Combination of Heterogeneous Criteria for the Automatic Detection of Ethical Principles on Health Web sites

Arnaud Gaudinat, MS, Natalia Grabar, PhD, Célia Boyer, MS
Health on the Net Foundation, SIM/HUG, 24 rue Micheli-du-Chrest, Geneva, Switzerland

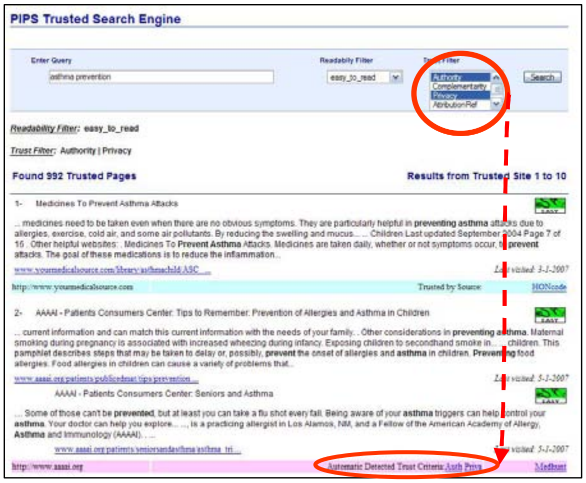
Abstract

The detection of ethical issues of web sites aims at selection of information helpful to the reader and is an important concern in medical informatics. Indeed, with the ever-increasing volume of online health information, coupled with its uneven reliability and quality, the public should be aware about the quality of information available online. In order to address this issue, we propose methods for the automatic detection of statements related to ethical principles such as those of the HONcode. For the detection of these statements, we combine two kinds of heterogeneous information: content-based categorizations and URL-based categorizations through application of the machine learning algorithms. Our objective is to observe the quality of categorization through URL's for web pages where categorization through content has been proven to be not precise enough. The results obtained indicate that only some of the principles were better processed.

Introduction

The detection of ethically correct health web sites is an important issue as the quality and reliability of online health documents are uneven. The concern is more so due to the growing number of users accessing online health information. In fact, out of ten Internet users, eight access healths related information[1]. Moreover, such searches are related to the health condition of the users themselves or of their family members. Thus, the information found can have a direct impact on the users' health and well-being, their healthcare and their communication with medical professionals. It becomes essential therefore, to standardize the available information by the application of an ethical code and indicate clearly, all sites respecting this code. A search engine with quality criteria filters would then be used to distinguish reliable information from the information pool. Such a prototype is presented in Figure 1, which displays the result of a search for 'Asthma Prevention'.

- The main ethical principle emphasized below is *Privacy* (the provision of a privacy policy on a web site). As shown, if a privacy policy has been provided for the site, the corresponding page/s are indicated under '*Automatic Detected Trust Criteria*.' On clicking this link, the user is then able to access the privacy policy of the site.
- As was done before, the key words used (*asthma prevention*) and other words relevant to this query are shown in bold.



Readability Filter: easy_to_read

Trust Filter: Authority | Privacy

Found 992 Trusted Pages **Results from Trusted Site 1 to 10**

1- Medicines To Prevent Asthma Attacks

... medicines need to be taken even when there are no obvious symptoms. They are particularly helpful in **preventing asthma** attacks due to allergies, exercise, cold air, and some air pollutants. By reducing the swelling and mucus... ... Children Last updated September 2004 Page 7 of 16 . Other helpful websites: . Medicines To **Prevent Asthma** Attacks. Medicines are taken daily, whether or not symptoms occur, to **prevent** attacks. The goal of these medications is to reduce the inflammation...

www.yourmedicalsource.com/library/asthmachild/ASC ... *Last visited: 3-1-2007*

http://www.yourmedicalsource.com Trusted by Source: HONcode

2- AAAAI - Patients Consumers Center: Tips to Remember: Prevention of Allergies and Asthma in Children

... current information and can match this current information with the needs of your family. . Other considerations in **preventing asthma**. Maternal smoking during pregnancy is associated with increased wheezing during infancy. Exposing children to secondhand smoke in... ... children. This pamphlet describes steps that may be taken to delay or, possibly, **prevent** the onset of allergies and **asthma** in children. **Preventing** food allergies. Food allergies in children can cause a variety of problems that...

www.aaaai.org/patients/publicedmat/tips/prevention ... *Last visited: 5-1-2007*

AAAAI - Patients Consumers Center: Seniors and Asthma

... Some of those can't be **prevented**, but at least you can take a flu shot every fall. Being aware of your **asthma** triggers can help control your **asthma**. Your doctor can help you explore... ..., is a practicing allergist in Los Alamos, NM, and a Fellow of the American Academy of Allergy, **Asthma** and Immunology (AAAAI). . ...

www.aaaai.org/patients/seniorsandasthma/asthma_tri... *Last visited: 5-1-2007*

http://www.aaaai.org Automatic Detected Trust Criteria: Auth Priva Medhunt

Figure 1. Automatic detection of pages addressing the privacy policy and authority within a search engine.

Various initiatives have previously been proposed for the quality control of health information on the Internet including referencing of information, accreditation of a site, popularity of a site, collaboration of users, self-regulation and the user's education. At the Health on the Net Foundation (*www.HealthOnNet.org*), we have adopted an accreditation program performed through third party evaluation of health web sites according to the Ethical Code of Conduct, the HONcode[2]. The Code is composed of eight ethical principles, namely *authority*, *complementarity*, *privacy*, *reference*, *justifiability*, *transparency*, *sponsorship* and *advertising*. Currently, the HONcode accredited database contains over 5,500 web sites (over

1,200,000 web pages) in 32 languages. Web sites requesting accreditation are evaluated and treated accordingly, while already accredited sites are reviewed at least once a year to confirm continuing compliance. A dated seal and certificate, both being valid for one year, support confirmation of accreditation. Each web site is evaluated by experts at HON who confirm the compliance of a site, for each of the eight HONcode principles. This evaluation is currently done manually, thus guarantying accurate results. However, this approach is time consuming, especially considering the ever-growing quantity of online health information.

The purpose of our work is to take advantage of the existing database with quality-annotated web sites (sites already evaluated by the HONcode Team) and to propose a method suitable for the automatic detection of reliable health web sites through the application of the HONcode ethical principles. Previous research has shown successful application of regular expressions[3] or detection of the HONcode label[4] within web pages. We chose to apply supervised learning methods as they allow better characterization of expected categories related to principles. Indeed, such methods allow the formulation of textual events with more precision and refinement, in addition to capturing events, which would otherwise have not been detected by manually defined regular expressions. Moreover, categorization methods are helpful in automatic detection of specific textual documents, e.g., the detection of hostile messages[5], documents with racial implications[6], readability level of documents[7] and spam[8]. In previous work, we proposed an automatic tool for the categorization of web pages according to the HONcode principles, which relies on the analysis of the content of web documents[9] and their corresponding URL addresses (the URL is the Uniform Resource Locator which indicates the location of a web page on the Internet). E.g. The URL name *anatome.ncl.ac.uk /tutorials/privacy.html* has a corresponding textual content that is unique to that specific URL. Thus it can be seen that these two variables (URL and Content) are closely linked although having different functions and status. The use of keyword information within a URL to improve precision of a web system is not new; it is largely practiced in the web search engine community such as Google.

We propose to combine these two heterogeneous data to further facilitate and provide more accuracy to the automatic categorization of pages related to the ethical HONcode principles. A study was carried out to practically test this theory. We present a description of this study, including material and methods used, results obtained and analysis performed.

Material and Method

We outlined two main steps of the method, which we propose for the categorization of web pages according to HON principles. It is based on two kinds of language models:

1. Creation of two separate language models further to the analysis of documents' content and URL names with learning algorithms;
2. Combination of scores obtained for the two categorization models

Creation of language models

The automatic categorization method considers documents as vectors within a vector space. The dimension of this space depends on the number of units (often words) within the whole collection of documents, whereas the size of each vector corresponds to the frequency of a given unit in a document.

Different learning algorithms were separately applied to the two variables (URL and Content) as was seen in a previous study, where we used two machine learning algorithms from several proposed by our learning framework[10] Naive Bayes (NB) and Support Vector Machine (SVM).

Features tested were the following: (1) with or without use of stop words; (2) with or without application of stemming algorithm[11] in order to lexically normalize words, e.g. {*treating*, *treat*}; (3) learning unit set up to the word combination (e.g. n-grams of 1 to 4 words); (4) learning unit set up to the word co-occurrences within a sentence, or a group of words; (5) learning unit set up to the character combination (e.g. n-grams of 1 to 5 characters). Features 1 to 4 were used for the categorization of document content and features 4 and 5 for the categorization of URL names. Categorization of content and URL are different and specific to the aimed units (i.e. tokenization, normalization) [12,13,14].

Different combinations of features and categorization algorithms have been applied to the processed data in English[15], where the material size is the most important. Evaluation has previously been performed with four measures in their micro and macro versions: namely precision, recall, F-measure and error rate. Macro precision (maP) is representative of the distribution of features in each category (principle), while micro precision (miP), in each processed unit (URLs).*

On the basis of this evaluation, models produced by SVM algorithm were chosen for both content (with word unit feature) and URL names (with 5-gram character unit feature) categorization. Following the production of these models by the SVM algorithm, they were then combined.

Combination

The test material was first categorized separately by the language models i.e. content and URL models. As the Content model has previously been applied in similar tasks, it was easily recognized and corresponded to the standard level in the processing of documents. However, it was more complex with the URL model where the exact role of URL names is yet to be established. In previous studies done, we observed that URL names do indeed play an important role in the identification of web pages related to HONcode principles. Our goal this time was to study in more depth. Here, scores obtained with each model were combined in three ways according to the weight given to URL-based categorization: URL's were (1) not taken into account; (2) they were taken into account with 30% ratio; (3) taken into account with 100% ratio. For all of these cases, the content-based categorization was fully considered. The objective was to find out which configuration produced the most precise results. As a result of this ambitious aim, we were able to show improved accuracy of search results through the combination of this heterogeneous information:

- The linguistic and semiotic nature of content and URL names of web pages are different. The role of content is to convey and transmit information on a given topic while the role of URLs is to indicate the location of a web page on the Internet. In the study, the textual content was analyzed by the sentences and phrases and the URL names was analyzed through name and each entity of the name.
- Content and URL names do not represent the same entity. While the URL model is represented by the whole web page, the Content model is extracted from the analysis of sentences on the web page. Information pertaining to ethical criteria may appear anywhere on the page, thus making the Content model a more suitable unit.

However, our intention was to find out whether these two variables (URL and Content) possess features relevant to the discrimination of ethical principles, and whether their combination provides complementary performance.

Evaluation

For the evaluation, though the proposed method was applied to all the HONcode accredited web sites, only a part of this extensive database was employed as learning material. Automatically produced classification was compared with the manual review of sites. More specifically, we evaluated the method with web pages which have the tendency to show a low rank compared to the manual review. For each site and criterion pair, the system proposed a list of pages ranked by the class score. In fine, the goal consists to find the manual selected page at the first rank for the system. We analyzed those web sites where, with the content-based categorization, the reference pages have the lowest scores among top nine ranks. Five web sites were randomly selected for each principle. We then observed the classification and scores of the manually selected pages, analyzing in particular how the combination of content and URL based categorizations allows improvement of ranking and so of classification of web pages and how such heterogeneous information could be applied for the detection of HONcode related web pages.

Results and Discussion

Our study was based on a set of 35 web sites where 351 pages were processed in total.

We analyze here average ranks and scores of classification. The two figures shown below (**Figure 2** and **Figure 3**) display from left to right, the three types of combinations used for each principle. i.e. (1) without URL-based categorization, (2) with URL-based categorization taken into account at 30% ratio, (3) with URL-based categorization taken into account at 100%.

Processing and evaluation were performed for seven HONcode principles: *h1-authority*, *h2-complementarity*, *h3-privacy*, *h4-reference*, *h6-transparency*, *h7-sponsorship* and *h8-advertising*. Two principles, *h5-justifiability* and *h4-date* were excluded from this evaluation, the reason being that detection of *h5-justifiability* appeared to be too difficult and *h4-date* (which along with h4-reference make up the *h4 principle*) have already produced near-perfect results with the content-based

* Further details can be found in previously presented work.

Average rank

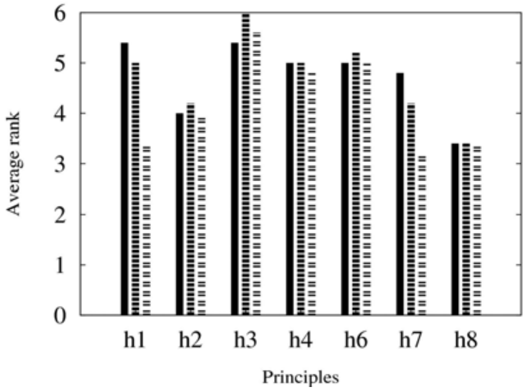


Figure 2: **Mean rank per principle.**

Figure 2 displays the mean rank obtained for each principle. The lower the rank, the better the position of classification (i.e. Rank one would be considered to have the best position). Expected outcome was a figure showing gradation with categorization without URL's at the lowest section and 100% use of URL's at the other end. However, the actual outcome was quite different. Principles *h1-autority*, *h7-sponsorship* and *h4-reference* to a certain degree, did in fact display the expected results. However, with principle *h8-advertising*, ranking was stable and with the remaining three (*h2-complementary*, *h3-privacy* and *h6-transparency*), there was no transitional gradation seen at all. This is due to the fact that web pages obtaining the same scores were given the same rank. Thus, it appears that the rank by itself appears to be an unreliable representative for the evaluation of the classification performances.

Average score

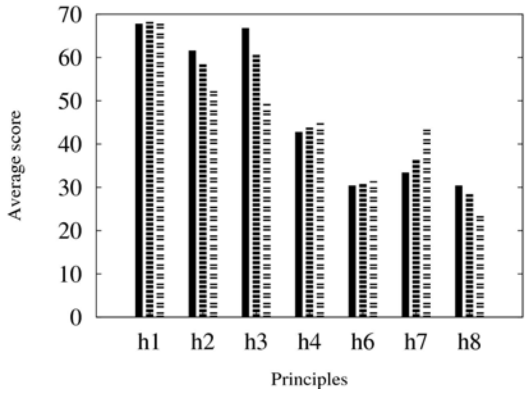


Figure 3: Mean score per principle.

Figure 3 presents the mean scores obtained for each principle from the manually selected web pages. A higher score indicates that the relevant web page has been classified with a better score. Here again, the expected outcome was a gradation pattern for the three combinations.

Unlike results seen for the rank, those obtained for the score was as expected. As can be seen in Figure 3, each principle shows a gradation between the three combinations, either positive or negative. For principles *h2-complementarity*, *h3-privacy* and *h8-advertising*, performance decreased as the percentage of URL scores taken into account increased, *h1-authority*, *h4-reference* and *h7-sponsorship* showed an increasing trend while *h6-transparency* was almost stable with a minute positive gradation.

After analyzing these results, it can be seen that only three principles (*h1-authority*, *h4-reference* and *h7-sponsorship*) show a definite advantage when content-based and URL-based categorization are combined. It was also observed that the same result was produced with ranking of the pages for the above three principles. We found these results to be surprising; as previous results had demonstrated the efficiency of URL-based categorization, with special emphasis on principles *h3-privacy* and *h6-transparency*.

Gold standard

A more detailed analysis of both manually selected pages and pages well classified by the automatic system show that the gold standard gives a non exhaustive view of relevant pages. It was seen that in a given web site, few or even several pages related to a given principle can exist, while only one of them is detected and recorded by reviewers. In this respect, the automatic system detects them more systematically and would facilitate browsing of sites by reviewers as a result of providing more in-depth and complete information about a site.

Conclusion and Perspectives

In this paper, we presented a novel approach for the categorization of web pages according to the quality and ethical policy/ies of web sites. We used the HONcode accredited database and a SVM machine learning algorithm. This algorithm was applied separately, to the content of web pages and their URL names. The two consequently generated language models were then combined in three ways according to the percentage given to the score of URL-based categorization: 0%, 30% and 100%. The purpose was to observe whether the use of URL-based scores would help to improve categorization of web pages that are otherwise not categorized well by the content-based language model. The results obtained show that when URL scores were taken into account, performances with principles *h2-complementarity*, *h3-privacy* and *h8-advertising* decreased, with *h1-authority*, *h4-transparency* and *h7-sponsorship* increased, and remain stable with *h6-reference*. The conclusion is that the combination of URL scores was only advantageous for three principles out of seven which was a surprising find, as previous studies done showed URL-based categorization to be particularly efficient with *h3-privacy* and *h6-transparency* principles. However, it could be argued that in this study, only pages categorized as difficult were used, thus leading to a fallacious result. It was also observed that the combination of these two language models is only one of the factors which influenced our results. Indeed, the results depended on several factors: (1) performance of the two language models when applied to web pages; (2) Ratio of combination (proportion given to URL score); (3) learning material; (4) the gold standard. In the future work done, we propose to employ other means of combining these two models and comparing results obtained and our ultimate goal is to develop a search engine sophisticated enough to automatically extract all the ethical information found on a site, with regards to the HONcode principles.

Acknowledgments

This work has been realized as part of the PIPS[16] (Personalized Information Platform for Life & Health) project funded by the European Commission program *Information society technology*, contract number 507019.

References

1. Fox S. Online Health Search 2006. Most Internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find. Technical report, Pew Internet & American Life Project, Washington DC, 2006.
2. Boyer C, Baujard O, Baujard V, et al. Health on the net automated database of health and medical information. Int J Med Inform 1997;47(1-2):279.
3. Wang Y and Liu Z. Automatic detecting indicators for quality of health information on the web. International Journal of Medical Informatics 2006.
4. Price S and Hersh W. Filtering web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web. In: AMIA 1999, 1999:9115.
5. Spertus E. Smokey: automatic recognition of hostile messages. In: American Association for Artificial Intelligence, 1997:105865.
6. Vinot R, Grabar N, and Valette M. Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In: TALN, e 2003:25784.
7. Wang Y. Automatic recognition of text difficulty from consumers health information. In: IEEE , ed, ComputerBased Medical Systems, 2006.
8. Carreras X and Márquez L. Boosting trees for anti-spam a email filtering. In: Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, BG. 2001.
9. Gaudinat A, Grabar N, and Boyer C. Machine learning approach for automatic quality criteria detection of health webpages. In: McCray A, ed, MEDINFO 2007, Brisbane, Australia. 2007. To appear.
10. Williams K and Calvo RA. A framework for text categorization. In: 7th Australian document computing symposium, 2002.
11. Porter M. An algorithm for suffix stripping. Program 1980;14(3):1307.
12. Salton G. Developments in automatic text retrieval. Science 1991;253:9749.
13. Singhal A, Salton G, Mitra M, and Buckley C. Document length normalization. Information Processing & Management 1996;32(5):61933.
14. Koller D and Sahami M. Toward optimal feature selection. In: International Conference on Machine Learning, 1996:28492.
15. Yang Y and Liu X. Re-examination of text categorisation methods. In: Proc of 22nd Annual International SIGIR, Berkley. 1999:429.
16. http://www.pips.eu. org