NLP-enhanced Content Filtering within the POESIA Project

Mark Hepple[⋆], Neil Ireson[⋆], Paolo Allegrini[†], Simone Marchi[†], Simonetta Montemagni[†] and Jose Maria Gomez Hidalgo[◇]

[⋆]University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield, UK. {m.hepple, n.ireson}@dcs.shef.ac.uk

[†]Istituto di Linguistica Computazionale, CNR, Area della Ricerca di Pisa Via Moruzzi 1, 56124 Pisa, Italy. {allegrip, simone.marchi, simonetta.montemagni}@ilc.cnr.it

[◇]Departamento de Inteligencia Artificial, Universidad Europea de Madrid, 28670, Villaviciosa de Odon, Madrid, Spain. jmgomez@uem.es

Abstract

This paper introduces the POESIA internet filtering system, which is open-source, and which combines standard filtering methods, such as positive/negative URL lists, with more advanced techniques, such as image processing and NLP-enhanced text filtering. The description here focusses on components providing textual content filtering for three European languages (English, Italian and Spanish), employing NLP methods to enhance performance. We address also the acquisition of language data needed to develop these filters, and the evaluation of the system and its components.

1. Introduction

POESIA (Public, Open-source Environment for Safer Internet Access: IAP 2117/27572) is a multisite project funded under the EU Internet Action Plan, which is developing an advanced internet filtering system, intended primarily for use in schools and other educational establishments, with the aim of providing safe and educationally appropriate internet access for young people. The system is *open-source*, providing a basis for its development and maintenance beyond the project's lifetime, and is freely available for download and installation.[1] In this paper, we will sketch the overall POESIA system, and then provide greater detail of the methods used for filtering on the basis of textual content. Considerable quantities of web-page data are required both for the development and evaluation of the system. We will describe the approach taken for collecting this data, and the available early results on evaluation.

2. The POESIA System

POESIA's approach is to use multiple filters, each of which addresses some source of evidence that is of potential use in identifying harmful pages. The evidence detected can then be combined by a Decision Mechanism (DM) component to produce an overall decision for each page. In this way, POESIA can best exploit whatever information is available to determine whether pages should be filtered. Work to date on the system has focussed on the filtering of pornographic content, but the same mechanisms could be reapplied to other domains.

At the heart of the POESIA architecture is a central controller, the Monitor, which interfaces with the internet caching proxy (e.g. Squid or Shweby), or other filtering client, to determine the pages that must be assessed for filtering, and to return accept/reject decisions. The Monitor also invokes the other system components, and facilitates the traffic of data and results between filters and the DM, and caches filtering results for recently seen pages.

The POESIA filters include some that implement widely used filtering methods, e.g. positive/negative URL lists and PICS.[2] These methods are fairly effective for the sites and pages explicitly addressed, but given the enormous size of the internet and its dynamic character, these approaches can only ever achieve partial coverage. This suggests the need for filtering based on the *content* of pages, and POESIA includes filters addressing both image and textual content. The image filter assesses the likelihood that the images within a page are pornographic, based on the proportion of image area corresponding to skin, and on the shape and orientational characteristics of major skin areas. The multiple filters of the system should be seen as operating in combination. For example, a page from a site which is not on the URL lists will be analysed for content. If the page contains a reasonable quantity of text, this alone might allow a reject decision, but if there is limited text, it might require the combination of image and text evidence for a decision to be made. The DM plays a crucial role in weighing the available evidence to produce an overall decision.

The POESIA architecture readily allows for the inclusion of additional or substitute filters, and so the open-source character of the project allows for the continuing development and relevance of the system into the future.

3. Filtering for Textual Content

Within POESIA, three language-specific text filters have been developed by different sites which specialise in

[1]See `http://www.poesia-filter.org` for a full listing of the project partners, additional information on the system, and a link to the open-source repository from which the system can be downloaded.

[2]PICS (Platform for Internet Content Selection) is a scheme by which web content providers can assign labels rating the content of their pages, which can be used directly by a suitably configured browser to prevent children accessing pages with inappropriate content.

NLP for the target languages, which are English, Italian and Spanish. The filters differ in the methods they employ, partly reflecting an attempt to optimise over the different NLP resources available for each language. However, the filters are alike in offering both 'light' and 'heavy' filtering modes. Light filtering, which uses little NLP, provides rapid assessment of content for straightforwardly classifiable pages. For other pages, heavy filtering, making greater use of NLP, is invoked to provide more sensitive detection of content indicators. This trade-off is important to the overall efficiency of the system. In the case of pages that contain insufficient text for a conclusion to be drawn, filters can return a special result unknown.

The POESIA system includes a language identifier component, which is required to ensure that page text is routed to the appropriate language-specific text filter. This component uses a standard approach based on character n-gram statistics, see e.g. (Cavnar & Trenkle, 1994)).

4. Data Acquisition

The language identifier was trained using a large (∼560Mbytes), publicly-available, parallel corpus, which covers 11 European languages, including English, French, Italian and Spanish.

The development and testing of the POESIA text filters requires a substantial quantity of pages for each language, which have been precategorised as pornographic and non-pornographic. Manual collection of this data would be infeasible. Instead, the data was automatically spidered from the WWW, using the Google directory structure (http://directory.google.com) to locate sites which fall into the pornographic or non-pornographic category. The spider tranverses links within identified sites to retrieve pages at varying depths. Pages are stripped of HTML and the text is analysed to ensure it is of the target language and to highlight potential misclassifications. Despite these checks, there will inevitably be some number of pages which are incorrectly classified, and this fact will be reflected in the final performance scores. The corpus collected for each language ranges in size between 5k and 20k pages.

5. Text Filtering for English

The English light filter employs a conventional statistical approach to text classification, using a bag-of-words representation, with stoplisting and stemming. Indexing terms are selected via a minimum threshold for document frequency in the training corpus. A model is constructed of each category consisting of a ranked frequency list of index terms. Classification is done using an out-of-place measure over term frequency rankings.

The English heavy filter focuses on pages that have been misclassified as non-pornographic by the light filter during training. A set of keywords is identified from these pages, which are the n highest-ranking terms according to the tf.idf measure. A value of n=10 was found to be suitable in this context. As might intuitively be expected, these terms commonly appear to be indirect indicators of pornographic content, e.g. *adult*, *explicit*. An instance-based approach is used to learn contextual differences for the use of these keywords between pages that have been correctly and incorrectly classified as pornographic by the light filter. The contextual pattern is determined by a window of words around the keyword. The learning process can generalise these patterns by replacing words with their stem, POS tag or a named entity (NE) label, or with a "wildcard" symbol. The pattern matching process can either consider the absolute position of the adjacent words with respect to the keyword, or consider the preceeding and following words as an ordered list or unordered set. The best predictive patterns were produced by a 6-word window: a smaller window did not provide enough context to differentiate keyword use and a larger window did not improve the prediction. In addition, our experiments showed that no significant benefit resulted from allowing generalisation during learning to either POS tags or to NE categories of the kind produced by standard NE recognition systems, i.e. person, company, date, etc. However, benefit was found for generalisation by stemming and by a special case of NE recognition in which person names are categorised for gender. An approach of representing contexts as a list was found to perform better than one representing them as a set. At runtime, any documents that are classified as non-pornographic by the light filter, but which contain keyword occurrences, are passed to the heavy filter which applies the contextual patterns to determine the predicted document class.

The underlying approach of the English heavy filter can be seen as one of using local contextual cues to disambiguate between alternative uses or senses of key terms, as is relevant to particular categorisations. As such, the approach can be likened to that of (Riloff & Lorenzen, 1999), except that they use a linguistically richer representation, for example including aspects of syntactic structure. The simpler approach we have used has potential advantages in terms of portability (i.e. to other domains and languages), and robust application, since the contexts in which the keywords appear in html-stripped web pages may not correspond to grammatical sentences, and yet may exhibit regularities facilitating category prediction.

6. Text Filtering for Italian

The Italian light filter works at two levels. The first level employs a statistical word-based categorization, using local term counts rather than global frequencies. Text is tokenised and segmented into windows of 100 words. Each window is assigned a score based on the maximum local frequency of domain relevant words (markers). For each text, the filter outputs the maximum cumulative word score over different text windows. For efficiency, given the morphological richness of Italian, the morphological variants of ∼40 unambiguous marker lemmata, extracted from a linguistically annotated training courpus, are precomputed. The second level consists of recognition of relevant regular expressions, extracted from the training corpus, mostly associated with warnings (e.g. "adult content", or "download the dialler program"), with all possible lexical variations. Even though this recognition is implemented in the light filter, the identification of these expressions has required the use of some advanced NLP techniques, for the extraction of multi-word terms and the detection of semantic similar-

ity. Thresholds map text scores to low/medium/high values; low/high values are notified directly to the DM. For medium values, the heavy filter is invoked.

The heavy filter operates on morpho-syntactically tagged and lemmatized texts. For this purpose, we used a tool combining ILC's morphological analyzer MAGIC, and an optimized version of the Brill tagger. Filtering is based on recognition of $\sim$ 2400 domain relevant lemmata (including ambiguous words). Category learning uses an entropy-based classifier: CASSANDRA (Complex Analysis of Sequences via Scaling AND Randomness Assessment), which computes the rate of information increase generated by salient lemmata. Shannon's information S for the probability $P(x;l)$ of finding a fixed number x of "salient" lemmata in a moving window of length l was recently shown to give a maximal entropy change $dS/d(\log l)$, when genre-salient lemmata are selected (Allegrini, et al., 2004). A major role is played by the concept of scaling, defined by

$$p(x,l) = \frac{1}{l^{\delta}} F(\frac{x}{l^{\delta}}). \quad (1)$$

Complex systems, obeying Zipf's law, are expected to generate a departure from the condition of ordinary statistics, where $\delta = 0.5$ and $F(y)$ is a Gaussian function of y. The computation of the Shannon's information functional, in the case when the property of Eq. (1) applies, is easily proved to yield $S(l) = A + \delta \ln(l)$, where A is a constant whose explicit expression is of no interest here. It is evident that with this method the scaling parameter is easily evaluated by plotting $S(l)$ in a linear-log representation.

The CASSANDRA method works as follows. We study a time series that is not stationary. Then, we supplement the Shannon entropy method with the introduction of a big window of size L, which has to be considered as a sequence of its own, and we move it along the sequence being analysed, for the purpose of assessing its local properties. The size of this window has to be large enough as to make it possible to make a statistical analysis (in practice, we choose $L = 100$ words). For some positions of the big window we evaluate the quantity

$$C_j(\lambda) = \frac{\sum_{l=1}^{\lambda} [S_j(l) - S_j(1) - 0.5 \ln(l)]}{\lambda} \quad (2)$$

where $S_j(l)$ and $S_j(1)$ denote the Shannon entropies corresponding to small windows of size l and 1, respectively, moving within the big window with position j. Eq. (2) means comparing the actual entropy change to the ideal change occurring with an infinitely fast (Poissonian) transition to randomness, corresponding, to the entropy increasing as $0.5 \ln(l)$. The validity of Eq. (2) rests on the mathematical inequality $\lambda << L << N$. With the condition $L << N$, we can locate the big window in different positions of the text, identifying *where* the domain relevant (i.e. erotic) lemmata are meaningful, with a large information increase. The condition $\lambda << L$ makes it possible for us to use enough data to reach a conclusion about the statistical property of the small region under observation. This is why the CASSANDRA classifier is able to perform well in difficult tasks, detecting pages containing *erotic stories* vs, for instance, pages of *sexual education*, making a "wise" use of ambiguous terms. On the other hand, for pages with only a small amount of text, performance does not improve significantly over that of the light filter.

7. Text Filtering for Spanish

The Spanish light filter uses state-of-the-art text categorization techniques (Sebastiani, 2002). Text in Web pages is firstly tokenized, stoplisted, and stemmed. The top 1% Information Gain (IG) scoring terms of the training data are used to represent pages as term-weight vectors according to the Vector Space Model (VSM), using binary weights. A linear Support Vector Machine (SVM) classifier is trained over this representation, to classify new pages as either Porn or non-Porn. The Spanish heavy filter uses the same machine learning approach, but with two additional, linguistically motivated, multi-word input features: Noun Phrases and Named Entities.

- Noun Phrases are recognised via part of speech tagging and regular expression matching according to a compact noun phrase grammar. The part of speech tagger follows a Maximum Entropy approach (Ratnaparkhi, 1998) trained on the Spanish CONLL'02 corpus. The Maximum Entropy tagger has been iterated until an accuracy of 96% is reached on the training collection. The phrases found in the training phase are normalized by stoplist filtering, stemming individual words and alphabetical ordering.
- Secondly, Named Entities in the training collection are recognized using a subset of the attributes suggested in (Carreras et al., 2002), and the decision tree learner C4.5 trained on the CONLL'02 Spanish Corpus. Attributes considered in our approach include the actual words in a 5-word window around the target word, and capitalization properties of these words. The current version reaches a F_1=0.828 on the CONLL'02 test collection when considering only Named Entities but not their type (locations, persons, organizations and miscelanea). Again, Named Entities are normalized as Noun Phrases.

Named Entities and Noun Phrases are taken as additional features to stoplisted, stemmed words in a VSM binary representation. We retain the 10% top IG scoring features, and learn a linear SVM classifier over the training collection, as with the light filter. The evaluation of this approach has not yet been completed, but we believe these additional features will improve the effectiveness of learning, producing a more effective, if also more time-consuming, classifier.

8. Results and Discussion

To evaluate the various language-specific text filtering components, we have tested them in direct use as classifiers of pornographic vs. non-pornographic web pages.[3] The re-

[3]The filters normally provide an assessment of content as input to the DM, so this direct use as binary classifiers is not their normal context of use. The English and Italian filters may return a result 'unknown', when the decision for a page is unclear. In the results reported here, such pages are given a default assignment to either the porn (for Italian) or non-porn (for English) category.

ALL TEXTS			Light Filter					Light+Heavy Filters				
Language	Category	Pages	Prec	Rec	F_1	Eff	OvB	Prec	Rec	F_1	Eff	OvB
English	Porn	5090	.969	.938	.953	93.8%	3.2%	.967	.952	.960	95.2%	3.4%
	Non-Porn	4840	.937	.968	.952			.951	.966	.958		
Italian	Porn	3500	.948	.963	.955	96.3%	4.4%	.975	.953	.964	95.3%	2.0%
	Non-Porn	4195	.968	.956	.962			.961	.980	.971		
Spanish	Porn	1000	.995	.916	.953	91.6%	1.9%					
	Non-Porn	4000	.999	.981	.989							

Figure 1: Performance results for text filters

sults are given as precision and recall scores for each category (i.e. porn, non-porn) together with the corresponding F-measure scores (F_1). In addition to these familiar metrics, percentage scores are also given for two additional metrics which are widely used in a filtering context. These are *effectiveness* (Eff), which is the proportion of harmful pages blocked (here corresponding to recall for the porn category), and *overblocking* (OvB), which is the proportion of harmless pages that are incorrectly blocked (equivalent here to one minus recall for non-porn). Results are provided both for light filters alone, and for where the light and heavy filters of a language are used together.

Scores for the combined light/heavy filter for English and Italian indicate benefits for both languages of including the heavy filter, as shown by increased F_1 values. (The corresponding scores for Spanish were not available at the time of completing the paper.) However, the key benefit observed differs between the two languages. For English, we see a reduction in error rate for porn of 22.6%, i.e. so that the number of harmful pages incorrectly allowed through is reduced by nearly a quarter. For Italian, the key benefit is a reduction in overblocking, such that the number of harmless pages that are incorrectly blocked is reduced by $\sim$55%, although this is accompanied by some reduction in the effectiveness score.

Not surprisingly, the performance of text filters is significantly affected by the quantity of text within files. To take the case of English (although similar observations could be made for the other languages), excluding files that contain $\leq$20 distinct terms produces the following results:

$>$20 *terms*	Light filter			Light+Heavy		
Category	Prec	Rec	F_1	Prec	Rec	F_1
Porn	.979	.959	.969	.977	.976	.977
Non-Porn	.959	.979	.969	.976	.977	.976

Comparing to Figure 1, we see for light filtering that the F_1 rises from around .953 to .969 for both porn and non-porn, and for light+heavy filtering, F_1 rises from around .96 to around .977. For these higher content pages, the heavy filter reduces the error (misclassified pages) for porn by over 40%. Performance for the omitted low text content pages is accordingly lower (with F_1's around .90). However, we would expect pornographic pages with low text content to have high image content, and hence to be identified by the image filter, so that the combination of image and text filtering can perform more effectively than either alone. Evaluation of such combined filtering (i.e. image+text) is at a preliminary stage, but early results do suggest that this synergy of content filters does occur.

A question that might be raised regarding the overall POESIA system is whether it will allow for filtering of other languages, given that specialist NLP techniques have been used in the filters developed for the key target languages of English, Italian and Spanish. It should be noted, however, that the light filter systems developed for the three languages all employ generic text classification approaches, that can readily be reused to produce light filters for other languages, provided that a sufficient quantity of categorised training data can be acquired. This portability has been demonstrated within the project by the creation of a light filter for French using the code developed for the English light filter. In addition, the flexible architecture and open-source character of POESIA allow that heavy text filters for additional languages can be incorporated into the system should there by groups willing to develop them.

9. References

Allegrini, P., Grigolini, P. and Palatella, L. (2004). Intermittency and scale-free networks: a dynamical model for human language complexity, *Chaos, Solitons & Fractals*, v. 20, pp. 95-105.

Carreras, X., Márques, L. and Padró, L. (2002). Named Entity Extraction using AdaBoost. In *Proceedings of the Sixth Conference on Computational Natural Language Learning*.

Cavnar, W.B. and Trenkle, J.M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175.

Ratnaparkhi, A. (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Riloff, E. and Lorenzen, E. (1999). Extraction-based text categorization: Generating domain-specific role relationships automatically. In T. Strzalkowski, ed., *Natural Language Information Retrieval*. Kluwer AP.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.