# Text filtering at POESIA: a new Internet content filtering tool for educational environments*

**José María Gómez Hidalgo**
**Enrique Puertas Sanz**
**Manuel de Buenaga Rodríguez**
**Francisco Carrero García**
Departamento de Inteligencia Artificial
Universidad Europea CEES, Madrid, Spain
{jmgomez,epuertas,buenaga,fcarrero}@dinar.esi.uem.es

**Resumen:** Internet proporciona a los niños acceso a la pornografía y otros contenidos poco adecuados de formas mucho más expeditas que otros medios. Con el propósito de mejorar la efectividad de los filtros actuales, presentamos el proyecto POESIA, que pretende desarrollar y evaluar una herramienta de código abierto para el filtrado de material accesible por Internet en ámbitos educativos.
**Palabras clave:** Filtrado de texto, Código Abierto, Análisis de contenido textual, Procesamiento de Imágenes

**Abstract:** Internet provides to the children an easy access to pornography and other harmful materials. In order to improve the effectiveness of existing filters, we present POESIA, a project which objetive is to develop and evaluate an extensible open-source Internet filtering software in educational environments.
**Keywords:** Text Filtering, Open-source, Text Analysis, Image Processing

## *1. Introduction*

POESIA (Public Open-source Environment for a Safer Internet Access) is a new, in-development open-source tool for filtering inappropriate Internet content in educational environments. POESIA will filter several domains (including, at least, pornography, violence and racism), Internet channels (including, at least, Web and e-mail) and languages (English, Italian, Spanish, French), by using several technologies: text analysis, image processing, script code analysis, etc.

POESIA is being developed by a consortium with ten organizations: Istituto di Linguistica Computazionale (Italy), Commissariat à l'Energie Atomique (France), Ecole Nouvelle d'Ingénieurs en Communication (France), M.E.T.A. S.r.l. (Italy), Universidad Europea CEES (Spain), University of Sheffield (United Kingdom), Fundació Catalana per a la Recerca (Spain), PIXEL Associazione (Italy), Liverpool Hope University College (United Kingdom), and Telefónica Investigación y Desarrollo (Spain).

The system will run on a Internet access server at places where browsing and other Internet activities are undertaken, i.e. classrooms and libraries. POESIA aims at being more effective and flexible than current (commercial or not) tools.

Two are the main strategic elements at POESIA: the integration of knowledge sources and technologies, and its open-source nature. First, it is expected that, while isolated information sources lead to limited effectiveness filters, the effectiveness of the system will high by exploiting the combination of them. Secondly, the open-source nature of the software will promote external contributions (more effective filters, more languages and domains covered, etc.), and will allow a faster development process (by reusing available open-source tools as Squid, JigSaw, WEKA, and others). In fact, any external collaboration is welcomed.

## *2. System Architecture and Operation*

The approach is learning single information source classifiers, and a general classifier based on them. Each single classifier is

* This project is partly funded by the European Commission, Information Society, under the Safer Internet Action Plan. See http://www.europa.eu.int/information_society/programmes/iap/index_en.htm.

based on leading techniques: text categorization and (some) understanding for text filtering, abstract interpretation for scripts code analysis, etc. For each information source, two filters will be developed: a light filter based on superficial analysis techniques, and a heavy filter based on deep analysis techniques. The heavy filter will be used when the light filter is not able to take a decision. Text based filters will be language dependent, with an additional language recognition component.

## 3. Text Filtering at POESIA

Regarding Spanish text filtering, we follow a text categorization (Sebastiani, 2002) approach:

- For the light filter, a Vector Space Model (Salton, 1989) text representation has been selected, with stemming and stoplist filtering. A cost sensitive learning process based on Instance Weighting and Support Vector Machines has lead to high effectiveness for spam filtering in our previous work (Gómez, Maña, y Puertas, 2000; Gómez, 2002), so we will follow this approach for Web pornographic pages.
- For the heavy filter, a better representation based on Natural Language Processing is being designed. Techniques used for this task will possibly include: automatic extraction from the corpora of significant "terminology" (single words, cue phrases, fixed multi-word expressions, frozen text patterns, etc); construction of domain relevant thesauri/semantic lexicons; and shallow linguistic analysis techniques, facilitating identification of variable multi-word expressions and text patterns (named entity recognition, chunking, functional analysis, etc.).

## 4. First Results

A prototype of the Spanish light filter is readily available, in the form of a Muffin proxy filter. The prototype has been coded by reusing the learning environment WEKA[1] (Witten y Frank, 1999), the proxy system Muffin[2] and the library HTMLParser[3], and includes 26 Java classes and 2700 code lines. It is based on a binary vector text representation with stemming and stoplist, Information Gain (Sebastiani, 2002) term selection, and Support Vector Machines (Joachims, 1998) as learning approach. With a small sample of 55 training web pages (27 X rated pages), it has been evaluated by ten fold cross validation with 0.941 X precision and 0.806 non-X precision. In our opinion, the results are promising, and we expect to get much better results with a better text representation and a high scale Web page sample.

[1] Available at `http://www.cs.waikato.ac.nz/ml/weka/`

[2] Available at `http://muffin.doit.org`

[3] Available at `http://www.isacat.net/2001/code/`

## Bibliografía


Gómez, J.M., M. Maña, y E. Puertas. 2000. Combining text and heuristics for cost-sensitive spam filtering. En *Proceedings of the Fourth Computational Natural Language Learning Workshop, CoNLL-2000*, Lisbon, Portugal. Association for Computational Linguistics.

Gómez, J.M. 2002. Evaluating cost-sensitive unsolicited bulk email categorization. En *Proceedings of the the ACM Symposium on Applied Computing*.

Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. En Claire Nédellec y Céline Rouveirol, editores, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, páginas 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1398.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley.

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

Witten, Ian H. y Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann Publishers.