Content-aware proxy applications for the Mobile Web

Position Paper

Konstantinos Chandrinos,
Internet Content Filtering Group, IIT - NCSR "Demokritos"

November 18-19, 2004,
Barcelona, Spain

NCSR "Demokritos" is the biggest public research centre in Greece comprising eight institutes, which span the entire spectrum of research in physical science, from nuclear technology to computational intelligence, nanotechnology and molecular biology.

In the past few years, a group of researchers at the Institute of Informatics and Telecommunications have geared their expertise in semantic modelling and indexing towards Internet content. As a result, the Internet Content Filtering Group (i-config) has produced technologies for automatic resource discovery on the Web based on ontologies, multimedia content classification relying on machine learning, as well as user-tuneable content filtering, relying on probabilistic models. Underlying our work is the notion of cost-sensitive filtering and appropriate evaluation metrics, which proved very critical in minimising false positives, particularly in spam filtering.

Our obscene content analysis and filtering proxy, FilterX, has been selected by the EU, under the Internet Action Plan for pilot deployment and testing in various schools in Europe, ranging from Reykjavik, Iceland to Thessalonica, Greece, where students were offered filtering based on their age-group via the use of smartcards for age-verification. This allowed us to showcase the ability of the underlying technology to adapt its filtering strength automatically to varying user-types, content languages and user cultures.

Additionally, *FilterX* has been tuned to produce ICRA-compatible PICS pseudo-labels, with respect to nudity and sexual content, and is currently available as an *ICRAplus* plug-in, i.e. a filtering module of the latest ICRA platform (co-developed by ICRA, NCSR "Demokritos" and Optenet SA). *FilterX for ICRAplus* analyses HTTP content on-the-fly in the absence of ICRA labels (or lack of trust thereof) and returns a pseudo-label so that the platform can apply user-settings seamlessly. Extensive testing indicates an exceptionally high accuracy.

Currently, our team is working closely with ICRA in preparation of a module, which will take advantage of the newly proposed ICRA vocabulary expressed in RDF. In the same strand, our team is a technology provider for the "Quality Assurance and Content Description" (QUATRO) project (launched Nov. 1st, 2004). In QUATRO we intend to use ontology-based machine learning for the discovery and classification of Web resources, with respect to quality and trust. We are also working on IST-STREP project SHARE on the development of a semantic model for rescue-team operations, to be used for indexing incoming multimedia material on mobile platforms, under extreme conditions. Project SHARE will capitalise on and extend Push-to-Talk technology.

Our immediate interest and potential offer to the Mobile Web Initiative move along the axes of:

Automatic mobile-friendly resource discovery. We intend to transfer our expertise from leading EU-IST project "Cross-lingual multi agent retail comparison" (CROSSMARC), where we produced generic technology to discover Web resources combining expert ontologies and structural analysis to classify them. However, we would be interested to see how this effort can avoid the danger of Mobile Web segmentation, where users may find themselves confined to a virtual library.

Automatic content filtering and adaptation. Capitalising on the work of the Device Independence Working Group, we intend to look into methods for automatic content filtering and adaptation by a content-aware proxy. Given that RDF provides a framework for describing both content and delivery formats, we are keen to explore various paths of combining adaptation with filtering, be it label-based or relying on intelligent content analysis. Along with our expertise on real-time pseudo-label creation, we are interested in defining label trust issues and propose remedies.