Automatic retrieval of web pages with standards of ethics and trustworthiness within a medical portal: What a page name tells us

Arnaud Gaudinat, Natalia Grabar, Célia Boyer

Health on the Net Foundation, SIM/HUG, Geneva, Switzerland
name.surname@healthonnet.org

Abstract. The ever-increasing volume of health online information, coupled with the uneven reliability and quality, may have considerable implications for the citizen. In order to address this issue, we propose to use, within a general or specialised search engine, standards for identifying the reliability of online documents. Standards used are those related to the ethics as well as trustworthiness of websites. In this research, they are detected through the URL names of webpages by applying machine learning algorithms. According to algorithms used and to principles, our straightforward approach shows up to 93% of precision and 91% of recall. But a few principles remain difficult to recognize.

1 Introduction

The issue related to the quality of online information is important, specially in the medical area, as eight Internet users out of ten look for health information [1] and as often such searches are directly linked to their own health condition or to their relatives. But the quality and reliability of proposed online health documents are uneven and we assume that this should be clearly indicated to users. Various initiatives exist for the quality control assessment of health information on Internet [2]. At the Health on the Net Foundation (*www.hon.ch*), we have adopted an accreditation program through the third party evaluation of health website's reliability done according to the Ethical Code of Conduct HONcode [3]. The Code is composed of eight ethical principles, namely *authority*, *complementarity*, *privacy*, *reference*, *justifiability*, *transparency*, *sponsorship* and *advertising*. Each website, which asks for the accreditation, is evaluated by HON's experts in order to check whether it provides clear statements for these principles. Up to now, the HONcode accredited database contains over 1'200'000 webpages in 32 languages. When perfomed manually the accreditation process guarantees high quality results but must cope with the increasing number of online health information. In this work, we want to take advantage of the database with quality annotated websites and to propose a method and data suitable for the automatic detection of health websites' quality.

We apply supervised learning methods: they allow to better characterise and constrain expected categories related to the eight HONcode criteria. In previous

Table 1. Learning data: numbers of URLs used for generation of the language model in English.

Principle	Meaning	Total	Learning	Evaluation
HC1	Authority	2843	2571	272
HC2	Complementary	2470	2218	252
HC3	Privacy	2374	2115	259
HC4	Reference	1855	1674	181
HC5	Justifiability	460	407	53
HC6	Transparency	2539	2275	264
HC7	Sponsorship	2088	1893	195
HC8	Advertising	1545	1389	156
HC9	Date	1545	1378	167

research, regular expression [4] or presence of HONcode label [5], have been used. Comparing to these, supervised learning methods allow to formalize textual events with more precision, and to capture events which would be not detected by humans. Moreover, categorisation methods shown to be helpful in automatic systems working with textual documents [6, 7]. In previous work, we proposed an automatic tool for the categorization of webpages according to the HONcode principles on the basis of documents' content [8]. We propose now to apply similar method for the categorization of documents through their URL addresses.

2 Material

A key component of any system for the automatic text categorisation is a knowledge base with positive examples. In this work, we use the name of URL webpage. URL is the *Uniform Resource Locator* which indicates the location of a webpage on Internet. Each URL is unique. URL begins with the scheme name that defines its namespace, while the remaining part of the URL corresponds to the hierarchical structure of website and to the name of file. The reason to use URLs as material for the categorisation of webpages is that they can be composed with keywords related to the HONcode principles. Here, a few examples of URL names registered for the *privacy* principle within the HONcode accredited database:

anatome.ncl.ac.uk/tutorials/privacy.html
www.wmcnet.org/workfiles/media/noticeofprivacyplan.pdf
parathyroid.com/disclaimer.htm, *www.vh.org/welcome/help/vhpolicies.html*

The learning dataset is composed of over 12'623 URLs of some HONcode accredited English sites. Table 1 indicates number and title of principles, and number of the URLs used in our work. As the principle HC4 *reference* covers heterogeneous information (reference to date and reference to statement on clinical trials, etc.), it has been separated into two sets, and *date* has been exported. Additionally, notice that some of the URLs recorded can cover up to 5 principles.

3 Method

Pre-processing of material. For the detection of significant parts of URLs, for instance, *www.hon.ch/confidentiality_page/privacy_disclaimer.html*, we split them two parts, *inurl* and *endurl*:

- *inurl*, when exists, corresponds to the directory names in which the file is located. It includes the entire directory pathway except the domain and file names. In the example above, *inurl* is *confidentiality_page*
- *endurl*, corresponds to the file name: *privacy_disclaimer*

inurl and *endurl* are segmented on non alphanumeric characters, *ie.* _ - ? / =

Training step. Machine learning algorithms used are those proposed by our learning framework [9]: Naive Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbours (kNN) and Decision Tree (DT). Different combinations of features and categorisation algorithms have been applied to data in English. *Features* tested within the learning process are the following: (1) word combination (*e.g.* n-grams of 1 to 4 words); and (2) character combination (*e.g.* n-grams of 1 to 5 characters). *Unit weight* is defined by three elements [10]: term frequency, inverse document frequency and length normalisation. *Features selection*, which aims at reducing vector-space dimension through selection of the most discriminatory features, is performed with document frequency (DF) [11].

Evaluation. We used 10% of our corpora for the evaluation task, the 90% being used exclusively for the learning task. These two corpora are independent. Evaluation is performed with four measures in their micro and macro versions: precision, recall, F-measure and error rate. Macro precision (maP) is representative of the distribution of features in each category (principle), and micro precision (miP) in each processed unit (URLs).

4 Results and Discussion

The method has been applied to three sets of material:

- *inurl*: learning and evaluation performed on *inurl* parts of URLs;
- *endurl*: learning and evaluation performed on *endurl* parts of URLs;
- *red*: learning and evaluation performed on reduced set of 6 principles: *authority*, *complementary*, *privacy*, *transparency*, *sponsorship* and *advertising*.

Among all the methods, features and weightings indicated, we present only those which show significant differences between them, namely: two learning methods (*NB* and *SVM*); two features (single words *w1* and 5-grams of characters *c5*). Figure 1 presents figures for the average precision and recall obtained with these three sets, and we can distinguish three clusters: (1) *inurl* set, which provides the less performing results for both recall and precision; (2) *NB* method generates good recall figures (with *endurl* and *red* sets); (3) *SVM* method generates good precision figures (with *endurl* and *red* sets). We assume that merging these two methods, SVM and NB, can be interesting: NB can guarantee better recall and

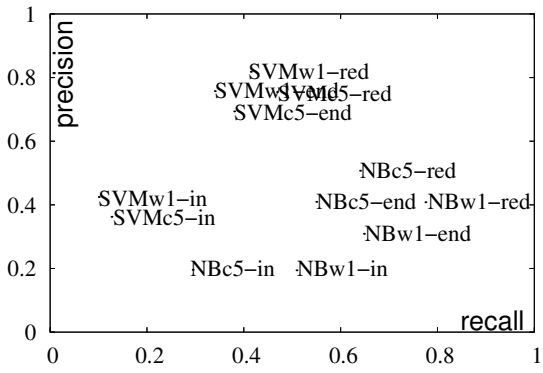

Fig. 1. Micro precision and micro recall of various methods applied.

Table 2. Categorisation of URL names (*endurl*). Recall/Precision contingency of quality criteria. System setting: method *SVM*, language *English*, single word *w1*

Auto/Man	Authority	Compl.	Privacy	Refer.	Justif.	Trans.	Spon.	Adver.	Date
Authority	**68/84**	7/6	2/1	2/5	0/0	7/5	11/36	2/1	1/9
Compl.	5/4	**64/38**	7/4	1/3	5/29	5/3	1/3	12/33	0/0
Privacy	2/2	2/2	**93/87**	0/0	0/0	0/1	1/3	2/7	0/0
Refer.	4/2	4/1	4/1	**51/62**	9/29	0/0	2/3	0/0	24/48
Justif.	25/1	25/1	0/0	0/0	**25/7**	0/0	0/0	0/0	25/4
Trans.	1/2	1/1	3/3	1/5	1/7	**91/91**	1/3	1/3	1/9
Spon.	8/2	4/1	0/0	8/5	0/0	4/1	**76/53**	0/0	0/0
Adver.	8/2	15/3	12/2	8/5	4/7	0/0	0/0	**50/47**	0/0
Date	5/1	5/1	5/1	26/14	16/21	5/1	0/0	0/0	**37/30**

SVM better precision. Furthermore, we can observe that, not surprisingly, *red* set allows to generate better results than *endurl*.

Table 2 indicates the precision/recall confusion matrix obtained with *SVM* algorithm applied to single words *w1* from *endurl* with no weighting. We can observe that out of nine criteria the following ones could be processed with good results: *transparency* (91%/91%), *privacy* (93%/87%), *authority* (68%/84%), and *sponsorship* (76%/53%). Webpages related to principles *complementarity* and *advertising* are categorised with mean performances. As for three remaining principles (*justifiability*, *date* and *reference*) they show scarce performances, and other approaches should be applied for their detection.

5 Conclusion and Perspectives

In this paper, we presented a novel approach for the categorisation of webpages according to the quality and announced ethical policy of the websites. We exploit for this the HONcode accredited database and two machine learning algorithms (SVM and Naive Bayes). These algorithms have been applied to URL names of webpages and show competitive results, up to 93% of precision and 91% of recall according to principles. URL-based categorisation can be thus run separately or in combination with the content analysis. In our further work, our special attention should be given to the combination of both approaches (URL and content based), to two hard to modelise principles (*reference* and *justification*), and to the visualisation of the quality information within a search engine.

Acknowledgement. This work has been realised as part of the PIPS (Personalised Information Platform for Life & Health) project funded by the European Commission programme *Information society technology*, contract nb 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. Risk, A., Dzenowagis, J.: Review of internet information quality initiatives. Journal of Medical Internet Research **3**(4) (2001) e28
3. Boyer, C., Baujard, O., Baujard, V., Aurel, S., Selby, M., Appel, R.: Health on the net automated database of health and medical information. Int J Med Inform **47**(1-2) (1997) 27–9
4. Wang, Y., Liu, Z.: Automatic detecting indicators for quality of health information on the web. International Journal of Medical Informatics (2006)
5. Price, S., 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) 911–915
6. Vinot, R., Grabar, N., Valette, M.: Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In: TALN. (2003) 257–284
7. Wang, Y.: Automatic recognition of text difficulty from consumers health information. In IEEE, ed.: Computer-Based Medical Systems. (2006)
8. Gaudinat, A., Grabar, N., Boyer, C.: Machine learning approach for automatic quality criteria detection of health webpages. In McCray, A., ed.: MEDINFO 2007, Brisbane, Australia (2007) To appear.
9. Williams, K., Calvo, R.A.: A framework for text categorization. In: 7th Australian document computing symposium. (2002)
10. Salton, G.: Developments in automatic text retrieval. Science **253** (1991) 974–979
11. Koller, D., Sahami, M.: Toward optimal feature selection. In: International Conference on Machine Learning. (1996) 284–292