# TABLE OF CONTENTS

**Introduction**  
Technological Platform  
Global Data Center  

**Internet Content Filtering**  
Overview  

**Spam Protection**  
Spam Protection Technologies  

**Filter Database Services**  

**Summary**  

**Appendix**
INTRODUCTION

Effective Web and e-mail filtering relies on three components: a robust content analysis process, a vast amount of data, and the means to analyze that data. Internet Security Systems' (ISS) content analysis technology uses all three of these components to provide the foundation for all ISS content security products and solutions.

TECHNOLOGICAL PLATFORM

The content analysis process is built on a platform that classifies millions of Web pages and e-mails every day. This platform uses intelligent algorithms and massive parallel computing to run fully automated Web crawlers and analyze Web sites, e-mails, images and other content. In addition, the platform manages multiple database clusters to cache and store Web site and e-mail information, data signatures, hyperlink structures, images, Web site text and other important content.

GLOBAL DATA CENTER

The Global Data Center is the heart of the platform. As the largest facility of its kind in the world, it processes up to 50 million Web pages, e-mails and images every day. The Global Data Center is powered by a clustered server architecture with 1000 CPUs and provides the massive processing power for the Global Data Center’s content analysis operations and high-speed database searches.

Built for scalability, the Global Data Center can be readily expanded to meet any level of processing power. The Global Data Center is staffed around the clock by highly skilled professionals for process control and system monitoring.
INTERNET CONTENT FILTERING

OVERVIEW
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, and 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.

ISS uses a new approach to Internet filtering. The ISS content filtering process automatically scans the complete Internet and categorizes each Web site by its content using a proven combination of different technologies for intelligent text classification, superior image recognition and structural analysis.

Fully automated Web crawlers scan the Internet and inspect millions of new and updated Web sites every day. All Web sites are categorized automatically using advanced technologies and a super-computing infrastructure which provides the power necessary for this process. The result is a fresh and daily updated database of the Internet.

ISS analyzes and independently categorizes Internet content into 61 categories. Currently, ISS provides customers with a filter database that contains 60 million entries. This knowledge is based on the inspection and categorization of 3.9 billion Web pages and images from the Internet.

ISS started this process in 1999. Since then it has improved the quality of the content filtering process, expanded the computing infrastructure and implemented new technologies to provide better and fresher data every day.

Web Crawling
Web crawling is a challenging and ongoing task. It must deal with thousands of issues such as non-standard Web sites, unavailable Web servers, server performance, spamming sites, sites that are optimized for search engines, multiple languages, parking domains and so on. Web crawling is a very sensitive process because it involves interacting with millions of Web servers that are all beyond the control of the system. Beyond the technical challenges, classifying the Web also involves social and cultural issues about morals and ethics.

ISS operates a fast, distributed crawling system that is capable of visiting millions of Web servers each day. Web crawling is based on a chain-reaction metaphor. Starting at one Web site, a crawler downloads all of the HTML text and images from that particular site and stores this content for further analysis in the cache. Crawlers also follow all the hyperlinks (URLs) that are included in that first Web site to other Web sites, and then follow all the hyperlinks contained in these other Web sites until they have accessed all possible hyperlinks and downloaded all content.

The crawling strategy for following hyperlinks is adapted dynamically. For instance, the crawlers' first priority is to visit newly discovered hosts and domains instead of going deeper on the same host. Also, crawlers do not download massive amounts of data from the same host in a single visit. Web crawlers visit one host several times and perform multiple downloads.

To cover unconnected islands in the Internet (i.e., Web sites that are not linked or referred to on any other Web site), ISS systematically feeds the crawler fresh information about new Web sites, domains and hosts based on public host lists, domain registry information, automated customer feedback and other external sources.

In addition to the crawling process, the crawling system also performs update and maintenance processes. These processes run in parallel. One part of the crawler searches for new content while the other part constantly updates the content of known Web sites. Web sites that change more often are visited more frequently. This process is adapted dynamically to keep up with the ever-changing nature of the Internet. For example, the crawler visits Web sites that contain link lists and news on a daily basis for fresh information.

The current platform processes up to 15 million Web pages and images daily and adds about 5 million new links every day.
Web Site Categorization

After the crawlers have downloaded all the Web site and image content, this content must be analyzed and categorized. For ISS, Internet filtering means more than just a simple keyword search or URL and filename analysis. ISS runs multiple analysis processes on each Web site to get the highest level of accuracy for the categorization.

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

Text Classification

ISS uses both keyword search and intelligent text classification to analyze a Web site’s textual content. The keyword search determines which category the words belong to, depending on the occurrence of certain words. One disadvantage of this procedure is that many words have different meanings (e.g., “sex”) and therefore are difficult to categorize. The advantages of this procedure include high performance and easy configuration. Also, this method operates well when only a few words are available, for example when classifying a URL.

In addition to classifying on the basis of single words, intelligent text classification also classifies on the basis of frequency and word combination. Text is classified using word heuristics and word combinations together with support vector machines for the final decision process. The ISS text classification technology has a very high reliability; therefore, basically no errors occur if the number of words is large enough.

Both text analysis methods can be applied to different sources. For sources with only a few words, the keyword classifier is preferred, e.g., OCR, URL, or filename. 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, ISS also uses face detection technology. If a face is detected in an image, a color sample is taken from that area as reference for the skin color. This type of recognition is very reliable because the information about the skin tone is taken directly from the face. This also decreases overblocking, since portrait images are not rated as pornographic. If a face is not present in an image, the algorithm uses a statistical representation of the skin tone as an approximation.

Visual Object Recognition

This technology analyses each image for special signs, symbols, trademarks, and so on. The method is currently used to recognize forbidden symbols, such as the swastika in Germany. All major credit card logos, sport brands, car brands and other brands are also detected.

Visual Optical Character Recognition

A lot of textual information on a Web site is found in images. ISS performs Optical Character Recognition (OCR) on each image and processes this information with the abovementioned text classification methods. Because the text in an image is highly relevant to the image content, this method improves overall accuracy.

Structure and Link Analysis

The Web consists of millions of interconnected Web sites that relate to each other in different ways. Given the knowledge of interconnectivity that is stored in the Global Data Center, a detailed structural analysis of how Web sites link to each other yields a new layer of classification. For example, if a Web site has 10 consecutive links to other Web sites, and 9 of those 10 are pornographic, then the probability that the 10th Web site is also pornographic is higher.

Spyware

As a major security threat, spyware is specially handled in the Global Data Center. Directly integrated into the crawling and analysis process, all binaries and installation packages are subject to spyware inspection. Using a combination of matching known signatures and software behavior analysis, malicious and suspect files are detected, and the related URLs (the origin of the files) are categorized accordingly as spyware sites. The Proventia® filter database also contains the Internet hosts that known spyware programs use for communication (“phone home” sites).

Overall Classification

The final classification combines the results of each analysis method with a finely tuned weighting. This is a very important step in overall categorization. The combination of multiple methods can resolve ambiguities: for example, when a Web site contains images showing nude content but the surrounding text contains medical or educational information. In this situation, ratings from the other analysis methods become very important for the final classification. Therefore, a combined weighted rating of multiple analysis methods is essential to achieve the highest quality in the categorization process.

Support for Multiple Languages

The overall crawling and classification process is designed to be completely independent from the language used on a Web site. Only a fraction of the analysis algorithms require language-specific tuning (text classification and visual optical character recognition). For each Web page, language is determined automatically, and language-specific modules are used for text classification. For the training of language-specific modules, ISS has linguistic experts who currently cover the following languages: English, Spanish, French, German, Portuguese, Italian, Russian, Polish, Chinese, Japanese, Korean, Arabic and Hebrew.

Filter Database Production

The production process of the filter database is the essential step in building a highly accurate Internet filtering list. Various pieces of information from multiple sources are combined using a proprietary weighting scheme. Each information source contributes to the final categorization at 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 billions of Web sites by a team of experts.
This process also resolves conflicts: for example, when sub-sites are categorized differently from the home page. Especially large Web sites such as portals provide a variety of offerings including search, finance, news and mail, each requiring a different categorization.

As the final step, the filter database is tuned for performance, as related information is combined in efficient data structures and optimized for high-speed access. After this, the incremental filter database updates files are published on the update servers.

**What is Contained in the Filter Database?**

The core information stored in the filter database is the URL and its categorization information. Every URL stored in the filter database belongs to one or more of the predefined filter categories. Also, the filter database correctly supports different categorizations for subpages (www.this.com/media vs. www.this.com/erotic) and multiple hosts on the same domain (games.yahoo.com vs. finance.yahoo.com).

The database contains multiple information about domains, hosts and URLs of the following types:

- **Domains:** sex.com
- **Hosts:** www.sex.com vs. images.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
- **Protocols:** http:// vs. ftp://
SPAM PROTECTION

Spam is one of the most common abuses of e-mail today. Spam brings in unwanted content to users and also wastes valuable bandwidth and system resources. ISS spam protection technologies combine a service model, which provides daily updates with fresh spam data (signatures, URLs, etc.) through the Global Data Center, and local e-mail scanning technologies for highly efficient spam filtering.

ISS operates hundreds of spam collectors worldwide (e-mail accounts that attract spammers, also known as honey pots) that receive hundreds of thousands of definite spam mails every day. All spam collectors feed their data into the Global Data Center for processing. ISS has established a network of trusted partners who provide fresh spam data for the Global Data Center as an additional source.

In the Global Data Center, spam processing consists of multiple steps including spam signature extraction, structure signature extraction, spam URL analysis and others. All relevant data is stored in the ISS filter database and is updated several times a day. End-user feedback about unknown spam is provided through the SpamLearn service, which provides updates to the spam signature database.

Currently, the filter database contains more than 40 million relevant spam signatures and millions of spam URLs.

SPAM PROTECTION TECHNOLOGIES

Spam Signatures

Every e-mail is broken into several logical parts (sentences, paragraphs, etc.), and a unique 128-bit signature is computed for each part. The signature is invariant to minor modifications in the e-mail, but still accurate enough to uniquely identify a known spam through a matching signature in the filter database.

In the Global Data Center, spam signatures are computed for all known spam (from spam collectors and other sources) and stored in the filter database.
Structure Analysis
The structure analysis module examines the HTML structure of the e-mail and computes two signatures based on the structure. For example, some spam typically has a bold headline followed by one or more paragraphs in a different color, and then some random text at the bottom. Such layout structures are invariant to the actual text in the e-mail and are therefore an excellent addition to the textual spam signatures mentioned above.

Structure signatures are computed for all known spam (coming from spam collectors and other sources) and are stored with spam signatures and URLs in the filter database.

Bayesian Classifier
The Bayesian classifier is a system which determines whether an e-mail is spam based on e-mail statistics. To train the classifier, thousands of examples of spam and regular e-mail are presented to the system and relevant data is extracted and stored in a statistical model. Through this training, the classifier is able to learn the difference between spam and regular e-mail. ISS offers an updated, pre-trained Bayesian database which is trained using thousands of different spam types coming from the spam collectors and through end-user feedback.

The advantage of the Bayesian classifier is the ability to recognize new types of spam, whereas the signature technology is better in detecting identical and nearly identical spam.

URL Checker
More than 80 percent of all spam e-mails contain URLs and links to related Web offerings. All relevant URLs that appear in spam e-mail are stored in the filter database together with the stored spam signatures. Spam classification also checks to see if e-mails contain links to pornographic Web sites and images, banking sites (see “Phishing Detection”) and medications, which are very popular targets for spamming.

Keyword Analysis
This classifier covers standard keywords and patterns (regular expressions) that are typically found in spam e-mail. ISS has extracted relevant keywords and patterns from known spam and weighted individual relevancy for additional spam protection.

Spam Heuristics
This classifier is based on heuristics that are typical for spam e-mail. It employs an internal scoring system with each heuristic receiving either positive or negative points, depending on whether the heuristic is designed to match spam or ham (“normal” e-mail). If the point count reaches a predetermined threshold, the e-mail is classified as spam. For example, the following information is used for heuristic analysis:

- Message-ID field characteristics
- Received field invalid or missing
- Checks for “Apparently-To:” or “X-Apparently-To” fields
- Checks for mailing list fields
- Checks for multiple recipients and alphabetic recipient patterns like a@, b@, c@
- Checks for missing fields like “From” and “To”

Flow Control
The flow control module analyzes e-mail flow within a specific time frame. If the same e-mail (based on a number of similarity measures) is received more than a threshold number of times within the time frame and has different sender domains, then the e-mail is a classified as spam. This technology can detect completely unknown types of spam based on the way spam is typically created and sent.
Black/White Sender Check
The filter database contains entries for e-mail addresses marked as “Black” or “White.” The sender e-mail address is checked against the database. If the address is marked as “Black,” the e-mail is considered to be spam. If the address is marked as “White” and the Signature Database has not classified the e-mail as spam, the e-mail is considered to be ham, regardless of what results the other classifiers give. The signature database check exists to avoid the possibility of spammers using a “White” e-mail address to send spam.

Phishing Detection
Phishing e-mails are a very special type of spam that is intended to retrieve personal information from potential victims. Typically, phishing e-mails look as if they are coming from an individual’s bank, favorite shopping sites and so on, but the intention is to steal that person’s account information, passwords etc. In many cases it is very difficult for the average end-user to distinguish a real e-mail that was sent to him by his bank from a phishing e-mail.

For phising detection, ISS combines a variety of methods in the Global Data Center. The URL checker is able to detect links to banking and other commercial sites in all spam coming from the spam collectors. Phishing mails also show typical heuristics compared to regular spam. This way, phishing e-mails are categorized separately from regular spam in the filter database.

FILTER DATABASE SERVICES
Update Process
The ISS filter database is updated several times a day. To ensure that customers always receive suitable updates, ISS uses the following process:

Each product and service that depends on the filter database can be configured to check for updates many times a day. During that process, a secure HTTP connection is established to one of the ISS update servers. The update process determines which update files customers are required to download, based on the product and the subscribed services. Once the download is complete, the local filter database is updated in the background.

The update workflow is transaction based. If the process is interrupted for any reason (for example, if the server is disconnected), the update will continue automatically at the right point.

Updates are always incremental. Multiple update files are published every day. On average, the total daily update volume is about 10 MB and contains approximately 250,000 new, updated or deleted entries for the filter database. The entire filter database can be updated or only certain parts of it (such as URLs or spam signatures), depending on the product and the subscribed services.

As an optional feature, ISS offers anonymous, automated reporting of unknown URLs back to ISS’ Global Data Center. This service helps ISS and its customers to increase coverage and provides immediate categorization of previously unknown Web sites by feeding these URLs directly into the crawling process.

Updating Spam Data
Spam data has different update schemes. The spam signature database is updated every hour with fresh signatures extracted from the worldwide spam collector network. The ISS Bayesian database is typically updated on a monthly basis to incorporate new types of spam. Depending on new spam attacks, new tricks and outbreaks, update files for spam heuristic data and keyword patterns will be available.

SUMMARY
As an established worldwide leader in information security, Internet Security Systems delivers products and services that protect against Internet threats. With the acquisition of Cobion AG in 2004, ISS added the content security products Proventia® Network Web Filter and Proventia® Network Mail Filter to its portfolio, taking advantage of the leading technology for URL and spam filtering. Today, ISS offers a strong portfolio of content security products for Web filtering, E-mail filtering, spam protection and anti-phishing. With more than 60 million categorized Internet sites, ISS provides the world’s largest URL database for Web filtering.
FILTER DATABASE CATEGORIES

Pornography/Nudity
Pornography: Includes Web sites containing the depiction of sexually explicit activities and erotic content unsuitable for persons under the age of 18.

Erotic/Sex: Includes Web sites containing erotic photography and erotic materials that are the same as those seen on television or obtained free of charge from magazines. Sex toys are also in this category. Sexually explicit activities are not included.

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

Ordering
Online Shopping: Includes Web sites with online shops that offer a range of products which can be ordered.

Auctions/Classified Ads: Includes Web sites with online/offline auction sites, auctions houses and online/offline advertisements.

Society/Education/Religion
Governmental Organizations: Includes Web sites 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 the Web sites of non-governmental organizations such as clubs, communities, non-profit organizations and labor unions.

Cities/Regions/Countries: Includes Web sites with regional information, Web performances of cities, regions, countries, city maps and city magazines.

Education: Includes the Web sites of universities, colleges, public schools, schools, kindergartens, adult education, course offerings, dictionaries and encyclopedias for any topic.

Political Parties: Contains the Web sites of political parties and those sites that provide information about a particular political party.

Religion: Includes Web sites with religious content, information about the five main religions and religious communities that have emerged from these religions.

Sects: Contains sites about sects, cults, psycho-groups, Occultism, Satanism, etc.

Criminal Activities
Illegal Activities: Includes activities that are illegal according to German law, such as instructions and manuals for murder, manuals for bomb building, instructions for illegal activity, child pornography, etc.

Computer Crime: Includes the illegal manipulation of electronic devices, data networks, procedures, password encryption, manuals for virus programming and credit card misuse.

Political Extreme / Hate / Discrimination: Contains Web sites with extreme right- and left-wing groups, sexism, racism and the suppression of minorities.

Hacking/Warez/Illegal Software: Contains sites with software cracks, license key lists and illegal license key generators.

Violence/Extreme
Includes Web sites that are normally assigned to other categories, but that are particularly extreme in their content (e.g., violence).

Games/Gambling
Gambling: Includes lottery organizations, casinos and betting agencies.

Computer Games: Classifies the Web sites of computer games, computer game producers, cheat sites and online gaming zones.

Toys: Contains information about dolls, modeling, scale trains/cars, board games, card games and parlor games, etc.
Entertainment/Culture
Cinema/Television: Includes Web sites in the area of cinema, television, program information, and video on demand.
Recreational Facilities/Amusement/Theme Parks: Contains organizations for recreational activities, such as public swimming pools, zoos, fairs and amusement parks.
Art/Museums/Memorials/Monuments: Includes Web sites for theaters, museums, exhibitions and opening days.
Music: Includes Web sites for radio stations, online radio, MP3, Real Audio, Microsoft Media, homepages of bands, record labels and music vendors.
Literature/Books: Contains literature such as novels, poems, specialized books, cookbooks, advisories and many more.
Humor/Comics: Contains sites with jokes, sketches and other humorous content.

Information/Communication
General News/Newspapers/Magazines: Contains Web sites that give information about general themes such as youth magazines or newspapers.
Web Mail: Contains Web sites that enable Internet users to send or to receive e-mails through the Internet (mailbox). All providers of Web mail services are categorized here as well.
Chat: Contains Web sites that allow users to have a direct exchange of information with another user from a different place. This category also includes chat room providers.
Newsgroups/Bulletin Boards/General Discussion Sites: Contains Web sites that enable the sharing of information such as on a black board, and includes a variety of topics.
SMS (Short Message Service)/Mobile Phone Accessories: Includes Web sites that enable a user to send short messages via SMS through the Internet to a mobile phone. It also includes providers and services for mobile phone accessories that are useless as games, ring tones and covers.
Digital Postcards: Includes Web sites that allow people to send digital postcards via the Internet, and also the providers of these services.
Search Engines/Web Catalogs/Portals: Contains search engines, Web catalogues and Web portals.

Information Technology (IT)
Software/Hardware/Distributors: Includes hardware producers in a specific area of information, measuring and modular technology, software vendors and distributors that provide hardware and software.
Communication Services: Includes Web sites such as Web hosting and Internet Service Providers, as well as providers of broadband services.
IT Security/IT Information: Contains Web sites that inform people about security, privacy, data protection for the Internet and for other broadband services such as telecommunications.
Web Site Translation: Contains Web sites that enable the translation of parts or the entire content of a Web site into another language.
Anonymous Proxies: Includes Web sites that allow the user to anonymously view Web sites.

Drugs
Illegal Drugs: Contains Web sites about LSD, heroin, cocaine, XTC, amphetamines, marijuana (pot) and the utilities for drug use (e.g., water pipes).
Alcohol: Includes Web sites that deal with alcohol as a pleasurable activity (e.g., wine, beer, liquor, breweries) and the Web sites of alcohol distributors.
Tobacco: Includes Web sites that deal with tobacco and smoking (cigarettes, cigars, pipes), and Web sites of tobacco vendors.
Self-Help/Addiction: Contains Web sites of self-help groups, marriage guidance counseling and help for addiction problems.
Lifestyle
Dating/Relationships: Contains Web sites that promote interpersonal relationships.
Restaurants/Bars: Contains Web sites about bars, restaurants, discos and fast food restaurants.
Travel: Includes Web sites about monuments, buildings, sights, travel agencies, hotels, resorts, motels, airlines, railways, car rental agencies and tourist information.
Fashion/Cosmetics/Jewelry: Contains Web sites about fashion, cosmetics, jewelry, perfume, modeling and modeling agencies.
Sports: Contains Web sites about resort sports, fan clubs, events (e.g., Olympic Games, World Championships), sport results, clubs, teams and sporting federations.
Building/Residence/Architecture/Furniture: Contains Web sites about property markets, furniture markets, prefabricated houses, design, etc.
Nature/Environment/Animals: Contains Web sites about pets, market gardens, environmental protection, etc.

Private Homepages
Contains private Web sites and homepage servers.

Job Search
Includes Web sites about job offerings, job searches, job agencies, labor exchanges, temporary work, etc.

Finance/Investment
Investment Brokers/Stocks: Includes Web sites that handle stock exchanges rates and deal exclusively with the main stocks like finance, brokerage and online trading.
Financial Services/Investment/Insurance: Contains Web sites about real estate, insurance and construction financing.
Banking/Home Banking: Contains Web sites about resort bank offices, credit unions and online bank accounts.

Vehicles/Transportation
Contains Web sites about automobiles, car tuning, car exhibitions, motorbikes, airplanes, ships, submarines, bikes, railways, etc.

Weapons
Deals 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: Contains Web sites about hospitals, doctors, drugstores, psychology, nursing, health food stores and medicine.
Abortion: Contains all Web sites that deal with abortion.

Spam
Spam URLs: Contains Web sites that are solicited in spam e-mails.
Phishing URLs: Includes Web sites that are contained in phishing e-mails.

Spyware
Contains Web sites that install data-transmitting programs without the user’s knowledge.