Named Entity Recognition for Web Content Filtering*

José María Gómez Hidalgo, Francisco Carrero García, and Enrique Puertas Sanz

Universidad Europea de Madrid, Villaviciosa de Odón, 28670, Madrid (Spain),
jmgomez,francisco.carrero,epuertas@uem.es,
WWW home page: http://www.esi.uem.es/~jmgomez

Abstract. Effective Web content filtering is a necessity in educational and workplace environments, but current approaches are far from perfect. We discuss a model for text-based intelligent Web content filtering, in which shallow linguistic analysis plays a key role. In order to demonstrate how this model can be realized, we have developed a lexical Named Entity Recognition system, and used it to improve the effectiveness of statistical Automated Text Categorization methods. We have performed several experiments that confirm this fact, and encourage the integration of other shallow linguistic processing techniques in intelligent Web content filtering.

1 Introduction

In the recent years, we have witnessed an impressive growth of on-line information and resources. The self-regulating nature of Web publishing, along with the ease of making information available on the Web, has allowed that some publishers make offensive, harmful or even illegal contents present in Web sites across the world. This fact makes the use of filtering and monitoring systems a necessity in educational environments, and in the work place, to protect children and prevent Internet abuse.

There is a number of filtering solutions available in the market, including commercial products like CyberPatrol or NetNanny, and open source systems like SquidGuard or DansGuardian. According to in-depth evaluations of these products (e.g. the one performed in the European Project NetProtect [1]), their filtering effectiveness is limited by the use of simple techniques, like URL blocking, or keyword matching. There is the need of more sophisticate and intelligent approaches to increase effectiveness of filtering solutions, and thus, to improve children's protection and Internet abuse prevention.

Our goal is to investigate to what extent shallow linguistic processing can improve the effectiveness of current filtering software. This kind of analysis is

* The work described in this paper is partly funded by the European Commission under the Safer Internet Action Plan, contract POESIA - 2117 / 27572. POESIA stands for Public Open-source Environment for a Safer Internet Access, and more information about it can be fount at http://www.poesia-filter.org.

suitable for that Web pages that have been considered suspicious by simpler techniques, being the majority of Web pages tagged as harmful or harmless by faster and more simple methods. We focus on Spanish language pornographic Web pages, but the methods discussed here can easily extended to other European languages.

In this work, we are concerned with the possibility of using Named Entity Recognition to improve statistical classification of pornographic Web pages. According to recent evaluations ([2]), there are a number of well known and relatively effective methods for detecting Named Entities (NEs) in running text. Some NEs can be highly indicative of pornographic content (e.g. names of famous porn stars, or manufacturers of sex toys), while others suggest harmless content (e.g. names of politicians, corporations, or locations).

We test this intuition by conducting a set of experiments that compare a statistical pornographic detection approach based on state-of-the-art Text Categorization techniques, with an improvement of this one that considers also NEs as attributes of Web pages. The NEs are detected using an simple approach based on lexical evidence and learning. The results of our experiments show that the usage of NE Recognition (NER), and thus, of shallow linguistic analysis, can improve the accuracy of otherwise highly accurate methods.

2 Web Content Filtering

In this section, we review the main approaches to Web content filtering, focusing specially on those that use intelligent content analysis based on text classification. Also, we present our basic approach for pornographic Web content detection based on Automated Text Categorization state-of-the-art methods.

2.1 Review of Recent Work

The Web Content Filtering (WCF) approaches have been classified in four major groups [3]:

- Self or third party ratings, specially the usage of Platform for Internet Content Selection (PICS) or Internet Content Rating Association (ICRA) ratings. Authors or reviewers label Web pages according to several types of content and levels, which are used by the filtering system to allow or block the pages according to the settings defined by the user or administrator. Unfortunately, only a small fraction of Web pages are labeled, and authors can inadvertently (or intentionally) label their pages with incorrect tags.
- Uniform Resource Locator (URL) listing, that is, maintaining a list of blocked and/or allowed web sites. A Web page is blocked if its URL contains a blocked URL, or its outgoing links point to blocked URL addresses. These kinds of lists can be automatically or manually built, but they are difficult to keep updated, and do not account for domain aliasing.

– Keyword matching, in which a set of indicative keywords or key-phrases ("sex", "free pics") are manually or automatically derived, form a set of pornographic Web pages. A Web page is blocked if the number or frequency of keywords occurring in it, exceeds a predetermined threshold. This approach is prone to over-blocking, that is, blocking safe Web pages in which these keywords occur (e.g. sexual health, etc.).
– Intelligent content filtering, which involves a deeper understanding of the semantics of text and other media items (specially pictures), by using linguistic analysis, machine learning, and image processing techniques. The heavy cost of building linguistic analyzers and image processing components, their domain dependence (e.g. technique for detecting nudes are quite different to those for recognizing Nazi symbols), and the delay caused by in-depth analysis, limit the applicability of these techniques.

The first three approaches, widely used in current filtering solutions, have proved quite ineffective, and have serious limitations [1]. We argue that intelligent content analysis is feasible, as far as the system design deals with delay issues, and linguistic and image processing are kept as shallow as possible. In particular, the project POESIA [4, 5] is designed for having two levels of filtering: Light filtering, for those Web pages that are not suspicious, or clearly pornographic; and heavy filtering, for those Web pages in which light filters are not able to give a clear judgment. Linguistic and image processing techniques in the light filters are very limited and efficient, while heavy filter use more advanced (but shallow, anyway) methods, giving a more accurate but delayed answer.

We focus here on shallow linguistic processing of Web pages textual items, in order to take a sensible decision about the pornographic orientation of the pages, when simpler methods fail. Our intuition is that, while Automated Text Categorization can be very effective on the task, it might be improved with shallow text analysis involving e.g. POS-tagging, Noun Phrase chunking, and in particular, Named Entity Recognition. We discuss the basic Text Categorization approach in the next section.

2.2 Web Content Filtering as Text Categorization

Automated Text Categorization (ATC) is the assignment of text documents to predefined categories. Text documents are usually news items, scientific reports, e-mail messages, Web pages, and so. Categories are often thematic, and include library classifications (e.g. the Medical Subject Headings), keywords in Digital Libraries, personal e-mail folders[1], Web directory categories (like Yahoo!'s), etc. Automated Text Categorizers can be built by hand (e.g. by writing rules for e-mail messages filing in personal folders), or they may be constructed automatically, by learning a text classifier on a set of manually labeled documents. This latter, learning-based, approach has become dominant, and current techniques allow building ATC systems as accurate as human experts in a domain [6].

[1] Some e-mail folders of an average e-mail user may be not thematic at all, but based e.g. on the sender (for instance, my brother's messages).

Pornography detection has been approached as learning-based ATC in several recent works, including [7–9, 3, 10], and ours [5]. From these works, we can model pornography detection as a 2-class learning problem: learn a classifier that decides if Web page is pornographic or not. In the learning phase, given two sets of pornographic (P) and safe (S) Web pages (the *training collection*), the following steps are given:

1. Each Web page in P or S is processed to extract the text it includes (pieces of text inside <TITLE>, <H1>, <P>, <META>[2]; or, just all tags are stripped out), defining its text content. The text is tokenized into words, which may be stemmed and/or stop listed (ignoring those occurring in a function word list), producing a list of text tokens.
2. Each page is represented as a term-weight (or attribute-value) vector, being the terms the previous text tokens. The weights can binary (a token occurs in the Web page, or not), Term Frequency (number of times the token occurs in the page), TF.IDF (the previous one times the Inverse Document Frequency), Relative Term Frequency, etc. [6].
3. Optionally, a number of tokens is selected according to a quality metric like Information Gain or χ^2 [11]. This step allows to reduce the dimensionality of the problem, speeding up learning and even increasing accuracy. The resulting set of tokens is the final *lexicon*.
4. Finally, a *classifier* is induced using a training algorithm over de set of training vectors and its associated class labels. Algorithms used in this problem include the probabilistic Naive Bayes algorithm [7] and Bayesian Networks [8], variants of lazy learning [9, 10], semi-supervised Neural Networks [3], and linear Support Vector Machines [5].

The two first steps define the text representation model, which in this case is often called the *bag of words* model. It corresponds to the traditional Vector Space Model in Salton's work [12]. In other works focused on Web page categorization according to thematic classes, there has been propose to enrich the text in the page with the text of nearby Web pages in the link graph structure (from incoming or outgoing links) [13].

The classification phase involves, given a new Web page which class is not known, its representation as term-weight vector similar to those in the training collection, and its classification according the model generated in the learning phase. This phase must be extremely efficient, avoiding long delays in Web pages delivery when they are allowed (classified as safe).

2.3 Our Text Categorization Approach

The previous approach has been proved quite successful in the literature, but current experimental results are not perfect, and filters often miss-classify web pages in which there is nearly no text, or those concerned with sexual education.

[2] Usually, restricted to attributes NAME and CONTENT.

Also, experimental results may be different to those obtained in operational environments, which have been not test yet. We believe that this model can be improved in the uncertain cases, by using linguistic techniques that provide a more meaningful insight of the page topic.

As a pilot study, we have included a NER system in our baseline ATC approach. The details of our method are:

- Text is extracted from training Web pages with an HTML parser.
- The extracted text is tokenized, each token converted to lowercase, and stemmed using an Spanish stemmer.
- We select the text tokens with an Information Gain score (that is, those stems that provide any indicative information in the training collection). The resulting tokens are the vocabulary.
- Each training Web page is represented as a term-weight vector, in which terms are vocabulary tokens, and their weights are their Term Frequency in the page.
- We learn a linear classifier (a linear function of the weights of the stems in the vocabulary) using linear Support Vector Machines[3].

We provide details on the Web page training collection in the section 4.1. Provided enough and representative training data, this approach leads to fast and accurate text classifiers in most cases, letting anyway some space for effectiveness improvements for the case of difficult Web pages.

3 Spanish Named Entity Recognition

In this section, we review the recent work in Spanish and Language Independent NER, along with our approach to the problem. We also provide effectiveness results on standard data collections, and how we have integrated the NER system in the previous ATC method.

3.1 Recent Work in Spanish NER

NER has been considered as an important task in the wider Information Extraction field, and it is nowadays fully integrated in typical text analysis tasks for learning-based Information Extraction applications [14]. It has been also the focus of recent Computational Natural Language Learning (CoNLL) Shared Task competitions (2002, 2003) [2], giving the area an important impulse. Currently, and given enough training data, it is possible to build a NER system that reaches high levels of accuracy (e.g. an F_1 value over .80 for Spanish, and near .90 for the English language).

[3] These steps have been performed by using the following open-source packages: the HTMLParser (`http://htmlparser.sourceforge.net/`), the SnowBall Spanish stemmer (`http://snowball.tartarus.org/`), and the WEKA Machine Learning library (`http://www.cs.waikato.ac.nz/~ml/weka/`).

Most top performing NER systems in CoNLL Shared Task competitions[4] follow a learning approach, sometimes enriched with the use of external resources as e.g. gazetteers. The task is approached as a Term Categorization problem, in which each word must be tagged as the beginning (B) of a Named Entity, an inner word in a Named Entity (I), or as other (O). The types of entities addressed in CoNNL competitions are persons, organizations, locations and miscellanea. As an example, the expression "Robinson lived in a island near South America" should be tagged as "Robinson/B-PER lived/O in/O a/O island/O near/O South/B-LOC America/I-LOC".

We discuss the top performing system presented for the Spanish language NER competition [15], as it illustrates the dominant method in the field. In this work, the NER is a two level learning system, the first level detecting the limits of a NE, and the second finding out its type. Its main characteristics are:

- The set of features for each word is defined in terms of the context of the word, including the word itself, its Part-Of-Speech (POS), orthographic features of the word (involving capitalization, hyphenation and others), the form of the words in a fixed length contextual window (lexical features), left predictions, and others.
- The learning method for both levels is the meta-learner Adaboost applied to small fixed-depth decision trees. This meta-learner has the property of improving a base learner, by iteratively focusing on those examples (words) that are incorrectly labeled by the trees learned in previous iterations.
- Two external resources are also used, a gazetteer that includes a number of NEs for Spanish, not see in the training phase, and a set of hand-crafted trigger words. These resources lift accuracy in a 2% for the type of entity classifier.

This work both shows the main characteristics of current learning-based NERs, and it demonstrates that high levels of accuracy can be achieved in the Spanish NER task. It has also partly (along with [16]) guided the design of our lexical NER method, which has been designed as a *knowledge poor* approach for the pornographic Web content detection pilot study.

3.2 A Lexical Spanish NER

Our NER system is designed to use only the most reliable features in the surrounding context of the target word, given the unstructured nature of Web pages (quite different from news items, used in previous experiments). In fact, we call our NER *lexical* because we found by trial and error that only this kind of information can be robustly extracted from Web pages text. In particular, we consider a set of features that includes: binary orthographic features for the words in a fixed-length window surrounding the target word, a list of frequent words and

[4] These can be compared in the tasks web pages, for CoNNL 2002 (`http://cnts.uia.ac.be/conll2002/ner/`) and 2003 (`http://cnts.uia.ac.be/conll2003/ner/`).

punctuation symbols in the window, and the predicted class for the previous words in the window.

The number of words (lexical items) considered, the width of the window, the binary or numeric nature of attribute values, and the utilization or not of previous tags, are parameters of our system. We have tested a wide number of parameter settings, in a wrapper approach: given a configuration of parameters (window size = 2, etc.), we test its accuracy by 5-fold cross validation on the training set, by using a decision tree learner (C4.5), and assessing its classification accuracy.

The best results have been obtained using 44 attributes: in a ± 2-size window, if the word has an initial capital letter (5 attributes, one per word in the window), if the word is all uppercased (5 attributes), if the target word is only a capital letter (1 feature), a capital letter or a period (1 feature), starts with one capital letter or it is a period (1 feature), and it uppercased or it is a period (1 feature). Last 30 features are the position of each of the 30 most frequent tokens (either words, lowercased, or punctuation marks) in the window, if they occur within the window.

We have after tested a representative range of learning algorithms on this feature configuration, including a Naive Bayes classifier, the C4.5 decision tree learner itself, Adaboost applied to C4.5 (ABC4.5), linear Support Vector Machines (SVM), and the lazy learner k-Nearest Neighbors with $k = 1, 3, 5$ values. The results of these experiments, along with the evaluation of the selected learners in the CoNLL experimental framework, are discussed in the next section.

3.3 Experimental Results on the CONLL Framework

Since in this pilot study, we are interested on detecting NEs, but not classifying them according to its type, we present results only for B, I and O tags, and for entire NEs.

First we present the three top performing learners results, obtained by 5-fold cross validation on the training set. In the Table 1, we show the F_1 measure for the three types of tags (B, I and O), and the overall accuracy. F_1 is a kind of average of recall (the proportion of items detected over the real number of items) and precision (the proportion of correctly retrieved items over the number of retrieved items). The accuracy (Acc) is the number of correct items over the total number of items.

The results shown demonstrate that the three learners are roughly equivalent, in terms of effectiveness. We are concerned with the less frequent B and I tags, that mark NEs in the texts. For these tags, ABC4.5 is slightly better than C4.5, but the decrease in speed (C4.5 with boosting is two orders of magnitude slower than C4.5 alone) does not deserve using it.

This analysis is confirmed by the results obtained on the CoNLL test data, using the standard evaluation scripts for the task. In the Table 2, we show the accuracy (Acc), recall (Re), precision (Pr), and F_1 scores for the C4.5, ABC4.5 and SVM learners. On this dataset, the C4.5 learner performs a bit better than ABC4.5.

Table 1. Results on training data for the top performing learners.

Algorithm	F_1/B	F_1/I	F_1/O	Acc
C4.5	0.889	0.825	0.989	0.972
ABC4.5	0.886	0.831	0.988	0.972
SVM	0.886	0.815	0.988	0.971

Table 2. Results on CoNLL test data for the top performing learners.

Algorithm	Acc	Re	Pr	F_1
C4.5	0.969	0.816	0.818	0.817
ABC4.5	0.969	0.800	0.821	0.810
SVM	0.969	0.806	0.812	0.809

We must note that the results of our NER system are not comparable with those by participants in the CoNLL Shared Tasks competitions, because they are forced to predict the type of the NE detected. This fact makes their results better than ours, because a decrease of effectiveness is expected on this phase. Also, we have detected a subtle tendency of our NER to classify any starting token in a sentence as a NE, which should be corrected in an operational implementation of the overall filtering approach.

On the basis of these experiments, we use a NER system based on the 44 attribute vector representation described above, and the C4.5 decision tree learner, for our following work.

3.4 Integrating NER in ATC-based Web Content Filtering

We have performed a straightforward integration of the NER method in the ATC-based Web page filtering approach. We have applied the NER system to the training collection Web pages text, and added to the lexicon the extracted NEs with an Information Gain score over 0. For these tokens, we also make use of Term Frequency weights.

The NER system has been trained on the training collection of the CoNLL Shared Task competition for Spanish (2002). However, it has been applied to the text extracted from Web pages, violating a basic assumption in Machine Learning: test (or working) data must resemble training data. We view this as a *domain transfer* of the NER system, and given this violation, we do not expect it to be as effective as it is on news items.

4 Experiments

In this section, we describe our data collection, and the experiments we have performed on it, in order to confirm that shallow linguistic processing (specifically, NER) can improve statistical ATC methods.

4.1 Data Collection

The data collection used in our experiments is the POESIA project Spanish Text Corpus. This corpus contains 6,463 pornographic Web pages and 29,133 non-pornographic Web pages, in HTML format, and they have been collected the following way:

1. An initial set of URL addresses have been obtained from the Spanish section of the Open Directory Project Web page[5], containing a list of around 1,000 pornographic URLs and 100,000 non-pornographic URLs. These latter list has been randomly sub-sampled to get around 5,000 URLs.
2. An initial corpus has been built by downloading those Web pages with less than 10 seconds answer. These Web pages have been filtered to delete frame based and Error 404 Web pages. Also, the files have been processed to get in-site links, providing a second set of URLs. The resulting lists have been sub-sampled, and the process has been repeated one more.

Essentially, current contents of the corpus include front to second level Web pages from a representative sample of pornographic and non-pornographic Web sites.

For our experiments, we have divided the corpus into a training set containing 2/3 of the corpus Web pages, and a test set with the remaining pages. Since the HTML parser fails to extract text on strongly unstructured Web pages, a portion of the pages containing less than 10 bytes of text have been removed from training and testing sets. This leads to 4,188 pornographic and 18,446 safe training Web pages, and 2,094 pornographic and 9,200 safe test Web pages.

After performing text extraction, tokenization, and stemming, we have collected a set of 17,859 word stems with an Information Gain score greater than 0 in the training collection. Also, we have applied the NER system to the text in training Web pages, deriving a set of 8,491 NEs with an Information Gain score over 0. The combined lexicon has 26,350 units, being NEs a 32.22% of them. However, there are only 17 NEs among the 200 top scoring units, and NEs are often section names with capitalized words (e.g. "Fotos" – the pictures section of the Web site).

4.2 Results and Analysis

The results of our experiments on Spanish pornographic Web content detection are summarized in the Table 3. In this table, we show the performance of linear

[5] Available at `http://dmoz.org`.

Support Vector Machines on three text representation models: one using only word stems, one using only NEs, and one using both. For each of the classes, pornographic Web pages (P) and safe Web pages (S), the recall (Re), precision (Pr) and F_1 scores are shown, along with the overall accuracy (Acc).

Table 3. Results of SVM on the three kinds of lexicon used.

Lexicon	Re/P	Pr/P	F_1/P	Re/S	Pr/S	F_1/S	Acc
Stems	0.867	0.983	0.921	0.997	0.971	0.983	0.972
NEs	0.812	0.960	0.880	0.992	0.959	0.975	0.958
Both	0.891	0.982	0.934	0.996	0.976	0.986	0.976

The results of these experiments are encouraging. It is clear that NEs are even good indexing units by themselves, although this may be due in part to the fact that many of them are ordinary words with some kind of capitalization (that is, they correspond to NER over-detection mistakes). The combination of both sources of evidence, stems and NEs, clearly outperforms NEs in isolation, and noticeably improves an stem based representation.

In a detailed analysis, the stem-based approach miss-classifies 31 safe Web pages, and 279 pornographic Web pages, while the combined approach miss-classifies 34 safe Web pages, and 229 pornographic Web pages. There is a relative improvement of the accuracy on pornographic pages, at the expense of a affordable decrease of performance on safe pages.

4.3 Additional considerations

In an operational environment, a classifier like those proposed here must be able to determine the class (porn or safe) of a high number of Web pages per minute. The stem based representation is easy and quick to compute, and the linear SVM perform linearly on the number of attributes (terms). This configurations warrants fast response times, even in the Java language, according to our experience in the POESIA project.

An important concern is if shallow linguistic analysis in general, and NER in particular, may slow down the filtering operation, even making it unfeasible. This point has been addressed in POESIA by using two level filtering. A light filter, based on statistical methods (e.g. SVM) and simple text representations (e.g. word stems) is first called, allowing a quick answer for most of the requests. If this filter classifies the Web page as "unsure", it is passed to a heavy filter that performs a deeper (and slower) processing of the request. According to our experience, this filter can employ even seconds to process a request, implying a delay for the user. However, this delay is accepted by most educational institutions, where Internet resources are shared by hundreds to thousands of students,

and there is always a delay due to bandwidth limitation. So in short, the framework for shallow linguistic analysis of Web pages proposed here, is acceptable in operational environments.

It is remarkable that commercial filters tend to block health Web pages, according to the study for the Kaiser Family Foundation [17]. Our model is specially suitable for this kind of contents, on which the cost of the delay is acceptable if classification accuracy is significantly improved.

5 Conclusions and Future Work

In this paper, we argue that shallow linguistic analysis in general, and Named Entity Recognition in particular, can be used to improve the effectiveness of text classification in the framework of intelligent Web content filtering. We have implemented a lexical NER system, and used it to demonstrate how NEs discovered on a training phase, can enrich a traditional, stem based vocabulary, leading to a more accurate filter.

The results of our experiments are encouraging, and motivate us to improve our NER system, and to include more shallow linguistic processing techniques (like e.g. text chunking) in our Web content filtering method. We believe that filtering methods can benefit from a deeper understanding of the meaning of text in Web pages.

References

1. Brunessaux, S., Isidoro, O., Kahl, S., Ferlias, G., Rotta Soares, A.: NetProtect report on currently available COTS filtering tools. Technical report, NetProtect Deliverable NETPROTECT:WP2:D2.2 to the European Commission (2001) Available: `http://www.netprotect.org`.
2. Roth, D., van den Bosch, A., eds.: Proceedings of CoNLL-2002, Taipei, Taiwan, Association for Computational Linguistics, Special Interest Group on Natural Language Learning (2002)
3. Lee, P., Hui, S., Fong, A.: A structural and content-based analysis for web filtering. Internet Research **13** (2003) 27–37
4. Gómez, J., de Buenaga, M., Carrero, F., Puertas, E.: Text filtering at POESIA: A new internet content filtering tool for educational environments. Procesamiento del Lenguaje Natural **29** (2002) 291–292
5. Hepple, M., Ireson, N., Allegrini, P., Marchi, S., Montemagni, S., Gómez, J.: NLP-enhanced content filtering within the POESIA project. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). (2004)
6. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47
7. Chandrinos, K.V., Androutsopoulos, I., Paliouras, G., Spyropoulos, C.D.: Automatic Web rating: Filtering obscene content on the Web. In Borbinha, J.L., Baker, T., eds.: Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, PT, Springer Verlag, Heidelberg, DE (2000) 403–406 Published in the "Lecture Notes in Computer Science" series, number 1923.

8. Denoyer, L., Vittaut, J.N., Gallinari, P., Brunessaux, S., Brunessaux, S.: Structured multimedia document classification. In: DocEng '03: Proceedings of the 2003 ACM symposium on Document engineering, ACM Press (2003) 153–160
9. Du, R., Safavi-Naini, R., Susilo, W.: Web filtering using text classification. In: Proceedings of the 11th IEEE International Conference on Networks, Sydney, IEEE (2003) 325–330
10. Su, G.Y., Li, J.H., Ma, Y.H., Li, S.H.: Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model. Journal of Zhejiang University SCIENCE **5** (2004) 1106–1113
11. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, 14th International Conference on Machine Learning. (1997)
12. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison Wesley (1989)
13. Ghani, R., Slattery, S., Yang, Y.: Hypertext categorization using hyperlink patterns and meta data. In Brodley, C., Danyluk, A., eds.: Proceedings of ICML-01, 18th International Conference on Machine Learning, Williams College, US, Morgan Kaufmann Publishers, San Francisco, US (2001) 178–185
14. Zhang, T., Damerau, F., Johnson, D.: Text chunking based on a generalization of winnow. J. Mach. Learn. Res. **2** (2002) 615–637
15. Carreras, X., Màrques, L., Padró, L.: Named entity extraction using adaboost. In: Proceedings of CoNLL-2002, Taipei, Taiwan (2002) 167–170
16. Chieu, H., Ng, H.: Named entity recognition: A maximum entropy approach using global information. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). (2002) 190–196
17. Richardson, C., Resnick, P., Hansen, D., Holly A. Derry, Rideout, V.: Does Pornography-Blocking Software Block Access to Health Information on the Internet? Journal of the American Medical Association **288** (2002) 2887–2894