Content Collection for the Labelling of Health-related Web Content

K. Stamatakis[1], V. Metsis[1], V. Karkaletsis[1],
M. Ruzicka[2], V. Svátek[2], E. Amigó[3], M. Pöllä[4], C. Spyropoulos[1]

[1] National Centre for Scientific Research "Demokritos"
{kstam, vmetsis, vangelis, costass}@iit.demokritos.gr
[2] University of Economics, Prague
{ruzicka, svatek}@vse.cz
[3] Universidad Nacional de Educacion a Distancia
enrique@lsi.uned.es
[4] Teknillinen Korkeakoulu – Helsinki University of Technology
mpolla@cis.hut.fi

Abstract. As the number of health-related web sites in various languages increases, so does the need for control mechanisms that give the users adequate guarantee on whether the web resources they are visiting meet a minimum level of quality standards. Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labelling, the MedIEQ project, integrates existing technologies and tests them in a novel application: the automation of the labelling process in health-related web content. MedIEQ provides tools that crawl the web to locate unlabelled health web resources, to label them according to pre-defined labelling criteria, as well as to monitor them. This paper focuses on content collection and discusses our experiments in the English language.

Keywords: content labelling, health information quality, web content collection, focused crawling, spidering, content classification, machine learning.

1 Introduction

The number of health information web sites and online services is increasing day by day. Different organizations around the world are currently working on establishing quality labelling criteria for the accreditation of health-related web content [9, 10]. The European Council supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Web Sites" [8]. However, self-adherence to such criteria is nothing more than a claim with little enforceability. It is necessary to establish rating mechanisms which exploit such labelling criteria.

Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labelling, the EC-funded project MedIEQ[1] aims to pave the way to-

[1] MedIEQ: Quality Labelling of Medical Web Content using Multilingual Information Extraction. Project site: http://www.medieq.org/.

wards the automation of quality labelling process in medical web sites by: a) adopting the use of the RDF[2] model for producing machine readable content labels (at the current stage, the RDF-CL[3] model is used); b) creating a vocabulary of criteria, re-using existing ones from various Labelling Agencies; this vocabulary is used in the machine readable RDF labels; c) developing AQUA (Assisting Quality Assessment) [11], a system through which a labelling expert will be able to identify unlabelled resources having health-related content, visit and review the identified resources, generate quality labels for the reviewed resources and monitor the labelled resources.

Our approach necessitates a robust web content collection mechanism with powerful classification skills. A substantial amount of previous work in the area includes various Web crawling or spidering techniques. A Focused Crawler (term introduced by Chakrabarti et al. in 1999 [2]) is a hypertext resource discovery system, which has the goal to selectively seek out pages that are relevant to a pre-defined topic or set of topics. Aiming to enhance crawling, methods that combine link-scoring (ranking of hyperlinks) with reinforcement learning (InfoSpiders [6], "Intelligent crawling" [1]) or others that link the crawler to domain specific linguistic resources, have been proposed. The latter approach was implemented in two, slightly different, ways. First, for the Crossmarc focused crawler [7], a domain specific ontology, linked to several language specific lexicons, provides the crawling start points, defining thus the subset of the web to be crawled. Second implementation: a domain specific ontology [4] or glossary [5] (MARVIN[4] of HON [10]), gives the crawler filtering capabilities: every accessed resource's relevance is estimated and irrelevant resources are excluded.

Section 2 outlines the AQUA system and describes its web content collection methodology, while section 3 discusses our evaluation methodology and experimental results. Section 4 gives our concluding remarks and suggests the future steps.

2 AQUA and its Web Content Collection subsystem

As already said, MedIEQ develops AQUA, a system designed to support the work of the labelling expert by providing tools that help the identification of unlabelled web resources, automate a considerable part of the labelling process and facilitate the monitoring of already labelled resources. AQUA is an enterprise-level, web application, which supports internationalization and implements an open architecture.

This paper focuses on the Web Content Collection subsystem of AQUA which involves the following components:

a) the Focused Crawler (identifying health related web sites),
b) the Spider (navigating web sites) with link-scoring and content-classification capabilities (the Spider utilizes a content classification component which consists of a number of classification modules, statistical and heuristic ones),
c) tools assisting the formation of corpora (to train/test classification algorithms),
d) a mechanism producing trained classification/scoring models (to be used by the Spider).

---

[2] http://www.w3.org/TR/rdf-schema.
[3] RDF-CL will be refined by the W3C POWDER WG (http://www.w3.org/2007/powder/).
[4] http://www.hon.ch/Project/Marvin_specificities.html

3 Evaluation methodology and results

A first set of 11 criteria, to examine our methodology and test our tools, was decided, by the Labelling Authorities participating in the project consortium. This set of criteria will soon expand to include additional quality aspects[5]. From the initial 11 criteria, the classification mechanism our Spider exploits has been examined using statistical classification techniques for all criteria depicted in Table 1. In addition, for the last criterion, a method using heuristic detection was examined.

Table 1. The MedIEQ criteria upon which our classification components were evaluated

Criterion	MedIEQ methodology
The target audience of a web site	Classification among three possible target groups: adults, children and professionals
Contact information of the responsible of a web site 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 web site that offer such services during the spidering process
Presence of advertisements in a web site	Detection of parts of a web site that contain advertisements during the spidering process

For the statistical classification, pre-annotated corpora were used and three different classifiers provided by the Weka[6] classification platform have been tested: SMO (Weka implementation of SVM), Naïve Bayes and Flexible Bayes. The HTML pages were pre-processed and tokenized in two different methods: a) all HTML tags were removed and only the clear text content of the document was used for the classification and b) both HTML tags and textual content were used. The performance of all classifiers was evaluated using 1-grams and 1/2/3-grams (our results in Tables 2, 3).

Heuristic classification was investigated only for the advertisement detection (our results in Table 4). A large part of current advertising in internet is associated with a reasonably small group of domains; a simple advertisement detection test can be performed by extracting all links on a web page and matching these to a known list of advertisement-providing domain names.

The classification performance for web pages of specific type seems satisfactory, especially if we consider the fact that the corpora were relatively small and the structural information of the pages was not used for the classification. Regarding the performance of the tested classifiers, the obtained values are generally balanced. According to our needs for better precision or recall we can vary the threshold between 0 and 1. All the results presented below use 0.5 as threshold.

The usage of 1/2/3-grams, once the HTML tags removed, gives better results in Target audience classification. A combination of HTML tags and 1/2/3-grams seems to be more helpful in the classification of Contact info and Virtual consultation pages. This latter seems reasonable if we consider that n-grams like “email address”, “phone number”, or sequences of HTML tags which indicate the existence of a communica-

[5] The final set of criteria will be announced through the project website: http://www.medieq.org
[6] http://www.cs.waikato.ac.nz/ml/weka/

tion form may boost the classification performance. On the contrary, we have indications that in Advertisements case such standard sequences of tokens do not occur.

Table 2. Target audience (Adults: 102 / Children: 98 / Professionals: 96), F-measure values.

		1-grams, no tags	1-grams, with tags	1/2/3-grams, no tags	1/2/3-grams, with tags
Adults	NB	0.77	0.71	0.77	0.83
	FB	0.75	0.76	**0.84**	0.76
	SMO	0.80	**0.84**	0.83	0.79
Childr.	NB	0.90	0.83	**0.93**	0.86
	FB	0.88	0.82	0.91	0.86
	SMO	0.92	0.90	0.91	0.91
Prof.	NB	0.90	0.83	0.91	0.91
	FB	0.88	0.82	**0.95**	0.91
	SMO	0.92	0.90	0.89	0.87

Table 3. CI: Contact info (109 pos. / 98 neg.) – VC: Virtual Consultation (100 pos. / 101 neg.) – AD: Advertisements - *Statistical classification* (100 pos. / 104 neg.), F-measure values.

		1-grams, no tags	1-grams, with tags	1/2/3-grams, no tags	1/2/3-grams, with tags
CI	NB	0.72	0.83	0.83	**0.88**
	FB	0.81	0.80	0.83	0.86
	SMO	0.84	0.81	**0.88**	0.87
VC	NB	0.83	0.85	0.85	**0.87**
	FB	0.83	0.83	0.83	0.83
	SMO	0.86	0.84	0.85	0.84
AD	NB	0.88	0.86	0.86	0.84
	FB	**0.89**	0.85	0.88	0.81
	SMO	**0.89**	0.83	0.85	0.82

Table 4. Advertisements - *Heuristic classification*

Precision	Recall	F-measure
0.84	0.72	0.78

As for the heuristic classification method used for the detection of web pages that contain advertisements, it gives moderate performance results when used independently. The not-so-high precision value is owed to the fact that the known lists of advertisement-providing domain names contain also domain names that provide various "tracking" services instead of advertisements. A potential filtering of those domains would enhance the performance.

4 Conclusions and Future work

MedIEQ employs existing technologies in a novel application: the automation of the labelling process in health-related web content. Such technologies are semantic web technologies to describe web resources and content analysis technologies to collect domain-specific web content and extract information from it.

Effective spidering using content classification is a vital part of the web content collection process. Our experimental results, investigating the performance of different learning and heuristic methods, clearly indicate that we are in the right direction. They also make appear even more feasible one of the big challenges of the MedIEQ project, that is, provide the infrastructure and the means to organizing and support the daily work of labelling experts by making it computer assisted. Such a system or platform aims to become AQUA.

Nevertheless, there is follow-up work to be done on content collection. In particular, it would be interesting to combine machine learning with heuristics and examine whether classification accuracy is boosted. At the same time, to scale-up AQUA, we should test our methodology in more languages and evaluate our mechanisms in additional quality criteria.

Acknowledgements
The authors would like to particularly thank the people from AQUMED and WMA for collecting the corpora necessary in our experiments, as well as I. Nieminen, V. Rentoumi, I. Kostopoulou, D. Bilidas and A. Skarlatidis for their support at the realization of the experiments.

References

1. Aggarwal C., Al-Garawi F. and Yu P: Intelligent Crawling on the World Wide Web with Arbitrary Predicates. In Proceedings of the 10th International WWW Conference, pp. 96-105, Hong Kong, May 2001.
2. Chakrabarti S., van den Berg M., and Dom B.: Focused Crawling: a New Approach to Topic-Specific Web Resource Discovery. Computer Networks, 31(11-16):1623 -1640, 1999.
3. Curro V., Buonuomo P. S., Onesimo R., de R. P., Vituzzi A., di Tanna G. L., D'Atri A.: A quality evaluation methodology of health web-pages for non-professionals. Med Inform Internet Med 29(2) (2004), 95-107.
4. Ehrig M. and Maedche A.: Ontology-focused crawling of Web documents. Proc. of the 2003 ACM symposium on Applied computing, pages 1174–1178, 2003.
5. Gaudinat A., Ruch P., Joubert M., Uziel P., Strauss A., Thonnet M., Baud R., Spahni S., Weber P., Bonal J., Boyer C., Fieschi M., Geissbuhler A.: Health search engine with e-document analysis for reliable search results, Int J Med Inform. 2006 Jan; 75(1):73-85.
6. Menczer F. and Belew R. K.: Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. Machine Learning, 39(2/3):203-242, 2000.
7. Stamatakis K., Karkaletsis V., Paliouras G., Horlock J., Grover C., Curran J., Dingare S.: Domain Specific Web Site Identification: The CROSSMARC Focused Web Crawler, Proc. of the 2nd International Workshop on Web Document Analysis (WDA), 2003.
8. European Commission. eEurope 2002: Quality Criteria for Health related Web sites. europa.eu.int/information_society/eeurope/ehealth/doc/communication_acte_en_fin.pdf
9. WMA, Web Mèdica Acreditada. http://wma.comb.es/
10. HON: Health on the Net Foundation. http://www.hon.ch
11. Stamatakis K., Chandrinos K., Karkaletsis V., Mayer M.A., Gonzales D.V., Labsky M., Amigo E., Pöllä M.: AQUA, a system assisting labelling experts assess health web resources, Proc. of the 12th International Symposium for Health Information Management Research (iSHIMR), Sheffield, UK, 18-20 July 2007 (to appear).