

# Content Collection for the Labeling of Health-related Web Content

K. Stamatakis<sup>1</sup>, V. Metsis<sup>1</sup>, V. Karkaletsis<sup>1</sup>,  
M. Ruzicka<sup>2</sup>, V. Svátek<sup>2</sup>, E.A. Cabrera<sup>3</sup>, M. Pöllä<sup>4</sup>

<sup>1</sup> National Centre for Scientific Research "Demokritos"  
{kstam, vmetsis, vangelis}@iit.demokritos.gr

<sup>2</sup> Vysoká Škola Ekonomická v Praze  
{ruzicka, svatek}@vse.cz

<sup>3</sup> Universidad Nacional de Educacion a Distancia  
enrique@lsi.uned.es

<sup>4</sup> Teknillinen Korkeakoulu – Helsinki University of Technology  
mpolla@cis.hut.fi

**Abstract.** As the number of health-related web sites in various languages increases, it is more than necessary to implement control mechanisms that give the users adequate guarantee that 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 labeling, the AQUA system, designed for the EC-funded project MedIEQ, aims to support the automation of the labeling process in health-related web content. AQUA provides tools that crawl the web to locate unlabelled health web resources in different European languages, as well as tools that traverse websites, identify and extract information and, upon this information, propose labels or monitor already labeled resources. Two major steps in this automated labeling process are web content collection and information extraction. This paper focuses on content collection. We describe existing approaches, present the architecture of the content collection toolkit and how this is integrated within the AQUA system, and discuss our initial experimental results in the English language (six more languages will be covered by the end of the project).

**Key words:** Web content collection, focused crawling, intelligent spidering, content classification, machine learning.

## 1. Introduction

The number of health information web sites and online services is increasing day by day. It is known that the quality of these web sites, published by various authorities, is very variable and difficult to assess. At the same time, the necessity to implement control measures that give the consumers adequate guarantee that the health web sites they are visiting meet a minimum level of quality standards and that the professionals offering the information on the web site are responsible for its contents, is increasing. Different organizations around the world are currently working on establishing quality labeling criteria for the accreditation of health-related web content [21, 22, 24-26]. The European Council supported an initiative within eEurope 2002 to develop a core set of "Quality Criteria for Health Related Websites" [23]. 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 labeling criteria.

There are two major mechanisms in medical quality labeling. The first one is based on third party accreditation: a web site is assessed by a labeling agency and, if certain criteria are met, a label is assigned and added to the web site. The second mechanism is based on classification and filtering: medical web sites are reviewed by experts and characterized against certain

criteria; some of them are filtered depending on their characterization; the rest are organized into web directories to facilitate access by health information consumers. Both mechanisms, as currently applied, present drawbacks. As for the first mechanism, the added label is not machine-processable (such that a web browser or a search engine could locate, parse, “understand” and display its characteristics in a human readable way). Moreover, both mechanisms require considerable human effort from labeling experts in order to inspect, characterize and monitor a large number of web sites.

To summarize the current situation in quality labeling of health web content:

- Various labeling authorities, issuing quality assessments, exist,
- Considerable human effort is necessary to manage an everyday increasing number of online health content resources,
- Existing labels are not machine readable and no specific technology has been proposed into this direction,
- No system aiming to update and maintain content labels, assisting thus the work of quality experts from different Labeling Agencies, has yet been designed.

Based upon state-of-the-art technology in the areas of semantic web, content analysis and quality labeling, the EC-funded project MedIEQ<sup>1</sup> aims to pave the way towards the automation of quality labeling process in medical web sites by:

- Adopting the use of the RDF model, for producing machine readable content labels;
- Creating a vocabulary of criteria, re-using existing ones from various Labeling Agencies; this vocabulary is used in the RDF labels;
- Developing AQUA<sup>2</sup>, a system through which a Labeling Authority will be able to generate, read, compare and update its RDF labels.

AQUA develops tools that crawl the Web to locate unlabelled medical web sites in seven different European languages in order to examine their content using a set of machine readable quality criteria, reducing this way the amount of work that labeling experts need to carry out. AQUA tools will monitor already labeled medical sites alerting labeling experts in case the sites’ content is updated against the quality criteria.

Two major steps in this automated labeling process are web content collection and information extraction. This paper focuses on content collection. It describes existing approaches, presents the architecture of the content collection toolkit of MedIEQ and how this is integrated within the AQUA system, and discusses our initial experimental results in the English language (six more languages will be covered by the end of the project).

AQUA utilizes a combination of traditional crawling techniques and existing infrastructure provided by general purpose search engines in order to make the content collection process more efficient. A combination of statistical machine learning techniques and heuristic methods is used whenever content classification is required in order to help the labeling process and speed up the information extraction task. A number of initial experiments have been carried out for the evaluation of the performance of the classification modules used by AQUA.

Related work in online content collection is described in section 2. Section 3 outlines the AQUA system. Section 4 describes the web content collection methodology in MedIEQ, while section 5 discusses our evaluation methodology and initial experimental results. Section

---

<sup>1</sup> MedIEQ: Quality Labelling of Medical Web Content using Multilingual Information Extraction.

Project site: <http://www.medieq.org/>

<sup>2</sup> AQUA: Assisting Quality Assessment

6 gives our concluding remarks, discusses the difficulties encountered so far in content collection and describes the future steps.

## 2. Related work

Several methodologies on web content collection or other suggesting improvements on existing ideas have been proposed in the past years. Different techniques have been combined with traditional crawling/spidering approaches willing to deal better with the different aspects of the web content collection process, including link analysis, linkage sociology (who-link-to-who), context graphs, machine learning techniques, query refinement, etc. And whether it is called just crawling, focused crawling, intelligent crawling or spidering, the concept aiming to be tackled is to efficiently collect web content.

A web crawler is a program which automatically traverses the web by downloading documents and following links from page to page. A general-purpose web crawler normally tries to gather as many pages as it can from a particular set of sites. In contrast, a Focused Crawler (the term “focused crawling” was introduced by Chakrabarti et al. in 1999 [5]) is a hypertext resource discovery system, which has the goal to selectively seek out pages that are relevant to a pre-defined set of topics. Rather than collecting and indexing all accessible web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the web. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date.

There is a substantial amount of work about the methodology of Web crawling. Many crawler architectures and prototypes (e.g. Mercator [11], PolyBot [18], UbiCrawler [3]) were proposed in the literature. The main focus of prior work lies on aspects like crawler scalability and throughput [11], distributed architecture [3], or implementational aspects in connection with particular programming languages (e.g. Java) [18]. However, these solutions do not address the demands of thematically focused Web retrieval applications.

More recent methods use information related to the structure of the Web graph, in order to perform more efficient focused crawling. Some of these methods take advantage of the Topical Locality of the Web (the property of pages with similar topic being connected with hyperlinks [4]) and use it to guide the focused crawler [5]. Moreover, the “backlink” information (pages that link to a certain document), provided by search engines like Google can be used to generate a model of the Web-graph near a relevant page, such as in the case of Context Graphs [8].

The concepts of a thematically focused Web retrieval framework were studied by Menczer in [16]. The idea of focused crawling was used for a variety of Web retrieval scenarios, including exploration of user-specific topics of interest on the Web [6, 19].

There are also methods that combine link-scoring with reinforcement learning in order to deal with focused crawling. In InfoSpiders system [15], a multi-agent focused crawler, the process is initialized by a set of keywords and a set of root pages. Each agent starts with a root page and performs focused crawling by evaluating the link value and following the most promising links. Link value is assessed using a reinforcement learning method, using contextual words as input. Reward values are calculated online, by the reward that the agent receives when following a link. The user can provide relevance feedback to assist the learning process. Another methodology, called “Intelligent crawling” [1] presents a significant improvement of the focused crawling approach. In contrast to the focused crawling method, it uses a combination of evidence, in order to rank the candidate hyperlinks by their level of interest and learns the relevant weight of these factors as it crawls. Making the assumption that the

initial set of starting points can lead to all interesting pages, very central sites should be used as starting points for the crawl (e.g. Yahoo, Amazon, etc.).

Another recent approach is linking the crawler to domain specific linguistic resources. This was implemented in two, slightly different, ways. First, for the Crossmarc focused crawler [20], 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 [9] or glossary (in the case of MARVIN<sup>3</sup>), gives the crawler filtering capabilities: every accessed resource's relevance is estimated and irrelevant resources are excluded. It's worth mentioning that MARVIN is the most known and successful content collection solution applied in the healthcare domain: MARVIN supports effectively and during many years, HON (Health on the Net) foundation's needs in the identification and collection of health related content in real-time. [2] provides more information on MARVIN architecture, while [10] describes how the collected data can be explored and displayed to the end-user.

### 3. The AQUA system outline

As already said, MedIEQ, aiming to make labels machine readable, develops a labeling schema, which is based on the RDF-CL model<sup>4</sup>, issued by the EC-funded project QUATRO<sup>5</sup>. RDF-CL will be refined to a new model by the W3C POWDER working group. At the same time, MedIEQ develops AQUA, a system designed to support the work of the labeling expert by providing tools that help the identification of unlabelled web resources, automate a considerable part of the labeling process and facilitate the monitoring of already labeled resources.

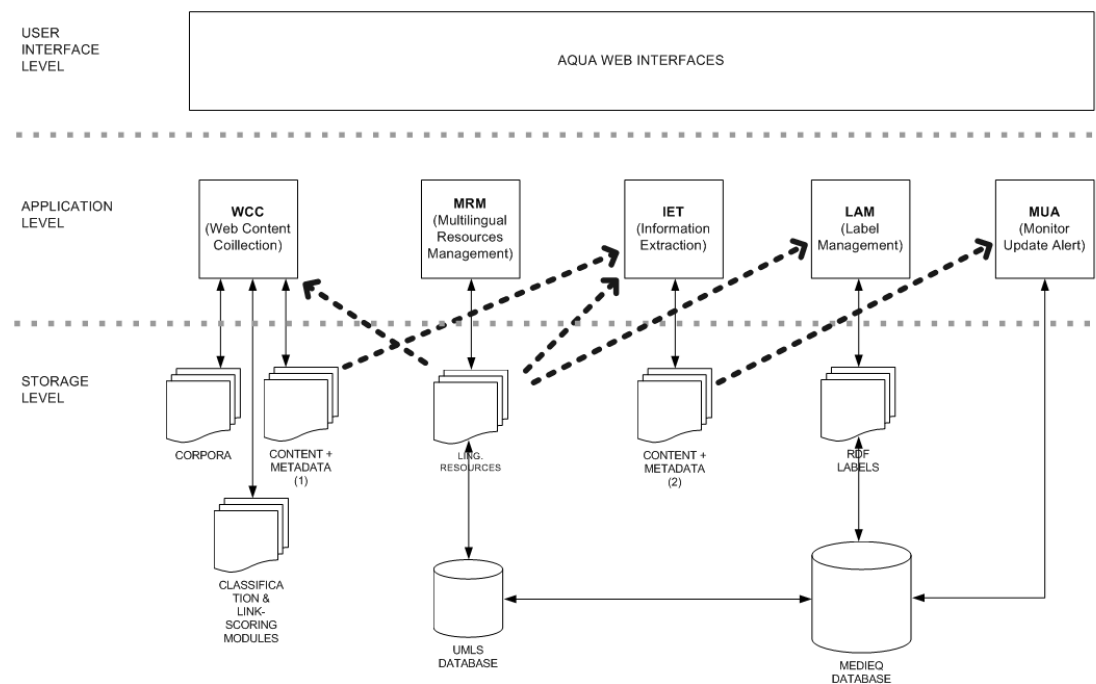


Figure 1 – Architecture of the AQUA system

<sup>3</sup> [http://www.hon.ch/Project/Marvin\\_specificities.html](http://www.hon.ch/Project/Marvin_specificities.html)

<sup>4</sup> [link]

<sup>5</sup> <http://www.quatro-project.org/>

AQUA incorporates several tools and functionalities for the labeling expert. The main characteristics of its implementation include:

- Accepted standards have been adopted in its design and deployment;
- It is a large-scale, enterprise-level, web application;
- Supports internationalization;
- Has an open architecture.

The workflow in AQUA simulates the current workflow in labeling agencies:

1. Identify unlabelled resources having health-related content,
2. Visit and review the identified resources,
3. Generate quality labels for the reviewed resources,
4. Monitor labeled resources.

Thinking that content collection is actively involved in tasks 1, 2 and 4, we understand the necessity of a well optimized such system.

AQUA's final version will cover a set of quality criteria in seven (7) European languages. Our early evaluation results, on a subset of the final criteria and just the English language, presented in the next section, are, however, encouraging. We tried both learning and heuristics methods, various learning algorithms and feature selection techniques in different classification tasks.

#### **4. The web content collection methodology in AQUA**

A handful of tools participate in content collection in AQUA:

- the Crawler: identifies unlabelled resources;
- the Spider: navigates web sites identified by the Crawler, storing locally only pages classified as interesting (interesting pages are then forwarded for information extraction);
- the Content Classification Component: manages the best performing classification models from different classification tasks; classifies every page visited by the Spider;
- the Corpus Formation Tool: assists the expert user in the formation and organization of the corpora necessary for training and testing in different classification tasks;
- the Trained Module Generator: given the corpora, employs different algorithms and trains several models; after testing, promotes the best performing ones.

Considering that content classification services are utilized during the crawling, the spidering and the corpus formation processes, we understand that the performance of our classifiers is crucial. It is the AQUA Crawler and Spider which are evaluated here (results later in this paper) and, therefore, some additional information on them seems necessary.

<WCC architecture – schema?>

##### **How does the AQUA Crawler work?**

The Crawler (or Focused Crawler) searches the Web for health related content, which doesn't already have a quality label (at least not a label found in MedIEQ records). It is a meta-search-engine, exploiting results returned from known search engines and directory listings from known Web directories.

To configure crawling for a specific topic, the user provides two types of start points: sets of keywords and sets of URLs of Web directories. The more relevant to a given topic these start points are, the more focused the crawling will be.

On one hand, keywords are used to query the supported general purpose search engines. Their results are parsed and URLs are collected. In order to be more focused, the Crawler filters the extracted URLs by visiting the HTML page to which each URL points and classifying the content of the page as relevant or non-relevant. If the probability that the HTML page is relevant exceeds a specified threshold then the respective URL is included to the returned set of URLs.

On the other hand, Web directories are explored (sub-trees are visited by the Crawler) and the contained URLs are collected.

The totalities of collected URLs from all sources are merged and a final URLs list is returned. Merging process minds to a) remove possible duplicates and b) ignore sub-paths of URLs already in list. Finally, URLs having already a quality label (Crawler consults the MedIEQ repository for this) are also removed.


### **How does the AQUA Spider work?**

How does a spider fetch all web pages? The only way to collect links to new pages (URLs) is to scan already collected pages for hyperlinks that have not been collected yet. This is the basic principle of crawlers/spiders. They start from a given set of URLs, progressively fetch and scan them for new URLs and then fetch these pages in turn, in an endless cycle.

In MedIEQ, the Spider investigates only specific web sites collected by the Crawler (to ensure spidering of web sites only from health domain) and it will follow only internal links (links pointing to new pages on the same site). Unlike general spiders this process is finite and potentially pending work is not so much hardware consuming.

Spider's first version follows only explicit links (text and image links, image maps), omitting links inside or generated by javascript, flash graphics and web forms (such cases will be examined in future Spider versions). The Spider examines sites from the Crawler output one-by-one in several independent threads. Unreachable sites/pages are revisited in next run.

Because not all web pages of a web site are interesting for the labelling process, the spider utilizes a content classification component which consists of a number of classification modules which in turn decide which pages contain interesting information and which not. Each of these modules relies on a different classification method according to the classification problem on which it is applied. The "interesting" information can be directly extracted during the classification process (e.g. whether the document contains advertisement or not, or which is the target audience of the document) or the resource can be stored locally in order to be used by the Information Extraction Component.

In very similar way spider interacts with link-scoring component. Modules in this component are analysing links or rather link-objects (link text, its attributes, context elements etc.) even re visiting target page. In some cases link-scoring component is able to decide, that target page contains no "interesting" information just according to content of link-object and the link is therefore not visited. This could save HW and time resources.

## **5. Evaluation methodology and separate results**

### **The evaluation methodology**

As far as statistical classification in spidering is concerned, three different classifiers provided by Weka<sup>6</sup> classification platform have been tested. These are SMO<sup>7</sup>, Naïve Bayes<sup>8</sup> and

---

<sup>6</sup> <http://www.cs.waikato.ac.nz/ml/weka/>

Flexible Bayes<sup>9</sup> (Naïve Bayes with kernel estimation). These classifiers have been tested for two-class classification problems. In case where the problem was multi-class by nature it has been reduced to two-class problem by considering as positive class one of the categories and as negative class the remaining categories.

Analytically, the training and testing procedure has as follows:

1. An annotated corpus is created manually.
2. The HTML pages<sup>10</sup> are preprocessed and tokenized.
3. Each document is ultimately represented as a vector  $\langle x_1, \dots, x_m \rangle$ , where  $x_1, \dots, x_m$  are the values of attributes  $X_1, \dots, X_m$ , and each attribute provides information about a particular n-gram of the document. We have performed the experiments with 1-grams and 1/2/3-grams<sup>11</sup>. The value of each attribute in the vector is the normalized term frequency, which we get if we divide term (n-gram) frequencies by the total number of n-gram occurrences in the document. Following common text classification practice, we do not assign attributes to tokens that are too rare (we discard tokens that do not occur in at least 5 pages from the training set). We also rank the remaining attributes by information gain<sup>12</sup>, and use only the 1000 best.
4. The Weka classifiers are trained and tested using the vector representations of the documents.

A number of different classification modules are used for the classification during the spidering process according to the type of the classification task and the performance of each classification method at each task.

Four of the quality labeling criteria that MedIEQ examines, and the approach we followed in order to detect their presence in web sites, are presented bellow:

Criterion	MedIEQ approach
The target audience of a web site or web document should be clear	Classification among three possible target groups: adults, children and professionals
Contact information of the responsible of a website or the author of a resource should be present	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

All the above criteria have been examined using statistical classification techniques. In addition the presence of advertisement in websites was examined using a heuristic detection method which is described later in this section.

---

<sup>7</sup> *SMO classifier*: Implements John C. Platt's [17] sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels.

<sup>8</sup> *NaiveBayes classifier*: Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. [12]

<sup>9</sup> *FlexibleBayes (NaiveBayes -K) classifier*: Class for a Naive Bayes classifier using kernel estimation for modeling numeric attributes rather than a single normal distribution.

<sup>10</sup> In future versions other document formats (e.g. pdf, doc) will be also used.

<sup>11</sup> All 1-grams, 2-grams and 3-grams are produced and the best of them according to information gain are selected. The final list of the selected n-grams may contain 1-grams, 2-grams and 3-grams.

<sup>12</sup> [explanation]

For the statistical classification pre-annotated corpora were used. The HTML pages<sup>13</sup> were preprocessed and tokenized in two different ways. In the first way, all HTML tags were removed and only the clean textual content of the document was used for the classification. In the second way, both HTML tags and textual content were used. Punctuation marks and white characters are used as delimiters for the tokenization process.

Heuristic classification was used only for the advertisement detection. 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.

Finally, in order to do the classification in crawling, we utilized a content classification module which had been previously trained on an annotated corpus. We have tested the classification performance of the crawler using a corpus of about 2600 English medical web sites which have been gathered by our crawler and have been manually annotated as health related or not. 1600 of them were used for training and the rest of them for testing. For the classification the Weka SMO classifier was used and during the tokenization process all the HTML tags of the HTML documents were removed.

## Classification evaluation results

### Crawling evaluation results

The evaluation results are presented in the following table.

<Table of results goes here>



### Spidering evaluation results

The classification performance results are presented below.

		1-grams						1/2/3-grams					
		Tags removed			Tags not-rem.			Tags removed			Tags not-rem.		
		Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.
Adults	Naïve Bayes	0.66	0.91	0.77	0.58	0.92	0.71						
	Flex. Bayes	0.71	0.78	0.75	0.74	0.79	0.76						
	SMO	0.81	0.79	0.80	0.84	0.84	0.84						
Childr	Naïve Bayes	0.94	0.87	0.90	0.76	0.90	0.83						
	Flex. Bayes	0.96	0.82	0.88	0.77	0.89	0.82						
	SMO	0.93	0.91	0.92	0.97	0.84	0.90						
Prof.	Naïve Bayes	0.94	0.87	0.90	0.76	0.90	0.83						
	Flex. Bayes	0.96	0.82	0.88	0.77	0.89	0.82						
	SMO	0.93	0.91	0.92	0.97	0.84	0.90						

Table 1 – Target audience (Adults: 102 / Children: 98 / Professionals: 96 samples )

		1-grams						1/2/3-grams					
		Tags removed			Tags not-remov.			Tags removed			Tags not-rem.		
		Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.
Naïve Bayes	0.75	0.90	0.72	0.78	0.90	0.83							
Flex. Bayes	0.73	0.90	0.81	0.74	0.87	0.80							
SMO	0.84	0.84	0.84	0.85	0.77	0.81							

Table 2 – Contact info (109 positive samples / 98 negative samples)

<sup>13</sup> In future versions other document formats (e.g. pdf, doc) will be also used.



	1-grams						1/2/3-grams					
	Tags removed			Tags not-remov.			Tags removed			Tags not-rem.		
	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.
Naive Bayes	0.78	0.87	0.83	0.82	0.88	0.85						
Flex. Bayes	0.75	0.92	0.83	0.78	0.87	0.83						
SMO	0.90	0.82	0.86	0.88	0.81	0.84						

Table 3 – Virtual Consultation (100 positive samples / 101 negative samples)

	1-grams						1/2/3-grams					
	Tags removed			Tags not-remov.			Tags removed			Tags not-rem.		
	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.	Prec.	Rec.	F-m.
Naive Bayes	0.93	0.83	0.88	0.88	0.85	0.86						
Flex. Bayes	0.88	0.91	0.89	0.87	0.83	0.85						
SMO	0.89	0.89	0.89	0.87	0.81	0.83						

Table 4a (Statistical classification) – Advertisements (100 positive samples / 104 negative samples)

	Precision	Recall	F-measure
Heuristic classification	0.84	0.72	0.78

Table 4b (Heuristic classification) – Advertisements

## 6. Conclusion and Future work

Future work:

- Examine more classification algorithms
- Use structure properties and meta-information of HTML documents for classification
- Handle different types of documents (e.g. pdf, doc, etc.)
- Incorporate more heuristic methods for classification
- Use Link Scoring to speed-up the spidering process
- Test classification in other languages
- Experiment with larger corpora
- Test the classification performance in other quality criteria.

....

## Acknowledgements

The authors would like to thank Vassiliki Rentoumi, Irene ..., Dimitris Bilidas and Tasos ... for their support at the realization of the experiments upon which this paper has been based.

## 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] Baujard O., Baujard V., Aurel S., Boyer C. and Appel R. D.: Trends in medical information retrieval on internet, Computers in Biology and Medicine, Volume 28, Issue 5, September 1998, pages 589-601.
- [3] Boldi P., Codenotti B., Santini M., and Vigna S.: UbiCrawler: a Scalable Fully Distributed Web Crawler. Software -Practice and Experience (SPE), 34(8):711- 726, 2004.
- [4] Brin S. and Page L.: The Anatomy of a Large Scale Hyper-textual Web Search Engine. 7th International World Wide Web Conference (WWW), Brisbane, Australia, 1998.

- [5] 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.
- [6] Chakrabarti S.: *Mining the Web: Discovering Knowledge from Web Data*. Morgan Kaufmann, 2003.
- [7] Curro V., Buonomo 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.
- [8] Diligenti M., Coetzee F., Lawrence S., Giles C.L. and Gori M.: Focused Crawling Using Context Graphs. 26th International Conference on Very Large Data Bases (VLDB), Cairo, Egypt, pages 527-534, 2000.
- [9] Ehrig M. and Maedche A.: *Ontology-focused crawling of web documents*, 2003.
- [10] 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.
- [11] Heydon A. and Najork M.: Mercator: A Scalable, Extensible Web Crawler. *World Wide Web*, 2(4):219-229, 1999.
- [12] John G. H. and Langley P.: Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. pp. 338-345. Morgan Kaufmann, San Mateo, 1995.
- [13] Kohler C., Darmoni S. D., Mayer M. A., Roth-Berghofer T., Fiene M., Eysenbach G.: MedCIRCLE - The Collaboration for Internet Rating, Certification, Labelling, and Evaluation of Health Information. *Technology and Health Care, Special Issue: Quality e-Health. Technol Health Care* 10(6) (2002), 515.
- [14] Mayer M. A., Leis A., Sarrias R., Ruíz P.: Web Médica Acreditada Guidelines: reliability and quality of health information on Spanish-Language websites. In: Engelbrecht R et al. (ed.). *Connecting Medical Informatics and Bioinformatics. Proc of MIE2005* (2005), 1287-92.
- [15] 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.
- [16] Pant G., Bradshaw S. and Menczer F.: Search Engine -Crawler Symbiosis: Adapting to Community Interests. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Trondheim, Norway, pages 221-232, 2003.
- [17] Platt J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.
- [18] Shkapenyuk V. and Suel T.: Design and Implementation of a High-Performance Distributed Web Crawler. 18th International Conference on Data Engineering (ICDE), San Jose, USA, pages 357-368, 2002.
- [19] Sizov S., Biwer M., Graupmann J., Siersdorfer S., Theobald M., Weikum G. and Zimmer P.: The BINGO! System for Information Portal Generation and Expert Web Search. 1st Conference on Innovative Systems Research (CIDR), Asilomar, USA, 2003.
- [20] Stamatakis K., Karkaletsis V., Paliouras G., Horlock J., Grover C., Curran J., Dingare S.: Domain Specific Web Site Identification: The CROSSMARC Focused Web Crawler, In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, Edinburgh, UK, 2003.
- [21] Winker M. A., Flanagan A., Chi-Lum B.: Guidelines for Medical and Health Information Sites on the Internet: principles governing AMA web sites. *American Medical Association. JAMA* 283 (12) (2000), 1600-1606.
- [22] CISMef: Catalogue et Index des Sites Medicaux Francophones. <http://www.chu-rouen.fr/cismef/>
- [23] European Commission. eEurope 2002: Quality Criteria for Health related Websites. [http://europa.eu.int/information\\_society/europe/ehealth/doc/communication\\_acte\\_en\\_fin.pdf](http://europa.eu.int/information_society/europe/ehealth/doc/communication_acte_en_fin.pdf).
- [24] Hi-Ethics, Inc.: Health Internet Ethics. Ethical Principles for offering Internet Health services to consumers. <http://www.hiethics.com/Principles/index.asp>
- [25] HON: Health on the Net Foundation. <http://www.hon.ch>

[26] URAC: Health Web Site Accreditation.  
<http://webapps.urac.org/websiteaccreditation/default.htm>