Automatic detecting quality criteria on health web pages

Célia Boyer and Arnaud Gaudinat

Quality and Ethics, Health on the Net Foundation, Geneva 14, Switzerland

Abstract

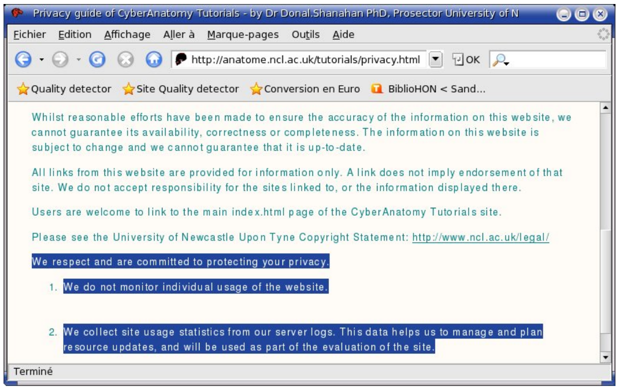
Quality and transparency of information on the web is one of major issues in the medical and patient safety area. HON foundation has defined a set of eight ethical principles which correspond to the first step in quality control of medical websites. Network of experts is working in order to manually define if a given website satisfies required principles. As amount of information on the web is going increasingly, manual expertise becomes unsufficient and automatic systems should be used in order to help medical experts. In this paper we present design and evaluation of automatic system conceived for the categorisation of medical and health documents according to HONcode ethical principles. We justify different choices made when designing the system. First evaluation shows promising results. Currently the system shows 0.78 of micro precision and 0.73 of F-measure, with 0.06 of errors. Several improvements remain a perspective to this work.

Keywords:

Medical web, Quality of health information, Automatic categorisation of documents, Natural language processing

Introduction

Web proposes the ever-increasing number of medical and health sites, and makes it possible to access quickly and easily several billion pages devoted to health information on the web. However, the quality of information on these sites shows the great variation. Among criteria which allow to state about the quality of a medical and health website, we focuse particularly on the ethical code. The principles underlying the ethical code and conduct have been initiated by HON[1], and have been widely adopted as HONcode [1] by publishers of medical and health websites. Eight ethical principles are then proposed: *authority*, *complementarity*, *privacy*, *reference*, *justifiability*, *authorship*, *sponsorship*, *advertising*. On one hand, each principle is clearly defined by HON. On other hand, websites candidate to accreditation must clearly state on these principles. For instance, *privacy* principle means that *Confidentiality of data relating to individual patients and visitors to a medical/health website, including their identity, is respected by this website. The website owners undertake to honour or exceed the legal requirements of medical/health information privacy that apply in the country and state where the website and mirror sites are located.* Figure 1 above shows that given website gives clear statement about its conduct according to this principle: *We respect and are committed to protecting your privacy. 1. We do not monitor individual usage of the website. 2. We collect site usage statistics from our server logs. This data helps us to manage and plan resource updates, and will be used as part of the evaluation of the site.*


Privacy guide of CyberAnatomy Tutorials - by Dr Donal.Shanahan PhD, Prosector University of N
Fichier
Edition
Affichage
Aller à
Marque-pages
Outils
Aide
http://anatome.ncl.ac.uk/tutorials/privacy.html
OK
Quality detector
Site Quality detector
Conversion en Euro
BiblioHON < Sand...
Terminé

Whilst reasonable efforts have been made to ensure the accuracy of the information on this website, we cannot guarantee its availability, correctness or completeness. The information on this website is subject to change and we cannot guarantee that it is up-to-date.

All links from this website are provided for information only. A link does not imply endorsement of that site. We do not accept responsibility for the sites linked to, or the information displayed there.

Users are welcome to link to the main index.html page of the CyberAnatomy Tutorials site.

Please see the University of Newcastle Upon Tyne Copyright Statement: http://www.ncl.ac.uk/legal/

We respect and are committed to protecting your privacy.

1. We do not monitor individual usage of the website.

2. We collect site usage statistics from our server logs. This data helps us to manage and plan resource updates, and will be used as part of the evaluation of the site.

Figure 1: Statement on privacy HONcode principle.

Thus, only if information required by a principle is clearly indicated on the website, this principle is considered as satisfied. A network of experts is working on the checking out if websites candidates to the HON accreditation clearly indicate information required by HONcode principles. This approach guarantees a high quality reviewing of websites and consequently this allows to certificate reliable medical and health pages. However, to remain effective in the face of accelerating growth in the number of online documents, manual compliance reviewing needs to be complemented and systematically achieved by automated means. In the contexte of quality control, such methods are expected to help making distinction between websites which satisfy ethical HONcode principles and those which do not. Further this information is usefull to detemine the reliability of medical/health documents and to ease the manual reviewing process.

Two kinds of methods can be applied for the automatic identification of quality principles: supervised and unsupervised ones. Supervised methods, or categorisation, are based on a learning or description step, while unsupervised methods,

[1] HON stands for *Health on the Net* foundation which proposes service of accreditation of medical and health websites: *www.hon.ch*

Principle	English	French	Spanish	Italian
Authority	1685	188	123	230
Complementary	1738	182	119	190
Privacy	1561	128	106	187
Reference	1039	112	71	128
Justifiability	323	25	17	28
Authorship	1813	177	120	201
Sponsorship	1473	163	101	163
Advertising	1030	103	86	142

Table 1: Learning data

or classification, are driven by internal properties of the processed data. In the first case expected categories are known, while in the last case they emerge. We have choosen to apply supervised learning methods as they allow to better characterise and constrain expected categories related to the eight HONcodes. Categorisation methods seem to be indeed helpful in automatic systems working with textual documents, *ie.* when detecting hostile messages [2] or racist content [3], or when filtering spams [4]. The objective of this paper is to propose an automated system for the detection of websites which satisfy ethical HONcode principles and contain reliable medical information. In the following of this paper we describe first material used, then methods defined. We then present and discuss obtained results, and conclude.

Material

A key component of any system for the automatic text categorisation is a knowledge base with positive examples which satisfy the above mentioned principles. In this work, the learning dataset is composed of over 5,000 HONcode accredited sites created in 72 countries accross the world and representing over 1,200,000 webpages through Google. This unique database is fruit of long experience in the field of health website certification. As indicated above, in the HONcode accreditation process, medical experts verify that a website complies with each of eight principles of the HONcode, as each principle needs to be found and checked for accuracy. To make the training dataset even more relevant for the learning process and text categorisation, human experts were asked to extract paragraphs which deal with eight HONcode ethical principles. We thus obtain separate datasets related to current principles: *authority*, *complementarity*, *privacy*, *reference*, *justifiability*, *authorship*, *sponsorship*, *advertising*. Table 1 indicates number of documents[2] from which information about different principles could be extracted for four processed languages (English, French, Spanish and Italian). Notice that we have distinguished one principle more, *ie.* *data* principle, which have been extracted from *reference* principle dataset. As result, we have nine learning datasets: eight HONcode principles and *data* principle. These datasets and `urls` of source webpages are recorded in `mysql` database.

Data extracted from English material is the most complete as the number of accredited websites is more important in this language. Thus, number of paragraphs in training sets is ten times larger than in other processed languages. Size of learning data in table 1 indicates also that, in all the languages, *justifiability* principle receives the less of statements. *Reference* principle is globally also less populated than remaining principles. We are aware that learning on sets with smaller data will give worse results comparing to those where the data is larger and more complete, although it is difficult to define the optimal size of training data.

Methods

Automatic categorisation methods consider documents as vectors within a vectorspace. Dimension of this space depends on the number of units (often words) of the whole collection of documents, and size of each vector corresponds to the frequency of a given unit in a given document. In our application, we aim at categorising sentenses and not documents. Indeed, sentences are more suitable for our purpose: (1) statement about principles can be located into one or more sentences, and (2) information contained in each sentence is expected to be more homogenous than information contained in the whole paragraph. Segmentation of paragraphs into sentences is performed by regular expressions based on punctuation marks and `html` tags.

Features tested within the learning process are the following: (1) with or without use of stopwords (prepositions, determinants, etc.); (2) with or without application of stemming algorithm based on [5] in order to lexically normalise 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 cooccurencies within sentence, or bag of words.

Unit weight within sentences is defined by three elements [6,7]: term frequency, inverse document frequency and length normalisation.

Machine learning algorithms used are those proposed by our learning framework [8]: Naive Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Decision Tree (DT). Different combinations of features and categorisation algorithms have been applied to the English data which is the most complete. Combinations which showed the most satisfying results have been applied to other languages.

Feature selection aims at reducing vectorspace dimension through selection of the most discriminant features, and thus at obtaining more relevant results [9]. We performed feature selection with document frequency (DF) criterion, which favors units distributed in the largest number of sentences. DF is quick and efficient, and usually allows to reduce about 80% of features [10].

Learning and test sets are composed of 90% and 10% of available documents respectively.

Evaluation is performed with the following measures in their micro and macro versions: precision, recall and F-measure. Macro precision (maP) is representative of the dis-

[2] ajouter nb d'occurrence par corpus dans la table 1 ?

tribution of elements in each category, and micro precision (miP) in each sentence.

Results

Table 2 shows global results obtained with different algorithms and features. First column *Lang* indicates the language processed and next three columns indicate learning system setting up: *w* stands for the segmentation of sentences into units (single word *w1*, bi-gram *w2*, three-gram *w3* ... w_n, cooccurence *cooc* and stemmed single word *s1*); *meth* stands for the learning algorithm used (Naive Bayes *NB*, Support Vector Machine *SVM*, k-Nearest Neighbors *kNN* and Decision Tree *DT*); *weight* stands for the weighting of units (first character indicates term frequency: natural *n*, logarithmic *l* or augmented *a*; second character indicates if inverse document frequency is taken into account *t* or not *n*; third character indicates if document length is normalised with cosine *c* or not *n*. Following columns indicate evaluation figures for precision, recall, F-measure in their macro and micro versions, and errors rates. This table allows particularly to observe precision figures which are the most important when evaluating how helpful this system can be for human reviewers. These figures are included in the interval between 0.59 and 0.78 points.

Table 3 presents contingency between precision and recall figures for all nine sets of data considered. It allows to observe how successfull is the categorisation of data according to principles during the test step. The best results are observed when the contingency between precision and recall is high, i.e. *privacy* principle with the contingency 0.92 / 0.90.

Discussion and Perspectives

All the combinations of features (some of them are actually presented in first four columns of table 2) have been tested with English data but all these experiments show no significant differences. Thus, *nnn* setting up of *SVM* algorithm, which shows most interesting results, has been applied to datasets of other languages.

We can consider that precision figures, which correspond to the percentage of correct categorisations among all the results, are the most relevant as for the jugement about performance of the system in context of its use by reviewers in daily work. Furthermore, we consider that micro precision (*miP*) is more suitable to be taken into account as it corresponds to the precision with which a sentence is assigned into a given category. Figures in table 2 indicate that *SVM* algorithm with unique word as processed unit and *nnn* weighting shows the best results: 0.78 of micro precision. Recall of this setting is 0.69 which is one of lowest figures, while F-measure is one of highest (0.73). Error rate of this setting is one of lowest (0.06). We can thus expect that applying *SVM* algorithm with such setting would give results which relevance is closer to the human categorisation. Indeed, its application to other languages datasets, ...

Analysis of contingency figures from table 3 indicates that the principle the better recognised is the *privacy* principle. Indeed, its figures are highest: 0.92 / 0.90. It means also that on lexical level, which gives basic data for the categorisation system, this principle statements are formulated with specific lexicon. For instance, among units with highest frequencies we can find *identity*, *personal*, *respected*, *individual*, *confidentiality* or *privacy*. The principle which is the most difficult to recognise is the *justifiability* principle, showing the precision/recall contingency of 0.45 / 0.33. Moreover, it appears to be highly ambiguous with the *complementarity* principle. Other couples of ambiguous principles are *reference / authority* and *advertising / sponsorship*. Concerning *justifiability* and *reference* principles, they are sub-populated and don't represent large enough learning set. As for the *advertising / sponsorship* couple, the main reason of confusion facing them is that, on lexical level, system detects some similarities. For instance, for both principles there are mentions of *funding*, *maintenance*, *acceptance*. Furthermore, statements on these principles can be located at the same pages or paragraphs.

As expected, due to small number of documents stating on two of searched principles (*justifiability* and *reference*) these principles could be hardly processed by the system. In this regard, other methods should be tested. For instance, similarity measures between documents [6] as inspired by information retrieval field. Thus, the similarity can be computed directly with principle definitions, which exist in various languages, and small size of the reference data would not be a limitation for the system. Furthermore, the role of `url` analysis can be important as they often convey indications about principles as well. For instance, when webpage is named *privacy.html* or *policy.html*, this offers direct indication about the nature of processed pages and their expected content. Combination of these different approaches and clues with currently used machine learning system is a perspective.

Currently we aim at categorisation of sentences while in reality entire documents should be processed. Indeed, information concerned by quality principles can be distributed among different sentences within a document or even among different pages of a website. Moreover, this information is get bogged down in all the content of pages and categorisation system must deal with this. Taking into account whole documents and websites enhances difficulty of the categorisation process and should decrease performances of the system. But first evaluations showed that our learning system acquires necessary database for the categorisation of entire pages and websites.

As discussed, our system tries to categorise sentences according to principles, but it is not sensitive to detect in which kind of context, positive or negative, the statement occurs. For instance, a webpage can indicate that *privacy policy is not respected on the site*, while the system would just detect that this sentence is concerned with the *privacy* principle. This categorisation is correct but managing *nuances* occurring with the main information is even more important. Detection of such details remains a challenging perspective [11] and would give interesting complementary indications and weighting of categorisation results.

Another limitation is related to the fact that statements on principles can be hardly verified, *e.g.* web publishers can declare to respect the *privacy* principle while unofficially they sell informations on users. Notice that in this regard, HON proposes a complaint solution within which three parties (user, web publisher and HON) can anonymously communicate and try to resolve such situations.

Currently, evaluation of feature selection is based on document frequency criterion. It could be interesting to use for this purpose Mutual information or Chi2 criteria. We suppose this would enhance feature selection efficiency and improve categorisation results.

In order to detect webpages which satisfy HONcode principles, we have used database of positive examples of medical and health pages as training set. But negative examples can also be used if we want to detect pages and sites which don't satisfy these principles. Each page or site could thus be weighted according to its positive and negative scores, and the global jugement about it further computed.

HONcode ethical principles are currently translated into 32 languages and the accreditation process is being adopted all over the world. System currently trained for four languages (English, French, Spanish and Italian) can be adapted and applied to other languages. In the same way, problematics related to quality and transparency of information on the web is not only reserved to medical area. Our system can be trained on data from other areas as well. In the field of medical area, this system can be tested with other quality principles, which can be different from the ethical ones.

Conclusion

We have presented our work on designing an automatic system for the detection of quality and transparency HONcode principles of medical and health documents on the web. System is based on machine learning methods for document categorisation. First evaluations performed show promising results, *ie.* *SVM* algorithm with simple words as processed units and *nnn* weighting of features shows 0.78 of micro precision and one of the most high F-measure (0.73). Error rate of this setting is one of the lowest (0.06). These results seem to confirm the relevance of our approach for the categorisation of webpages according to HONcode ethical principles. We outlined several perspectives which, we believe, will bring some improvement to our system. Additional evaluation is nevertheless needed, *ie.* comparison of pages detected by our automatic system and those already manually categorised by experts. It would give indications about the suitability and reliability of automatically computed accreditation of documents according to HONcode principles.

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 number 507019.

References

[1] Boyer C, Baujard V, and Scherrer J. HONcode: a standard to improve the quality of medical/health information on the internet and HON's 5th survey on the use of internet for medical and health purposes. In: 6th Internet World Congress for Biomedical Sciences (INABIS 2000), 1999.

[2] Spertus E. Smokey: automatic recognition of hostile messages. 1997.

[3] Vinot R, Grabar N, and Valette M. Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In: TALN, 2003.

[4] Carreras X and Márquez L. Boosting trees for anti-spam email filtering. In: Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, BG. 2001.

[5] Porter M. An algorithm for suffix stripping. *Program* 1980;14(3):130–7.

[6] Salton G. Developments in automatic text retrieval. *Science* 1991;253:974–9.

[7] Singhal A, Salton G, Mitra M, and Buckley C. Document length normalization. *Information Processing & Management* 1996;32(5):619–33.

[8] Williams K and Calvo RA. A framework for text categorization. In: 7th Australian document computing symposium, 2002.

[9] Koller D and Sahami M. Toward optimal feature selection. In: International Conference on Machine Learning, 1996:284–92.

[10] Yang Y and Liu X. Re-examination of text categorisation methods. In: Proc of 22nd Annual International SIGIR, Berkley. 1999:42–9.

[11] Chapman W, Bridewell W, Hanbury P, Cooper G, and Buchanan B. Evaluation of negation phrases in narrative clinical reports. In: Annual Symposium of the American Medical Informatics Association (AMIA), Washington. 2001.

Address for correspondence

Célia Boyer, Arnaud Gaudinat
Health on the Net Foundation, HUG/DIM
24, rue Micheli-du-Crest
1211 Geneva 14, Switzerland
tel: +41 22 372 62 50
fax: +41 22 372 88 85
email: {celia.boyer ; arnaud.gaudinat}@healthonnet.org

Lang	w	meth	weight	maR	maP	maF1	miR	miP	miF1	Err
eng	w1	NB	nnn	0.72	0.67	0.66	0.81	0.65	0.72	0.07
	s1	NB	nnn							
	w1	NB	ann							
	w1	NB	ntn							
	w1	NB	nnc							
	w1	NB	atn							
	w1	NB	atc							
	w1	NB	lnn							
	w1	NB	ltn							
	cooc	NB	nnn							
	cooc	NB	atn							
	cooc	NB	atc							
	cooc	NB	ann							
	w2	NB	nnn							
	w2	NB	atn							
	w2	NB	ann							
	w3	NB	nnn							
	w4	NB	nnn							
	w1	SVM	nnn							
	cooc	SVM	nnn							
	cooc	SVM	ann							
		DT	nnn							
		kNN	nnn							
fre										
spa										
ita										

Table 2: Results and their evaluation according to macro and micro precision, recall and F-measure

	Authority	Compl.	Privacy	Reference	Justif.	Authorship	Sponsorship	Advertising	Date
Authority	0.64/0.72	0.05/0.05	0.01/0.01	0.19/0.34	0.01/0.09	0.04/0.13	0.04/0.09		
Compl.	0.05/0.05	0.80/0.82	0.05/0.03	0.01/0.02	0.06/0.44	0.00/0.00	0.03/0.05		
Privacy	0.02/0.03	0.02/0.04	0.92/0.90	0.00/0.01	0.00/0.03	0.01/0.02	0.01/0.02		
Reference	0.24/0.12	0.03/0.02	0.03/0.01	0.64/0.57	0.02/0.08	0.01/0.01	0.02/0.02		
Justif.	0.06/0.01	0.32/0.03	0.06/0.00	0.06/0.01	0.45/0.33	0.02/0.01	0.00/0.00		
Authorship	0.06/0.02	0.02/0.01	0.08/0.02	0.02/0.01	0.00/0.00	0.81/0.81	0.00/0.00		
Sponsorship	0.05/0.03	0.04/0.02	0.02/0.01	0.01/0.01	0.00/0.02	0.02/0.02	0.69/0.69		
Advertising	0.01/0.01	0.02/0.01	0.05/0.01	0.00/0.00	0.00/0.02	0.00/0.00	0.13/0.12	0.00/0.00	
Date	0.00/0.00	0.01/0.00	0.01/0.00	0.06/0.03	0.00/0.00	0.00/0.00	0.01/0.01		0.00/0.00

Table 3: Precision/Recall contingency of quality criteria. System setting: method *SVM*, language *English*, single word *w1*