# Text Categorization for Internet Content Filtering*

José M. Gómez, Ignacio Giráldez, Manuel de Buenaga
Universidad Europea de Madrid
Villaviciosa de Odón, 28670 Madrid, Spain
jmgomez,giraldez,buenaga@uem.es

**Abstract**

Text Filtering is one of the most challenging and useful tasks in the Multilingual Information Access field. In a number of filtering applications, Automated Text Categorization of documents plays a key role. In this paper, we present two of that applications (Hermes and POESIA), focused on personalized news delivery and Internet inappropriate content blocking, respectively. We are specifically concerned with the role of Automated Text Categorization in these applications, and how the task is approached in a multilingual environment. Apart from the details of the methods employed in our work, we envisage new solutions for a more complex task we have called Cross-Lingual Text Categorization.

**Keywords:** Text Filtering, Text Categorization, Cross-Lingual Information Retrieval

## 1. Introduction

Information Access is increasingly seen as an Information Society driving force. Information Access provides a lot of new opportunities but also generates new problems, that need new solutions. Intelligent services provide new solutions for quite different kinds of problems such as cognitive overhead or security. Intelligent services may include key elements such as advanced personalization or content analysis, among others.

Text Categorization – the assignement of documents to content-based categories – is a challenging and useful task in this context. Moreover, it is an important element for content filtering, specially in multilingual settings, demanding new approaches and paradigms. Content filtering is the core of many intelligent services.

*This work is partly funded by the Spanish Ministry of Science and Culture, through the PROFIT plan, and by the European Commision, under SIAP-2117 contract.

In this paper we discuss the role of some key concepts in the scenario sketched above, and the specific technology and integration of methods that we have developed for two R&D projects in which our institution is involved. First, the Hermes project had as main purpose intelligent news personalization in a multilingual setting exploiting automatic text content analysis and user modelling. Second, the POESIA project provides a safer environment for Internet access, so that inappropriate contents are avoided for users such as kids in the schools. In both projects, Text Filtering of documents (news stories, Web pages) and Text Categorization are essential elements for the solutions provided.

This paper is organized as follows. First, we discuss the role that Text Categorization plays in Text Filtering applications. Secondly, we describe the dominant learning-based approach to Automated Text Categorization that are applied in the systems described in this paper. Next, a number of alternatives in a multilingual setting are presented, and the Cross-

Language Text Categorization problem is introduced. Then, we describe the two R&D projects mentioned above, focusing in the role of multilingual Automated Text Categorization and how it is solved. Finally, we present some concluding remarks.

## 2. Filtering by Text Categorization

Text Filtering (TF) is a text classification task in which a "system sifts through a stream of arriving information to find documents relevant to a set of user profiles," in words of Hull and Robertson for the TExt Retrieval Conferences (TREC) series [20]. The arriving information can be news items, e-mail messages, etc. The user profiles express long-term and specific information needs, in contrast to other text classification tasks (e.g. Information Retrieval – IR involves short-term information needs; Text Categorization – TC involves rather static and community-defined interests; etc. – see [3, 23]).

Two basic approaches to filtering are defined in the literature [26]. In the first approach, called *content-based*, the selection of documents is made according to intrinsic properties of them (e.g. document text contents, structure, etc.). In the second one, called *collaborative*, the selection is made according to annotations or judgements made by other users (a kind of cross-marketing approach: users that like this item also like this other one, so the system will recommend similar items to similar users – e.g. [14]). Despite there is no doubt collaborative filtering helps to address multilinguality issues, we focus in this paper in multilingual content-based TF.

One of the most important issues in content-based TF is the way in which users specify their long-standing information needs, or in other words, how user profiles are defined [3]. The definition of the filtering task in the TREC series is specifically concerned with this issue. For example, all filtering task instances in TREC-9 are defined in terms of user profiles and possibly judged documents. The most simple definitions of profiles, such those used in pioneer systems like SIFT or InfoScope (see [26]), are made in terms of words manually entered by users to specify their information needs, much like IR search queries. Some current filtering mechanisms are often based in this kind of user profile: popular e-mail client systems like Eudora[1] or Outlook provide tools that allow users to define filters and profiles in terms of words ocurring in sections of email messages.

Other definitions of profiles are also possible. For instance, given a set of content based categories[2], users can specify their information needs in terms of these categories. For example, a user susbscribed to a library announcements e-mail service could ask the system to send her a message when new items arrive the library, if they belong to the "Computational linguistics" and the "Natural Language Processing" categories, but not to the "Information Science" category. When dealing with news stories filtering, the user profile is often based in previously defined categories. Current online newspapers allow readers to subscribe to the desired newspaper sections, getting an e-mail message with the headlines of news items daily put by editors in those sections. Web portals like Yahoo! provide personalised news services in which users subscribe to predefined, content-based categories.

Text Categorization is the assignment of documents to predefined categories [30]. Categorization makes sense as an intermediate task in Information Access, making easier to get information in document collections (either static, like when searching in Web directories or libraries, or dynamic, like in information filtering). Quite often, categories are assigned to documents by experts (library cataloguers, newspaper editors, etc.). This is a rather costly task, and may be even unfeasible. For instance, when combining different information sources (e.g. news stories from several news agencies or newspapers) for an information (news) delivery application, there is not a standard set of categories (each news agency uses their own set). A new set of categories must be built, to allow users to subscribe those which are of interest for them. New documents must be accurately and on-line classified according to the new set, a process that

[1] These names and other cited across this paper are trademark of their respective companies.

[2] Themes or topics like e.g. controlled keywords assigned to research reports in the ACM Digital Library, or book descriptors as the Library of Congress Subject Headings – LCSH

must be automatic, leading to what is called *Automated Text Categorization* (ATC).

ATC has attracted considerable attention from the research community in the latter years. In this paper, we focus on the role of ATC for two information filtering R&D projects, namely Hermes and POESIA. Hermes [13] is a multilingual news delivery system, in which users daily receive a personalized newspaper by e-mail. The personalized newspaper is built by selecting news stories from a British and a Spanish newspapers, according to a rich user profile defined in terms of Yahoo! categories (and other preferences). Since news articles are not originally categorized according to these categories, we have built an ATC module that performs this task.

POESIA [16] is an Internet filter that address inappropriate content blocking for kids and youngsters in schools and libraries. POESIA covers Web and e-mail for the English, Spanish and Italian languages, and is centered on pornographic, violent and gross-language content on-line detection. Currently in development, we describe our work in the Spanish Web pornography domain. We address pornography detection as TC, in which Web pages have to be classified according to two categories, PORNOGRAPHIC and SAFE. Therefore, the core of the filtering system is an ATC module able to classifiy Web pages with respect to those categories.

# 3. Learning Based Automated Text Categorization

ATC systems are able to autonomously classify documents into a set of predefined categories. To build such systems, two basic approaches are possible [30]. First, a team of classification experts (cataloguers, editors) and Computer Science professionals can build such systems by hand. In this so-called *knowledge-based* approach, classification knowledge is made explicit in the form of a set of classification "if-then" rules, much like those writen by users of e-mail filtering functions of popular e-mail clients. For instance, a system called Construe was developed by Carnegie Group for the reuters news agency, which used rules much like the following one:

$$(\text{“wheat”} \in D \wedge \text{“farm”} \in D) \rightarrow \text{WHEAT}$$

This rule, extrated from [2], means that if the words "wheat" and "farm" are found in a document $D$, then it must be classified in the category WHEAT (economic news reports dealing with wheat). Obviously, making classification knowledge explicit is laborious and expensive, affected by a *knowledge acquisition bottleneck* problem.

An alternative to this approach is making use of IR and Machine Learning (ML) techniques to semi-automatically build ATC systems. This *learning-based* approach starts with a set of manually classified documents in the target categories, called a *training collection*. Then IR and ML techniques are applied to learn a classification function called a *classifier*. This model is based on several elements:

1. The method for representing or *indexing* documents. The most usual one is representing documents as term weight vectors, according to the IR Vector Space Model (VSM) [28]. In this method, usually called the "bag of words" model in the literature, terms are words after filtering with respect to a stoplist and stemmed with a lemmatizer such as Porter's. The weight of each term in each document can be binary (1 if the term occurs in the document, 0 otherwise), TF (Term Frequency – the number of times the term occurs in the document), or TF.IDF (where IDF means Inverse Document Frequency, usually defined as $\log_2(n/df(t))$, in which $n$ is the number of training documents, and $df(t)$ is the number of training documents where the term $t$ occurs).

2. The method for selecting terms. In order to avoid over-fitting in the learning process, and to increase its efficiency and effectiveness, a subset of the original terms is often selected. This is done by computing a quality function to the terms, and selecting those that score higher. In [32], Yang and Pedersen empirically demonstrated that using the Information Gain (IG) and the

$\chi^2$ metrics, upto the 99% of the original terms can be deleted, getting an important efficiency improvement, and increasing the classification accuracy.

3. The learning algorithm. In the latter years, a wide variety of learning approaches have been applied to the problem, including Bayesian classifers as Naive Bayes (e.g. [24]), decision tree learners like C4.5 (e.g. [7]), rule learners as Ripper (e.g. [8]), kernel methods like Support Vector Machines – SVM (e.g. [21]), and many others (see [30] for a survey). Accuracy of learning methods is variable, being SVMs among the most effective ones.

The learning-based approach to ATC therefore involves two kind of processes, as presented in the figure 1. This figure is an adaptation of the ones by Belkin and Croft [3] for the IR and TF tasks. The previous techniques are used to learn a classifier in a batch process. When new documents have to be categorized, they are represented as term weight vectors, using the indexing vocabulary derived for training documents, and afterwards they are fed into the classifier to get their respective outcomes.

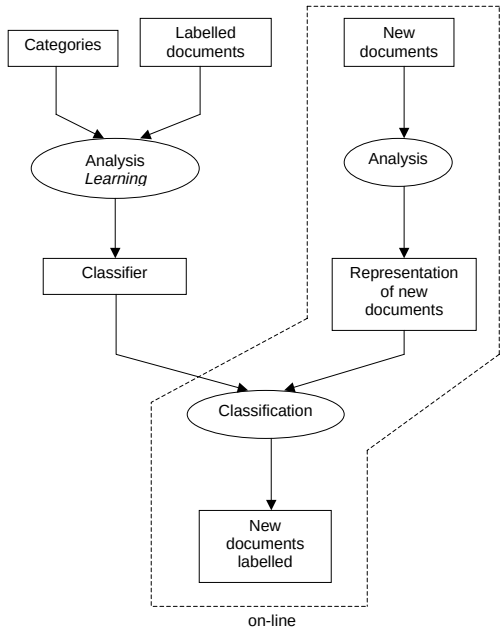


**Figure 1: Overview of processes in ATC.**

In order to evaluate the quality of a classifier, most researchers have been concerned with effectiveness or accuracy. Given a set of test documents, labelled with respect to the target categories, and called *test collection*, the classiffier is applied to them and the decisions made by it are examined by using IR metrics, specifically recall, precision and $F_1$. The learning-based approach is quite effective for thematic ATC. In [30], it is argued that nowadays learning-based classifiers are nearly as effective as human beings for TC. From the point of view of using learning based ATC systems for Information Filtering, the possibility of using autonomous and effective systems is quite promising and enables building many applications that would be unfeasible otherwise.

# 4. Multilinguality issues in ATC

An important issue for developing practical filtering applications is how multilinguality issues are addressed. It is remarkable that most of the techniques used in the learning-based model for ATC are language independent, perhaps excluding low-cost components like the stoplist and the stemmer. Therefore, a rather sensible approach is, given a set of training documents for each target language, to induce a language-dependent, learning-based classifier for each of the languages. If the language of an incomming document is not known in advance, these classifiers are complemented with a language identification system. Language identification is also a TC task, that has been solved in the literature (see e.g. [5]) with relatively simple techniques, reaching 99% accuracy for a range of languages. This approach is mentioned in e.g. [2], where the authors also suggest to avoid stemming and using a stoplist (only the most frequent words are considered as indexing terms in their work).

While in the broad Information Access field, much work has been devoted to address multilinguality, it has been often focused on IR and TF. ATC has been perhaps considered out of the arena, because of the previous consideration. In short, when enough training data is available for each language, the learning-based approach nearly solves the problem of categorizing multilingual collections of documents. However, enough training documents are not always available for all categories, a problem that is made harder when working with several

languages. It is sensible to consider scenarios in which training data is available only in one or two languages, and it is needed to categorize documents in several more. We call this task *Cross-Language Text Categorization* (CLTC).

Some of the work in Cross-Language IR (CLIR) is of application to CLTC. Three approaches to CLIR are considered predominant, according to the latest Cross-Language Evaluation Forum workshop [6]. Given a query expressed in one language, and texts in the same language and several others, these approaches consist of:

- Translating the user query to all target languages, retrieving documents for all languages separately, and combining the results into a single ranking.
- Translating all text documents into the query language, and operating as monolingual retrieval.
- Translating the user query to all target languages, append all queries into one multilingual query, and retrieving documents according to it all in the same run.

We would like to remark that these approaches require fast, accurate and domain-independent Machine Translation (which itself needs accurate Word Sense Disambiguation - WSD) for real world applications (e.g. Web search). This requirement is clearly far from the state-of-the-art methods. Also, except for the second, these methods are not easily mapped into the CLTC problem. There is a straightforward correspondence between user queries in IR and categories in ATC. However, it is not clear how models induced by some learning methods (e.g. decision trees, neural networks) can be translated from one language to the required languages when a multilingual text collection is to be labelled.

An alternative to this approach is using a controlled, language independent vocabulary for representing documents and queries in CLIR. This way, a classifier can be learned for documents in several languages, which is able to classify documents in terms of a language independent representation. Concepts in multilingual lexical databases like EuroWordNet (EWN) are specially suitable for this task. We believe that this is the most promising approach to CLTC, given the increasing work in using the monolingual lexical database WordNet for ATC (e.g. [4, 12, 22, 29, 31]), and some work on bilingual ATC [25] using a controlled bilingual thesaurus. Importantly, this approach also requires WSD, because you need to know which concepts are referenced by words and other collocations in actual texts. However, some works in IR using WordNet suggest that it is feasible to improve classification when WSD is not as accurate as expected (see e.g. [19]).

Finally, we would like to remark that in the two R&D projects discussed in this paper, we have followed the first of all approaches, given that we had training documents for all target languages and categories used for filtering.

# 5. HERMES: a Multilingual Personalized Newspaper

In this section we present Hermes, a multilingual news filtering system which allows users to receive personalized messages containing news extracted from digital versions of several European newspapers, using several languages.

## 5.1. Task Description

Nowadays, many newspapers offer web access to their contents. Moreover, users can subscribe to newspapers' services and receive daily news by e-mail. Unfortunately, most of them are simple transcriptions of their printed version. More advanced systems include user-profiling options which allow users to define what kind of information they want to receive.

In this context, there are two main approaches to define user interests about content. First, category-based systems list some categories –usually newspaper sections–, and users pick up categories considered interesting. After user profile definition, the service sends all news stories contained at selected categories at a daily basis. The second approach uses term-based descriptions to define user interests, so that users give interest-related keywords –and sometimes their interestingness degree–. Be-

fore sending daily messages to each user, the system selects which news items contain stored user-keywords, orders selected items by relevance, and includes the most relevant items in the final message to be sent by e-mail. Both approaches can be integrated, letting users define personalized categories by keywords (that is, each user-defined category is represented by a keyword list), in a way similar to Yahoo!'s stored searches. In this way, services are required to automatically categorize each news item, for each user-defined category, to build each user final message.

In the Mercurio [11] system, our efforts were oriented to offer personalized information access by integrating all previously defined user-profiling methods in a monolingual setting. However, there are many circumstances which favour multilingual information systems, specially in the current European Union context, where information flow involves interactions with documents in several languages. This kind of requirements promote Machine Translation and multilingual services –for instance, EU official institutions make use of translation systems like EC Systran, and multilingual Information Access like CELEX.

The synthesis of a multilingual personalized newspaper comprises the following subtasks:

- Storing data about the users of the system including a model describing the topics the user finds interesting.
- Finding the news stories (in several languages) discussing those topics.
- Integrating those news into a personalized multilingual newspaper.
- Sending to each user, through e-mail, his personalized multilingual newspaper.

## 5.2. System Operation

Hermes [13] has been designed for providing personalized news according to user interests, sending by e-mail a message with a set of news (title and a summary). For generating each message, both user and news information are retrieved, formally represented and properly processed to obtain the final resulting news representation for each user. News representation is based on the Vector Space Model (VSM), performing selection with a simplification of Rocchio algorithm that was previously applied with satisfactory results.

The first thing required for the system to work is the user information. In the sign-in process, each user fills in a form which collects preferences about news, interests and delivering settings (e.g. days to receive messages). Hermes collects each user information and stores it internally twice, a copy in English and another one in Spanish. Of course, users can also access again to this form to modify their preferences. The information about users topic interests is far richer than in other TF systems, since they are allowed to define keywords, and specify which sections and content-based categories they are interested in. We have selected two generally accepted categories systems (Yahoo! and Yahoo! Spain first level categories), in order to provide users an alternative to section-based selection.

Then, the system acquires news information. Everyday Hermes connects to e-newspapers, one in Spanish and another one in English, and gets textual content from each news item. These are processed to obtain their summaries and an internal representation according to the VSM: each news item is mapped to a vector which ponderates each term relevance. Each item is also categorized with respect to Yahoo! categories, to be sent to users who selected resulting category.

Once user prefences and news items are equally represented by term weight vectors, Hermes computes the relevance of news stories to user models. A ranking of the news items is obtained according to their relevance for a given user. Top items in the ranking are selected for delivery to the user according to the upper bound on the number of items per message specified in their profile. The selection process is based on a simmilarity formula between the representations of user profiles and news stories. In this formula, a linear combination of the relevance of user selected content-based categories and keywords is computed. Also all news items are processed to check if they belong to one of the sections selected in the user model.

When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the level of interest that the user assigned to each of the different reference systems. This implies that users looking for the same information but having chosen different methods to specify their interest may get different results. For the relevance values provided to the user to be easy to interpret, they are normalised over the number of selection methods involved in obtaining them. In this way, the system can quote a final relevance value in the $0 - 100$ range to every user, regardless of the number of selection methods that he chose. Selected news summaries for each user are included in his e-mail message.

There is a feedback easy-to-use interface in each item, which allows each user to specify his interest degree on that item. The user profile is modified according to his judgements of news items, and promotes the smooth evolution of profiles.

## 5.3. System Architecture

The Hermes system is composed by the following modules:

- **Information Server Agent:** This agent accesses the news items of the newspapers (in English as well as in Spanish) and selects those news items relevant for every user.
- **User Modelling Agent:** This agent creates a model of every user of the system. This model contains:
  - Management information: name, login, password and e-mail address of the user.
  - Messages information, including the days the user wants to receive the messages, a maximum and minimum of news per mail, a way to indicate that no more messages should be delivered (e.g. for holydays), preferred language for the model, and from which sources the user wants to obtain the news.
  - Information interests, including which sections of the newspaper and which categories the user is interested in. Also the terms chosen by the user are stored in the model. These terms are words, written in the language defined in the model, that the user considers interesting.
- **Translator Agent:** This agent provides the translation of terms required to compute the relevance of a news story written in a language different from that of the user model.

## 5.4. Text Categorization Approach

Each news item is categorized by the Information Server Agent using a generally accepted categories system (Yahoo! and Yahoo! Spain first level categories). This way, users can specify which categories (and to which degree) they are interested in, apart from classic newspaper sections.

The ATC method we have appplied in Hermes, fully discused in [18], is based on a VSM representation of documents and categories, in which terms are stemmed words previously filtered using a stoplist, and weights are binary. The learning method is based on a simplified Rocchio formula [27], which computes the weights of each term for each category. The simmilarity between a news item and a category is computed using the cosine formula of the VSM [28]. In ATC, it is often required a binary (yes or not) decision for each document and category, and thus a threshold is selected. This way, all documents whose simmilarity to a category is above the threshold are assigned to the category. That is not an issue in Hermes, because simmilarity between news items and Yahoo! categories is directly taken into account in the news selection formula.

A categorization submodule has been developed for each of the languages, which is only different in the stemmer and the stoplist used. The category system is unique, in the sense that there exists a nearly one-to-one correspondence between first level categories in the Spanish and English versions of Yahoo!.

An important issue is the collection of training data, because news stories are not originally categorized in newspapers according to Yahoo! categories. Instead of manually classifing a set of training news according to these categories for each language, we have taken the Yahoo Web pages as training data. For each category, its Web page and the Web pages of inmediate subcategories are taken as documents in the category. This is a strong violation of one of the most important requirements of learning-based ATC, in which training and new documents are assumed (and required) to be similar. However, according to the indirect evaluation described below, the system is as accurate as required to perform an effective news selection.

## 5.5. Evaluation

In order to evaluate the system, we have to consider two aspects: evaluation of the performance achieved by the system, analyzing measurable parameters, and an evaluation that considers user global satisfaction (usability) with the system [10].

The evaluation process involved a group of 23 users taken from the research team, a group of students (Computer Science and Journalism) and some users that are neither related with computers nor journalism. Evaluation takes into account quantitative and qualitative aspects of the system, made by the selected users. Qualitative analysis have been made using data taken from objective parameters of the system, and a final valuation that reflects user satisfaction with the system and its results. Then, a report is done by evaluators showing positive and negative aspects of the system.

In order to make the quantitative analysis, we designed a test with different groups of questions about the interface, categories and sections, summaries, and the bilingual features, which allows us to evaluate the system. Once we processed the answers, we used the results to support the qualitative analysis of the system, reducing or reinforcing the evaluation given by the evaluators when describing the system. The study of the results taken from both types of analysis, allows us to select the positive features and to identify its deficiencies.

### 5.5.1. Evaluation results

The overall satisfaction of users is very high. They show a high degree of satisfaction with the interface; the existence of a tutorial and an introduction manual to the system is very helpful, as well as the method for helping the user when entering the profile data. However, this last issue may be improved.

Profile configuration has also been graded very well. Users are highly satisfied by the feedback system and with the possibility of choosing the days of the week in which they receive the news. Also, high satisfaction has been reported regarding the way the categories, sections and terms can be selected when building the profile. Abstracts are considered very good, specially regarding the way they are shown and the information provided in them.

Another well appreciated aspect is the possibility of choosing the language. However, users think that the translation of the selected terms on the user profile could be improved. Another feature suitable for improvement is the correspondence between received news in both languages according to the terms selected in the profile.

The overall evaluation of retrieved documents analyzed is very high, especially regarding the quality of the contents and the relevance to the interests of the user. Evaluators reported that contents of the final documents satisfy their information needs. They were also satisfied with the way the received news offered new knowledge about other documents related to them. As a final observation, evaluators consider the system as a very interesting tool.

We also performed an empirical evaluation of the quality of the summaries, that aims to test if they show the most important information present in the original news items. We have defined three profiles and manually found the most relevant news items for each of them. We have compared the ability of the system to select right news items according to those profiles by using the full text, and by using only the automatically generated summary. The results of this evaluation, are promising and support our claim that it is better to present user adapted summaries instead of generic ones (see [1] for

more details).

#### 5.5.2. Evaluation of Text Categorization

We have not performed an empirical evaluation of the ATC method used in Hermes. To do so, a significative number of categorization decisions should have been collected and reviewed, which was out of the scope of our project. Instead, the user satisfaction showed above demonstrates that the ATC method is as accurate as required to make the suitable selection of news according to user profiles. This is a kind of indirect evaluation, because it focuses on meeting the global requirements of the system, instead of getting empirical results for the ATC module.

However, results from the direct (empirical) evaluation of our core ATC module in other research projects show that the methods employed are as accurate as required, although they might be improved [18].

# 6. POESIA: a Multilingual Inappropriate Internet Content Filter

In this section, we describe a multilingual inappropriate Internet content filter called POESIA[3], focused on on-line detection and blocking of several kinds of information in English, Spanish and Italian, for kids accesing to Internet in schools and libraries.

## 6.1. Task Description

Current information facilities for kids, at schools and libraries, include Internet access, which brings a big opportunity for them to get new and up-to-date knowledge with several educational purposes. However, dangerous or inappropriate behaviour is also possible, ranging from giving personal information, to surfing inappropriate Web sites (with pornographic or violent content), and even performing illegal activities such as infringing copyright laws in massive music downloads (with a practical consideration, that is consuming the possibly narrow bandwidth in inappropriate activities). The society is greatly concerned with this problems, and the problem is being addressed through a number of innitiatives, including the Safer Internet Access Plan of the European Community. In this plan, awareness campaigns are in development, hotlines have been created, and filtering products are being funded to address these problems.

POESIA [16] is one of the filtering products being developed under this Safer Internet Action Plan. POESIA, currenly under development, is an opensource online filtering solution, intended for schools and libraries, and focusing on Web and incoming e-mail filtering. POESIA will cover pornographic, violent and gross-language Web content, and pornographic junk-email, for the English, Spanish and Italian languages. Filtering will be performed in an "on-the-fly" fashion, coping with the evolutive nature of the Internet content. Also, security of filtering is promoted by a client-server architecture, in which the product is installed in a proxy-cache server, which is hard to hack by advanced students.

One of the most important goals of POESIA is achieving high accuracy by the integration of a number of technologies, including advanced image processing, Language Engineering, PICS and JavaScript analysis, etc. In this context, our institution is responsible for the analysis and processing of Internet Spanish text content, in coordination with the University of Sheffield (USH) for the English language, and the Istituto di Linguistica Computazionale (ILC) for the Italian language. We are approaching inappropriate Internet content detection and blocking as an ATC task, for which there are essentially two categories: INAPPROPRIATE and SAFE (with, of course, degrees and a division in kinds of contents – pornography, violence, etc.). In the next sections, we describe the operation and architecture of the system, and how ATC is being solved by us for the Spanish language.

## 6.2. System Operation

The operation of filtering systems focusing inappropriate content is quite simple. Our working scenario is the following one. The typical POESIA system runs on a PC under Linux

[3] See `http://www.poesia-filter.org/`.

(POESIA will not be targeted to Windows during the project), and has two network connections: an inside one to the classroom local area network, and an outside connection to the Internet. The POESIA box can be the teacher's station; it can also be remotely administered through Web interfaces; the teacher or responsible adult can explicitly permit or deny access to any content. For Web filtering, a content request (e.g. an HTTP GET request to an HTML page) is handled as follows:

- POESIA checks if the requested page is already in its cache. If it is in the cache, the cached filtering scores are read from the disk, are fed into a decision mechanism, and produce quickly a decision to reject or accept the request. If the request is accepted, the cached content is also read from the disk and returned to the user's browser (on the internal LAN connection).

- Otherwise (content is not cached) POESIA needs to fetch the page from the requested Web site (through the external Internet connection) and stores it into the disk (i.e. caches the content); it may also communicate with other near POESIA systems to retrieve - if it is available - the content and the filtering scores from them.

- Then POESIA uses the common filtering library; this library contains a set of specialized filters, and a decision mechanism. The filters are run, their filtering scores are written onto the disk (caching of filter scores).

When a Web content is rejected, the browsing student gets the usual 403 Forbidden HTTP reply in an HTML page containing some explanations (e.g. scores with a message like "suspicion of harmful content"), a form to explicitly ask permission to his teacher. The student can then either give up accessing the page or ask authorization. When the responsible adult explicitly gives permission (or confirms prohibition), this permission is cached and the content may be filtered again to adapt the POESIA system (i.e. to tune the decision mechanisms).

## 6.3. System Architecture

The system is being designed to meet a number of requirements: modularity (for promoting the extension to other languages and types of information), security (with a client-server architecture), efficiency (which is hard to get with online content processing), standards accomplishment (in order to plug it in standard tools), and, of course, accuracy.

According to these requirements, a number of modules have been defined and are being under development, including:

- A module for **image processing**, responsible of analyzing and rating pictures included in Web pages and e-mail messages.
- A module for **text processing** in each covered language, which analyzes and scores the textual part of the Web pages and messages. This module is supplemented with a **language identification** module, that automatically detects the language of the text in the analyzed content.
- A module for **Platform for Internet Content Selection (PICS) ratings analysis**, which uses standard self-assigned content ratings.
- A module for **URL and JavaScript processing**, which focuses on outgoing URLs included in the text and JavaScript code part of Web pages and messages.
- A module for **decision making**, which combines the scores given by the other modules into one rating, in order to take a single {yes, no} decision.

Given that Web pages have to be served as quickly as possible, POESIA filtering is two-layered: a first level of crude filters is first activated when scores are not stored in the cache. This "lite" filters will compute quickly some filtering scores. This should permit quick filtering of most contents. A second layer of elaborate or "heavy" filtering agents is activated when crude filtering is not effective, i.e. for contents perceived as complex by the POESIA system. This last layer performs more elaborate but more time consuming filtering.

The crude filtering (which combines several quick filtering technologies, e.g. light image processing, light text processing, PICS and light URL filtering, with some quick decision taking mechanism) is organised into a common filtering library, which also provides hooks to more elaborate filtering agents and to a more complex decision mechanism.

Clearly, technologies used in light and heavy filters are different, specially for text processing. Also, the experience of different partners in POESIA is leading to a variety of solutions regarding text filtering. For instance, the USH and our University are developing lite and heavy filters with, to some extent, the same techniques, and starting with lite filtering; the ILC has instead started with the heavy filter, cutting expensive functions to allow quick responses in the lite filter.

Regarding standarization, the POESIA filter is being developed as an Internet Content Adaption Protocol (ICAP)[4] module. This makes easy to integrate it in comercial applicationas like Network Appliance products, or other opensource systems like Squid. This latter one is specifically addressed in POESIA.

### 6.4. Text Categorization Approach

Here we describe the ATC approach we are following to develop the Spanish text filtering agents, focusing in the pornography Web content domain. Most of the work is similar to the one USH is doing for their agents.

Our previous experience on Unsolicited Bulk Email (UBE) or *spam* filtering has guided our current approach to ATC for pornography detection [15]. A fact that makes this problem special from the point of view of the ATC techniques employed is that it is a *cost sensitive* problem. As in UBE detection, there are two kinds of possible mistakes: blocking a safe content, and allowing an inappropriate one. According to user studies (on parents, teachers, etc.), they prefer the filter to make the first mistake than the second one, because this latter is more dangerous. Given that perfect accuracy is unreachable with current technologies, a method has to be developed in order to promote the system making low cost (i.e. less dangerous) mistakes.

Most of the solution being developed by us in the lite filter is based on well stablished and effective techniques, like a binary representation of the text, selection of terms according to the Information Gain metric, and using Support Vector Machines as learning method (see section 3). However, we are conducting research on cost-sensitive learning, evaluating methods that are able to make learning algorithms cost-sensitive (including e.g. MetaCost, Instance Weighting and others – see [15] for a review).

Heavy filtering is approached through more expensive, but (due to efficiency requirements) shallow Natural Language Processing methods. These include the following:

- 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.
- Shallow linguistic analysis techniques, facilitating identification of variable multiword expressions and text patterns, including tokenization, morphological analysis and lemmatization, named entity recognition, "chunking" (i.e. segmenting a text into non recursive phrasal nuclei (e.g. 'base' Noun Phrases)), identification of other (non-phrasal) collocations, and functional analysis (i.e. annotation of grammatical relations (such as subject, object etc.)).

More precisely, we are currently performing experiments with a Maximum Entropy Part of Speech Tagger for Spanish, that allows to make regular expression matching for Noun Phrases detection. This way, multiword but well motivated phrases will be considered as terms for more accurate classification. Also, we are developing a Named Entity Recognizer using current top-performing methods, according to the most recent work in the Computational Natural Language Learning workshops shared task. We expect that some popular names in the porno-

[4] See `http://www.i-cap.org/`.

graphic domain will be excellent indicators of this kind of content.

Most work in lite and heavy filtering will be based on a corpus of pornographic and safe Web data pages we have collected from the Web. In order to speed up the development of the filters, we are reusing a number of opensource systems and resources. On one side, we are using quite sophisticate software libraries like the Waikato Environment for Knowledge Analysis (WEKA) for Machine Learning, the General Architecture for Text Engineering (GATE), and the OpenNLP package[5]. On the other one, the building of our text corpus is partly based on an Yahoo!-like opensource Web directory, called the Open Directory Project (ODP)[6].

Finally, we would like to remark that most of the techniques used for lite filtering are language-independent, following the approach to multilingual ATC we described in section 4. However, and even while we are trying to make our work as language-independent as possible, most of the heavy filtering will be based on language-dependent data and techniques. Nevertheless, even the most promising approach to Cross-Lingual Text Categorization, based on concept indexing as described above, implies building expensive linguistic resources as multilingual lexical databases like EuroWordNet.

### 6.5. Evaluation

In the next months, a beta version of the whole prototype system will be available to end-users and external developers through the project website[7]. End-user partners are responsible for conducting an evaluation on it, but given that the beta version is focused on building the overall architecture of the system, we do not expect it to make very accurate filtering.

In the meantime, we are performing experiments with our Spanish text pornographic lite filter. Initial results using a binary representation and SVMs on a very small collection of tens of URLs are very promising (0.941 precision on pornographic content and 0.806 on safe content). We are building a training collection using the ODP data, that now is composed of 5,000 safe URLs and 1000 pornographic URLs. We plan to extend it to around 100,000 URLs. Next evaluation will be based on the Receiver Operating Characteristic Convex Hull method, as described in [15].

[5] These packages are available at `http://www.cs.waikato.ac.nz/ml/weka/`, `http://gate.ac.uk/` and `http://opennlp.sourceforge.net/`, respectively.

[6] See `http://dmoz.org/about.html`.

[7] See `http:\\www.poesia-filter.org`.

## 7. Conclusions

In this work, we have analyzed the role that personalization and content analysis may play in the context of two intelligent Information Access services: multilingual news personalization and inappropriate contents filtering; we have specially discussed the integration of elements to make specially useful Text Categorization for this services in a multilingual setting. Also, we have experienced the quality improvement of services offered to the users considering the specific domain and purpose.

The role of ATC in the two applications described above is critical, and virtually makes them possible. The multilingual approach that we have applied to ATC in this context, while straightforward, has proven effective. However, we have taken here the oportunity to sketch a more challenging multilingual scenario for ATC, that we have called Cross-Lingual Text Categorization, in which research is just emerging. In our opinion, research in this area may be better focused to using concept-based indexing vocabularies as language-independet text representation method. Although in a monolingual setting, we are performing substantive research in this area [4, 9, 17], and looking forward to practical situations in which this scenario makes sense. One promising area is inappropriate content filtering, because we have a preview of the difficulties we will have for hard domains like violence detection.

## References

[1] I. Acero, M. Alcojor, A. Díaz, J.M. Gómez, and M. Maña. Generación automática de resúmenes personalizados. *Procesamiento del Lenguaje Natural*, 27, 2001.

[2] C. Apté, F. Damerau, J. Fred, and S.M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.

[3] N. J. Belkin and W. B. Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM*, 35(12):29–38, 1992.

[4] M. de Buenaga, J.M. Gómez, and B. Díaz. Using wordnet to complement training information in text categorization. In *Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97*, volume 189 of *Current Issues in Linguistic Theory (CILT)*, pages 353–364. John Benjamins, 2000.

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

[6] A. Chen. Cross-language retrieval experiments at CLEF-2002. In *Results of the CLEF 2002 Cross-Language System Evaluation Campaign: Working Notes for the CLEF 2002 Workshop*, 2002.

[7] W.W. Cohen and H. Hirsh. Joins that generalize: text classification using WHIRL. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173. AAAI Press, 1998.

[8] W.W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2):141–173, 1999.

[9] J.C. Cortizo, M. Ruiz, and J.M. Gómez. Concept indexing based automated text categorization. In preparation, 2003.

[10] A. Díaz, P. Gervás, and A. García. Evaluating a user-model based personalisation architecture for digital news services. In *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries*, 2000.

[11] A. Díaz, P. Gervás, J.M. Gómez, A. García, M. de Buenaga, I. Chacón, B. San Miguel, R. Murciano, E. Puertas, M. Alcojor, and I. Acero. Proyecto mercurio: Un servicio personalizado de noticias basado en técnicas de clasificación de texto y modelado de usuario. *Procesamiento del Lenguaje Natural*, 2000.

[12] F. Fukumoto and Y. Suzuki. Learning lexical representation for text categorization. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 2001.

[13] J.I. Giráldez, E. Puertas, J.M. Gómez, R. Murciano, and I. Chacón. HERMES: Intelligent multilingual news filtering based on language engineering for advanced user profiling. In *Proceedings of the Multilingual Information Access and Natural Language Processing Workshop, VIII Iberoamerican Conference on Artificial Intelligence (IBERAMIA)*, pages 81–88, 2002.

[14] D. Goldberg, D. Nichols, B.M. Oki, and D.Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[15] J.M. Gómez. Evaluating cost-sensitive unsolicited bulk email categorization. In *Proceedings of SAC-02, 17th ACM Symposium on Applied Computing*, pages 615–620, Madrid, ES, 2002.

[16] J.M. Gómez, M. de Buenaga , F. Carrero, and E. Puertas. Text filtering at POESIA: A new internet content filtering tool for educational environments. *Procesamiento del Lenguaje Natural*, 29:291–292, 2002.

[17] J.M. Gómez, M. de Buenaga, L.A. Ureña, M.T. Martín, and M. García. Integrating lexical knowledge in learning-based text categorization. In *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data*, 2002.

[18] J.M. Gómez, R. Murciano, A. Díaz, M. de Buenaga, and E. Puertas. Using linear classifiers in the integration of user modeling and text content analysis in the personalization of a web-based spanish news service. In *Proceedings of the 1st International Workshop on New Developments in Digital Libraries (NDDL) , International Conference on Enterprise Information Systems (ICEIS)*, 2001.

[19] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[20] D. Hull and S. Robertson. The TREC-9 filtering track. *SIGIR Forum*, 33(1), 1999.

[21] T. Joachims. A statistical learning model of text classification with support vector machines. In *Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval*. ACM Press, 2001.

[22] M. Junker and A. Abecker. Exploiting thesaurus knowledge in rule induction for text classification. In *Proceedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing*, pages 202–207, 1997.

[23] D.D. Lewis. *Representation and learning in information retrieval*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, US, 1992.

[24] D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, pages 4–15. Springer Verlag, 1998.

[25] N.V. Loukachevitch. Knowledge representation for multilingual text categorization. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval Electronic Working Notes*, 1997.

[26] D.W. Oard. The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, (7):141–178, 1997.

[27] J.J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system: experiments in automatic document processing*, pages 313–323. Prentice-Hall, 1971.

[28] G. Salton. *Automating text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley, 1989.

[29] S. Scott. Feature engineering for a symbolic approach to text classification. Master's thesis, Computer Science Department, University of Ottawa, Ottawa, CA, 1998.

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

[31] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi. Event tracking using WordNet meronyms. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 2001.

[32] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, 1997.