of human-understandable information, to a global knowledge repository, where information should be machine-readable and processable, enabling the use of advanced knowledge management technologies [6]. This change is based on the exploitation of *semantic web* technologies. The Semantic Web is "an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation" based in metadata (i.e. semantic annotations of the web content) [7]. These metadata can be expressed in different ways using the Resource Description Framework (RDF) language. RDF is the key technology behind the Semantic Web, providing a means of expressing data on the web in a structured way that can be processed by machines. In order for the medical quality labeling mechanisms to be successful, they must be equipped with semantic web

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece, email: {vangelis, pythk, kstam}@iit.demokritos.gr
[2] University of Economics, Czech Republic, email: {labsky, ruzicka, svatek }@vse.cz
[3] ETSI Informática, UNED, Spain, email: enrique@lsi.uned.es
[4] Adaptive Informatics Research Centre, Helsinki University of Technology, Finland, email: matti.polla@tkk.fi
[5] Web Médica Acreditada, Medical Association of Barcelona (COMB), Spain, email: {mmayer.wma, mleis.wma}@comb.es
[6] Agency for Quality in Medicine (AquMed), Germany, email: villarroelgonzales@azq.de

technologies that enable the creation of machine-processable labels as well as the automation of the labeling process.

In this paper, we address the problem of automating the accreditation of medical web content. To this end, we present the AQUA system, developed within the MedIEQ[7] project, which provides the infrastructure and the means to organize and support various aspects of the daily work of labeling experts by making them computer-assisted. More precisely, we describe the challenges addressed in AQUA development and the results achieved so far. In Section 2, we present and discuss the design principles of AQUA. In Section 3, we focus on AQUA's components/modules responsible for automating the accreditation process. In Section 4, we present the methodological steps for extending AQUA to support new languages, labeling criteria, as well as, labeling authorities. Finally, in Section 5 we present experimental results on the use of these components, discuss our findings and the conclusions that can be offered.

2 THE AQUA SYSTEM OVERVIEW

By analyzing the two main approaches of medical quality labeling (filtering portals and third party accreditation), we have identified the following key tasks, followed entirely or partially by most labeling agencies:

- *Identification of new web resources*: this could happen either by active web searching or by voluntary application from the information provider.
- *Labeling of the web resources*: this could be done with the purpose of awarding an accreditation seal or in order to classify and index the web resources in a filtering portal.
- *Re-reviewing or monitoring the labeled web resources*: this step is necessary to identify changes or updates in the resources as well as broken links and to verify if a resource still deserves to be awarded an accreditation seal.

As a result, the AQUA system [14] was designed to support the main tasks of the web content accreditation process, that is: Identification of unlabeled resources having health-related content; Visit and review of the identified resources; Generation of content labels for the reviewed resources, and Monitoring the labeled resources.

Compared to other approaches that partially address the assessment process [8, 9], the AQUA system is an integrated solution. AQUA aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts by making them computer-assisted. More specifically, AQUA supports labeling experts in:

- *Creating machine readable labels,* by adopting the use of the RDF model [10] for producing machine-readable content labels; at the current stage, the RDF-CL model [11] is used. In the final version of AQUA, the POWDER model, introduced by the W3C Protocol for Web Description Resources (POWDER) working group [24], will be supported.
- *Automating the accreditation process* by helping in the identification of unlabeled resources, extracting from these resources information relative to specific accreditation criteria, generating content labels from the extracted information and facilitating the monitoring of already labeled resources.

2.1 System Architecture

AQUA incorporates several subsystems (Figure 1) and functionalities for the labeling expert. The *Web Content Collection* (WCC) component identifies, classifies and collects online content relative to the labeling criteria.

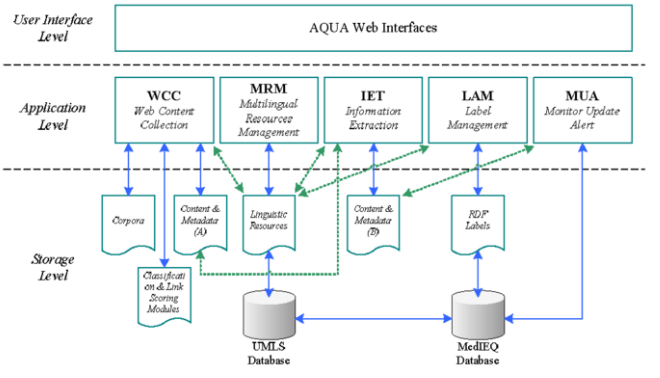

Figure 1. Architecture of the AQUA system.

The *Information Extraction Toolkit* (IET) analyses the web content collected by WCC and extracts attributes for the content labels. The *Label Management* (LAM) component generates, validates, modifies and compares the content labels. The *Multilingual Resources Management* (MRM) subsystem gives access to health-related multilingual resources; input from such resources is needed in specific parts of the WCC, IET and LAM toolkits. Finally, the *Monitor-Update-Alert* (MUA) tool handles the configuration of monitoring tasks, database updates, and alerts to labeling experts.

3 AUTOMATING THE ACCREDITATION PROCESS

3.1 Locating Unlabeled Web Resources

The AQUA *crawling* mechanism is part of the *web content collection* environment (WCC). The Crawler searches the

[7] MedIEQ (www.medieq.org) is an EC-funded project, under the DG-SANCO Programme "Public Health". MedIEQ stands for "Quality Labeling of Medical Web content using Multilingual Information Extraction.

Web for health-related content that does not have a content label yet. It is a meta-search engine that exploits results returned from known search engines and directory listings from known Web directories. All collected URLs from all sources are merged and filtered, and a pre-final URLs list is returned.

Apart from two well-known general-purpose search engines (Google, Yahoo!), the AQUA Crawler exploits two searching services specialized to the health domain, one from HON [15] and a second from Intute's Health and Life Sciences branch [16]. Crawler's open architecture guarantees that additional search engines can be added, at any moment, if this is needed for a specific application.

At the same time, the Crawler is enhanced with a content classification mechanism, which can be trained to distinguish health from non-health content. For the initial training, the user manually classifies as Pos or Neg a part of the initial results. A quick preview-assess-characterize interface is available for this purpose. A classification model is trained which is then used for the automatic classification of the results returned in next search iterations. At the end of a search iteration, the new results automatically get a classification score (pos, neg or uncl) proposed by the model trained from user's feedback. The user can continue by manually verifying the automatically classified URLs, as well as, by checking the unclassified ones. Then, the model can be again re-trained and so on.

Interesting similar approaches exist, for example [21], where a context-based adaptive personalized web search to adapting search results according to user's information needs is proposed. However, this approach depends on domain specific ontologies, whereas our Crawling mechanism is domain agnostic, enabling the user to classify the results and improve the tool performance.

3.2 Spidering

The AQUA *Spider* examines individual pages of a website via following links. The web resources whose URLs are obtained from the Crawler are processed by the Spider one-by-one in several independent threads. Unreachable sites/pages are revisited in next run.

Since not all web pages of a web site are interesting for the labeling process, the Spider utilizes a content classification component that consists of a number of *classification modules* (statistical and heuristic ones). These modules decide which pages contain interesting information. Each of them relies on a different classification method according to the classification problem on which it is applied. Pages identified as belonging to classes relevant to some of the labeling criteria are stored locally in order to be exploited by the Information Extraction subsystem (for instance, contact pages in order to extract contact information from them).

3.3 Extracting Information Relative to Accreditation Criteria

The present work continues and builds upon the work of previous projects in the area of *information extraction* (IE) [12, 13]. The AQUA IE toolkit (IET) employs a set of components responsible for the extraction of elementary information items found in each document and for the integration of these items into a set of semantically meaningful objects called *instances*.

The core IE engine currently used within IET is the *Ex* system [13], which relies on the combination of the so-called *extraction ontologies* with exploiting the local *HTML formatting regularities* and embedding *trainable classifiers* for specific low-level tasks. The output of IET is proposed to the labeling expert through the AQUA LAM user interface.

3.4 Monitoring Already Described Resources

Another part of AQUA, called MUA (from Monitor-Update-Alert), handles problems such as the *configuration* of *monitoring tasks*, the necessary MedIEQ *repository updates* and the *alerts* to labeling experts when important differences (relative to the quality criteria) occur during the monitoring of previously labeled sites. MUA thus extends the functionality of the content collection and extraction toolkits by shifting from a one-shot scenario to that of continuous monitoring.

4 EXTENDING AQUA

AQUA addresses a complex task. However, various design and implementation decisions help MedIEQ partners keep AQUA extensible and easy to maintain. The main characteristics of its implementation include: a) open architecture, b) accepted standards adopted in its design and deployment, c) character of large-scale, enterprise-level web application, and d) internationalization support.

AQUA has been designed so as to be able to support addition of new languages, labeling criteria and labeling authorities. Figure 2, presents the methodological workflow for extending AQUA to support a new language. This process consists of the following main steps:

- *User Interface Localization.* This is in practice the translation of a text messages file.
- *Spider Model Training.* The training of Spider's classifiers is facilitated using a specialized tool called Corpus Formation Tool for the collection and annotation of corpus.
- *IET Model Training*. The training of extraction models is supported using the BOEMIE Annotation Tool [22], which enables the annotation of named entities and relations.
- *Topic Categorizer Configuration.* AQUA's topic categorizer employs the Automatic Ontological

Concepts Extraction Tool (POKA) [23], a general purpose tool for automatic extraction of ontological concepts (MeSH in the current implementation of AQUA).

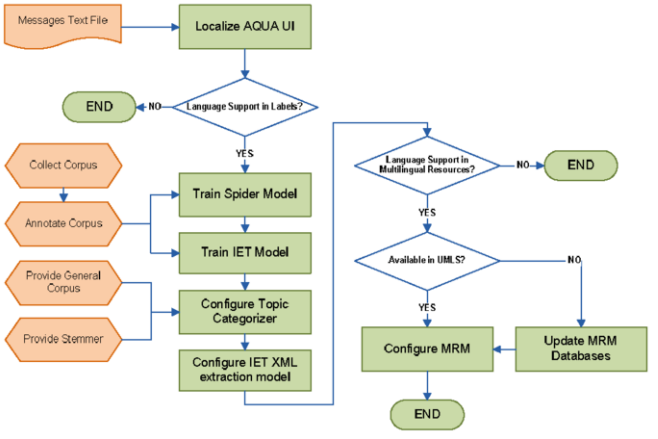

Figure 2. Extending AQUA to support a new Language.

Concerning the extension to support new labeling criteria, a well-defined process is used, which involves the following key steps:

- Adaptation of the LAM User Interface,
- Definition of the desired criterion as an RDF element of the supported Vocabulary,
- Specification of the methodology for the extraction of information relevant to the new criterion, e.g. use of regular expressions, heuristics or machine learning.

Finally, the extension to support a new labeling authority is an iterative process that extends AQUA for the new labeling criteria and language(s) of this labeling authority.

5 EXPERIMENTAL RESULTS

So far, the 1st version of AQUA has been developed and evaluated by labeling experts. The scope of this evaluation was the performance evaluation of AQUA on supporting the labeling process, as well as, the usability evaluation of AQUA interface. The evaluation showed that by using only the proposed by AQUA links, it was possible for the labeling experts to identify the right value in more than 80% of the different labeling cases. In this section we present evaluation results, for locating unlabeled web resources, spidering and information extraction.

5.1 Locating Unlabeled Web Resources

In this section, we summarize evaluation results on Crawler's content classification component. For this evaluation, we used an English corpus, consisting of 1976 pages (944 positive & 1032 negative samples), all manually annotated. Three different classifiers have been tested (SVM, Naïve Bayes and Flexible Bayes). Best performance was achieved with 1-grams and HTML tags removed (see Table 1).

Table 1. Classification performance results for content classification

	1-grams (Tags removed)		
	Prec.	**Rec.**	**Fm.**
NB	0.75	0.63	**0.68**
FB	0.73	0.55	**0.62**
SMO	0.75	0.61	**0.67**

The relatively low performance of the content classifiers is justified by the fact that is difficult, even for humans, in various cases to assess whether a website has health-related content or not. In addition, the health domain covers a wide set of topics.

5.2 Spidering

The Spider classification mechanism has been examined for the accreditation criteria listed in Table 2.

Table 2. A sub-set of the AQUA supported accreditation criteria

Accreditation Criterion	***AQUA approach***
The target audience of a website	Classification among three possible target groups: adults, children and professionals
Contact information of the responsible of a website must be present and clearly stated	Detection of candidate pages during the spidering process and forwarding for information extraction
Presence of virtual consultation services	Detection of parts of a website that offer such services during the spidering process
Presence of advertisements in a website	Detection of parts of a website that contain advertisements during the spidering process

Table 3. SVM performance.

Category	**Precision**	**Recall**	**Fm**
CI	0.84	0.96	0.90
AD	0.87	0.80	0.83
VC	0.87	0.87	0.87
Adults	0.78	0.75	0.77
Children	0.80	0.78	0.79
Professional	0.77	0.81	0.79

Several learning schemes were tested. The performance of the SVM classifier which provides the best results is presented in Table 3, for English corpora. The most difficult criterion for classification purposes seems to be the target audience, as being a highly subjective one. Although not reported in this paper, experiments have been also performed for other project languages, namely, Czech, Greek, Finish and Spanish. The results received vary a lot between corpora in different languages. Changing the experimental settings, with respect to the use of HTML tags, stemming, etc., does not affect significantly the results indicating that the characteristics of each language specific corpus is the determinant factor.

5.3 Extracting Information Relative to Accreditation Criteria

The IET serves as a container for IE engines which exposes a uniform API used by the AQUA system.

Currently it integrates two IE engines: the first engine is based on extraction ontologies [16], and the second is a machine learning based one. Both engines can be combined within one extraction task; however the possible benefits are not yet fully exploited. Preliminary F-measure results are presented below for the extraction of contact information using extraction ontologies for three languages. In Table 4, the first scores are stricter since they require the extracted values to be exactly the same as the gold standard. The second scores also give some credit to partial matches (proportional to the correct field's word length). Results marked with a dash are not yet available.

Results are presented for named entity extraction. Corpora sizes were 109 contact HTML pages for English, 200 for Spanish and 108 for Czech. The collections contained roughly 7000, 5000 and 11000 named entities, respectively. One contact extraction model was developed per language (with shared common parts) based on seeing 30 randomly chosen documents from each dataset and evaluated using the remaining documents. The system exploits manually encoded extraction knowledge in the form of cascaded pattern matching rules, axioms and HTML formatting regularities induced over the analysed documents.

Table 4. Extraction performance for contact information

Field	Fm English		Fm Spanish		Fm Czech	
Person name	0.77	0.83	0.76	0.81	0.78	0.80
Person degree	0.77	0.82	-	-	0.88	0.90
Street	0.56	0.75	0.56	0.71	0.71	0.75
City	0.57	0.59	0.59	0.61	0.68	0.77
Zip	0.86	0.89	0.89	0.93	0.94	0.94
Country	0.70	0.71	0.72	0.72	0.74	0.78
Phone	0.89	0.91	0.87	0.92	0.88	0.89
Email	1.00	1.00	0.96	0.97	0.98	0.98
Organization	0.45	0.64	-	-	-	-
Department	0.46	0.62	-	-	-	-
Overall	0.73	0.79	0.76	0.80	0.82	0.84

6 CONCLUSIONS

In this paper, we address the problem of automating the accreditation of medical web content. To this end, we present the AQUA system, which provides the infrastructure and the means to organize and support various aspects of the daily work of labeling experts. In general the evaluation results were promising for the further development of AQUA as an assisting system for the labeling experts. However, for the evaluation of the final version of AQUA towards its integration within the day-to-day activities of a labeling organization, the application in a real day-to-day practice scenario will be performed.

ACKNOWLEDGEMENTS

The work presented in this paper is supported by the EC-funded project MedIEQ (www.medieq.org), under the DG-SANCO Programme "Public Health". The authors of this paper would like to thank the labeling experts participating in AQUA evaluation.

REFERENCES

[1] Eysenbach G. Consumer health informatics. BMJ 320 (4) (2000), 1713-16.
[2] Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients´use of the Internet for medical Information. J Gen Intern Med 17(3) (2002), 180-5.
[3] Soualmia LF, Darmoni SJ, Douyère M, Thirion B. Modelisation of Consumer Health Information in a Quality-Controled gateway. In: Baud R et al. (ed.). The New Navigators: from Professionals to Patients. Proc of MIE2003 (2003), 701-706.
[4] Analysis of 9th HON Survey of Health and Medical Internet Users Winter 2004-2005, 2005. Available Online at: http://www.hon.ch/ Survey/ Survey2005/ res.html
[5] http://europa.eu.int/information_society/eeurope/ehealth/doc/ communication_acte_en_fin.pdf.
[6] Eysenbach G. The Semantic Web and healthcare consumers: a new challenge and opportunity on the horizon?. J Healthc Techn Manag 5 (2003), 194-212.
[7] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American, May 2001.
[8] Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. J Med Internet Res. 2005 Dec 30;7(5):e59.
[9] Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the Web. Int J. Med Inform. 2006 May 31.
[10] http:// www.w3.org/TR/rdf-schema/
[11] http://www.w3.org/ 2004/12/q/doc/content-labels-schema.htm
[12] Karkaletsis V, Spyropoulos CD, Grover C, Pazienza MT, Coch J, Souflis D. A Platform for Crosslingual, Domain and User Adaptive Web Information Extraction. In Proceedings of the European Conference in Artificial Intelligence (ECAI); 2004; Valencia, Spain; p. 725-9.
[13] Labsky M., Svatek V., Nekvasil M., Rak D.: The Ex Project: Web Information Extraction using Extraction Ontologies. In: Proc. PriCKL'07, ECML/PKDD Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery. Warsaw, Poland, October 2007
[14] Stamatakis K, Chandrinos K, Karkaletsis V, Mayer MA, Villarroel D, Labsky M, Amigó E, Pölla M. AQUA, a system assisting labelling experts assess health web resources. In Proceedings of the 12th International Symposium on Health Information Management Research (ISHIMR); 2007; Sheffield, UK, 75-84.
[15] http://www.hon.ch
[16] http://www.intute.ac.uk/healthandlifesciences/
[17] http://www.cismef.org/
[18] http://wma.comb.es
[19] http://www.aezq.de or http://www.patienten-information.de
[20] http://www.urac.org/
[21] Pan, Xuwei; Wang, Zhengcheng; Gu, Xinjian, "Context-Based Adaptive Personalized Web Search for Improving Information Retrieval Effectiveness," Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on , vol., no., pp.5427-5430, 21-25 Sept. 2007
[22] Fragou P., Petasis G., Theodorakos A., Karkaletsis V., Spyropoulos C.D, BOEMIE ontology-based text annotation tool. In Proc. of the Language Resources and Evaluation Conference (LREC-2008), Marrakesh, 28-30 May 2008.
[23] http://www.seco.tkk.fi/tools/poka/
[24] http://www.w3.org/2007/powder/