Technical Whitepaper

Astaro Content Filtering Process

Last changed: May 2002
Version: 1.06

Introduction

Traditional Internet filtering methods depend on manually compiled blocking lists, individual ratings or online applied heuristics algorithms. These methods are, for the most part, inadequate, cannot keep up with the growth of the Internet or result in high numbers of false positives. As a consequence, inappropriate content is often allowed through the filter while acceptable content is blocked.

Cobion instead uses a new approach to Internet filtering. The Cobion Content Filtering process, methodically and automatically scans the complete Internet and categorizes each website by its content using a proven combination of intelligent text classification and superior image recognition methods.

The scanning of the Internet is performed by Cobions Supercrawler, which inspects millions of new and updated websites every day. The categorization of all websites is done automatically using advanced technologies and is powered by a super computing infrastructure which provides the computing power that is necessary for this process. The result is a fresh and daily updated database of the Internet.

Cobion analyzes and independently categorizes Internet content into 58 categories. Currently, Cobion provides customers with a URL Database that contains more than 13 million categorized web page entries. This knowledge is based on the inspection and categorization of more than 1.8 billion web pages and images from the Internet. Cobion started this process in 1999 and since then has improved the quality of the content filtering process, expanded computing infrastructure and implemented new technologies to become more accurate and more up-to-date every day.

The service that Cobion provides to its customer is available under the name COFS (Content Filtering Service). This service contains the fresh and daily updated Cobion URL Database with 58 categories and is provided along with content security products for customers worldwide.

Content Technology Platform

Cobions Content Technology Platform is the underlying platform for classifying Internet content with millions of web pages every day. This platform operates Cobions Supercrawler, performs the analysis and inspection of web sites, images and other content using massive parallel computers and manages multiple database clusters to cache and store web site information, hyperlink structures, images, web site text and other important content. Furthermore, the process of updating the URL Database and the maintenance of the COFS is also handled through Cobions platform.

Crawling the Internet ...

Crawling the Internet is a challenging and ever ongoing task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with millions of web servers which are all beyond the control of the system. Cobion operates a fast distributed crawling system that is able to visit millions of web servers every day. The crawling of the Internet is based on a kind of "snowball"-principle. Starting at one website, the crawler downloads all HTML text and all images of this website and stores this content for further analysis on Cobions ContentStorage.

Then, the crawlers follow all hyperlinks (URLs) to other websites that are contained in the current website. This way, the crawlers follow deeper and deeper all hyperlinks and download all content until no more unknown hyperlinks are found.
The crawling strategy on how to follow hyperlinks is adapted dynamically. For instance, the crawlers have the priority to first visit newly discovered hosts and domains instead of going deeper on the same host. Also, crawlers do not download massive amount of data from the same host on one single visit. Instead, crawlers visit one host multiple times and perform multiple downloads.
To cover ‚islands' in the Internet (i.e. parts of the Internet, where no links from other websites refer to), Cobion systematically feeds the crawler system with fresh information about new websites, domains and hosts based on public hostlists, domain registry information and other external sources.
Beside the crawling process, the crawling system also performs an update and maintenance process. Both processes run in parallel, so that a part of the crawlers are searching new content and the other part is updating. This is important for constantly updating the content of known websites. Websites that change more often are being crawled more often. This process is also adapted dynamically to keep up with the ever changing nature of the Internet. With the current system, the crawlers download up to 3 million new images, 3 million new pages and approximately 6 million new hyperlinks every day.

Website Content Analysis

Once the crawlers have downloaded website and image content, all content needs to be analysed and categorized. For Cobion, Internet filtering means more than just a simple keyword search or URL and filename analysis. Cobion runs multiple analysis processes on each website to get the highest level of quality for categorization.

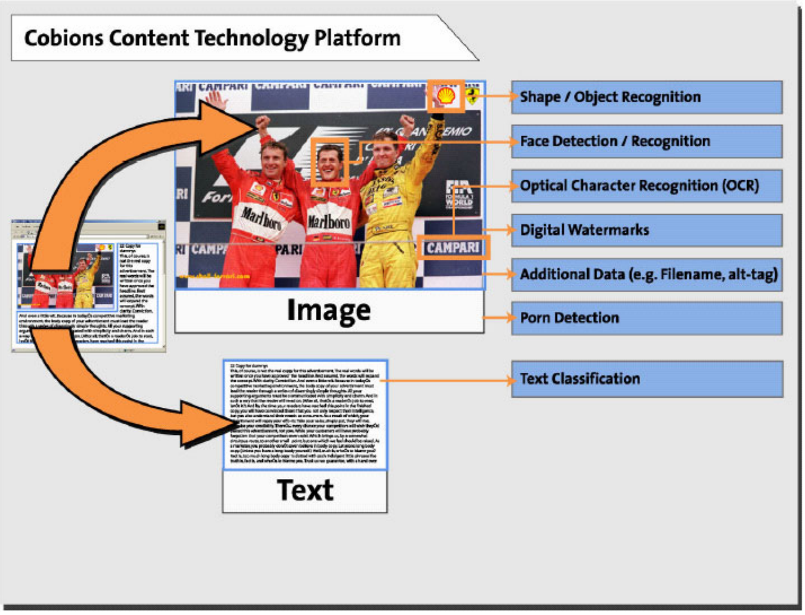
Cobions Content Technology Platform
Image
Text
Shape / Object Recognition
Face Detection / Recognition
Optical Character Recognition (OCR)
Digital Watermarks
Additional Data (e.g. Filename, alt-tag)
Porn Detection
Text Classification

Figure 1 Applying Multiple Content Analysis Technologies

As shown in the figure above, the analysis consists of the multiple steps that produce important information and metadata for the final categorization of a website.

Text Classification

Both, keyword search and intelligent text classification are used to analyze textual content of a web site. The keyword search decides depending on the occurrence of certain words, to which category the words belong to. Disadvantage of this procedure is that many words occur in different meanings (e.g. sex) and therefore are difficult to categorize. Advantages are high performance and the easy configuration. Also, this methods operates well when only a few words are available, for example when classifying a URL.

Intelligent text classification not only classifies by single words, but on the basis of frequency of occurrence and combination of words. Text classification is done using word heuristics and combinations hereof together with Support Vector Machines for the final decision process. The Cobion text classification technology has a very high reliability, therefore basically no errors occur if the number of words is large enough.

Both text analyses methods can be applied to different sources. Whereas for sources with only a few words, the keyword classifier is preferred (OCR, URL, filename, etc.) For longer text information, e.g. whole web sites, text classification is chosen.

Visual Porn Detection

Visual Porn Detection is an image analysis technology that is able to detect a high concentration of flesh tones in images. For increased accuracy, Cobion uses face detection. If a face is detected in an image, a sample color is taken from the skin. Then it is checked if from the head large portions of skin is attached, whereas the size of the head can also be taken into consideration. This kind of recognition is very reliable as the information about the skin tone are directly taken from the face. This also decreases overblocking, since portrait images are not rated as pornographic. If no face was found in an image, statistical assumptions of the skin character are used. This procedure is clearly less reliable, as for the skin detection no information from the image itself can be extracted.

Visual Object Recognition

This technology analyses each image for special signs, symbols, trademarks etc. The method currently is used to recognize forbidden symbols like the swastika in Germany. Furthermore major credit card logos, sport brands, car brands and others are detected. This method is very important for some categories.

Visual Optical Character Recognition

A lot of textual information on a web site is found in images. Cobion performs Optical Character Recognition on each image and processes this information with the above mentioned text classification methods. Since text information within an image has a high relevancy regarding the image content, this method improves overall accuracy.

Overall classification

The final classification is performed by combining the results of all applied methods with a fine tuned weighting for each method.
Since multiple methods contribute to the overall classification, erroneous classification by a single method is eliminated by other classification processes.

Global Data Center

The Global Data Center is the heart of the Content Analysis of Cobion. The system of the Global Data Center searches the Internet and analyzes all content. It is the largest facility of its kind in the world and processes up to 50 million web pages and images each day. The Global Data Center consist of two major operational parts, the Supercomputer and the Control Center.

Figure 2 Supercomputer

The Supercomputer is a clustered PC architecture with 1000 CPU´s and provides the massive processing power for the Global Data Center's content analysis operations and high speed database searches. The systems performs multiple different jobs, like web crawling, text analysis, image analysis, website categorization and many more. Architected for scalability, it can be readily expanded to meet any level of demand.

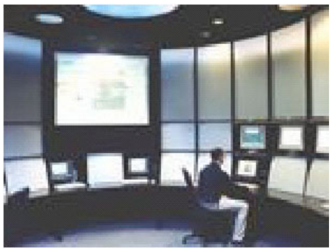

Figure 3 Cobion Control Center

The Control Center is the main operational facility for process control and system monitoring of Cobions Global Data Center. All essential information is routed through the Control Center that is staffed 24x7x365 by highly skilled professionals.

URL Database Production

The production of the final URL Database for the Content Filtering Service is the essential step in building a high accurate Internet filtering list. Various information from multiple sources are combined using a proprietary scheme. Each source of information contributes to the final categorization in a certain level with a defined reliability.
The probability and reliability of each source and each analysis method have been derived from the categorization process of millions of websites by a trained team of experts over the past years.

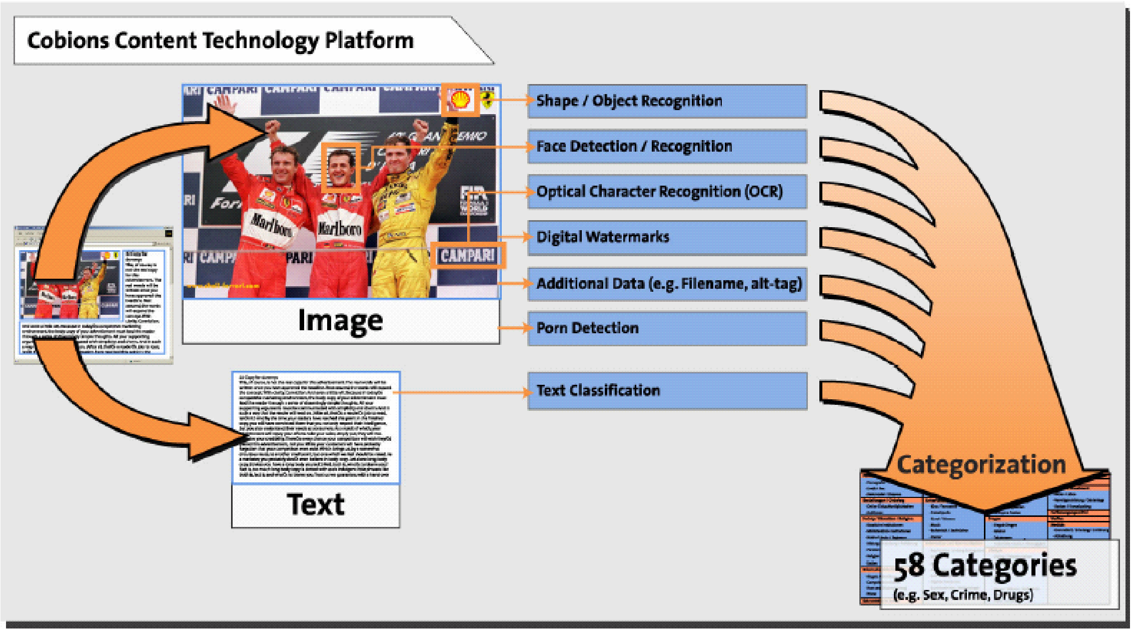
Cobions Content Technology Platform
Image
Text
Shape / Object Recognition
Face Detection / Recognition
Optical Character Recognition (OCR)
Digital Watermarks
Additional Data (e.g. Filename, alt-tag)
Porn Detection
Text Classification
Categorization
58 Categories
(e.g. Sex, Crime, Drugs)

Figure 4 Website Categorization using multiple information sources

In addition to automatically analyzed website and image content, Cobion also uses managed link lists, newsgroups, search engines and other additional sources.

What is contained in the URL Database?

The database contains multiple information about domains, hosts and URLs of the following types:
Domains: sex.com
Hosts: www.sex.com
Directories: www.sex.com/pics/
HTML pages: www.sex.com/pics/index.html
Image URLs: www.sex.com/pics/001.jpg
IP addresses: http://194.12.2.3

Database Update Process

The COFS URL Database is updated daily. To get always the suitable updates the following process is used:
Within a configured time interval the COFS server establishes a secure HTTP connection to Cobions COFS master server. The COFS Database server retrieves an XML document with information about the actual available update files on the master server. With the knowledge about its own update state, the COFS server can determine which update are required to be up to date. The COFS server finally downloads the appropriated update files and merges this information into the database. This is a background process and does not interfere with the normal activity of the COFS server in any way.
The update workflow is transaction based. If the process is interrupted at any state (for example if the connection to the master server breaks down) the update will continue automatically at the right point. It is also important to know, that the update is a 'pull'

process. Under no circumstances Cobion will send any data to remote COFS Database servers unless your COFS Database server requests updated data and downloads it.
The average update file is about 2.5 MB and contains roughly 50.000 new, updated or deleted entries for the COFS URL Database.
As an optional feature, Cobion offers anonymous, automated reporting of unknown URLs back to Cobion. This service helps Cobion and its customers to increase coverage and provides immediate categorization of previously unknown content by feeding these URLs into the normal crawling-process.

Benchmarks

Benchmark URL Database Category 1: Pornography

- 1750 URLs provided by http://www.net-protect.org/
- 1300 pornographic URLs (German, English)
- 450 non pornographic URLs (German, English)

The URLs were put together (i.e. via search engines) by Net Protect.

German Tool	Blocking	Overblocking
BizGuard	66%	13%
Cobion	84%	4%
Cyber Patrol	62%	7%
CYBERsitter	56%	2%
Cyber Snoop	82%	27%
Internet Watcher 2000	47%	2%
Net Nanny	5%	0%
Norton Internet Security	38%	6%
Optenet	88%	31%
SurfMonkey	68%	10%
Webwasher	63%	5%
X-Stop	73%	16%

English Tool	Blocking	Overblocking
BizGuard	63%	18%
Cobion	85%	2%
Cyber Patrol	73%	2%
CYBERsitter	64%	4%
Cyber Snoop	66%	31%
Internet Watcher 2000	38%	0%
Net Nanny	44%	11%
Norton Internet Security	73%	10%
Optenet	86%	30%
SurfMonkey	80%	15%
Webwasher	62%	2%
X-Stop	80%	1%

This benchmark shows that Cobion provides the best blocking rate while maintaining a very low overblocking rate.

Benchmark Text Classification

This benchmark measures the quality of the Cobion text classification engine. The text classifier was trained using the training data set given below. The test data set for the benchmark is not part of the training data set.

Set of training data

Hacking/Warez:	Yes-data: 1890 HTML-Files	No-data: 6871 HTML-Files
Illegal drugs:	Yes-data: 1511 HTML-Files	No-data: 2682 HTML-Files
Weapons:	Yes-data: 1263 HTML-Files	No-data: 3381 HTML-Files

The HTML-Files were investigated via search engines.

Set of test data

Hacking/Warez:	Yes-data: 108 HTML-Files	No-data: 285 HTML-Files
Illegal drugs:	Yes-data: 113 HTML-Files	No-data: 95 HTML-Files
Weapons:	Yes-data: 65 HTML-Files	No-data: 174 HTML-Files

The HTML-Files were investigated via search engines.

Blocking (100% - false negatives)			
	Hacking / Warez Illegal	**drugs**	**Weapons**
Cobion	97,78%	97,1%	98,46%

Overblocking (false positives)			
	Hacking / Warez Illegal	**drugs**	**Weapons**
Cobion	0,3%	3,16%	0,57 %

Appendix

URL Database Categories

Nudity

Pornography
Includes websites containing the depiction of sexually explicit activities and erotic content unsuitable to children or persons under the age of 18.

Erotic / Sex
Includes websites containing erotic photography and erotic material, as it can be found on television or obtained from magazines free of charge. Sex toys are also in this category. Sexually explicit activities are not listed here.

Swimwear / Lingerie
Includes websites containing nudity, but with no sexual references. Includes bikini, lingerie and nudity.

Ordering

Online Purchasing
Includes websites with online shops, where there is a possibility to select from a product range and order online.

Auctions / Small Advertisements
Includes websites with online/offline auction sites, auction houses and online/offline advertisements.

Society / Education / Religion

Governmental Organizations
Includes websites with content for which governmental organizations are responsible (e.g. government branches or agencies, police departments, fire departments, hospitals) and supranational government organizations such as the United Nations or the European Community.

Non-Governmental Organizations
Includes websites of non-governmental organizations such as clubs, communities, non-profit organizations and labor unions.

Cities / Regions / Countries
Includes websites with regional information, web sites of cities, regions, countries, city maps and city magazines.

Education / Enlightenment
Includes websites of universities, colleges, public schools, schools, kindergartens, adult education, course offerings, dictionaries and encyclopedias of any topic.

Political Parties
Includes websites of political parties and those sites that provide information about a particular political party.

Religion
Includes websites with religious content, information about the five main religions, and religious communities that have emerged out of these religions.

Sects
Includes websites about sects, cults, psycho-groups, occultism, Satanism etc.

Criminal Activities

Illegal Activities

Includes websites describing illegal activities according to German law, such as instructions for murder, manuals for bomb building, manuals for murder, instructions for illegal activity, child pornography, etc.

Computer Crime

Includes websites describing illegal manipulation of electronic devices, data networks, methods and also password encryption, manuals for virus programming and credit card misuse.

Hate and Discrimination

Includes websites with extreme right and left-wing groups, sexism, racism and the suppression of minorities.

Hacking

Includes websites with software cracks, license key lists and illegal license key generators.

Extreme

Includes websites that are normally assigned to other categories, but are particularly extreme in their content (e.g. violence).

Games / Gambling

Gambling

Includes websites of lottery organizations, casinos and betting agencies.

Computer Games

Includes websites of computer games, computer game producers, cheat sites and online gaming zones.

Toys

Includes websites containing information about dolls, modeling, scale trains/cars, board games, card games and parlor games, etc.

Entertainment / Culture

Cinema / Television

Includes websites ranging from cinema, television, program information, to video on demand.

Amusement / Theme Parks

Includes websites containing organization for recreational activities, e.g. public swimming pools, zoos, fairs and amusement parks.

Art / Museums

Includes websites from theatres, museums, exhibitions, and opening days.

Music

Includes websites from radio stations, online radio, MP3, Real Audio, Microsoft Media, homepages of bands, record labels and music vendors.

Literature / Books

Includes websites containing literature such as novels, poems, specialized books, cooking books, advisories and many more.

Humor / Comics

Includes websites with jokes, sketches and other humorous content.

Information / Communication

General News / Newspapers / Magazines

Includes websites that inform about general topics such as youth magazines or newspapers.

Web Mail

Includes websites that enable internet users to send or to receive e-mails via the internet (mailbox). All providers of web mail services are categorized here as well.

Chat

Includes websites that allow users to have a direct exchange of information with another user from place to place. Also listed are chat room providers.

Newsgroups / Bulletin News Boards / Discussion Sites

Includes websites that enable sharing information such as on a pin board, including a variety of topics.

SMS / Mobile Phones Fun Applications

Includes websites that enable users to send short messages via SMS via the internet to a mobile phone. It also includes providers and services for mobile phone accessories that are not necessary for daily use such as games, ring tones and covers.

Digital Postcards

Includes websites that allow people to send digital postcards via the internet, and also the providers of these services.

Search Engines / Web Catalogs / Portals

Includes websites containing search engines, web catalogues and web portals.

IT

Software and Hardware Vendors / Distributors

Includes websites of producers of hardware used for information, measuring and modular technology, vendors of software, and distributors that provide hardware and software.

Web Hosting

Includes websites such as web hosting and Internet Service Providers as well as providers of broadband services.

Information Security Sites

Includes websites that inform people about security, privacy, data protection in the Internet and in other broadband services as telecommunications.

URL Translation Sites

Includes websites that enable the translation of parts or the entire content of a website into another language.

Anonymous Proxies

Includes websites that allow users to anonymously view websites.

Drugs

Illegal Drugs

Includes websites about illegal drugs such as LSD, heroine, cocaine, XTC, pot, amphetamines, hemp and the utilities for drug use (e.g. water pipes).

Alcohol

Includes websites dealing with alcohol as a pleasurable activity (e.g. wine, beer, liquor, breweries) and the websites of alcohol distributors.

Tobacco

Includes websites about tobacco and smoking (cigarettes, cigars, pipes), and websites of tobacco vendors.

Self Help / Addiction

Includes websites from self-help groups, marriage guidance counseling, and help for addiction problems.

Lifestyle

Dating / Relationship

Includes websites that promote interpersonal relationships.

Restaurant / Bars

Includes websites about bars, restaurants, discotheques, and fast food restaurants.

Travel

Includes websites about monuments, buildings, sights, travel agencies, hotels, resorts, motels, airlines, railways, car rental agencies and tourist information.

Fashion / Cosmetics / Jewelry

Includes websites about fashion, cosmetics, jewelry, perfume, modeling and model agencies.

Sports

Includes websites such as resort sports, fan clubs, events (e.g. Olympic Games, World Championships), sport results, clubs, teams and sporting federations.

Building / Residence / Furniture

Includes websites such as property markets, furniture markets, prefabricated houses, design, etc.

Nature / Environment

Includes websites about pets, market gardens, environmental protection etc.

Private Homepages

Includes private websites and homepage servers.

Job search

Includes websites of job offerings, job searches, job agencies, labor exchanges, temporary work, etc.

Finance / Investing

Brokerage
Includes websites displaying stock exchanges rates, and deal exclusively with the main stocks like finance, brokerage and online trading.
Investing
Includes websites about real estate, insurance, and construction financing.
Banking
Includes websites of resort bank offices, credit unions, and online bank accounts.

Transportation

Includes websites from the resort automobiles, car tuning, car-exhibitions, motorbikes, airplanes, ships, submarines, bikes, railway etc.

Weapons

Includes websites dealing with guns, knives (not including household or pocket knives), air guns, fake guns, explosives, ammunition, military guns (tanks, bazookas), guns for hunting, and swords.

Medicine

Health / Recreation / Nutrition

Includes websites about hospitals, doctors, drugstores, psychology, nursing, health food stores and medicine.

Abortion

Includes websites about abortion.

Contact Information

Astaro AG
Jahnstrasse 1
D-76133 Karlsruhe
Germany

Tel. + 49 721 49 00 69 - 0
Fax + 49 721 49 00 69 - 55
E-Mail: info@astaro.com
Web: www.astaro.com