

# Effective Content Analysis for Email Inspection & Control

A Whitepaper By:

**Nemx Software Corporation**

14 Poplarwood Avenue  
Ottawa, Ontario K2S 1V3  
Canada

**Effective Content Analysis  
for Email Control**

**TABLE OF CONTENTS**

<b>Email Control—Understanding the Issue .....</b>	<b>3</b>
<i>Content-based vs. Access-based Rules .....</i>	<i>4</i>
<b>Effective Email Content Analysis .....</b>	<b>6</b>
<i>The Secret to Accurate Content Analysis.....</i>	<i>6</i>
<i>Shortcomings of Common Content Analysis Techniques .....</i>	<i>6</i>
Key Words/Phrases .....	6
Bayesian Analysis .....	7
<b>Key Features Of Intelligent Content Analysis (ICA) .....</b>	<b>7</b>
Semantic Inference.....	8
Positional Analysis .....	8
Linguistic Analysis .....	8
Term Weighting.....	9
Information Concepts.....	9
<b>Other Considerations .....</b>	<b>10</b>
<b>Conclusion .....</b>	<b>11</b>

## Effective Content Analysis for Email Control

*This white paper will describe the techniques and capabilities required to provide intelligent, effective content analysis for email monitoring, compliance and control applications and discuss the shortcomings of traditional key word/key phrase solutions.*

***There will be 103 billion corporate email messages per day in 2008.<sup>1</sup>***  
***Nearly 97% of most companies' communications are via email.<sup>2</sup>***  
***Over 75% of the documents created by the enterprise get circulated by email.<sup>3</sup>***

At the same time as this burgeoning growth in email volume and use has occurred enterprises are facing numerous strict new regulations, both externally and internally imposed, governing the disclosure, safekeeping and distribution of personal, private, financial and other corporate sensitive information.

No wonder email content control has become a major concern for virtually every organization.

### Email Control—Understanding the Issue

Effective management and control of the information flowing through corporate email systems is imperative, in some cases mandatory. As the statistics above clearly illustrate, organizations have seen their corporate email system become a de facto repository and distribution mechanism for the vast majority of their data and corporate knowledge. What's more, vast amounts of personal, non-business related, and other types of possibly inappropriate information is flowing through (and stored within!) corporate email networks.<sup>4</sup> Once inside the email environment it is extremely difficult to protect and control the information.

In recent years new rules and laws have been enacted such as the Sarbanes-Oxley Act (SOX), Gramm-Leach Bliley Act (GLBA), Health Insurance Portability & Accountability Act (HIPAA), SEC Rule 17a, Personal Information Protection & Electronic Documents Act (PIPEDA, Canada), and others that strictly govern the protection, use and sharing of specific types of information. All these regulations recognize email as a valid record of corporate communication, conversation and behaviour. This has had far reaching implications in terms of corporate risk and liability associated with the use of email. Email is frequently targeted and subjected to serious scrutiny during any compliance investigation, legal proceeding, or dispute of any kind.<sup>5</sup>

Corporate enforcement of regulatory controls is mandatory, but organizations have also been quick to recognize the "best practices" embodied by these legislated regulations and this has sparked renewed emphasis on implementing and enforcing an organization's own internal corporate policies and practices, in addition to their legal or regulatory obligations. The challenge, of course, is how to enforce these internal rules (as they apply to email) when they can be so diverse and potentially unique within any given company.

<sup>1</sup> Source: Radicati Group Inc., "Active Policy Management – Third Generation Compliance for Today's Corporate Environment" Whitepaper February 2005

<sup>2</sup> Source: Gartner Group – survey 2002

<sup>3</sup> Source: Gartner Group – Wireless Messaging Requirements, March 2001.

<sup>4</sup> "Nearly 50% of corporate email users have sent or received inappropriate content" – Source: Harris Interactive

<sup>5</sup> "20% of businesses surveyed have had email subpoenaed in the course of a lawsuit or regulatory investigation." Source: American Management Institute (AMA) and the ePolicy Institute joint 2004 Workplace Email survey

## Effective Content Analysis for Email Control

The real problem, from a management and control perspective, is that these regulations and corporate policies define rules for policing *information* not people. Technically, this is a much bigger challenge.

Consider the following typical company scenarios;

- o all email containing improper content (i.e. offensive, harassing or discriminatory language) should be blocked;
- o current product information can be sent internally or externally but product *plans* must be restricted to certain employees
- o corporate financial information can be sent between company executives and to external board members only
- o sales may be allowed to send customer names or lists to prospects as references but you don't want them going to competitors
- o distribution of personally identifiable data and other sensitive or confidential corporate information must be strictly controlled in a manner that complies with both regulatory and corporate rules and policies

The point is clear, the determining factors on what should or shouldn't happen to an email is not just who is sending or receiving it, but also, the *type* of information it contains.

### Content-based vs. Access-based Rules

For generally all other corporate applications control is achieved by associating rules and security policies to individuals that control whether or not they can access a given application or data. The focal point for traditional security policy definition, and enforcement, is the person—what s/he can access, manipulate or share—based on their individual corporate and functional responsibilities. Applied to usual corporate database and formal application systems, where the nature of the data/information they contain or manage is always known in advance, the normal procedure is to prevent all access by default then define "permissions" that govern whether specific individuals, groups or departments can access, manipulate or share the information. This model allows for very sophisticated and comprehensive management and control, but it is predicated on well defined user roles and on having prior knowledge concerning the contents, or information concept(s), contained within the system for which access is being granted.

A typical example of access-based control might be a human resources database. The database would contain all employment information for all company employees including sensitive information like salaries, performance reviews and financial and banking data (for payroll) etc. Knowing the nature of the information in the HR system, access to this database would be restricted to HR staff and perhaps senior executive management only. In fact, even within the HR department there might be additional restrictions on information like salaries where only HR managers would have access to salaries of other managers, directors and executive officers. This example highlights typical security policies that are based on "who you are and what your role in the organization is."

*You don't provide blind, uncontrolled access to your corporate databases — so why would the information flowing through or stored within your corporate email system be completely open and uncontrolled?*

Just the opposite is true of email. What has made it such an indispensable tool is the fact that everyone, by default, has access. In addition, you can send and receive almost any kind of data or information concerning virtually any subject to anyone with an email address. Unlike the majority of corporate databases, email has evolved into a non-subject matter-specific repository. The content of each email message is not something that's predefined and accessed, rather it's created 'on-the-fly', and sent impulsively. If you can type it, or attach it, you can send it. The completely spontaneous nature of email makes it impossible to predetermine a

## Effective Content Analysis for Email Control

message's content and, therefore, to impose any presumptive controls or restrictions like most database applications do. The real problem here is that by the time anyone but the author knows what the message content is, it's too late—the email has already been sent. We know from experience, supported by numerous studies, that it's impractical to expect and rely on users to classify and appropriately process every message they send or receive based on its content – even if such policies are defined and documented.<sup>6</sup>

Using the HR example again, while employees' salary information is protected within the HR database, what prevents an HR staff member, with legitimate access to salary information, from sending an email containing an employee's salary information to someone not authorized to have it? The answer for most organizations is—Nothing!

This unpredictable, ubiquitous, and content-generic nature of email prevents effective control being established solely through rules related to people and their roles or responsibilities. This is because the use of, and information contained in, email is, of necessity, not always directly related to an individual's specific corporate role. Consequently, the traditional kind of security and control measures utilized for other applications are ineffective because a) it is too easy to circumvent or violate them (deliberately or inadvertently), and b). there are simply too many valid variations and exceptions to handle. For example, while it's possible that the author has quite legitimate access to a particular document s/he has attached to their email it does not follow that all the intended recipients necessarily do. The average email system has no inherent checks and balances to guard against such potential unauthorized information dissemination.

*An organization with 500 employees sending only 10 external emails each per day would typically have up to 250 emails per day sent externally that contain confidential or sensitive corporate information!*

From a security officer's perspective email is a sieve – moving information, potentially sensitive information, is as easy as sending or receiving a message upon which there are few if any restrictions. Imagine a corporate file server where literally anyone, including non-employees, had access to any and all of the information contained on it. It's an untenable thought and would never be allowed to happen. But, without effective information-centric controls that's pretty much what most corporate email environments are like today!

From a compliance officer's perspective email is a largely uncontrolled, unmanaged risk exposing the corporation to costly liabilities and potentially severe penalties.

The only practical way to effectively control information distribution via email is through an information-centric approach to policy definition and enforcement. In effect, the content of each email must be inspected *before* it is delivered and any applicable actions or restrictions (such as block delivery, quarantine, reroute, copy or encrypt) enforced.

Of course, the real challenge arises in trying to accurately determine the subject matter (content) of the message, or its attached document so that the appropriate rules/policies can be applied.

<sup>6</sup> "Less than half of email users always comply with corporate email policies" – Source: Harris Interactive

## Effective Content Analysis for Email Control

### Effective Email Content Analysis

#### The Secret to Accurate Content Analysis

The first step in email content compliance enforcement, therefore, is to understand whether or not the message (and its attachments) contains information that is subject to any internal or external controls, policies or restrictions. Only then can the rules associated with that particular type of information be applied—is the sender allowed to send this information, is the recipient(s) allowed to receive it, should someone else be notified, should the email be archived, redirected or copied, should the message be encrypted, etc.

#### ***The Key to Accurate Content Analysis***

- 1) *well defined descriptions of the various types of information that needs to be detected and controlled*
- 2) *understanding the actual context of the information in an email message or attachment, and*
- 3) *recognizing underlying information concept(s) contained within the message and its attachments, and not simply looking for words*

Accurate detection of controlled content is crucial to the success of all compliance efforts. Errors cost the organization in terms of:

- severe financial penalties (fines) and/or potential liability
- damaged corporate reputation and trust
- possible litigation
- lost business (revenue, customers, partners)
- user frustration and lost productivity, and
- increased administrative and operational costs associated with email

if either compliance violations (i.e. failure to detect controlled content and take the appropriate actions) or too many false positives (i.e. messages erroneously identified as restricted) occur.

#### Shortcomings of Common Content Analysis Techniques

Various technologies and techniques have been employed by vendors in their email scanning or filtering products. Unfortunately, many of the most popular approaches are far less than ideal.

##### ***Key Words/Phrases***

The most prevalent method for email filtering is scanning for specified words or phrases. Despite its widespread use, it is one of the least accurate and effective approaches for compliance applications. It is truly a scanning operation and not content analysis.

The most significant drawbacks to the key words/phrases approach are:

- it's easily defeated (deliberately or innocently) as this method looks for an exact match – a simple misspelling or word variation goes undetected
  - 'flavor' could be detected but 'flavour' would not (unless it was specifically added to the word list)
  - plurals and varying tenses are not recognized i.e. test ≠ tests ≠ tested ≠ testing
- it lacks any contextual understanding that can be derived from the word's location within the email or document or from semantic analysis of the words in the same sentence or paragraph
  - 'breast' may be a key word that triggers a sexual harassment policy, but when 'cancer' or 'disease' or 'x-ray' or 'chicken' are in the same sentence it is most probably a legitimate use of 'breast' in a non-harassing context

## Effective Content Analysis for Email Control

- it generates a high false positive rate (usually because there is no contextual understanding – see point above)
- it generates exceptionally high false negatives and fails to detect significant numbers of messages that should have been controlled
  - scanning for 'financial report' will not find 'financial statement'
  - 'trigger' will not detect the misspelled 'triger'
  - 'formula' will not detect 'recipe' or 'equation'
- developing, maintaining and administering the word lists is an extremely time consuming, costly and usually manual exercise
  - there are over 230,000 words in the English language of which 20,000 – 30,000 have common usage, as just the few examples above illustrates it is impossible to second guess and maintain all the potential variations, combinations and alternatives that might be used by your users even if partial word lists are provided in the product

### **Bayesian Analysis**

Bayesian analysis is a frequently used technique for anti-spam filters and has been applied by some vendors to general purpose content monitoring for email compliance applications. While the method has been helpful for certain spam applications its effectiveness for corporate compliance-based content control suffers from some notable shortcomings.

- limited accuracy (although an improvement over pure key words alone)
- it is a statistical probabilistic approach based on word frequency that still ignores any linguistic analysis and as a result any contextual knowledge that can be derived from the content
  - a Bayesian pattern of related words could represent a tree species and words like oak, pine, poplar and birch would be linked together, but an article about the Toronto Maple Leafs NHL hockey team could easily be misinterpreted as being about trees!
- the complex patterns of linked words created by Bayesian networks is not easily modified or updated even by admin users and these learned patterns must often be re-trained
  - some industries, disciplines and fields of technology have evolved patterns of word use that take on different meaning within their specific fields of reference than these same words might have in other more common use—for instance is the occurrence of terms like 'load', 'channel', 'delivery', 'customer', 'overseas', 'distribution', 'network' and 'ship' referring to the shipping/transportation industry or a software company's indirect sales model?

### **Key Features Of Intelligent Content Analysis (ICA)**

To achieve the requisite level of detection accuracy for all content control operations the shortcomings identified earlier must be overcome. This can be achieved by adding both *context* and *conceptual* meaning to the content analysis operation.

A variety of measures and techniques can be implemented to aid in determining contextual knowledge as well as conceptual meaning. Some of these are briefly described below.

## Effective Content Analysis for Email Control

### *Semantic Inference*

By considering the proximity of the terms to each other the context in which they are used is more easily determined. The content analysis engine should understand grammatical structure and have the ability to identify or exclude terms within:

- the same sentence
- the same paragraph
- the document (anywhere), or
- a user specified distance of each other (usually stated as within 'x' characters)

This capability allows the content analysis operation to differentiate the inappropriate statement 'wow, she's got huge breasts' from the quite legitimate 'she may have breast cancer' enabling the email control solution to block delivery of the former message and allow the latter to be delivered normally.

### *Positional Analysis*

Very often the location of where a term is found can provide clues as to its significance to the overall concept and context of the content. For example, is the occurrence of 'confidential' in the middle of an email message or a paragraph in a document mean that the message or document is itself confidential, or is it just a reference to some other confidential material? That's difficult to know. On the other hand if 'confidential' appears in the footer of the document or in the subject line of the email message there is a very high probability that it indicates that the email and/or document should be considered confidential and controlled accordingly.

Intelligent content analysis should be able to infer significance based on location or position within the target content. The system should differentiate between terms found:

- in the subject line of an email message
- near the beginning of the message
- near the end of the message
- in a title
- in a header
- in a footer
- in a bullet point

### *Linguistic Analysis*

As stated earlier the biggest problem with key word scanning is that such systems look for exact matches of just the specified term(s), Word variations and similar words with the same meaning are ignored. This results in a huge number of false negatives – controlled or restricted information that simply slips through the system undetected!

Consider this example. You want to identify email discussing smart action trigger technologies. You create a rule looking for: 'smart' and 'action' and 'trigger'. A key word-based solution will identify the following message:

"We need to build in a smart action trigger capability."

This message: "We need to include smart action triggers."

will be undetected simply because of the plural use of triggers. Triggers ≠ Trigger in a key word system.

If 'confidential' appears in an attached Word document does that mean the document is confidential?

But, what if 'confidential' appears in the footer of the document?

## Effective Content Analysis for Email Control

Natural language processing and other linguistic techniques solve these kinds of problems. Effective content analysis requires a solution that *automatically* expands user specified terms to include plurals, different tenses and similar words. Features that enable this capability include:

- o stemming (prefix/suffix stripping) – this deconstructs a term to its ‘root’ word
- o root word expansion – this effectively re-appends all the potential suffix and prefix combinations (which accounts for plurals, tense variations, etc.)
- o dictionary lookup – this technique adds terms that may be used to describe the concept or meaning of the search term used (it’s useful as well in spam filters to identify words that do not exist as spam emails often contain strings of randomly generated characters)
- o thesaurus lookup – this technique adds terms that have the same meaning as the specified term (i.e. are synonyms)
- o thresholding – searches for similar misspelled words, such that ‘locaton’ produces a match for the term ‘location’

The combination of these techniques, together with positional and semantic analysis, provides the ability to define and monitor for *information concepts* (i.e. a subject) rather than simple search terms. The power of these techniques is evident in the following example. Using the same rule defined above – search for: ‘smart’ and ‘action’ and ‘trigger’, a message containing:

“This system has a very clever triggering mechanism and reacts well”

will be identified by an intelligent content analysis engine as being about essentially the same subject or concept as a smart action trigger.

Similarly, the phrase ‘company proprietary’, ‘eyes only’ and ‘not for external distribution’ would be related to the concept of ‘confidential’ and messages or documents containing these phrases would trigger the ‘confidential’ policy.

### ***Term Weighting***

The ability to associate a weighting factor, or score, to specific terms or concepts is a key method that enables fine tuning for content analysis. The combination of automatic system assigned weightings and the ability for users to assign scores associated with terms and/or locations where terms are found provides significantly improved accuracy in the policy representation of the information concept being defined.

The approach works by setting a threshold score, say 100, that causes an action (block the message, quarantine the message, redirect it, etc.) to be triggered. Search terms and concepts can have weights assigned that determine whether or not the triggering criteria have been met. For instance, if the threshold value is 100 the following criteria could be used;

- o if ‘confidential’ appears in the footer its weight is 100 – thus triggering the rule
- o if confidential appears in the body of the email its weight is 35 – thus it requires that ‘confidential’ occurs at least 3 times in the message for the rule to be triggered

### ***Information Concepts***

An information concept is not so much a technique as it is the cumulative effect of an intelligent content analysis capability that performs the kind of functions and analysis described above. The ability to very precisely describe the ‘nature’ of the information you are looking for is far more powerful and user-friendly than any key word/phrase or purely statistical analysis-based approach.

The information concept model:

- 1) achieves significantly higher accuracy for all detection operations

## Effective Content Analysis for Email Control

- 2) dramatically reduces the number of false negatives
- 3) substantially reduces the number of false positives
- 4) provides greater flexibility for users to fine-tune their policies with corresponding improvement in results
- 5) considerably reduces the administrative and support burden associated with key word based solutions

### Other Considerations

In this whitepaper we have focused on describing the need for, and characteristics of, intelligent content analysis for any truly effective email content control solution. We have also explained the differences between scanning for key word search terms and monitoring for 'information concepts'— and the superiority of the latter.

Usability and ease of administration are two additional and important considerations. Because most email monitoring solutions are key word based and support generally limited message actions they express content control policies as simple, individual, flat rules. The criteria (what to look for), conditions (when to do something) and actions (what to do) are all embodied within a rule. Thus, any different combination of, or even slight variation in, the criteria, conditions or desired action(s) requires a new rule be created. It is not uncommon for an organization to have hundreds, even thousands of rules defined over time.

This is a very high maintenance approach, not only in terms of initial rule definition, but also, in terms of ongoing administration. Remember, each different combination of search terms (or conditions or actions!) constitutes a new rule. Now suppose the action, whenever certain content was found, is to copy the message to a specific person (the Compliance Officer). If that individual leaves, or someone else becomes responsible for reviewing the email, then all the rules that contained the action to copy the email to a particular email address have to be individually and manually updated. The simplest of changes could result in hundreds of rules requiring modification – a time-consuming and expensive exercise. Administration of the system can quickly become unmanageable.

A better approach is to provide the ability to define criteria (information concepts), conditions (events that trigger actions) and actions independently of each other. An organization's content policies are represented by combining these three policy elements. This 'policy' (not rule) model offers advantages such as:

- *reusability* — information concept (content) templates and actions can be defined and shared across multiple policies. Often times the difference between two policies may simply be a subtle variation to the concept (information) template. For example, Policy A may be based on looking for particular content defined in template B, while Policy D is monitoring for the same information as Template B and also additional content defined in Template C. Unlike a rules-based model, reusability allows Policy D to simply reuse Template B and add Template C.
- *hierarchical combinations* — concept templates can be stacked where one broader definition is created by combining two or more narrower concepts. This allows for fairly wide definitions that encompass the generic cases that can subsequently be narrowed to deal with exceptions.
- *ease of administration* — using the earlier example, in a true policy-based system an action called 'Copy to Compliance Officer' (for example) would have been defined as 'copy to abc-email-address', this action would have been associated with various policies. Now instead of modifying potentially hundreds of individual rules, you simply modify the single action to 'copy to xzy-email-address' and all the policies that used the 'Copy to Compliance Officer' action automatically inherit the change.

## Effective Content Analysis for Email Control

### Conclusion

Email content control is a major concern for government organizations and companies of every size. An information-centric approach to active email control is crucial to meeting increasingly sophisticated internal and external compliance requirements. Effective content analysis is the cornerstone of successful email monitoring and control operations.

The ability to represent contextual knowledge and information concepts when defining corporate email policies, then detect them within the messages and attachments flowing through the mail network requires flexible, powerful and sophisticated content analysis capabilities. Key word oriented solutions suffer severe limitations. Solutions with truly intelligent content analysis features, like those described in this whitepaper, are superior in terms of overall performance, accuracy and results.

- *'Information-centric', not role-based, policies are key to proper email control*
- *'Context' is crucial to effective content analysis*
- *Monitoring for 'Information concepts', rather than just key words or phrases, is more flexible and results in substantially more accurate detection, reducing both false positive and negative identifications*

SecurExchange, developed by Nemx Software, is the leading active email control solution for Microsoft Exchange Server environments. SecurExchange incorporates one of the most comprehensive and sophisticated intelligent content analysis engines that includes all the features outlined in this whitepaper.

SecurExchange provides organizations of every size effective, comprehensive and accurate monitoring and enforcement of email compliance policies enhancing security, mitigating corporate risk and liability, and safeguarding sensitive or confidential business information. Moreover, unlike most email monitoring products (particularly appliance-based products and managed services) that only monitor inbound and outbound email traffic (a mere 15% of most organizations total traffic!) SecurExchange provides 100% email coverage—monitoring and controlling inbound, outbound and internal email traffic.

To learn more about SecurExchange and how it can contribute to your organization's email compliance efforts visit Nemx at [www.nemx.com](http://www.nemx.com), email us at [info@nemx.com](mailto:info@nemx.com) or call us at (613) 831-2010 x230.

**SecurExchange**  
*Total E-mail Peace of Mind™*