Entrust® Securing Digital Identities & Information

Securing Your Digital Life

Email Compliance Through Advanced Policy-Based Content Scanning

September, 2004

Entrust®
Securing Digital Identities & Information

Table of Contents

1 Introduction

As the number of email messages and documents exchanged continues to grow exponentially, there is a critical need to examine the best approaches to filter, sort, and organize them. According to IDC, today, over one billion corporate email boxes will receive 17 Billion spam messages[1]. In addition, over 75% of all documents created in the enterprise circulate in email[2]. Furthermore, 20% of all enterprise documents are redundant, with multiple copies being circulated and stored in the repository and archive[3]. The result is that end users and enterprises are being flooded by too many emails and documents, creating an urgent requirement to mitigate risk and better manage information.

To help them more easily manage the flood of information, organizations are examining various types of email and security tools and technologies. Their requirements include the need to examine both inbound and outbound email as well as any document and message repositories, and the need to reduce ever-increasing infrastructure costs required to manage large amounts of information. For these reasons, information filtering solutions and their underlying technologies are being investigated for their value in helping to lower costs.

This whitepaper provides a summary of technologies that are designed to address email and document filtering and analysis. It also illustrates why the Entrust content filtering and analysis solution is a valuable solution for organizations struggling with information management for risk mitigation. The paper focuses on addressing both corporate governance and regulatory compliance as it pertains to email communication. It also examines the management of the ever-increasing volume of spam in the enterprise and its impact on infrastructure costs and risk mitigation. Filtering spam is viewed as a subset of the problem of risk mitigation for the enterprise. The reader should note that it is more effective to adapt filtering technology to help achieve compliance, rather than to focus it on dealing with spam, as many spam filtering approaches are not effective for ensuring compliance.

2 What is Content Filtering and What Are the Typical Approaches?

As a field in information technology, content filtering has existed as an area that intersects information retrieval, information management, database technology, artificial intelligence, expert systems and computer science for many decades.

2.1 The Filtering Process

There are two basic steps in any content filtering process. The first is to create or capture a set of patterns to filter against and the second is to use a content filtering engine to detect the patterns found within email and electronic text. The content filtering process is illustrated in Figure 1.

The creation or capture of patterns can be as primitive as putting together a word list to filter against or it can be a fairly sophisticated hierarchy of concepts that represent patterns that terms contained within email and documents are matched against, as described in the following section. The content filtering and analysis engine can be equally primitive, simply searching for the words contained in the word lists and outputting any matches. Alternatively, the engine can be more

[1] IDC Study: The True Cost of Spam and the Value of Antispam Solutions, May 2004.
[2] Wireless Messaging Requirements, Gartner, March 2001.
[3] Pitney-Bowes, Meta Group 2003 report

sophisticated and encompass a set of statistical and linguistic tools to determine a match and the degree of probability that the match is accurate based on techniques such as fuzzy matching and other probability metrics.

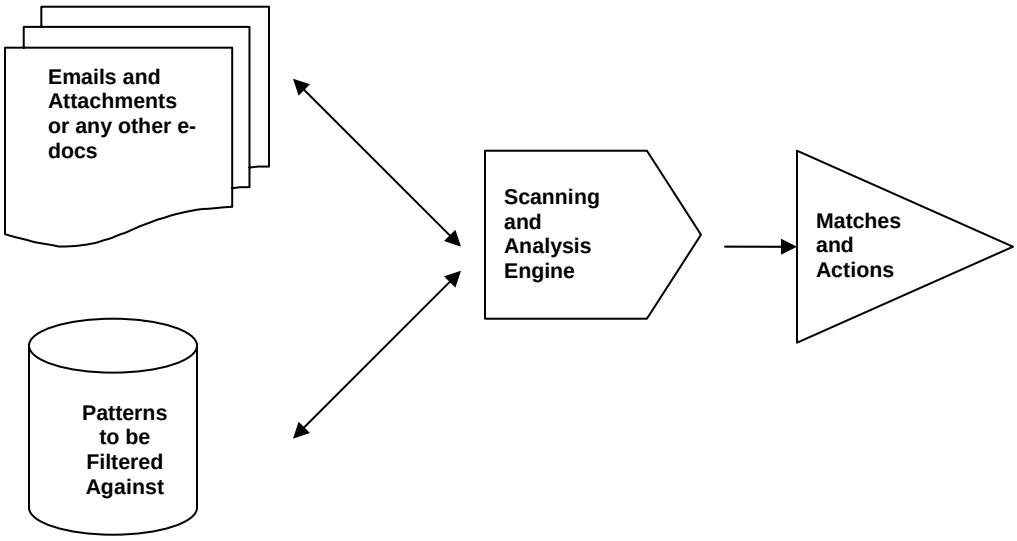


Figure 1: Filtering Process

.

2.2 Content Filtering Approaches

2.2.1 Word Lists

The earliest content filtering approaches relied upon search algorithms for one-on-one matches with words encoded as strings. These approaches have been in existence since the early days of computer science in the 1950's and words are typically encoded as strings of characters.

The following is an example of a word list created for blocking porn, a common form of spam:

Sex
Breast
Ass
.
.

Word lists can quickly grow to include thousands of words and provide a poor content filtering approach. They are particularly easy for spammers or abusive users to defeat by simply modifying words contained within the spam message by one letter such as adding an 's' or any other character such as ! or *.

Word lists have a number of disadvantages including the following:

- They rely on one-to-one exact word matches.
- You have to encode all word variations, which is an onerous task for the author. For example, to detect an email about mortgage rates in 'financial' spam or news about mortgages, a perfect match with the word 'mortgage' and the word 'rates' would have to occur. As such, if the email only included words such as mortgages and rate, no match would be found.
- They ignore phrases and look for only exact word matches making terms such as 'mortgage rates' fairly impossible to find.
- They ignore the semantics or the meaning behind the words which can be derived from looking at the words in context within the phrase, within the sentence, within the paragraph and finally within the email or document.
- They do not handle misspellings, soundex (a form of matching based on sounds-like) or word end variations which would all decrease the chance of a perfect match.
- They generate a high number of false positives as they assume a simple approach to matching the patterns.
- They are largely inaccurate and generate mostly false negatives as they ignore the underlying meaning of content and look too narrowly for a word match.

2.2.2 Rules-Based

The word list approach was extended in the 1970's to a rules-based approach that combined the use of a rules-based inference engine with a set of IF-THEN-ELSE conditions. In the case of spam, a rule is typically encoded as follows:

IF "breast" THEN SPAM

This is evolved to include exceptions to the rule:

IF ("breast") AND (NOT "breast cancer") THEN SPAM

This is evolved further as more exceptions begin to appear:

IF ("breast") OR ("breast enlargement" AND "sex") AND (NOT "breast cancer") AND (NOT "chicken breast") AND (NOT "breast reduction") AND (NOT "breast enlargement") THEN SPAM

In studies of Computer Science, it has been shown that for every rule, there is typically an exception[4]. According to the Oxford Dictionary, there are 231,000 current words and we can assume that only 20,000 are in common use. The IF-THEN rules-based systems have to encode 20,000 rules and are more than likely to encode an additional 10,000 to 20,000 exceptions. The process of creating and maintaining the rule base is far too onerous on IT Managers and Systems Administrators.

[4] William J. Clancey is considered one of the inventors of rule-based systems and has been widely published on the subject of rule-bases and diagnostic systems such as MIT's MYCIN which is a heart drug Q&A diagnostic system.

Rules extend the approach of one-to-one matches with words by including phrases and words in close proximity. This is the second oldest approach to content filtering, originating from the mid-1970's with the invention of expert systems that processed rules through an inference engine to make diagnoses based on symptoms of illness. For complete coverage, a rules-based engine has to encode the complete English Dictionary which has 231,000 entries with 20,000 in regular use. As such, rule bases become very large, with thousands of rules and at least one exception for each. Typically, a rule base will grow by 50% to manage the exceptions. Rules are very hard to manage and typically a large organization will have to employ full-time staff for their management. Used in isolation, rules generate a high number of false positives and can become inaccurate very quickly as word combinations change.

Rules-based filtering is the most prevalent form of content filtering. However, like word list filtering, it has a number of disadvantages including the following:

- It still relies on one-to-one perfect word matches for the conditional terms. As such, you still have to encode all word variations, which is onerous for the author of the rule base.
- Rule-based filtering handles phrases by stringing together a list of words and still looks for exact word matches in the phrases.
- It triggers on the first rule matched and ignores the subsequent rules in the chain. As such, rule order is critical, as one incorrect rule in a chain of rules will produce incorrect and unpredictable results. This is very difficult to manage for large rule bases.
- Rules are similar to word lists and they ignore the semantics or the underlying meaning behind the words which can be derived from looking at the words in context.
- Rules also typically generate a high number of false positives as they assume a simple approach to matching the patterns. Every false positive has to be handled by another exception rule thus adding to the complexity of managing the rule base.
- They are inaccurate and generate a lot false negatives as they ignore the underlying meaning of content and look too narrowly for a word match.

2.2.3 Natural Language Processing

Another filtering approach is based on a branch of AI known as Natural Language Processing (NLP)[5]. NLP as a field evolved from information theory and database search. The idea behind NLP was to allow end users to use language more natural to them in searching databases. As a field NLP has been around since the 1970s, if not earlier. Seminal work in the area was completed by people including Terry Winnograd and Roger Schank.

NLP is based on the linguistic analysis of text-to-verb phrases, noun phrases and speech acts. The goal of NLP is typically to analyze text based on grammar trees that are based on the rules of grammar. NLP has resulted in very sophisticated analysis based on speech acts and is used in a number of areas including text analysis. The importance of NLP is that it can produce highly accurate results.

The main disadvantage of NLP is that it can be NP-complete and sometimes no single solution can be found within a finite time period. Practitioners in NLP typically bind the processing time by limiting the analysis and reducing the number of grammar tree comparisons. Entrust's patented content analysis engine uses a hybrid partial NLP and statistical learning approach.

[5] Natural Language Processing (NLP) – one of the pioneers of NLP is Terry Winograd.

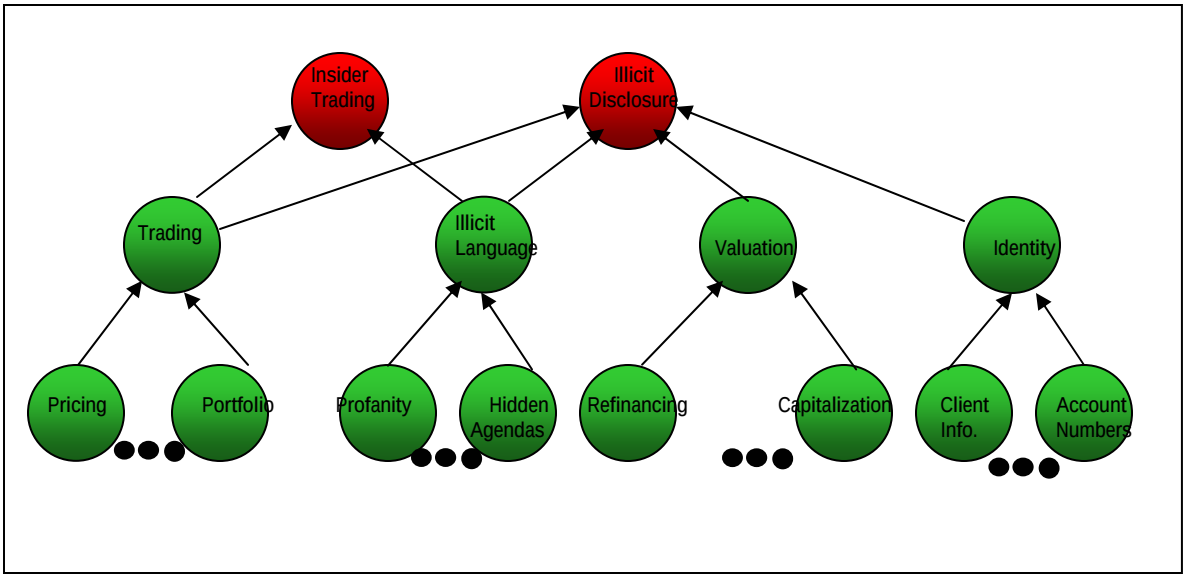
2.2.4 Concept-Based: Well Structured Knowledge

In the late 1980's and early 1990s, a new approach to organizing knowledge bases was invented[6]. This approach has dramatically changed the organization of knowledge bases. It is to content filtering and analysis what object oriented programming was to procedural programming. It greatly advances the manner in which content can be reasoned about and organized. This is especially useful in a broad suite of applications such as diagnostic systems, CRM tools, help desk applications and for compliance. Deep knowledge of terminology and context is essential in interpreting the difference between a compliance violation and a unique sales approach.

As an example, consider the fragment of a concept base for securities information, illustrated below. The intent of the concept fragment is to identify if the content of an email exchanged between a broker and a client provides the basis for non-compliance. The concept-based approach works in conjunction with a concept analysis engine that interprets the complex relations between concepts. The goal of a concept engine is to collect evidence in the way that a detective does to help interpret a case. Concepts are loosely organized into a hierarchy with parent and child nodes, as well as sibling nodes that represent "concepts." The concepts are not simply words, phrases or sentence fragments but rather represent a set of patterns that identify a concept with a certain degree of probability. The concept engine is fairly sophisticated in its matching capability and takes advantage of speech acts, stemming, fuzzy matches, distinction between verb and noun phrases and other linguistic relations used in daily personal and professional communication. The sophistication of the concept interpretation engine allows it to more accurately analyze and draw conclusions about potential compliance violations.

The trigger of concepts that are non-compliant results in actions being taken based on the policies and processes within an enterprise. In some cases, non-compliant emails are quarantined, silently audited, or sent to the end user for reconsideration in some cases, with a warning for education purposes.

[6] B. Chandrasekaran – Chandrasekara B. and Mittal S. Deep Versus Compiled Knowledge Approaches to Diagnostic Problem Solving. International Journal of Man-Machine Studies, Vol 19, No. 5, Nov 1983,
Abu-Hakima – RATIONALE – A tool for Developing Knowledge-Based Systems that Explain by Reasoning Explicitly, April 1988, Masters Thesis, Carleton University. See also PhD work on the Diagnostic Remodeler Algorithm, 1994.



The use of concept bases for content analysis allows the application of ***Context within Content***. This approach provides great advancement over the word list and rule-based approaches and is based on technology that originated in the mid-1990's and has advanced to today. The Context and Content Analysis approach is the most advanced of the content filtering approaches. It uses a hybrid of partial Natural Language Processing (NLP) to identify the meaning of content and places it in context. It also makes use of rapid statistical analysis to ensure that processing is optimized while maintaining linguistic fidelity. This is especially useful in providing automatic and deep analysis of all email content including: header, message body, closing/sign-off, and attachments. It makes use of speech acts, fuzzy matching, and natural and raw text search to provide the best pattern matching capability that is decades ahead of the rules-based approach. The well-structured knowledge base of concepts is organized as a set of objects to represent all aspects of a term and relates it to other terms. In comparison with a rules base, a concept base greatly reduces the complexity of the patterns to be matched against and is far easier for administrators to decipher and manage.

The main advantage of the hybrid partial NLP and statistical analysis approach that Entrust uses over the concept-based approach is its accuracy and speed of analysis. Entrust provides a number of concept bases, with an easy to use, sophisticated policy editor for corporate and regulatory compliance.

3 Key Challenge of Content Filtering: Pattern Generation

The patterns that are compared against for content filtering are a key challenge for word list, rule-based and concept based approaches. There are two types of pattern generation: manual and automatic.

The manual approach requires the IT Manager or Domain Expert to generate a set of words, rules or concepts. This is typically a task that should be taken on by someone knowledgeable in a particular domain, as it requires both background and foreground knowledge.

The automated approach is prevalent in the rule-based approach. A number of techniques from the Artificial Intelligence field have been used, as is discussed in the sections that follow.

3.1 Approaches to Generating Patterns

Word lists typically rely on electronic dictionaries. The dictionary has to include as many as 20,000 words. As those words are not matched, an additional list of words is added that could go as high as 230,000 words—the number of terms in the dictionary. This quickly becomes a non-viable approach.

Rule bases result in both rules and exceptions and as such, become fairly hard to manage. Rule bases can grow to include approximately 10,000-plus rules, with an equal number of exceptions. Automatic approaches discussed in section 3.2 are sometimes used to generate rules.

Concept bases are typically generated using hand-crafted expert knowledge that provides semantic knowledge to aid in the analysis of the text. Some aspects of concept-base knowledge can also be generated automatically based on sample patterns.

3.2 Automatic Generation of Rule Bases - Probabilistic Approaches: Classification Learning, Bayesian Networks, Neural Networks, Genetic Algorithms

To facilitate the automatic generation of rule bases, a number of probabilistic approaches are now commonly used to generate a set of patterns for content filters. These approaches are often referred to as Classification Learning, Bayesian Networks, Neural Networks or Genetic Algorithms.

These four approaches use probabilities that the words encoded in the patterns appear in the text to be matched against. They take advantage of statistical techniques to quickly analyze the text. The analysis is not very accurate, however, as it is based on the frequency of occurrence of the words. The focus of the analysis is how often the words appear in the text rather than their semantic meaning, and all approaches ignore the context and meaning of phrases within email.

While these approaches facilitate the creation of the rule base, they come with their own problems relating to accuracy and understandability for the end user. The approaches typically fall under the category of Artificial Intelligence (AI) and more precisely classified as Machine Learning algorithms, a branch of AI.

All the methods described below generate probabilities that relate words, phrases or word stems to others. As such, to match a document, the document content is first analyzed based on the occurrence of terms and the frequencies of the occurrences. A match is found if the terms and frequencies indicate that there is a match.

All these approaches require that the model that is 'learned' be re-trained. They are also purely statistical in their approach and as such, completely ignore contextual information that is gained from the analysis of the linguistic content.

3.2.1 Classification Learning

Classification learning is the main probabilistic approach to categorizing content based on the occurrence frequency of terms. Terms are expanded to include words and phrases. Some classification learners also attempt to bias the learning based on the proximity of terms such as the typical hand-coded rules-based systems do. Typical classifiers are based on published algorithms such as ID3, C4.5, C5[7]. A data set (for example a message base) is divided into a

[7] Ross Quinlan is considered a forefather of the Classification Learning Algorithms in the AI field.

training set and a test set. The training set is used to generate a set of patterns of words and phrases based on the occurrence frequency and sometimes proximity, to categorize a message against. Classification learning patterns are criticized for being difficult to understand and for problems with accuracy based on over-fitting the data to the model. As such, frequent re-training is required to continuously improve upon the accuracy. Iterative classifiers have been found to address this issue and allow users to modify the classification patterns to help improve performance. The Entrust categorization technology is capable of iterative classification.

3.2.2 Bayesian Networks

Bayesian networks[8] are used to classify words contained within a message or document. Bayesian networks typically generate words and a probability that relates a particular word to another. This set of patterns is used to compare against to categorize a new message. This, like all other machine learning approaches, is a statistical approach and typically ignores linguistic analysis. As such, Bayesian approaches miss the rich contextual knowledge that concept-based well-structured knowledge approaches have. The accuracy of Bayesian networks is often criticized for these reasons.

3.2.3 Neural Networks

Neural Networks[9] are another probabilistic-based approach used for classification. A single or multi-layered network is trained with test data to classify certain inputs. In turn, the training generates the probabilities that are used to tune the network. New data is input into to the network to categorize. Neural Networks have been successfully used to classify images but have resulted in poor accuracies for the classification of textual content.

3.2.4 Genetic Algorithms

Genetic Algorithms[10] (GAs) are another probabilistic-based approach for classification. They encode the patterns using genetic mapping of ones and zeros against the inputs and vary according to genetic variations. The problem with GAs is that they are even harder to decipher in the patterns they encode than typical classification algorithms and their classification results are often unpredictable.

[8] Bayes Theorem - Bayes, T. 1764. "An Essay Toward Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London* **53**, 370-418.
[9] Neural Network is an area of much study – see seminal work by Geoffrey E. Hinton.
[10] Genetic Algorithms – see work by John Holland considered to be the inventor of the field and and Tuevo Kohonen who has significantly contributed.

4 Summary of Content Filtering Approaches

The table below summarizes the content filtering approaches discussed in this white paper, from the most primitive approach of word list filters to the most sophisticated concept-based approach. Additional variations are introduced through the use of sophisticated machine learning approaches to automatically generate rule bases for matching against. All rule-based approaches, whether they use rule bases that are hand-coded or machine-generated through Bayesian, Classification, Genetic or Neural Network algorithms, quickly become cumbersome, hard to manage and inaccurate. Entrust's concept-based approach combines the sophistication of natural language processing and statistical analysis to help provide improved accuracy and consistently more accurate results.

Method	Description of Approach	Advantages	Disadvantages
Word List	I.T. Manager or user compiles a list of words to detect in email or document content.	Simple	Time-consuming to create & maintain Quickly becomes inaccurate
Rule Bases	I.T. Manager or user compiles lists of "If-then" conditions and exceptions to detect in email or document content.	First set of 100 rules easy to compile.	As rule lists grow to thousands of rules and exceptions, the lists become very time-consuming to create & maintain.
Natural Language Processing (NLP)	Natural Language breaks down the sentence into its basic grammatical components	Accuracy. Identifies the deep linguistic components.	Speed – very slow technique, not suitable for thousands or millions of messages.
Concept Bases	Knowledge base of hierarchy of concepts developed by human experts with sophisticated matching algorithms for content analysis. Concepts shared and applicable to multiple domains.	Consistently more accurate. Catches new waves of patterns in compliance and spam as content is analyzed using partial NLP and statistics rather than simplistic word matching. Improved Accuracy. Easier to maintain. Supports remote updates. Library of concepts provides a rich base for new concept hierarchies.	Requires human expert knowledge for creation. Makes use of natural language and statistical analysis to generate accurate matches.

4.1 Spam-Blocking Approaches

As spam continues to double the cost of infrastructure[11], it is important to examine the number of approaches that exist to battle spam. A number of these approaches are sometimes combined by vendors to improve spam-blocking accuracy and reduce the occurrence of false positives. The approaches to content filtering that apply to compliance and spam are highlighted in blue. Note that techniques like signature methods, RBLs, DCC, and permission-based methods do not transition to compliance detection. Furthermore, Bayesian methods can produce poor results for spam and do not transition effectively to compliance applications. The Entrust solution extends its concept-based approach to content filtering to block spam with improved reliably and accuracy.

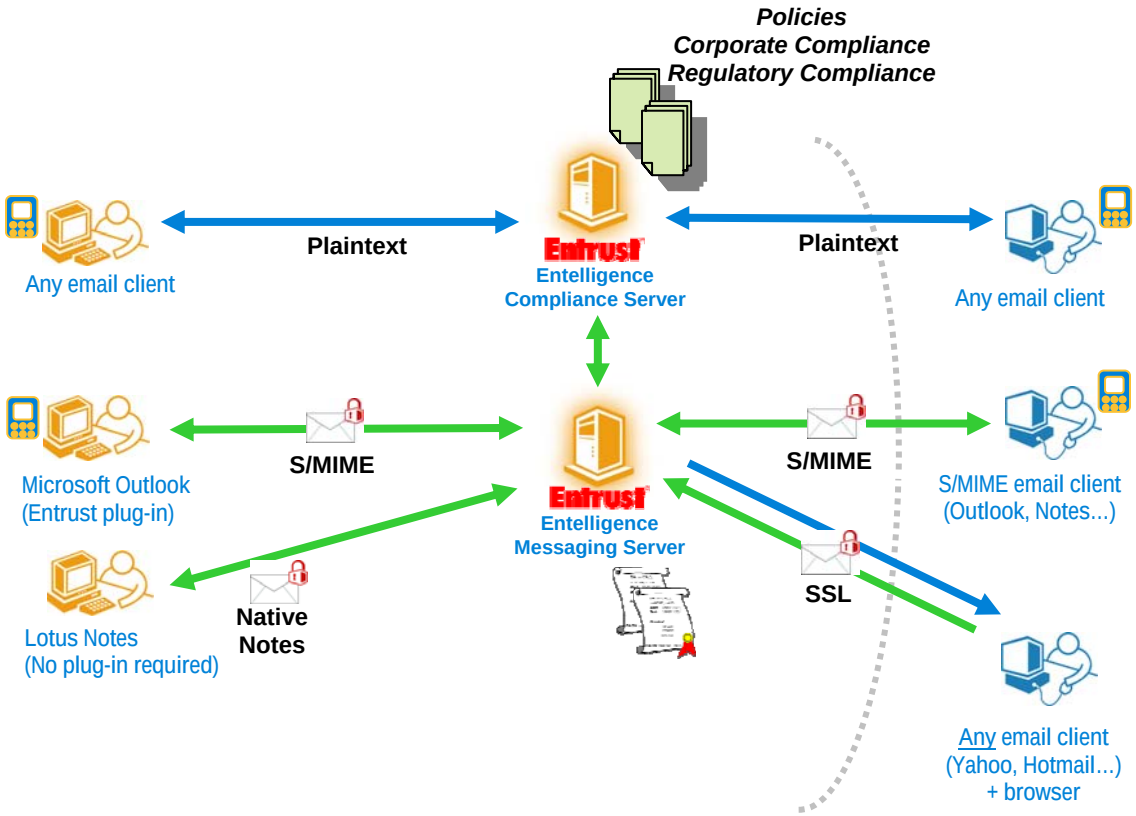
Technique	Method	Advantages	Disadvantages
RBL (Rule-Based Lookup)	IP blocking based on reported spam.	Simplicity.	Shotgun approach. Scalability. Trust.
DCC (Distributed Checksum Clearing houses)	Message blocking based on checksums seen.	Per-message. Simplicity.	Coverage. Latency. Network overhead. Whitelist required.
Other signature methods.	Like DCC, but reviewed by people commercial or Peer to Peer.	Per-message. Simplicity.	Latency. Aging out.
Rules-based Content filtering.	Rules-based keyword/phrase matching.	Per-message. Customizable.	Complex to maintain. Overhead on IT manager adding rules if not subscription-based.
Client-based whitelisting + Bayesian + no HTML	Address-book is Whitelist. Bayesian trained by user. Reject HTML-only email.	Per-message. Highly effective.	User is kept busy: address updates, re-training engine. False positives high.
Permission-based.	Sender is authenticated, either by mark or membership.	Per-message. Highly effective.	False positives. Irritation by senders and lack of coverage for billions of email users.
Concept-based Content filtering.	Uses highly organized knowledge bases of concepts and rich natural language and statistical analysis engine.	Per-message. Highly accurate in capturing spam and very few (if any) false positives.	Requires expertise for creation. Helps to removes overhead on IT manager through subscription updates.

[11] Radicatti 2004 Study

5 Entrust's Approach to Content Filtering

Entrust is a pioneer in concept-based filtering for compliance and anti-spam. Entrust's approach compares incoming emails and attachments to the patterns in the concept base. The analysis combines partial Natural Language Processing and faster statistical linguistic analysis. The analysis includes stemming, fuzzy matching and soundex capabilities.

The system also makes use of a patented structure extraction engine that extracts structure from basic components of an email or document: e.g. greeting, sign-offs, disclaimers, body, list of items, etc.



The Entrust Entelligence™ Compliance Server has a number of sophisticated policy modules. A set of these is bundled under the heading, of "Corporate Compliance," which includes modules for the detection of sensitive or confidential documents, spam, profanity, harassment and content that includes individual privacy-protected information. The regulated compliance modules include policy modules for specific securities rules (SEC, NASD, SOX, GBL), specific healthcare privacy rules (HIPAA) and individual privacy legislation.

The Entrust Entelligence Compliance Server is designed to monitor both inbound and outbound email. As emails enter the organization, they pass through the Entrust Entelligence Compliance Server, which can reject spam. The server can also tag the emails with categories that facilitate records management and archiving for the enterprise. It also makes them much easier for the end user to sort into relevant folders. As email leaves the organization, it can be examined for

sensitive content. If it requires encryption, it is encrypted and then released. The emails can also be analyzed and categorized for compliance purposes. If an email is non-compliant, it can be quarantined, silently audited, forwarded to a compliance officer or sent back to the end user for reconsideration.

The Entrust Entelligence Compliance Server offers the following advantages:

- It is adaptive to human dialogue used within email to help reduce false positive and false negative rates beyond the capabilities of statistical analysis techniques on their own.
- Provides enhanced speed and accuracy.
- Provides comprehensive regulatory compliance coverage.
- Provides audit capabilities that provide customers with the ability to react quickly to scan results, or to retrace activities and results after-the-fact and still take necessary action as required.

In conclusion, the Entrust Entelligence Compliance Server can help provide:

1. **Better network and productivity protection**
   - Exhibits fewer false positives
   - Provides speed and accuracy
   - Reduces total cost of ownership (TCO)
2. **Better compliance**
   - Provides customizable policies
   - Provided by the only vendor with out-of-the-box privacy rules for cross-border compliance
   - Offers audit capabilities
3. **Better risk mitigation**
   - Provides real-time compliance
   - Offers a robust secure email solution

6 About Entrust

Entrust, Inc. [NASDAQ: ENTU] is a world-leader in securing digital identities and information. Over 1,400 enterprises and government agencies in more than 50 countries rely on Entrust solutions to help secure the digital lives of their citizens, customers, employees and partners. Our proven software and services help customers achieve regulatory and corporate compliance, while turning security challenges such as identity theft and email security into business opportunities.

