



A WEBSense® WHITEPAPER

INFORMATION IDENTIFICATION: CRITICAL REQUIREMENTS FOR EFFECTIVE DATA SECURITY

BY DR. LIDROR TROYANSKY
RESEARCH FELLOW
WEBSense, INC.

Table of Contents

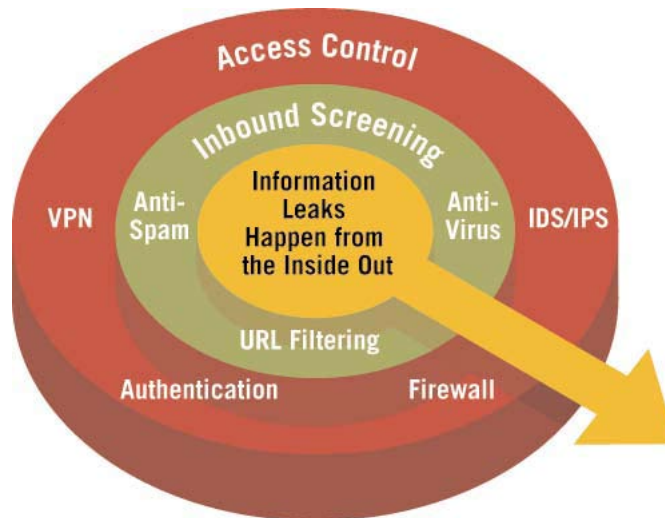
- Introduction: Information Leaks Start from the Inside Out**..... 3
- Identifying Sensitive Content**..... 4
- Information Classification** 5
 - Basic Classification: Keywords and Key Phrases 5
 - Advanced Classification: Machine Learning 6
- A Robust Combination of Information Identification & Classification Techniques**..... 7
 - 1. Global Filters..... 8
 - 2. Token-Based Filters..... 8
 - 3. Contextual Filters..... 9
- Information Fingerprints**..... 11
 - Websense PreciseID™ Fingerprinting Technology..... 11
- The Next Step—Effective Policy Enforcement**..... 14
- How Can I Start Protecting My Organization?**..... 15
- Conclusion: Websense Stops Information Leaks. Period**..... 16
- About Websense™**..... 16
- About The Author**..... 16

INTRODUCTION:

INFORMATION LEAKS START FROM THE INSIDE OUT

With the recent wave of security breaches in the news, executives are realizing that despite their best efforts, information leaks can affect organizations of all sizes and in any industry. Few organizations have found themselves immune to the insider threat: employees who expose sensitive information whether intentionally or accidentally. In the normal course of doing business, employees need the ability to communicate sensitive information with other employees, customer, partners, and other parties while maintaining data security.

The high cost of failing to prevent these information leaks is causing companies to reexamine how they defend themselves against potentially disastrous consequences. Even so, most companies expose themselves to significant financial and legal liability because they cannot reliably and accurately identify sensitive content in transit.



IDENTIFYING SENSITIVE CONTENT

The Importance of Identification

Information leaks occur, either accidentally or maliciously, because most firms have not guarded themselves against insider threats. This gap exists primarily because firms lack an efficient means to correctly recognize when a message containing sensitive content is headed to an unauthorized recipient. Therefore, a robust way to reliably and accurately identify information in real-time is a critical requirement for any solution that enforces distribution policies on sensitive content.

Without a high degree of accuracy, a content monitoring system will overwhelm the IT staff with an insurmountable burden of false positives. What's even worse is that any policy-based enforcement system plagued by false positives will cause business communications to be interrupted by a high rate of false positives or wrongly blocked messages. These false positives disrupt the normal flow of business and hinder productivity. While some solution providers offer policy enforcement platforms, simply blocking messages is not enough. The real challenge lies in identifying whether a message containing sensitive information is being sent to an inappropriate recipient, and then enforcing the appropriate security policy.

Without reliable and accurate identification, real-time policy enforcement can only be applied coarsely, rather than with a high degree of precision.

The Difficulty in Identification

Identifying sensitive content in transit is particularly difficult because:

- Fragments of sensitive content, such as credit card numbers, account numbers, customer records, and phone numbers, can be easily cut and pasted into other documents and messages, or simply posted to web sites.
- Sensitive information like contracts, employee offers, financial filings, and product specifications is often modified from an original document into derivations or excerpts.
- Multiple copies of sensitive information typically exist. Attempts to secure a particular file are thwarted by the reality that employees have access to identical or similar information else where on the network.
- Sensitive information can be kept in multiple unstructured file formats or structured databases with different versions and compatibility.
- Sensitive content can be communicated through a variety of channels, including e-mail, web-mail, instant messaging, FTP, fax and P2P applications.

To handle the variety of sensitive information, organizations need a set of identification technologies that are robust enough to identify both structured content (such as records in a database) and unstructured content (such as Word files, financial spreadsheets and Adobe PDF documents), as well as fragments and derivatives of either. These identification technologies must work in real-time or they risk rendering communication media ineffective. Finally, these identification algorithms must account for the context of the content to accurately and correctly recognize which messages must be handled by which policies. Without a complete set of identification technologies, real-time policy enforcement can not be applied with a high degree of precision and accuracy.

Sensitive content is difficult to identify in real-time because:

- Content can be cut and pasted into different documents.
- Content can be altered or reformatted from its original form.
- Multiple copies can exist.
- The same content can exist in multiple file formats.
- Content can be communicated through many different channels.

INFORMATION CLASSIFICATION

The Linchpin of an Information Security System

Classification defines the parameters by which information is protected or controlled. How information is classified determines both personnel and physical security.

In the classification process, information is divided into different classes like “confidential” or “public.” Classification is essential to provide a basic level of protection to the information assets based on specific policies associated with those classification levels. For example, an organization may decide that distribution of “secret” information is not permitted without explicit permission from an authorized manager.

Information classification is the linchpin on which entire security systems are based, granting access to and clearance for various documents.

Many organizations have implemented information security policies based on classification with only limited success. This is largely due to the low rate of accuracy. Enforcing information leak prevention using information classification requires a tool that can automatically determine the classification level of an item in the outgoing communications.

Protecting Information by Class

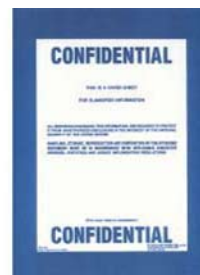
Organizations should use their own custom policies to define how to protect data by information class. A policy for top-secret documents could entail keeping them in a physically well-secured safe and never allowing them to leave the secured network. Another policy might specify that an unclassified document could be shared with third parties. Determining a document’s classification in real-time is a key requirement for protecting its distribution through different communication channels. Real-time protection can ensure organizations are preventing the problem rather than executing damage control in the face of a regulatory audit.

Websense helps organizations enforce their security policies by identifying any type of sensitive information (not just a specific file name) and classifying its content. Websense uses several classification methods which can identify document content and classify it in real-time.

Basic Classification: Keywords and Key Phrases

Simple methods of classification, such as keywords or key phrases, provide a first level of defense. Keywords or key phrases work on the assumption that the confidential content exists in its entirety and that classification holds for the entire document. The graphic below outlines three levels of classification that categorize sensitive content:

Unfortunately, classification techniques can only provide a coarse level of granularity, which creates issues when enforcing distribution policies on real-time communications.



Unfortunately, such classification techniques typically cause high levels of false positives and false negatives. Examples of these false positives and false negatives include:

- **Misleading words.** Using the word “confidential” causes false positives for any disclaimer of the type: “The content of this message may be confidential.” In this case, any message containing that disclaimer is automatically flagged for review.
- **Word manipulation.** If the keyword “black arrow” is used to designate confidential files for a particular project, it can easily be replaced with “yellow ribbon” and allow sensitive content or excerpts thereof to be released.
- **Misinterpretation of words.** Keywords may exist in another context with an innocent meaning. The name of a food company that is a candidate for M&A can appear in a simple recipe. A tax shelter can also be a resort location.

While info security users have been forced to accept the burden of false positives from monitoring solutions that use keywords and phrases, highly accurate identification is a requirement for moving from monitoring to real-time enforcement.

Advanced Classification: Machine Learning

More advanced methods for classification are based on machine learning. With machine learning, the system can learn to classify information using a limited set of previously classified information. System administrators need to provide the system two or more sets of information items, such as 1,000 “secret” information items and 1,000 “public” information items. The system extracts features or “tokens” that characterize the two sets and provides a function that allows discriminating new information items.

If a machine-learning solution is properly implemented, then the number of false positives and false negatives can be acceptable. Machine-learning solutions are often useful for spam detection and message sorting for e-mail response in customer-relationship management applications.

Advanced classification techniques like machine learning are intriguing, but require training time and proper resolution when matches occur.

Two fundamental drawbacks limit the effectiveness of machine learning solutions for Information Leak Prevention:

- **Training Time:** Because administrators typically devote substantial time and effort to “teach” the system, the actual cost of machine learning can be high. Many machine-learning applications have not matured past proof of concept and into full adoption because of the extensive training time involved.
- **Lack of Ownership and Proper Resolution:** Advanced classification using machine learning does not outline clear responsibility for who updates information and resolves issues when they arise. For information leak prevention to work, information assets must have an owner, authorized senders, and authorized recipients. This lack of ownership often causes too many combinations and too many classes for a machine-learning solution to handle

Classification methods provide a first step toward visibility into information leaks. However, they are ineffective when it comes to enforcement, because of the precision required when messages are blocked or quarantined.

To provide safe communication of sensitive information in real business processes, simple classification methods should be supplemented by accurate and robust identification capabilities that permit policy enforcement with a high degree of granularity.

A ROBUST COMBINATION OF INFORMATION IDENTIFICATION & CLASSIFICATION TECHNIQUE

Websense Data Security Suite is the industry’s first and only real-time enforcement solution based on ultra-precise information fingerprinting technology that is used to identify information beyond any doubt. Websense Data Security Suite uses PreciselD technology identifies content in the same way that a person can be identified with his or her unique fingerprint. Websense PreciselD technology uses a sophisticated and unparalleled combination of 27 patent-pending identification algorithms to quickly and accurately identify sensitive content.

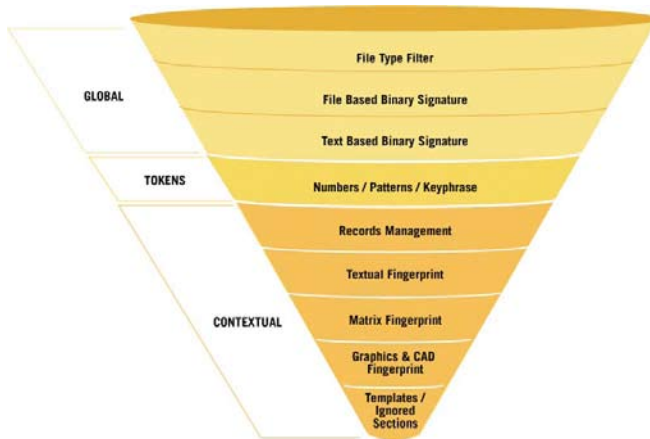
It includes a unique set of identification techniques operates similarly to how a police forensics investigator uses a combination of methods with various levels of accuracy to identify a suspect. A general description like “the suspect has brown hair” gives a relatively coarse level of identification. Adding details like “white male,” “six feet tall,” and “has a mustache” gives a more precise identification of an individual. Finally, with a fingerprint, a set of fingerprints, or DNA evidence a forensics investigator can identify the clearest, most reliable match.

Websense provides the most robust collection of information identification technology, delivering the most effective Data Loss Prevention solution available.

Like the forensics analogy, Websense delivers unparalleled identification capabilities with multiple levels of granularity. The most powerful techniques based on contextual filters include Websense’s revolutionary and patent-pending information PreciselD techniques.

Three Types of Identification Techniques

Websense achieves unprecedented accuracy and reliability by using a combination of state-of-the-art approaches, including global, token-based, and contextual techniques, in addition to the PreciselD technology, creating an advanced set of identification and classification capabilities. The diagram below highlights the three types of identification algorithms Websense employs.



Websense uses three types of information identification technology: global filters, token-based filters, and contextual filters.

Websense's revolutionary and patent-pending information PreciseID techniques.

1. Global Filters

Global ("Class 1") filters provide capabilities for basic monitoring and enforcement. Global filters can be categorized into three types:

File Type Filter

- Allows implementing a policy based on file type. An example is: "block .mp3 and .mpeg" or "convert MS-Word to PDF."
- Recognizes file type based on its content, not its extension, so it cannot be circumvented by renaming the extension, such as from .doc to .jpg.
- Recognizes nested compressed files recursively.

File-Based Binary Signature

- Assigns a number to a file that is a unique function of its content, thereby providing unique identification of any file, with a very high resolution. Small changes in the file completely change the signature.
- Provides very fast, but totally non-robust filtering.

Text-based Binary Signature

- Assigns a number (hash) to a file that is a unique function of its textual content.
- Provides very fast identification.
- Offers a little more robustness than the file-based binary signature. (It is robust to changes in the file metadata).
- Allows monitoring the integrity of the content.

Global filters provide a first line of defense by screening for certain file types and signatures.

Solutions using global classification provide basic monitoring and thus have a low rate of accuracy. They usually provide basic enforcement like blocking .exe or .src files. However, simple manipulations, such as changing one word and putting data into a .zip file or other company-allowed file format, easily thwarts this type of monitoring and enforcement.

2. Token-Based Filters

Token-based filters ("Class 2") provide another layer of protection and basic classification capabilities. Token-based approaches like e-mail filtering monitor content based on keywords, numbers, or patterns. Token-based filters are typically grouped into two types:

Pattern Recognition

- Uses regular expressions to identify numbers and strings in certain common formats like credit card numbers (xxxx-xxxx-xxxx-xxxx) or Social Security numbers (xxxx-xx-xxxx).
- Uses an advanced form of pattern recognition that contains special logic and flexible settings to mitigate false positives.
- Supplies a number of default patterns and template policies for pattern recognition.

Token-based filters provide more specific information identification, but lack the necessary granularity for proper policy enforcement.

Keyword and Key Phrase

- Allows detection of an unlimited number of numbers, keywords, and phrases.
- Enables policy application based on pre-defined dictionaries for HIPAA and Gramm-Leach-Bliley Act (GLBA) compliance.
- Contains "threshold policies" that apply a policy based on the accumulated number of words and numbers that were found, such as a message that contains more than five account numbers or more than 10 references to Social Security numbers.

Token-based filters offer good visibility into the magnitude and nature of information distribution when deployed in monitoring mode. However, these Class 2 filters often do not provide the granularity or resolutions required for true leak prevention and data integrity, and therefore produces a high rate of false positives and false negatives.

False alarms occur because these techniques cannot put commonly occurring words like “sensitive” and “confidential” in context. The rate of false alarms can limit token-based approaches’ reliability, because administrators may become conditioned to ignore alerts. False negatives occur simply because sensitive or confidential information does not contain the necessary patterns and keywords. To improve the accuracy of Class 2 filters, Websense’s patent-pending pattern recognition algorithms add context awareness and business logic to identified patterns, dramatically reducing false positives and negatives.

3. Contextual Filters

In addition to global and token-based techniques, Websense uses a state-of-the-art contextual and linguistic approach, which marries comprehensive monitoring with sophisticated enforcement. Websense combines multiple contextual identification algorithms, including records management, textual fingerprinting, matrix fingerprinting, graphics and CAD fingerprinting, and templates/ignored sections, to deliver accurate and reliable results. Information fingerprints are described in detail in the following section.

Contextual filters provide the highest levels of accuracy and allow for effective, real-time policy enforcement.

An effective Data Loss Prevention solution must address contextual (“Class 3”) filters. Each contextual filter is optimized to detect certain types of information, achieving extremely accurate information identification with top performance. These contextual filters can be organized into five major categories:

Records Management Filter

- Allows the application of Boolean logic to various fields within an individual record. For example, this allows Websense to quarantine the message if both a customer’s account number and date of birth are found in a single message or to encrypt the e-mail if the person’s name and the corresponding Social Security number appear in the same message.
- Greatly reduces false positives and false negatives by applying multiple criteria within a record.
- Applies intrinsic logic to detect instances that are more likely to result in damaging information leakage.

The Websense platform’s records management filter allows sophisticated rules to be set up against records organized in tables as seen in the diagram below. For example, a rule stating that information from no more than three rows may be sent in a single communication would stop an e-mail containing the account numbers 177355142, 123233486, and 342923776.

Records management permits very refined information identification by focusing on the combination of field elements within a record.

Name	Account Number	ID#	DOB
J. Clarke	177355142	19730806-5324353	5/23/68
F. Campbell	123233486	12349854-3083248	2/8/81
N. Lopez	342923776	19481119-1072491	9/12/57
L.Chen	288377464	19870622-8457582	7/2/79

Similarly, another rule might state that if more than two fields (columns) from a single record are sent in a message, that message should be quarantined. In this example, the Websense platform would quarantine a message containing L. Chen, Account Number 288377464 and DOB 7/2/79. The ability to detect multiple fields from a single record or multiple records within a single message greatly improves the ability to intercept truly suspicious messages.

Textual Fingerprint

- Allows extremely robust identification of content, including fragments or derivatives.
- The Websense platform is resilient to all types of data manipulation attempts, such as cutting and pasting, reformatting, and retyping.
- Converts unstructured text into a series of mathematical representations known as “information fingerprints.”
- Is based on a unidirectional process, which means that original content cannot be reverse engineered from a fingerprint.

Textual fingerprints not only detect whole documents or files, but also detect fragments and derivations of protected content.

Matrix Fingerprint

- Converts content from a tabular or spreadsheet format into a series of mathematical representations, while capturing its many idiosyncrasies.
- Is resistant to manipulation of content by applying certain proportionality checks against the content to ensure accurate identification of protected content. For example, it detects spreadsheets converted from dollars to euros.
- Utilizes a vectored-representation of the data that captures the original content's many idiosyncrasies.
- Is based on a unidirectional process, which means that original content cannot be reverse engineered from a fingerprint.

CAD/CAM Fingerprint

- Utilizes an approach that interprets the value associated with a diagram despite changes in its physical appearance like rotation or inversion.
- Is resilient to “reasonable” changes in the drawing. The Websense platform's CAD/CAM fingerprint filter solves this hard problem.
- Is based on a unidirectional process, meaning original content cannot be reverse engineered from a fingerprint.

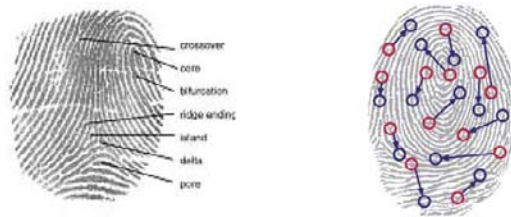
CAD/CAM fingerprints interpret the mathematical encoding of diagrams as well as their textual descriptions.

Template/Boilerplate Fingerprint

- Improves the accuracy of detection by accounting for false similarity and screens out commonly recurring text in similar documents, including boiler plates, disclaimers, template descriptions, forms, and contract terms.
- Websense is the only solution that employs sophisticated filters to account for templated content.
- This technique dramatically reduces the false positives associated with basic identification techniques, which often stumble over templated content.

INFORMATION FINGERPRINTS

Information fingerprints are highly optimized, mathematical representations of sensitive content that allow for extremely reliable and accurate identification of information. Just as human fingerprints include different elements that can be used to identify a person with great accuracy, information files can be threaded with the same concept as seen in the diagram below.



Websense PreciselD technology delivers robust, contextual information identification. Using a unidirectional process, Websense examines the content of documents or raw data and extracts a set of mathematical descriptors or “information fingerprints”. These fingerprints are compact and faithfully describe the underlying content. By assigning unique identities to each information asset, Websense PreciselD technology can track information in motion with great precision. Original content cannot be recreated or reverse engineered from Websense PreciselD information fingerprint.

The power of Websense PreciselD techniques is its ability to detect sensitive information despite manipulation, reformatting, or other modification. Fingerprints enable the protection of whole or partial documents, antecedents, and derivative versions of the protected information, as well as snippets of the protected information whether cut and pasted or retyped.

Information fingerprints are highly optimized mathematical representations of sensitive content.

The Websense PreciselD Technology

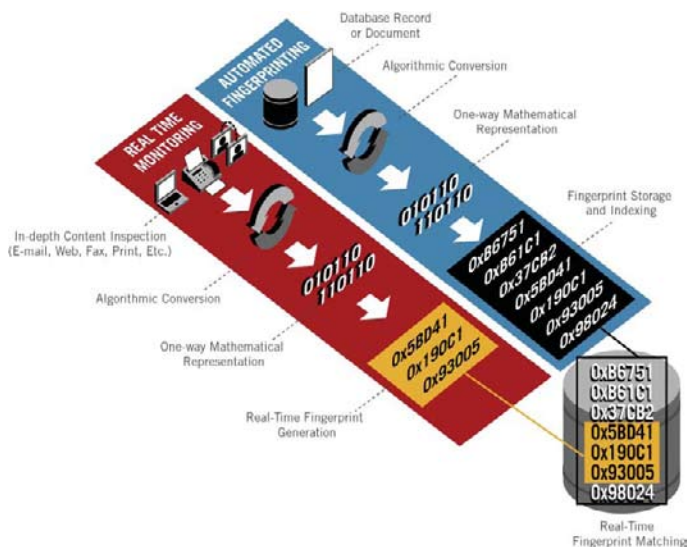
Websense PreciselD process works through a series of automated processes to create a fingerprint library and checks messages for sensitive content in real time. The system uses a compact and faithful numerical representation of the information, and supports multiple types of fingerprints for various information objects and business needs. Websense also offers a relevant measure of similarity from the business perspective to ascertain the uniqueness of content.

Websense PreciselD Technology works through two processes: an automated fingerprint creation process and a real-time matching process to reliably and accurately detect sensitive information.

Robust information identification is based on two main elements:

- The compact and faithful mathematical representation of information.
- A similarity measure that reflects the relevant similarity of the information items, using representations.

With Websense PreciselD , fingerprinting occurs in two processes: an automated process to extract fingerprints from originating content and a real-time process to match communications against known fingerprints. The diagram below outlines the steps at a high level.



Automated Fingerprint Creation

Based on a specific time interval, Websense PrecisID creates fingerprints using the following automated process:

- **Parsing:** Reading the document content from hundreds of supported file formats.
- **Normalization:** Ensuring that the information fingerprint is accurate and efficient.
- **Encoding:** Creating multiple, unique mathematical fingerprints that are robust to data manipulation.
- **Storage:** Storing and indexing each fingerprint in a highly optimized database with an associated rule that describes how the information may be used.

The automated PrecisID Technology creation process extracts sensitive data from existing sources, applies normalization and encoding to represent it, and stores it for real-time comparison.

Parsing: The textual (alpha-numeric) content of the information item is first extracted. The extraction is conducted from about 375 different formats, which make the encapsulation and format essentially transparent.

Normalization: The text is then “canonized” (or brought to a standard form) and preprocessed. Information that does not contribute to the identification process is removed, such as disclaimers, boiler plates, and some frequently used words. The canonized, pre-processed text is then transformed to the numerical domain using multiple hash functions, so there is a unique number for each segment of text. The length and structure of the segments and the overlap between the segments are carefully optimized, using Websense’s PrecisID technique. This set of numbers comprises a redundant representation of all the segments.

Encoding: To promote efficiency and security, a representative subset is thereafter selected from the redundant set, using a carefully tuned dilution scheme, which facilitates fast, robust efficient and accurate identification. This representative set is referred to as the primal textual fingerprint of information.

Storage: If time and storage are not a problem, one could simply store the information and compare two items using a standard comparison program. This, however, is not practical when there is a need to monitor intensive digital traffic and to decide if a certain message should be taken out of millions of confidential documents. In this case, a compact and faithful numerical representation of the information—a fingerprint—is required.

Fingerprinting an organization's confidential information is simplified with the Websense PreciseID file-system agent and user interface. The user can start by targeting directories with sensitive or confidential information and assigning the required policies to these directories. The file-system agent then recursively fingerprints all the information in these directories and stores the fingerprints together with the corresponding policies in a secure database. Finger-printing a very large file system can take days, but the process is automatic, and enforcement is operative from the start.

Real-Time Fingerprint Matching

Whenever Websense receives a message from a messaging server or application, the PreciseID fingerprinting engine creates a real-time fingerprint of that message and its associated attachments and stores it in memory. That real-time fingerprint is compared against the database of known fingerprints to identify any full or partial matches. Because the PreciseID algorithms are optimized for real-time performance, matching occurs in sub-second time in the same way anti-virus or anti-spam systems work, with no noticeable impact on messaging performance.

The PreciseID real-time fingerprint matching process creates fingerprints on the fly, compares them against the existing set of fingerprints, and identifies any full or partial matches.

To provide real-time accurate detection and identification, Websense has developed algorithms that allow fast comparison of the fingerprints of the analyzed traffic with fingerprints of multiple-millions of documents and to apply a context-sensitive similarity measure with adaptive thresholds. The similarity measure can detect, for example, a cut from a confidential document that was edited and then embedded in another large document, and can eliminate false positives that stem from non-relevant similarity.

Websense's PreciseID real-time fingerprint matching capabilities are agnostic to the communication channel it monitors. A Websense agent can be installed on any monitored communication channel and can extract fingerprints and other relevant information from the traffic in the channel and sends it for analysis. The agent then applies any relevant policy based on the results of the analysis.

How Are Contextual Identification Techniques Superior?

Contextual identification techniques provide significant advantages over less granular identification methods. While they can complement these earlier approaches, contextual identification techniques offer several key benefits:

- Significantly more accurate than global and token-based approaches alone.
- Extremely fast identification of sensitive content from millions of items indexed in a fingerprint library.
- Resilient to cut-and-paste attacks.
- Agnostic to the communication channel.
- Not file-specific, so the information itself is protected.
- Ability to identify information regardless of the format, encapsulation, and possible edits or changes to text.

Contextual identification techniques offer a wide range of benefits including: improved accuracy, resilience to data manipulation, and granularity to specific information, not just specific files.

THE NEXT STEP— EFFECTIVE POLICY ENFORCEMENT

Enterprises and financial organizations, as well as military and government agencies, are required to control and monitor the communications of sensitive information to protect customer data, confidential information and trade secrets. Unauthorized disclosure of this sensitive information can be prevented with robust information identification technology.

Less granular forms of information identification, such as detecting the binary signatures of files, can be rendered ineffective by any small change in a protected file. Robust information identification can identify the information, regardless of the format and reasonable edit changes.

These highly reliable and accurate information identification techniques are necessary to establish visibility into the magnitude and frequency of incidents. While some solutions provide detailed reporting and audit trails on leakage incidents, they typically do not prevent the actual transmissions of sensitive content. Simply monitoring for sensitive information in transit is insufficient.

While monitoring enterprise communications for sensitive information provides visibility into the policy violations, policy enforcement is required to stop unauthorized transmissions.

Real-time policy enforcement is the other critical component of a complete Data Loss Prevention solution. Real-time policy enforcement requires a high degree of granularity to enforce Data Loss prevention policies on real business processes.

Websense PreciseID is an advanced technology that permits truly effective Data Loss Prevention. The system's ability to assign identity to each information asset and to track information in motion is extremely powerful. In particular, Websense uses a combination of techniques that provide the high granularity required to enforce information distribution policies for real business processes.

Websense Data Security Suite can mirror specific internal policies to prevent unintended information leakage. For example, a rule may specify that document X, written by user Y, can only be sent by user Y or Z and only to recipients within the finance department. In addition, Websense is resilient enough to handle the normal modification of sensitive information. Administrators can easily define exactly who can send exactly what to whom under which circumstances in Websense. Real-life business processes often require that information to be edited, cut and pasted, or altered in some way, but the distribution of that information still needs to be controlled.

Ultimately, business-oriented policy enforcement should identify information regardless of the format and edit changes, and then apply the appropriate policies. Data Loss Prevention solutions must support, not hinder, existing business processes and should be transparent to users.

HOW CAN I START PROTECTING MY ORGANIZATION?

P³ Methodology: Prioritizing Information Identification Efforts

Many organizations grapple with how to address their Data Loss Prevention issues. Websense uses the P³ methodology for prioritizing information identification efforts. The P³ methodology consists of:

- **Principal**

Identify the principal business information in your organization. This 1% to 5% of information is the most proprietary and critical. Owners should know the exact whereabouts of principal information and what controls are in place to prevent its loss. Firms must fingerprint this critical information, assign information owners, and establish policies for the information.

- **Pareto**

Next, determine which 20% of the business information represents 80% of the value. Pareto represents the middle layer of information. This commonly used information typically resides in a few major data sources. This class of sensitive information should also be fingerprinted with specific owners designated and specific distribution policies outlined.

- **Progressive**

Finally, identify the lower-priority information assets that should be protected. Progressive information is often protected in phases, in which certain types of information assets are added in stages.

Starting an Data Loss Prevention project can seem overwhelming, but the P³ methodology helps organizations prioritize where to focus their efforts.

WEBSense STOPS DATA LEAKS. PERIOD.

Data protection success will be defined by those who understand how and what information is being communicated and who act quickly to deploy defenses that are designed to monitor and control this information when and where appropriate.

Websense Data Security Suite offers a comprehensive solution to stop information leaks reliably, accurately, and cost-effectively using patent-pending technologies and methods to detect sensitive content.

ABOUT WEBSense™

Websense, Inc. (NASDAQ: WBSN), a global leader in integrated Web, messaging and data protection technologies, provides Essential Information Protection(TM) for more than 42 million employees at more than 50,000 organizations worldwide. Distributed through its global network of channel partners, Websense software and hosted security solutions help organizations block malicious code, prevent the loss of confidential information and enforce Internet use and security policies. For more information, visit www.websense.com.

ABOUT THE AUTHOR

As a research fellow in Websense, Dr. Lidror Troyansky is leading the research of the DLP product-line in Websense Inc. Dr. Troyansky is engaged with DLP research for over 7 years and developed the algorithms of PreciseID fingerprinting and information classification. He is an Inventor and co-inventor of over 20 patents and patent-pending, related to PreciseID fingerprinting technology, DLP policies, signal and image processing, encryption and copyright protection. Lidror has also recently won the "Shaping Info Security" Award for his role in developing the core technology of the first DLP product with fingerprinting capabilities.

Dr. Troyansky is a published algorithms specialist with extensive experience in the fields of computational learning, computational complexity, pattern recognition and signal processing. He is a co-author of an important work regarding computational complexity, which was published in "Nature", with a follow-up report in the N.Y. Times science section, and has led a large number of research projects in various Hi-Tech industries.

Websense™
www.websense.com
+1 800.723.1166 Toll-Free