

# Online Content Filtering: Technology Overview

Konstantinos Chandrinos  
*i-sieve technologies ltd.*  
*chandrinos@i-sieve.com*

## Abstract

*In this paper we present an overview of online content filtering methods and technologies. We outline their strengths and weaknesses and discuss what is feasible and what's not.*

## 1. Introduction

Online content filtering is a collective name for a combination of technologies that enable users to select automatically or semi-automatically the type of information they will receive, according to preferences or needs.

At the time of writing, the prime medium of electronic information exchange is the Internet. We therefore analyse the possibilities of content filtering and information interception with the Internet in mind, although most of the remarks and conclusions are valid for other foreseeable information carrying networks.

We also need to point out the differences between illegal, offensive and undesirable content. Illegal content is material that violates the law in certain jurisdictions, such as material promoting paedophilia in most of the world, or pro-Nazi content in Germany. Offensive content is material, which, although legal to distribute to adults through some media (e.g. pornography), may be illegal to distribute through particular media such as broadcasting services, or be illegal to distribute to minors. Finally, undesirable content is material that users may consider irrelevant or unproductive (e.g. online gambling, day-trading sites, etc.) according to their preferences or needs.

Ideally, illegal content should be dealt by legal authorities and blocking of offensive or undesirable content should be passed on to filtering solutions. However, the current structure of online information exchange does not allow for the definition of clear jurisdiction and law enforcement. A striking example of the complications is the legal case against a major Web company in France, where the French-based

counterpart of a US company governed by French law was challenged to prohibit access to auctions of Nazi memorabilia to French citizens. A number of renowned technical experts testified that the technical implementation of such a decree would be close to impossible.

In the following paragraphs, we first provide an overview of the technical design issues pertaining to the Internet, the most widely used information exchange network, and then move on to analyzing techniques and methods that could be used to provide efficient content filtering, focusing on enabling the user's choice.

## 2. How is the Internet structured?

The generic term "Internet" is often used to describe either an electronic information delivery mechanism or services that can be delivered via this mechanism (Web, e-mail, etc.) or both. The basic characteristic of the Internet is that it is a "packet switched" network. This means that, unlike the telephone network, there is no dedicated real or virtual line that joins the sender and the receiver of information. Instead, the message being transferred from sender to receiver is broken to small chunks, or packets. These are transmitted independently from sender to receiver, with no predetermined route. A number of technology solutions, beyond the scope of this document, have been suggested and implemented to improve the robustness of such a scheme. To facilitate delivery and reconstruction at the receiver end, these packets include a minimal header defined by the TCP/IP protocol. The header typically includes the address of the sender, the address of the recipient and the ordinal number of the packet. These packets may be routed through alternative routes and some may even be duplicated on their way, all efforts aiming to their prompt arrival at their destination. It is important to point out that the Internet protocols cannot guarantee delivery. They rather adopt what is known as a "best effort" approach to information delivery, providing ways to detect and recover from delivery errors.

The Internet has been heralded as a novel concept in information communication. As such, it is hard for

anyone to provide a concrete analogy with known schemes of information flow. Still, we are tempted to suggest a functional analogy that will serve the purposes of identifying filtering opportunities: Consider a system of water delivery, full of complex piping interchanges, akin to the systems used to provide water to modern cities. We further consider that water comes into the system from a number of different sources, some of which may, either at certain times or always, introduce impure water. Turning on the tap will signify, for our purposes, a request for water from a particular source. As the possibly impure water runs through the piping system it gets mixed with clean water. In the eyes of the above analogy, filtering information on the Internet is close to filtering water in this water dispense system. One can filter impure water at the source, somewhere along the piping system, or at one's tap. A mental leap is required to imagine that water "entities" maintain information of their source, destination and order in the original water flow.

Already, this broad outline of how the Internet works, suggests three diverse approaches to content analysis and filtering: in principle, content can be identified and blocked at the content producer, at the distribution mechanism or at the content consumer. Before we explore technical possibilities and limitations of these approaches, we outline problems related to understanding online content in order to filter it.

### **3. Semantic labeling of information**

Electronic information is packaged in meaningful data groups called files. Unfortunately, the "meaning" of these files is only apparent to a human user. At the time of writing, there is no semantic information readily available in data files, less so in data chunks, which could be used by machines to interpret such information. We are left with intelligent information extraction techniques and artificial intelligence to combine extracted features so as to approach a semantic understanding of that information. One should bear in mind, that a machine need not actually "understand" information the way humans do. For our purposes, it should interpret it enough in order to perform filtering.

Lack of such "understanding", coupled by the semantic ambiguity inherent in human languages has led first generation search engines to return unacceptable results to keyword-based queries. Language engineering and machine learning have made significant progress in recent years (e.g. fact extraction) and a number of

surrogate techniques have been developed to side-step result ranking problems (e.g. link-analysis as a quality measure). However, even current search engines suffer from the lack of context placement of user queries.

Upon recognizing the expansion of the Internet and the necessity for machines to be able to determine the semantic attributes of online content, the World Wide Web Consortium (W3C) has introduced the Platform for Internet Content Standard (PICS) for Web based documents. PICS is a generic standard to define schemes which will allow the insertion of machine-readable metadata in the header of an HTML document. Such metadata can then be processed by software before taking further action. Current popular browsers are PICS compliant, in the sense that they can be configured by the end-user to accept or reject Web content, based on rules interpreting such metadata inserted as HTML head tags. It should be clear that this methodology relies on the content author and/or content host taking the effort to provide PICS compliant metadata. Alternatively, the W3C suggests methods to implement third-party rating, where trusted third parties provide the service of a ratings bureau. Clients can be redirected to consult such services before accessing content, with minimal network tax. It is reasonable to compare this self-labeling scheme with the system followed in many countries (but not all) with respect to movies or TV programs: a visible indicator ranks a TV program according to a predefined scheme, in a scale that starts from "All audiences" up to "Restricted for minors". In fact, the United States have gone one step further, to suggest that TV sets may optionally include electronic parts (the so-called "V-chip") which can recognize such metadata transmitted along with the signal at the beginning of the program and block it according to parent settings for the protection of minors.

With respect to harmful and offensive online content, the Internet Content Rating Association (ICRA) an independent industry body featuring as members a number of key industry players have been working on defining metadata schemes for Web content. In fact, they first adopted RSACi, a scheme suggested by the Recreational Software Advisory Council (RSAC) based on the paradigm of computer games. RSACi allows rating of individual Web pages or sites on a scale of 0-4 along the dimensions of sex, nudity, violence and language. A typical RSACi meta-tag would look like this:

"http://www.rsac.org/ratingsv01.html" 1 r ( n 3 s 3 v 0 1 4) gen true for

"http://www.site.com" r ( n 3 s 3 v 0 1 4)'

ICRA have been faced with the fact that such a poor granularity of categories cannot cope with the diverse Internet content. For example, how should nudity be treated when it is presented in a medical context? As a result, a second effort of ICRA attempted to rectify this by taking context into account: the meta-tag, produced after the interested author answers a longer online questionnaire, provides context-dependent tags. Recently ICRA has taken advantage of W3C's Resource Description Framework (RDF), a very powerful mechanism to provide semantic metadata to online content.

ICRA offers a free, browser-independent, software module for the MS Windows operating system called ICRAplus. ICRAplus can be extended with software modules and promises to combine the best of both worlds: self-regulation and filtering.

#### 4. Content filtering at the source

As the Internet stands, it is impossible to enforce filtering at the source. Users with a dialup connection can easily log into the network, acquire a temporary IP address and start spreading information without their ISPs ever getting a chance to know what kind of information is facilitated through their networks. In fact, in recent legal cases ISPs have claimed that they can be no more liable for the content they carry, without external notification, than a telephone company for discussions taking place in its voice networks. Moreover, the content creators have no foolproof way to know if the requesters of this information are entitled by law, age or preference to receive the information in question. The only resort left to good-willing content creators is to label their content in a pre-agreed way, which would allow recipients to screen it.

Content clients could be revamped to transfer a set of content rules along with their request and content servers could be forced to implement matching policies before dispatching the information. However, the disruption of the Internet and the costs incurred would be tremendous. Furthermore, the legal issues that would arise in an attempt to force ISPs all over the world to perform policy checking would be far too complicated.

#### 5. Content filtering at the distribution mechanism

Content on the Internet flows through unpredicted paths, determined spontaneously by routers. Although a complete technical description of routers is beyond the scope of this document, we outline below their chief characteristics focusing on their abilities and limitations to perform content filtering.

Routers are special-purpose hardware that examines the destination address contained in the header of Internet packets and acts in a number of ways, which include forwarding the packet to a destination nearer to the final destination, duplicating the packet or dropping it altogether due to network congestion. Eventually, packets arrive to the last router close to the destination and are streamlined towards the client, which bears the responsibility of assembling, re-ordering and perhaps requesting again missing packets of the information. Routers can be used for packet-based blocking of content as well. Since routers are expected to minimize the packet processing to avoid network congestion, the only viable solution to this day is to maintain "black lists" of URLs from which it is known that illegal or undesirable content originates. In such a scenario, a router can inspect the source address of a packet and compare it against a look-up table of forbidden URLs. If the packet is found to originate from such a URL, the packet is blocked. There are obviously two questions that naturally arise:

- how would a router know of the requester's laws, age or preference to determine if the black list applies?
- how can this black list of URLs be kept up to date?

Inability to answer these questions has dictated that router-based filtering solutions are implemented at the last router before the user. As this router is to be configured by the user or the appropriate network administrator, it is thought that this person can define and maintain an access control list of URLs to be off-limits for the user(s) in question. However, the general framework of router-based filtering is problematic: it is often the case that an entire site or even an IP subnet is blocked, just because sometimes, some packets coming from there are offensive or undesirable (e.g. www.[free-hosting-site]. com). It is not rare that legitimate e-commerce sites are (unwittingly) co-hosted with pornographic sites under virtual hosting which resolves to the same IP address. Keeping the list up to date is no small effort per se, given the dynamic IP

scheme, the unresolved and numeric IP duality as well as the IP spoofing/tunneling techniques currently available.

Packet-level blocking could be compared to current spam-prevention schemes, which maintain lists of suspicious mailservers: this could easily result to indiscriminately rejecting e-mail messages emanating from [famous-free-email-service].com. Until routers become more intelligent as to how they treat packet content, they cannot offer true content filtering.

## **6. Content filtering at the consumer side.**

It sounds reasonable that a value-adding service such as filtering should take place at the consumer (client) side. After all, it is the consumer of information that requires something extra, hence the producer should put no serious effort in satisfying all possible consumers. Filtering solutions in this category may address a large number of users if they are implemented on a proxy server or other gateway mechanism that caters for an entire organization, school or library. They could even apply to an ISP's special service implemented on a transparent proxy.

There is currently a wide range of options, when it comes to filtering on the client side, the simplest of which is the implementation of policies relying on the self-regulation described earlier. Self-regulation on the Internet requires agreement on a universal standard as well as a great amount of volunteer work. It has been realized that both these requirements delay widespread adoption. Internet gives practically anyone the opportunity to become content creator. Since our world is far from ideal, the responsibility required for self-regulation is absent and therefore the critical mass of labeled content for the system to be viable is simply not there yet. Although TV channels, barring satellite transmissions, can be forced by local legislation to adopt self-labeling of programs, the Internet cannot be treated the same way. Additionally, manual third party rating of a large proportion of the Web can be a costly effort, both in time and money. Last, whatever granularity a universally accepted standard entails, it is optimistic to believe that it could ever depict preferences or moral values of diverse communities. Of course, this is not to say that self-labeling and public awareness of the Internet will not be an important part of the solution in the years to come, but it is probably not sufficient on its own.

Another option offered to clients is the use of software that maintains local lists of appropriate and inappropriate URLs. This software usually traps the TCP/IP stack of a client PC and redirects all requests through a software module that consults a look-up table of allowable or forbidden URLs. Alternatively, if implemented on a proxy server for a group of users the system examines all specified requests (e.g. http, ftp, etc.)

As the Web is huge and dynamic, these lists face the problem of maintenance and validity, which is not alleviated seriously by the automatic update subscription schemes suggested by vendors. Also, the client-side list-based approach suffers from potential blocking of entire sites, in the same manner that this applies to router-based solutions.

To counterpart these inefficiencies, many vendors introduced surrogate shallow keyword matching. If a requested URL is not on the prescribed lists, then the URL and/or the textual content it contains are scanned for keywords drawn from a keyword list compiled by humans. These keywords are chosen to reflect certain undesirable categories (e.g. "sex", "adult", etc. under obscene content, or "bomb", "instructions" under terrorism) and if they are present in the content lead to rejection. Clearly, this is an unintelligent approach to the problem bringing one back to the very first search engines indexing the Web with a full text indexing mechanism. It has also brought a lot of derision to the filtering industry, by attracting media attention to "false alarms". Such false alarms include pages on the White House web server, Christian Associations Web sites and even pages of political candidates at past US elections, which were blocked due to "suspicious" words in a different context. What is perhaps more important is the human intervention in compiling the lists and choosing the keywords. Many people have complained that this obscure list-based strategy is not shy from straightforward censorship. Fear and speculation are enhanced by the fact that these blacklists are kept under encryption as a valuable company asset by filter vendors. When radical free speech activists and hackers managed to crack the encryption of certain software filters in the past, it was revealed that certain sites advocating non-mainstream beliefs about political or social issues were blacklisted under inappropriate categories, such as "sex and pornography". As the size of these lists grows to millions of URLs, even if we could take the vendor's good will on face value, it would be hard for someone

to know that the lists have not been maliciously tampered with.

## 7. Next generation filtering software

Recently, a number of technologies have been suggested that could lead to an automatic or semi-automatic analysis of online content without imposing the bias of human judgment. Most of these technologies rely on pattern recognition, attempting to extract textual information from the content. The information is used to compile a set of features that typically characterize a document category. Success of these systems has been varying and most of them have been included either optionally in filtering packages or used in the background to assist humans in populating lists of categorized sites.

In previous paragraphs we have presented the main problems of traditional filtering mechanisms based on lists: they present serious underblocking (i.e. they allow inappropriate content to slip through) and they are hard to maintain and often biased. In contrast, automatic methods, like keyword-based ones, suffer more or less from overblocking (i.e. they block legitimate content).

Efficiency of filtering systems can only be judged with the original intention in mind. If one wishes to make sure that no questionable content can skip the filter, one can employ a simple technique of keyword detection that scans the content against an extensive list of “suspicious” keywords and blocks if it finds even one of them. Obviously, this scheme will block a considerable amount of legitimate information, but the original intention would have been satisfied. However, overblocking can occur using even more sophisticated methods, due to flawed feature extraction or evaluation. Last, but not least, certain domains of filtering can spark controversy between human judges as to what constitutes overblocking or underblocking. To take one example, photographs by [famous-photographer], depicting nude men and women in leather restraints, are alternatively treated as art or pornography.

The research team behind i-sieve has worked in the past two years to develop an automatic methodology that minimizes the probability of overblocking in all filtering operations. More specifically, they designed a methodology for the automatic extraction of features and their ranking according to their filtering capabilities from (user-suggested) examples. Features are extracted from available media that comprise a complex document (e.g. text and images for a Web

page). Next, a machine-learning algorithm has been tuned to identify the contribution of each feature, irrespective of originating media. Last, an evaluation function has been crafted, which takes into account the features present in online content under examination and fuses their contribution under a probabilistic framework. The evaluation function returns a probability of the document belonging to a certain category. Thresholds upon this probability can be tuned by users to provide blocking of content or not.

The described methodology has been implemented in various application domains, ranging from Web filters for obscene content and e-mail spam filters, to e-news live feeds. Scientifically rigorous tests have been carried out which indicate that the systems present an underblocking behaviour below 3% (i.e. less than 3 out of a hundred items skip the filter) keeping overblocking at a few decimals distance from 0% (i.e. practically no legitimate items are wrongfully blocked). The implemented systems are fast and scalable. Furthermore, the systems can be augmented with content lists, preferably supplied by their own usage, to speed up decisions, sidestepping analysis of recently analyzed content or avoiding analysis of trusted content.

Nevertheless, it must be stated that due to the nature of the Internet it could be very challenging to achieve 100% blocking success of prescribed material, without total blocking of the entire network. To take an analogy from the computer security world, a 100% virus-safe computer is a stand-alone PC with no drive. A simple countermeasure to real-time filtering would be for the objectionable content producer to encrypt all traffic with strong cryptography and provide special clients for deciphering on the consumer side. But the objectionable material transmitted over the Internet is hardly worth the cost and effort of encryption and decryption.

Unlike contaminated water, objectionable material is not known to kill on the spot. As a result, a “best effort” approach could be adopted: organizations with extensive usage of the Internet, schools, libraries and other Internet facilities as well as individual users should exercise their right to access information they want to receive, without losing information rejected by obscure procedures. They should do so relying on automatic methods they can fine tune; otherwise they will be overwhelmed by uncontrolled and often undesired online content.