(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 May 2007 (24.05.2007)

PCT

(10) International Publication Number
WO 2007/059428 A2

(51) International Patent Classification:
H04Q 7/22 (2006.01)

(21) International Application Number:
PCT/US2006/060771

(22) International Filing Date:
10 November 2006 (10.11.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/736,121	10 November 2005 (10.11.2005)	US
11/383,347	15 May 2006 (15.05.2006)	US

(71) Applicant *(for all designated States except US)*: **SECURE COMPUTING CORPORATION** [US/US]; 2340 Energy Park Drive, St. Paul, MN 55108 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **JUDGE, Paul** [US/US]; 10090 Jones Bridge Road, Unit 3, Alpharetta, GA 30022 (US). **SCHNECK, Phyllis, Adele** [US/US]; 3221 Gables Way, Atlanta, GA 30329 (US). **YANG, Weilai** [CN/US]; 1185 South Bethany Creek Drive, Alpharetta, GA 30004 (US). **ZDZIARSKI, Jonathan, Alexander** [US/US]; 3069 Heritage Road, Milledgeville, GA 31061 (US).

(74) Agents: **VAN AACKEN, Troy, A.** et al.; Fish & Richardson P.C., P.O. BOX 1022, 3300 Dain Rauscher Plaza, 60 South Sixth St., Minneapolis, MN 55440-1022 (US).

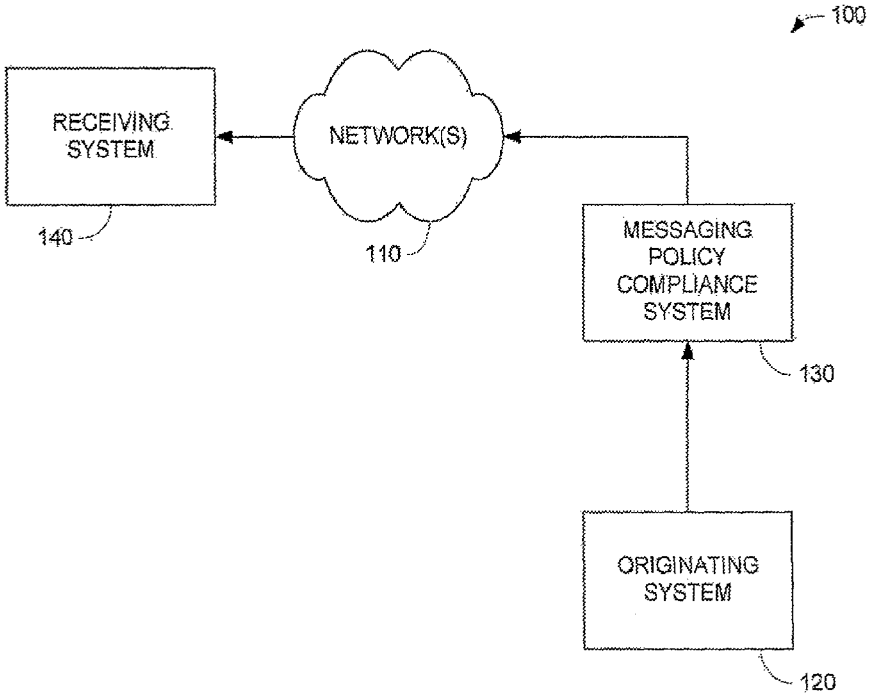
(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) **Title:** CONTENT-BASED POLICY COMPLIANCE SYSTEMS AND METHODS



(57) **Abstract:** Methods and systems for operation upon one. or -more data processors to filter communications of users in accordance with content-based policy.

CONTENT-BASED POLICY COMPLIANCE SYSTEMS AND METHODS

BACKGROUND AND SUMMARY

This document relates generally to systems and methods for processing communications and more particularly to systems and methods for filtering communications.

In the electronic mail filtering industry, most existing systems are aimed at filtering incoming messages. Content policy compliance (e.g., compliance with corporate or governmental policy) can be an important consideration for companies in view of the increasingly electronic character of important communications and availability of a variety of electronic communication techniques.

In accordance with the teachings disclosed herein, methods and systems are provided for operation upon one or more data processors to filter communications in accordance with content based policy compliance. For example, a method and system can include: defining a classification associated with the content of a class of files; receiving a set of characteristics distinctive to the classification; wherein the set of characteristics has been derived based upon the set of files; receiving a rule defining the treatment of content substantially similar to the set of characteristics; and, wherein the rule defines whether to forward a communication to a recipient based upon the classification of the content and at least one of the recipient or the sender.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram depicting a system for handling transmissions being sent over a network.

FIG. 2 is a block diagram depicting a compliance system that has been configured for classifying files and applying policies.

FIG. 3 is a block diagram depicting a compliance system operating on a local computer.

FIG. 4 is a block diagram depicting a compliance system that has been configured for classifying files based upon a combination of analysis techniques.

FIG. 5 is a block diagram depicting a compliance system that has been configured to use existing content to develop characteristics associated with a class.

FIG. 6 is a block diagram depicting a compliance system that has been configured to accept one or more content based policies from an administrator.

FIG. 7 is a flowchart depicting an operational scenario for allowing content based policy(ies).

FIG. 8 is a flowchart depicting an operational scenario for allowing content based policy(ies) whereby the characteristics of the content are automatically generated based upon a set of related files associated with a classification.

FIG. 9 is a flowchart depicting an operational scenario for generating content based policy compliance using access control rights to generate policy.

FIG. 10 is a flowchart depicting an operational scenario for filtering messages based upon content based policy(ies).

FIG. 11 is a flowchart depicting an operational scenario for converting communications from multiple formats and/or languages into a common format and/or language and distilling the communication into metadata describing the communication prior to parsing the message for any content match.

FIG. 12 is a block diagram depicting a server access architecture.

FIG. 13 is a block diagram depicting another server access architecture.

DETAILED DESCRIPTION

FIG. 1 depicts at **100** a system for handling transmissions received over a network **110**. The transmissions can be many different types of communications, such as electronic mail (e-mail) messages sent from one or more messaging entities **120**. The system **100** uses a messaging policy compliance system **130** to help process the communications from an originating system **120**. The messaging policy compliance system **130** examines characteristics associated with the communications from the originating system **120**, and based upon the examination, an action is taken with respect to the communications. For example, a communication may be determined to be legitimate and thus the communication should not be filtered by the messaging policy compliance system **130** and instead provided to a receiving system **140** for delivery to the intended recipient.

This disclosure relates to filtering the content of packets communicated over the network based upon a classification associated with the communication. It should therefore be understood that such communications can include e-mail, voice over internet protocol (VoIP) packets, instant messaging (IM), file transfer protocol (FTP) packets, hypertext transfer protocol (HTTP) packets, Gopher packets, and any other method whereby protected or sensitive content (e.g., trade secrets, privileged information, etc.) can be transferred over a network to another user.

It should be further understood that an organization often includes several departments which operate, to some degree, independently of one another. However, an organization may wish to prevent theft or disclosure of information based upon the person accessing the information, or based upon the person to whom the information is being sent. For example, an organization may not want engineering content disclosed to persons outside of the organization. Thus, the organization desires to limit the recipients of messages that include content related to engineering. However, traditional filtering systems do not provide an accurate classification of content being sent to/from users within an organization.

To increase the accuracy of classifying content associated with messages properly (e.g., engineering files, marketing files, legal files, etc., including text documents, voice recordings, images, drawings, among others), a messaging policy compliance system **200** can be configured with a message content classification program **210** as shown in FIG. **2**. A message content classification program **210** can use one or more message classification techniques or filters to classify message content.

The message content classification program **210** analyzes the content of a communication (intended to travel across the network to a recipient) in order to classify the content of the communication. However, it should be understood that the messaging policy compliance system can also inspect incoming communications before distributing the communications to a receiving system. The messaging policy compliance system **200** compares at **220** the classification produced by the message content classification program **210** to a set of one or more rules to determine whether the message is in compliance with policy.

If the communication is in compliance with the organization's policies, the communication is forwarded to a recipient via the network **230**. However, if the communication is not in compliance with the organization's policies, the communication is quarantined, dropped, notify an administrator or a sender/recipient, or take some other action, as illustrated by block **240**. Other actions can include, for example, stripping content and/or an attachment of the message before forwarding, automatically encrypting the message before forwarding, delay delivery of the message, or other appropriate actions in response to a compliance violation. It should be understood that automatic encryption can include requesting a user or administrator's approval to encrypt. Moreover, automatic encryption can further include completely automating the decision to encrypt content at the server or client level, in accordance with policy and without user or administrator approval.

If only a portion of the communication is not in compliance with the organization's policies (e.g., a message contains two attachments where one complies with the policy(ies) and the other does not), the communication may be blocked (e.g., dropped, quarantined, etc.). Alternatively, such a communication could be automatically redacted by the messaging policy compliance system **200**, such that it complies with the organization's policy(ies). Moreover, in the event that a message cannot be transferred to a recipient because of a policy violation, a notification can be delivered to the originating system. It should be noted that the originating system can notify a system administrator. Alternatively, a system administrator can be notified directly by the messaging policy compliance system. It should be understood that there are numerous ways to manage a response to policy violations, each of which is intended to be included within the scope of this disclosure.

Another example of a messaging policy compliance system is shown in FIG. 3. For example, the messaging policy compliance agent **260** is located on a user's computer **265**. In addition to the messaging policy compliance agent **260**, the local computer **265** can include an communication client **270**. It should be understood that the communication client **270** could be integrated with the messaging policy compliance agent **260**, in some examples.

Upon receiving a message from the communication client **270**, the messaging policy compliance agent **260** would use the message content classification program **275** to determine a classification associated with the content of the communication. The messaging policy compliance agent **260** at **220** compares the classification associated with the content of the communication with a content-based messaging policy which could be set by the user, or by a system administrator. Where the communication does not comply with a content-based messaging policy, the agent can drop the communication, quarantine the communication, etc. as shown by block **285**. It should be understood that such functionality could be integrated with the communication client. However, it should also be noted that the functionality of block **285** could be provided by the agent itself.

If the communication complies with the content-based policy(ies), the messaging policy compliance agent forwards the message to the originating system **290**. It should be understood that the functionality of the originating system **290** could be included on the local computer **265** itself. The originating system **290** then sends the message to a recipient system via network(s) **295**.

It should be noted that the messaging policy compliance agent can be used in conjunction with a messaging policy compliance server. Using such an architecture could provide multiple levels of content compliance checks. The agent/server architecture could allow the messaging policy compliance agent to record the user's activity and/or various events that occur on the computer (e.g., policy violations), and periodically provide updates of the user's activity to a messaging policy compliance server. The agent/server architecture could further allow the messaging policy server to periodically communicate updated content-based policy(ies) to the agent. It should be further noted that a messaging policy compliance agent **260**, where practicable, can include any of the functionality of a messaging policy compliance system as described in the present disclosure. As such, any of the functionality described with respect to a messaging policy compliance system can be used on a messaging policy compliance agent in accordance with the present disclosure.

The messaging policy compliance agent could further allow a user to request the addition of content-based policy(ies) at the local or server level. Where the requested content-based policy(ies) do not conflict with administrator content-based policy(ies), the local and/or server could apply the user requested content-based policy. Further, the messaging policy compliance agent could allow the user to request encryption on a communication via the communication client interface. Where the encryption request complies with content-based policy(ies) at the agent and/or server level, the requested encryption can be performed by either the server or the agent.

By way of example, a message content classification program **310**, as shown in FIG. **4**, can include a number of classification techniques **360**, **370**, **380**. Example message content classification techniques or filters **360**, **370**, **380** that a message content classification program **310** can use include:

- ***Contextual Analysis*** – a classification technique that performs a Markovian analysis of files to identify phrases and words which are unique to a classification of file, which can be done by analyzing the rarity of a word or phrase to a particular type of file, and treating such words or phrases as indicative of a group of files with some percentage of certainty.

- ***Fingerprint Analysis*** – a technique to identify copying between two electronic texts at multiple levels (e.g. whole file, paragraph, sentence, or unstructured alphanumeric components) by, for example: 1) Applying a normalization layer to remove whitespace and other noise; and, 2) utilizing a winnowing algorithm to generate a minimized, yet optimal number of hashes for each file, adding an ambiguity factor to identify files with very minimal, but significant duplications of data.

- ***Cluster Analysis*** – a classification technique that partitions the data into related subsets sharing a common trait that can be defined as a function of a defined distance measure (e.g., Euclidian distance) that marks a point as a part of at least one cluster.

- ***Adaptive Lexical Analysis*** - a classification technique which can be performed on electronic text or data which adaptively learns structures of sparse and nonsparse patterns by, for example: 1) Instantiating a series of Markov chains using components of the presented classification medium as members; and, 2) Applying a series of weights based on the complexity of the chain, factored with the learned appearance vectors of each chain to deduce a probability. This process allows for the learning and identification of sparse patterns, exact phrases, words, or binary patterns which have a probability of one disposition based on their historical occurrence across a continually building corpus, using the original medium as a process of continuing self-calibration.

It should be understood that these analysis techniques can be modified (sometimes significantly) based upon the desired results, and that all implementations of these analysis techniques are intended to be included within the present disclosure. For example, the cluster analysis filter **380** can use a number of different algorithms to identify clusters, such available techniques can include, for example, but not limited to: *k*-means clustering, quality threshold (QT) clustering, fuzzy *c*-means clustering, and spectral clustering, among others.

Thus, it should be recognized that using a combination of classification algorithms on the content passing through the messaging policy compliance system **300** can provide a classification **390** associated with the content, and make a determination, as shown by decision block **320**, whether the content of the message complies with content policy. Where the content complies with content policy the message is forwarded to a recipient system via a network **330**. Where the content does not comply with content policy, the content can be dropped, quarantined, etc. as shown by block **340**. Where the message is not forwarded to the recipient system, the messaging policy compliance system **300** can notify a sender, an originating system **350** and/or an administrator (not shown).

As an example, a policy could limit engineering-type information from being transmitted by human resources staff or to individuals outside of the company. The message content classification could identify engineering-type information, for example, by the inclusion of equations or words or phrases that are most commonly associated with engineering documents, presentations or drawings – and/or by one of the classification techniques previously listed in this application. Communications identified as including characteristics in common with engineering documents, presentations or drawings would be tested by examining a message header to determine whether the sender was a human resources employee, or whether the recipient domain was not associated with the company.

It should be understood that using this method, an administrator could identify an individual suspected of leaking information. This individual could be monitored for compliance with company policy. Moreover, the message content classification program **310** can detect situations where the user is sending subsets of files, or where the individual is summarizing or rewording documents to avoid detection.

As shown in FIG. 5, a message compliance system **400** could be configured to examine an existing set of related files **492**, as specified by an administrator **494**, to create identification characteristics associated with the set of related files **492**. The files **492** could be supplied to the message content classification program **410**. The message content classification program **410** could use each of the techniques **460**, **470**, **480** on the set of related files to determine what features or characterizations mark their relationship. For example, legal documents might often include Latin phrases such as *in re*, *ipso facto*, or *prima facie*. Such an inclusion could be discoverable by a message content classification program **410**.

A message content classification program **410** can generate a set of identifying characteristics for a class of content. The same techniques **460**, **470**, **480** are then used on communications entering the messaging policy compliance system **400**. The characteristics of the communication may then be compared to the identifying characteristics for a class of content to determine in which class (if any) the content of the communication belongs, thereby producing a content classification **490** for the communication. The messaging policy compliance system **400** then applies any policies related to the content classification, as shown by decision block **420**, to determine whether the communication will be delivered via network **430**, or dropped, quarantined, etc. as shown by block **440**. In the event that a communication does not satisfy policy, the originating system **450** can be alerted to the policy failure. The messaging content compliance system could also notify a system administrator and/or sender of the policy failure.

It should be recognized that content policy can be created in a myriad of ways. For example, as shown in FIG. 6, the messaging policy compliance system can accept content based policies **596** from a system administrator **594**. The administrator **594** can supply a content policy by supplying both the related content **592** for the message content classification program **510**, and supplying a set of policy rules **596** configured to be parsed by a policy compliance decision block **520**.

It should also be recognized that a messaging policy compliance system can be set up to inspect access control rights of users authorized to access a set of related files. These access control rights can be used to automatically analyze content-based policy, where the users (who are authorized) view and/or modify the set of related files also have the ability to send and/or receive such similar content as they are allowed to access.

Furthermore, it should be recognized that a messaging policy compliance system can be trained for recognizing content-based anomalous behavior associated with the users of the system. For example, a messaging policy compliance system can observe all communications sent through the system over a period of time. Then, upon detecting that a user is sending communications that include content that is abnormal with respect to the historical usage patterns of that user, the messaging policy compliance system can be configured to drop/quarantine the communication and/or notify a system administrator. In an adaptive manner, a messaging policy compliance system can generate content-based policy(ies) based upon historical usage of content.

FIG. 7 depicts a flowchart illustrating an operational scenario **600** for a messaging policy compliance system, whereby a system administrator can define content-based policy. At step **610**, a system administrator creates a classification of content. For example, classifications could include, engineering content, medical records content, human resources content, legal content, marketing content, accounting content, forecasting content, etc.

A messaging policy compliance system could then receive a set of characteristics associated with the created classification, as shown at step **620**. It should be noted that these characteristics could be internally generated, or received from another system. At step **630**, the operation scenario allows an administrator to define a rule or policy for communications that include content that matches the characteristics associated with the created classification, whereby a message filtering system could be configured to block messages that do not comply with the defined rule/policy.

FIG. 8 depicts a flowchart illustrating an operational scenario **700** for a messaging policy compliance system, whereby a system administrator can define content-based policy for communications by supplying a set of related files. At step **710**, the messaging policy compliance system receives a new classification from the administrator. At step **720**, the system administrator provides a set of related files which exemplify the new classification. The messaging policy compliance system generates a set of characteristics associated with the set of related files, as shown by step **730**. At step **740**, the messaging policy compliance system receives a rule for communications identified as belonging to the new classification.

Another example of an operational scenario **800** for a messaging policy compliance system is shown in FIG. **9**. At step **810**, an administrator provides a new classification to the messaging policy compliance system. At step **820**, the administrator provides a set of related files which correspond to the new classification provided at step **810**. The messaging policy compliance system then generates a set of characteristics that distinguish the set of related files from other types/classes of files, as shown by step **830**. The messaging policy compliance system then examines the access control rights of each of the related files in order to develop content-based policy, thereby allowing users with access to the set of related files to send content which shares distinguishing characteristics with the related files.

A messaging policy compliance system can filter messages, for example, as shown by the operational scenario **900** in FIG. **10**. At step **910**, a communication is received. At step **910**, the content of the communication is compared to existing classifications. This is done, for example, by using one or more techniques that attempt to match elements of the content to sets of characteristics associated with the existing classifications. At decision block **930**, the messaging policy compliance system determines whether a threshold match has been made to identify the communication content as being related to the existing classifications.

Where the messaging policy compliance system is unable to discover a threshold match between the content and the existing classifications, the communication is determined to contain no protected content as shown by step **940**. Communications which contain no protected content can be forwarded to the recipient(s), as shown by step **950**.

However, where the messaging policy compliance system determines there is a threshold match between the content of the communication and the existing classifications, the communication is examined to determine if content-based policy is satisfied, as shown by decision block **960**. Where the content-based policy is not satisfied, the communication is quarantined, dropped, or otherwise blocked by the system, as shown in step 970. Where the content-based policy is satisfied, the communication is forwarded to the one or more systems associated with the intended recipient(s).

FIG. 11 depicts an alternative operational scenario 980 used to parse communications prior to forwarding the message to a recipient. At step 982, a communication is received. At step 984, the communication is normalized. Normalization in various examples, can include converting the communication to a common protocol. For example, where the system receives a VoIP packet, the communication could be converted to another format (e.g., a text based format) for examination. It should be understood that communications in any format can be converted to any other format for parsing, and that the present disclosure is not limited to converting all varied protocols to any particular protocol, but that the choice of a common comparison protocol is merely a design choice to be made in light of the circumstances of a particular solution (e.g., where the primary communication mechanism is VoIP, the common comparison protocol may be chosen to be VoIP to reduce the resources used for protocol translation).

In various examples, normalization can also include translating a communication from a variety of languages into a common comparison language. For example, where a communication is in German, comparison techniques would not detect a classification match where the classification has been defined by English language documents. Thus, for a more complete analysis of all communications, communications can be translated to a common comparison language. It should be understood that this disclosure is not limited to a particular common comparison language. Moreover, it should be understood that the common comparison language may not even be a practiced language, but may merely be a language that is created by a user which has special characteristics that aid in classification of the communication. Further, the common comparison language in various examples may include a combination of several different languages, such as where discrete concepts used in different languages are not adequately described by a single language.

In step 986, the operational scenario 980 generates metadata related to the communication. The metadata can distill the files into identifying characteristics and reduce superfluous language which may not be helpful in associating the communication with any of the classifications. For example, definite and indefinite articles, pronouns, and various other linguistic devices are often irrelevant to classification of a file. At step **988**, the metadata associated with the communication is compared to existing metadata triggers to determine a classification associated with the communication. At decision block **990**, the messaging policy compliance system determines whether a threshold match has been made to identify the communication metadata as being related to the existing classification metadata.

Where the messaging policy compliance system is unable to discover a threshold match between the content and the existing classifications, the communication is determined to contain no protected content as shown by step **992**. Communications which contain no protected content can be forwarded to the recipient(s), as shown by step **994**.

However, where the messaging policy compliance system determines there is a threshold match between the content of the communication and the existing classifications, the communication is examined to determine if content-based policy is satisfied, as shown by decision block **996**. Where the content-based policy is not satisfied, the communication is quarantined, dropped, or otherwise blocked or delayed by the system, as shown in step **998**. Where the content-based policy is satisfied, the communication is forwarded to the one or more systems associated with the intended recipient(s).

The systems and methods disclosed herein are presented only by way of example and are not meant to limit the scope of the invention. Other variations of the systems and methods described above will be apparent to those skilled in the art and as such are considered to be within the scope of the invention. For example, a system and method can be configured to handle many different types of communications, such as legitimate messages or unwanted communications or communications violative of a pre-selected policy. As an illustration, a communication could include a type of content as recognized by the system, and a policy could include a corporate communication policy, a messaging policy, a legislation or regulatory policy, or an international communication policy.

As an example of an architecture the could be used in accordance with systems and methods disclosed herein, an originating system **1000**, a receiving system **1010**, and a messaging policy compliance system **1020** can each be connected via one or more networks, as shown by FIG. **12**. The originating system **1000** can send a communication to the receiving system **1010** via the messaging policy compliance system and network(s) **1030**. The messaging policy compliance system **1030** would then be operable to forward the message to the receiving system **1010** via network(s). It should be understood that network(s) **1030** can include many subnets including but not limited to wireless networks, local area networks, wide area networks, metropolitan area networks, corporate intranets, and combinations thereof.

It should also be noted that originating system **1000** and/or receiving system **1010** can include an electronic mail server and/or client, an instant messaging server and/or client, a voice over internet protocol (VoIP) server and/or client, a gopher server and/or client, a file transfer protocol (FTP) server and/or client, a hypertext transfer protocol (HTTP) server and/or client, and combinations thereof, among many other existing network communications protocols.

As another example of the wide scope and variations of systems and methods disclosed herein, the systems and methods may be implemented on various types of computer architectures, such as for example on different types of networked environments. As an illustration, FIG. **13** depicts a server access architecture within which the disclosed systems and methods may be used (e.g., as shown at **1100** in FIG. **8**). The architecture in this example includes a corporation's local network **1190** and a variety of computer systems residing within the local network **1190**. These systems can include application servers **1120** such as Web servers and e-mail servers, user workstations running local clients **1130** such as e-mail readers and Web browsers, and data storage devices **1110** such as databases and network connected disks. These systems communicate with each other via a local communication network such as Ethernet **1150**. Firewall system **1140** resides between the local communication network and Internet **1160**. Connected to the Internet **1160** are a host of external servers **1170** and external clients **1180**. It should be understood that the present disclosure can any variety of network, including, but not limited to an intranet, wireless network, wide area networks, local area networks, and combinations thereof, in order to facilitate communication between components.

Local clients **1130** can access application servers **1120** and shared data storage **1110** via the local communication network. External clients **1180** can access external application servers **1170** via the Internet **1160**. In instances where a local server **1120** or a local client **1130** requires access to an external server **1170** or where an external client **1180** or an external server **1170** requires access to a local server **1120**, electronic communications in the appropriate protocol for a given application server flow through "always open" ports of firewall system **1140**.

A system **1100** as disclosed herein may be located in a hardware device or on one or more servers connected to the local communication network such as Ethernet **1180** and logically interposed between the firewall system **1140** and the local servers **1120** and clients **1130**. Application-related electronic communications attempting to enter or leave the local communications network through the firewall system **1140** are routed to the system **1100**.

System **1100** could be used to handle many different types of e-mail and its variety of protocols that are used for e-mail transmission, delivery and processing including SMTP and POP3. These protocols refer, respectively, to standards for communicating e-mail messages between servers and for server-client communication related to e-mail messages. These protocols are defined respectively in particular RFC's (Request for Comments) promulgated by the IETF (Internet Engineering Task Force). The SMTP protocol is defined in RFC **1221**, and the POP3 protocol is defined in RFC **1939**.

Since the inception of these standards, various needs have evolved in the field of e-mail leading to the development of further standards including enhancements or additional protocols. For instance, various enhancements have evolved to the SMTP standards leading to the evolution of extended SMTP. Examples of extensions may be seen in **(1)** RFC **1869** that defines a framework for extending the SMTP service by defining a means whereby a server SMTP can inform a client SMTP as to the service extensions it supports and in **(2)** RFC **1891** that defines an extension to the SMTP service, which allows an SMTP client to specify (a) that delivery status notifications (DSNs) should be generated under certain conditions, (b) whether such notifications should return the contents of the message, and (c) additional information, to be returned with a DSN, that allows the sender to identify both the recipient(s) for which the DSN was issued, and the transaction in which the original message was sent.

In addition, the IMAP protocol has evolved as an alternative to POP3 that supports more advanced interactions between e-mail servers and clients. This protocol is described in RFC 2060.

Other communication mechanisms are also widely used over networks. These communication mechanisms include, but are not limited to, Voice Over IP (VoIP) and Instant Messaging. VoIP is used in IP telephony to provide a set of facilities for managing the delivery of voice information using the Internet Protocol (IP). Instant Messaging is a type of communication involving a client which hooks up to an instant messaging service that delivers communications (e.g., conversations) in realtime.

It is further noted that the systems and methods disclosed herein may use data signals conveyed via networks (e.g., local area network, wide area network, internet, etc.), fiber optic medium, carrier waves, wireless networks, etc. for communication with one or more data processing devices. The data signals can carry any or all of the data disclosed herein that is provided to or from a device.

Additionally, methods and systems described herein may be implemented on many different types of processing devices by program code comprising program instructions that are executable by one or more processors. The software program instructions may include source code, object code, machine code, or any other stored data that is operable to cause a processing system to perform methods described herein.

The systems' and methods' data (e.g., associations, mappings, etc.) may be stored and implemented in one or more different types of computer-implemented ways, such as different types of storage devices and programming constructs (e.g., data stores, RAM, ROM, Flash memory, flat files, databases, programming data structures, programming variables, IF-THEN (or similar type) statement constructs, etc.). It is noted that data structures describe formats for use in organizing and storing data in databases, programs, memory, or other computer-readable media for use by a computer program.

The systems and methods may be provided on many different types of computer-readable media including computer storage mechanisms (e.g., CD-ROM, diskette, RAM, flash memory, computer's hard drive, etc.) that contain instructions for use in execution by a processor to perform the methods' operations and implement the systems described herein.

The computer components, software modules, functions and data structures described herein may be connected directly or indirectly to each other in order to allow the flow of data needed for their operations. It is also noted that software instructions or a module can be implemented for example as a subroutine unit of code, or as a software function unit of code, or as an object (as in an object-oriented paradigm), or as an applet, or in a computer script language, or as another type of computer code or firmware. The software components and/or functionality may be located on a single device or distributed across multiple devices depending upon the situation at hand.

It should be understood that as used in the description herein and throughout the claims that follow, the meaning of "a," "an," and "the" includes plural reference unless the context clearly dictates otherwise. Also, as used in the description herein and throughout the claims that follow, the meaning of "in" includes "in" and "on" unless the context clearly dictates otherwise. Finally, as used in the description herein and throughout the claims that follow, the meanings of "and" and "or" include both the conjunctive and disjunctive and may be used interchangeably unless the context clearly dictates otherwise; the phrase "exclusive or" may be used to indicate situation where only the disjunctive meaning may apply.

What is claimed is:

1. A method for operation upon one or more data processors to filter communications based upon content based policy compliance, comprising:

defining a classification to be associated with a set of files, the classification generally classifying content of the set of documents as being associated with one of a plurality of business activities;

receiving a set of known identifying characteristics with respect to the classification;

wherein the set of known identifying characteristics has been derived based upon the set of files;

receiving a rule specifying treatment of content identified as associated with the set of known identifying characteristics;

wherein the rule defines whether to forward a communication including content to a recipient based upon the classification of the content and at least one of the recipient and the sender.

2. The method of claim 1, wherein the received set of known identifying characteristics associated with the set of files are generated by examining the content of the set of files for triggering characteristics included in the files.

3. The method of claim 2, wherein the examination of the files includes at least one of a fingerprinting analysis, a cluster analysis, a contextual analysis, and an adaptive lexical analysis.

4. The method of claim 1, wherein the rule is generated according to access control rights associated with at least one of the sender or the recipient, wherein the access control rights are derived from access control rights associated with the set of files.

5. The method of claim 1, wherein the rule is generated based upon access control rights associated with at least one of the sender or the recipient, content contained in a communication, usage of the content contained in a communication, or combinations thereof.

6. The method of claim 5, wherein access control rights are provided to groups of users, wherein the sender and the recipient are included in at least one group of users.

7. The method of claim 1, further comprising the steps of:

receiving a communication from an originating system;

extracting identifying characteristics associated with the communication;

comparing the communication's identifying characteristics with the set of known identifying characteristic to identify a determined classification to be associated with the communication;

applying a rule associated with the determined classification, the rule specifying whether the communication should be forwarded to its intended recipient.

8. The method of claim 7, further comprising the step of forwarding the communication based upon application of the rule.

9. The method of claim 7, further comprising the step of forwarding the communication responsive to the communication not being associated with any known identifying characteristics.

10. The method of claim 7, further comprising the step of automatically generating a rule responsive to learning communications patterns associated with groups of users and based on types of content being distributed among the groups of users.

11. The method of claim 7, further comprising delaying delivery of the communication based upon application of the rule.

12. The method of claim 11, further comprising notifying an originator of the communication responsive to delivery of the communication being delayed.

13. The method of claim 12, wherein the delay is at least one of storing the communication in a quarantine folder, dropping the communication, temporarily rejecting the communication, storing the communication until approval is received from an administrator to forward the communication, automatically encrypting the communication, notifying an administrator, notifying a recipient, or combinations thereof.

14. The method of claim 7, further comprising converting a communication from one of a plurality of mismatched formats to a comparison format prior to extracting identifying characteristics from the communication.

15. The method of claim 7, further comprising translating a file or communication into a common language or format prior to extracting identifying characteristics or generating the known identifying characteristics.

16. The method of claim 15, wherein the translating step creates metadata to be used in extracting identifying characteristics.

17. The method of claim 7, further comprising the steps of:

observing communication traffic;

identifying one or more patterns exhibited by observed communication traffic;

generating a rule based upon the identified one or more patterns.

18. The method of claim 17, wherein a communication falling outside of one or more identified traffic patterns is sent to a quarantine folder, dropped, temporarily rejected, stored until approval is received from an administrator to forward the communication, automatically encrypted, a recipient is notified, a sender is notified, or combinations thereof.

19. The method of claim 1, wherein the defined classifications are at least one of: management files, legal files, technical files, marketing files, financial files, information technology files, proprietary files, strategy files, sensitive files, or government classified files.

20. The method of claim 1, wherein a system administrator specifies the rule by selecting a classification of files for application of the rule, selecting a class of users who are permitted to send the selected classification of files, and selecting a class of users who are permitted to receive the selected classification of files.

21. The method of claim 1, wherein the steps of selecting a class of users comprises selecting individual users who have permission to send or receive the selected classification of files.

22. A content-based policy compliance system configured to filter messages based upon content and at least one of senders or recipients associated with the messages, the system comprising:

a messaging content classifier configured to receive a message and classify the message as associated with at least one of a plurality of content classifications based upon the content of the message and upon known identifying characteristics of the plurality of content classifications;

a messaging filter configured to receive the at least one content classification from the messaging content classifier and to apply a rule to the message based upon the at least one content classification and upon at least one of a sender or recipient of the message; and

forwarding logic configured to transmit the message responsive to output from the messaging filter.

23. The system of claim 22, wherein the messaging content classifier is further configured to receive a plurality of files associated with a specified class, and extract any identifying characteristics from the messages to generate the known identifying characteristics and associate the known identifying characteristics with the specified class of the plurality of content classifications.

24. The system of claim 23, further comprising a user interface configured to receive the plurality of files and the specified class from a user and provide the plurality of files and the specified class to the messaging content classifier.

25. The system of claim 24, wherein the user interface is further configured to allow a user to specify rules for the messaging filter.

26. The system of claim 25, wherein the rules specify which classes of individuals are permitted to send and receive a specified content class associated with the rule.

27. The system of claim 26, wherein the messaging content classifier is configured to use one or more of the following identification techniques to identify commonalities between the plurality of files as well as to classify messages with one of the plurality of classes: a fingerprinting analysis, a cluster analysis, a contextual analysis, and an adaptive lexical analysis.

28. The system of claim 22, wherein the forwarding logic is operable to forward the message to a recipient, quarantine the message, drop the message, or encrypt the message before forwarding the message to a recipient.

29. The system of claim 22, wherein the system is a messaging client, wherein the messaging client periodically receives updates from a messaging server comprising at least one of update rules, updated content classifications, or updated identifying characteristics for the content classifications.

30. The system of claim 22, wherein the message comprises an e-mail communication, an instant messaging communication, an HTTP communication, an FTP communication, a WAIS communication, a telnet communication, a Gopher communication, or a voice over internet protocol communication.

31. Computer readable storage media storing instructions that upon execution by a system processor cause the system processor to filter communications transmitted over a communication network based upon the content of a communication and upon the sender and recipient(s) of the communication, the media having stored instruction that cause the system processor to perform the steps comprising:

receiving sets of known identifying characteristics with respect to a plurality of content classifications from a message content classifier;

wherein the set of known identifying characteristics for the plurality of content classifications has been derived based upon sets of files provided to the message content classifier;

receiving a set of rules specifying treatment of content identified as associated with the sets of known identifying characteristics from a system administrator;

wherein the rules define whether to forward a communication to a recipient based upon the classification of the communication content and at least one of the recipient and the sender;

receiving a communication from a user, the communication containing an originating address, a receiving address or plurality of receiving addresses and content;

determining whether the content of the communication substantially matches any of the plurality of content classifications based upon the sets of known identifying characteristics associated with the content classifications, respectively;

forwarding the communication responsive to the rule associated with the substantially matched content classification, wherein the rule specifies an action to perform on the communication based upon the content classification of the communication and upon the originating address and the receiving address of the communication.

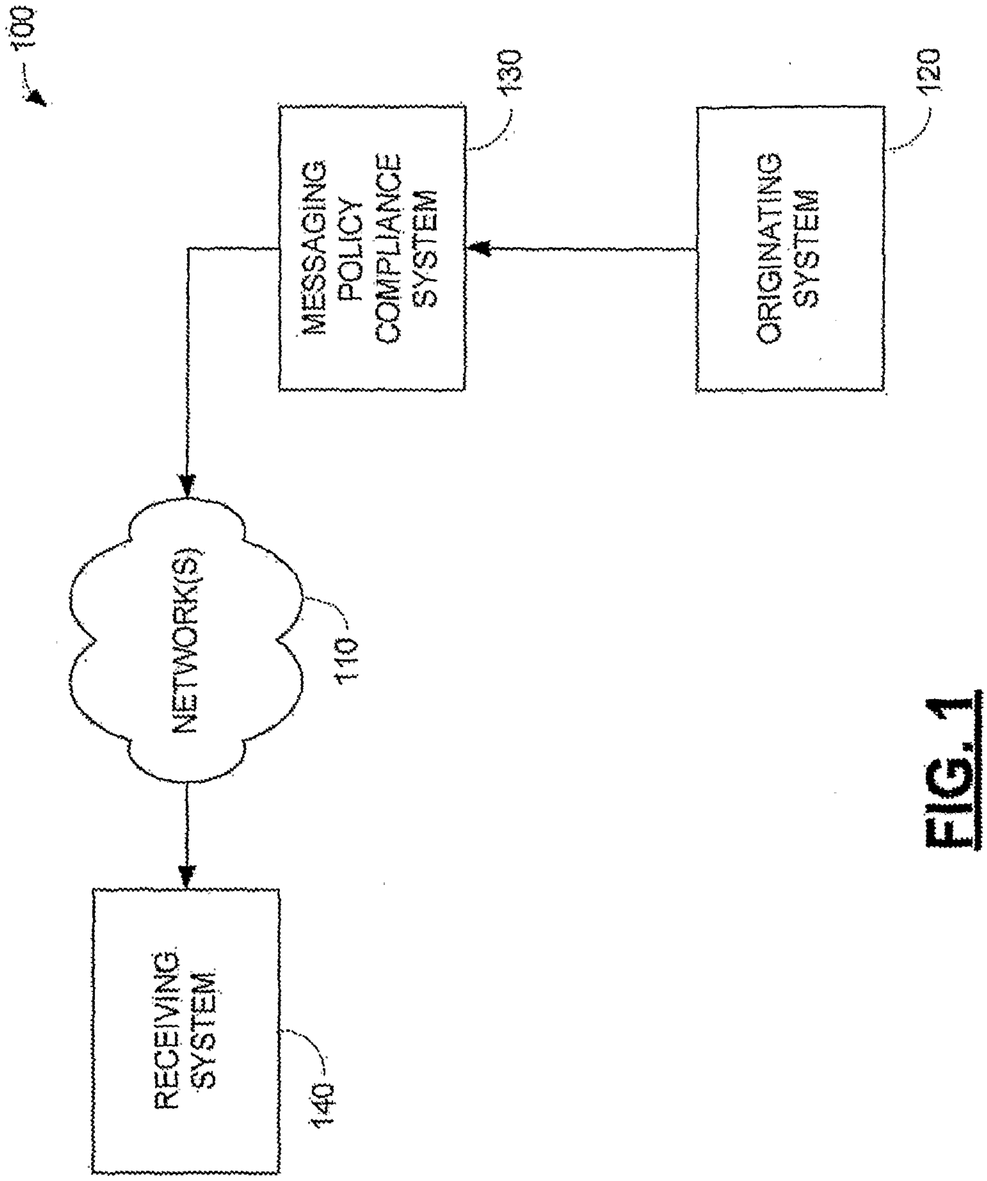

100
MESSAGING POLICY COMPLIANCE SYSTEM
130
ORIGINATING SYSTEM
120
NETWORK(S)
110
RECEIVING SYSTEM
140

FIG. 1

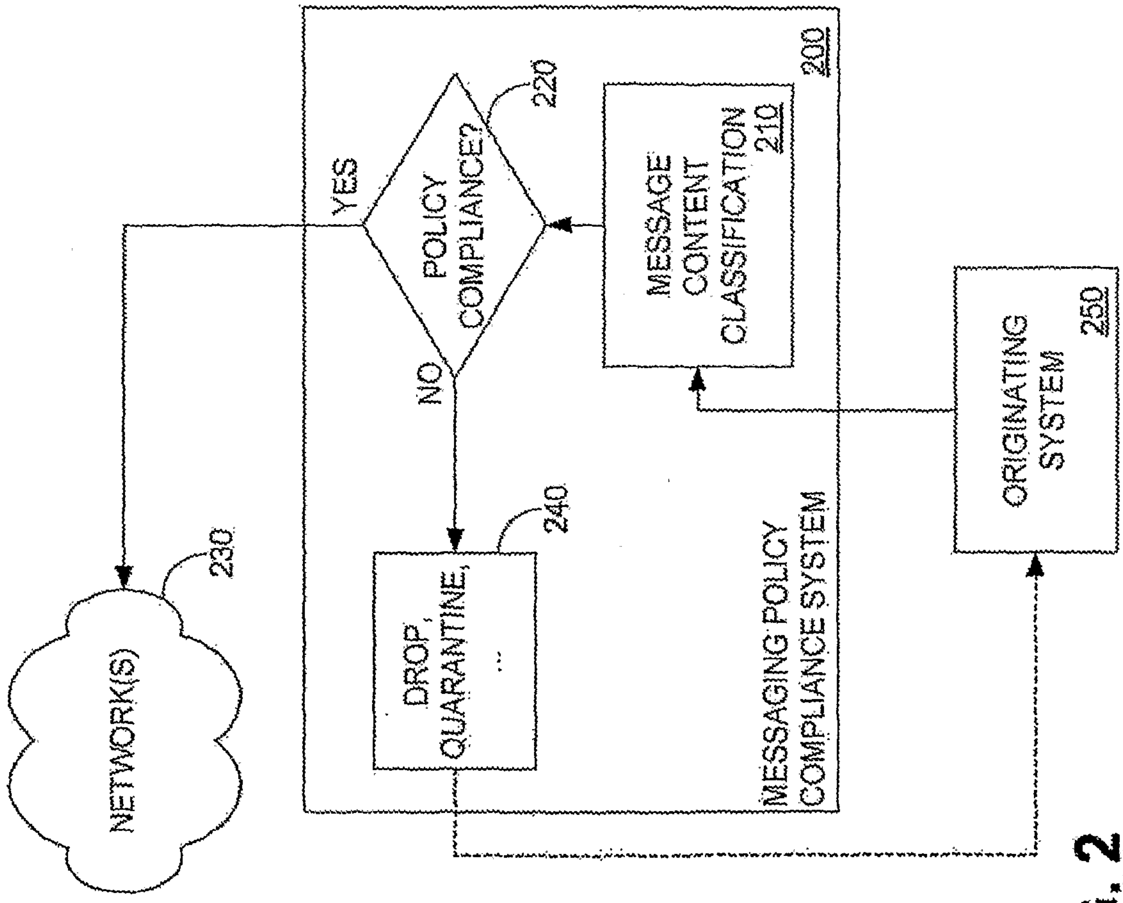

NETWORK(S)
230
DROP,
QUARANTINE,
...
240
POLICY
COMPLIANCE?
220
YES
NO
MESSAGE
CONTENT
CLASSIFICATION
210
MESSAGING POLICY
COMPLIANCE SYSTEM
200
ORIGINATING
SYSTEM
250

FIG. 2

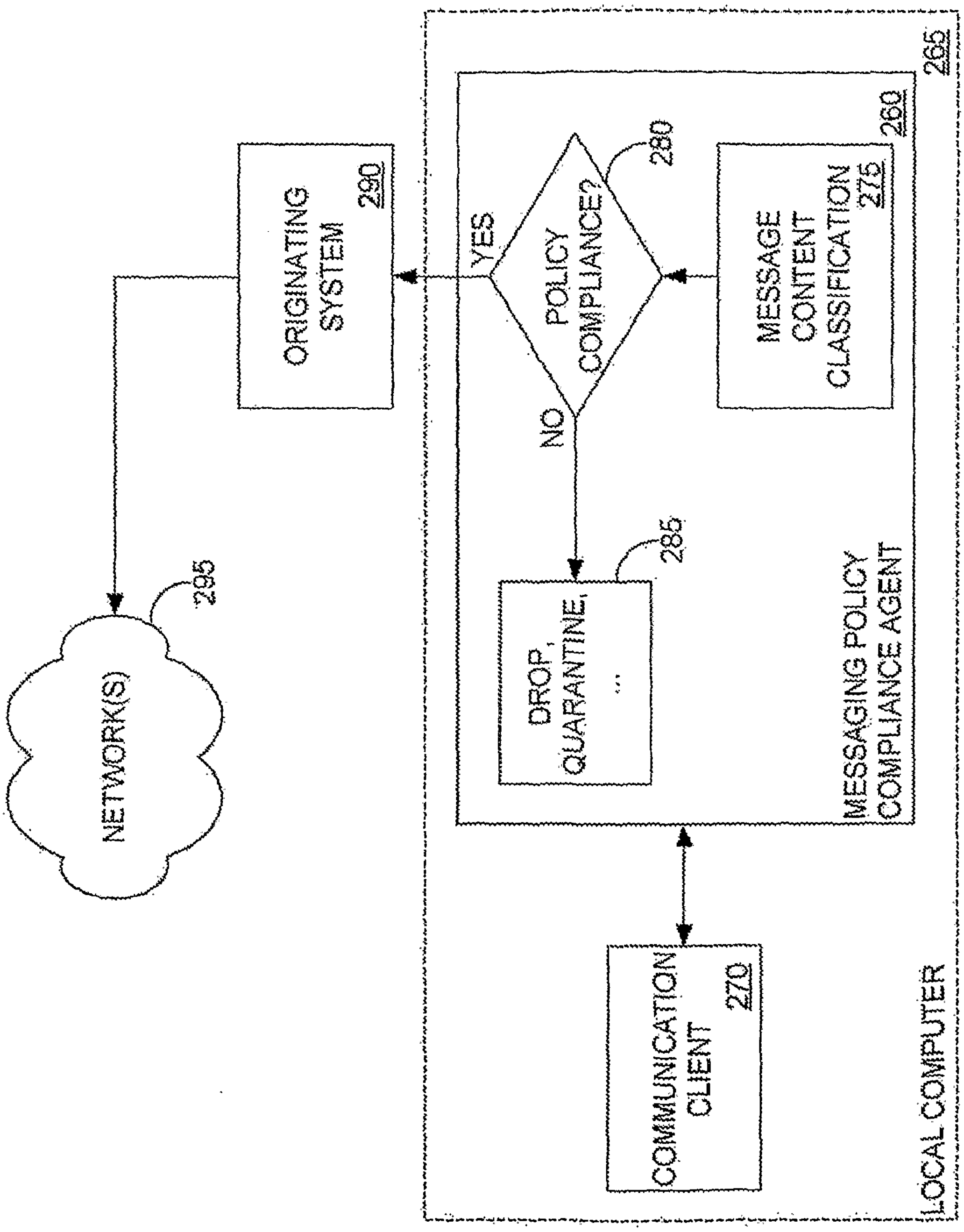

NETWORK(S)
295
ORIGINATING SYSTEM
290
YES
POLICY COMPLIANCE?
280
NO
MESSAGE CONTENT CLASSIFICATION
275
DROP, QUARANTINE, ...
285
MESSAGING POLICY COMPLIANCE AGENT
260
265
COMMUNICATION CLIENT
270
LOCAL COMPUTER

FIG. 3

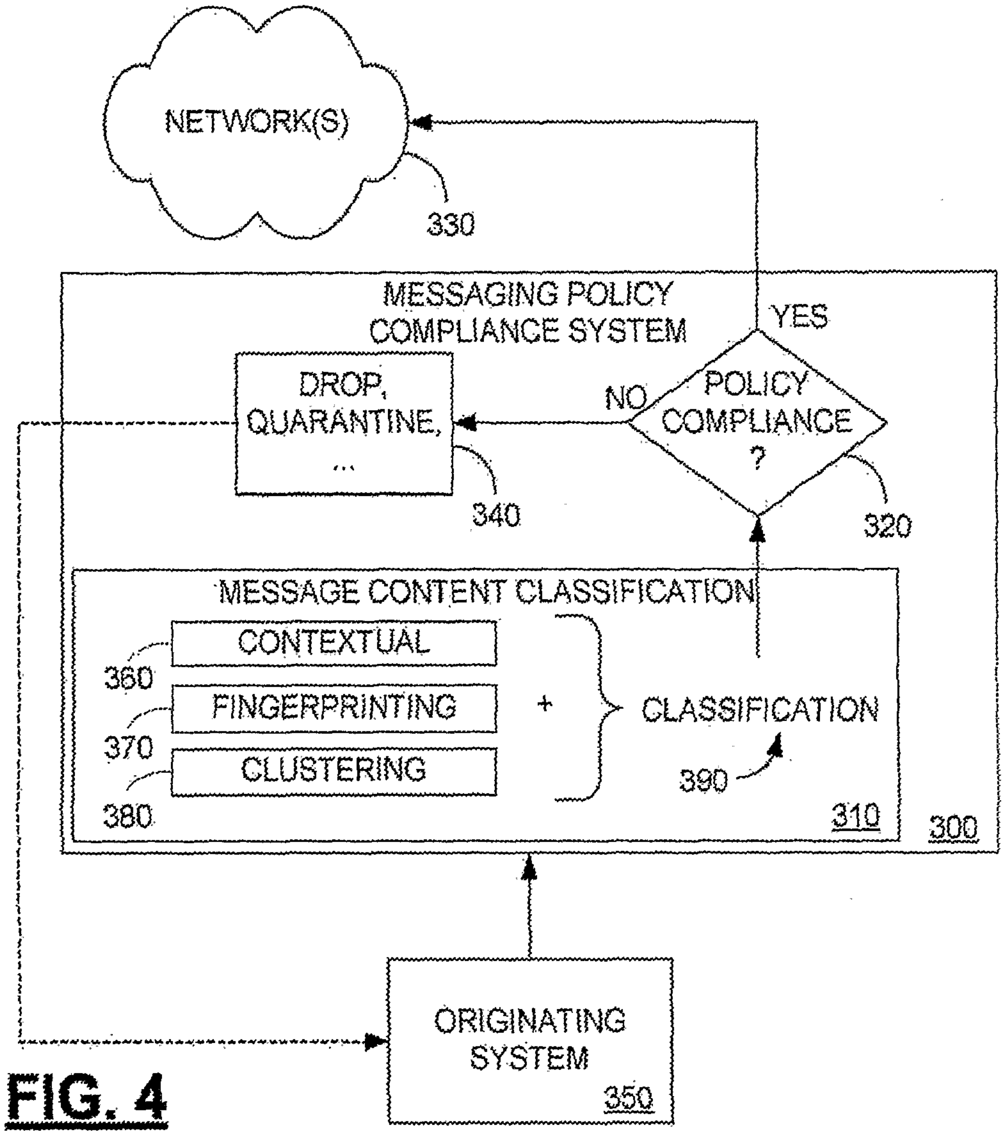

NETWORK(S)
330
MESSAGING POLICY COMPLIANCE SYSTEM
YES
NO
POLICY COMPLIANCE ?
320
DROP, QUARANTINE, ...
340
MESSAGE CONTENT CLASSIFICATION
CONTEXTUAL
360
FINGERPRINTING
370
CLUSTERING
380
+
CLASSIFICATION
390
310
300
ORIGINATING SYSTEM
350

FIG. 4

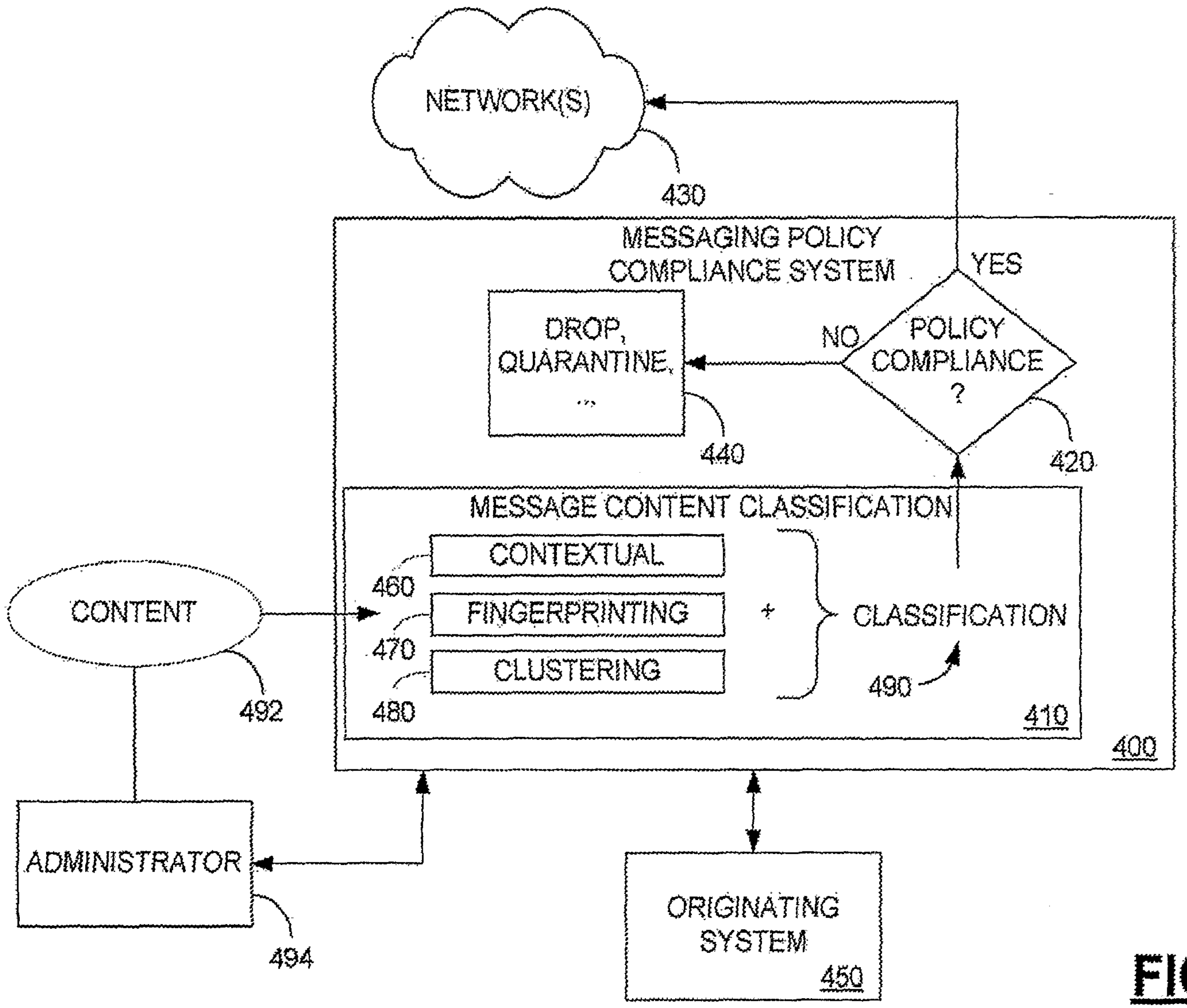

NETWORK(S)
430
MESSAGING POLICY
COMPLIANCE SYSTEM
YES
DROP,
QUARANTINE,
...
NO
POLICY
COMPLIANCE
?
440
420
MESSAGE CONTENT CLASSIFICATION
CONTEXTUAL
460
FINGERPRINTING
470
CLUSTERING
480
+
CLASSIFICATION
490
410
400
CONTENT
492
ADMINISTRATOR
494
ORIGINATING
SYSTEM
450

FIG. 5

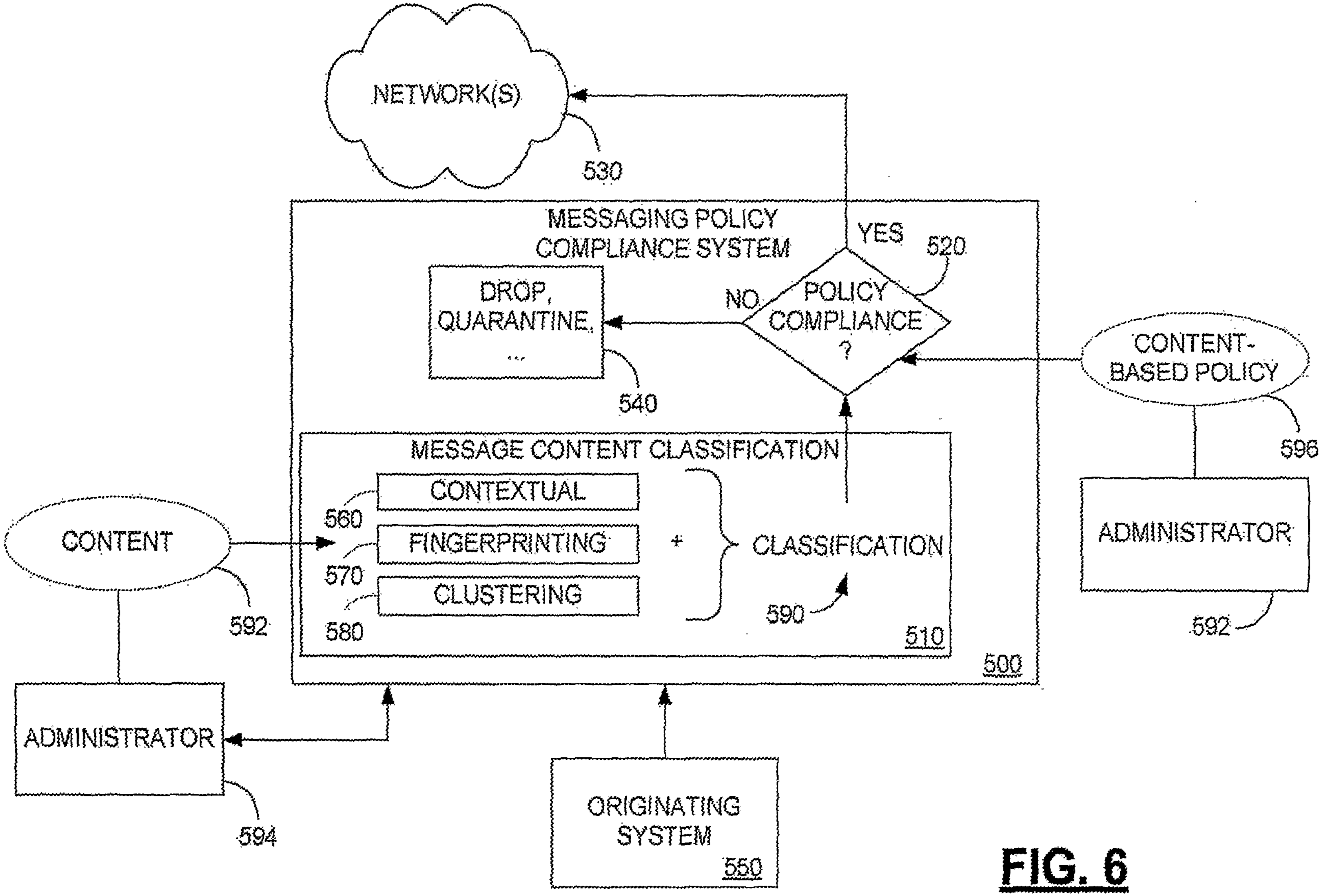

NETWORK(S)
530
MESSAGING POLICY
COMPLIANCE SYSTEM
YES
520
NO
POLICY
COMPLIANCE
?
DROP,
QUARANTINE,
...
540
CONTENT-
BASED POLICY
596
MESSAGE CONTENT CLASSIFICATION
CONTEXTUAL
560
FINGERPRINTING
570
CLUSTERING
580
+
CLASSIFICATION
590
510
500
CONTENT
592
ADMINISTRATOR
594
ADMINISTRATOR
592
ORIGINATING
SYSTEM
550

FIG. 6

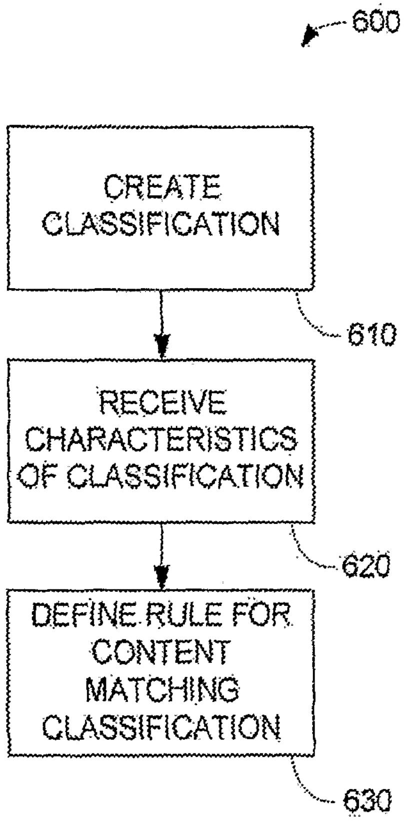

600
CREATE CLASSIFICATION
610
RECEIVE CHARACTERISTICS OF CLASSIFICATION
620
DEFINE RULE FOR CONTENT MATCHING CLASSIFICATION
630

FIG. 7

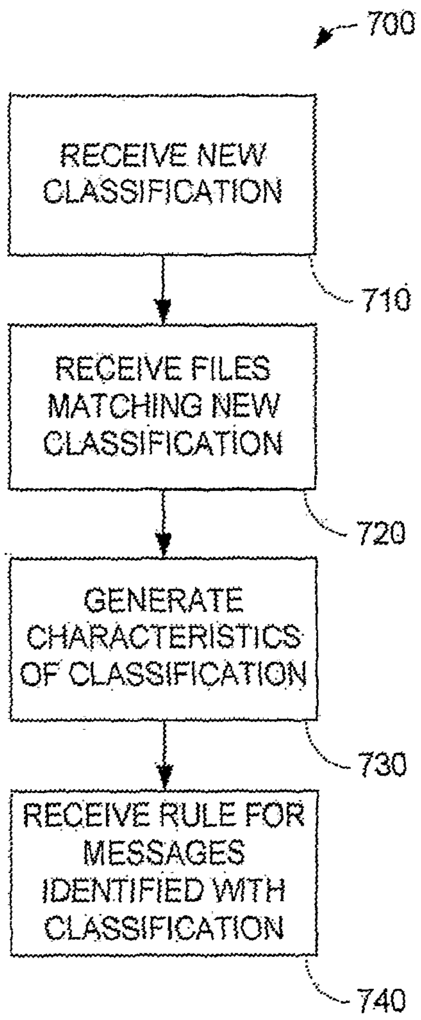

700
RECEIVE NEW CLASSIFICATION
710
RECEIVE FILES MATCHING NEW CLASSIFICATION
720
GENERATE CHARACTERISTICS OF CLASSIFICATION
730
RECEIVE RULE FOR MESSAGES IDENTIFIED WITH CLASSIFICATION
740

FIG. 8

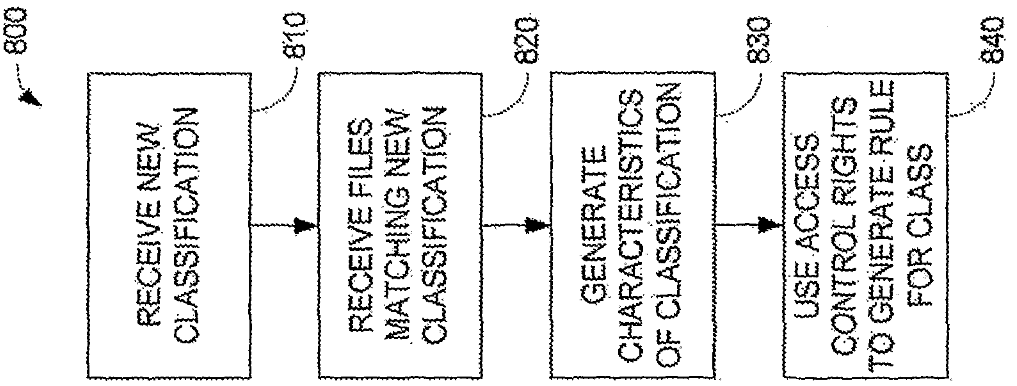

800
RECEIVE NEW CLASSIFICATION
810
RECEIVE FILES MATCHING NEW CLASSIFICATION
820
GENERATE CHARACTERISTICS OF CLASSIFICATION
830
USE ACCESS CONTROL RIGHTS TO GENERATE RULE FOR CLASS
840

FIG. 9

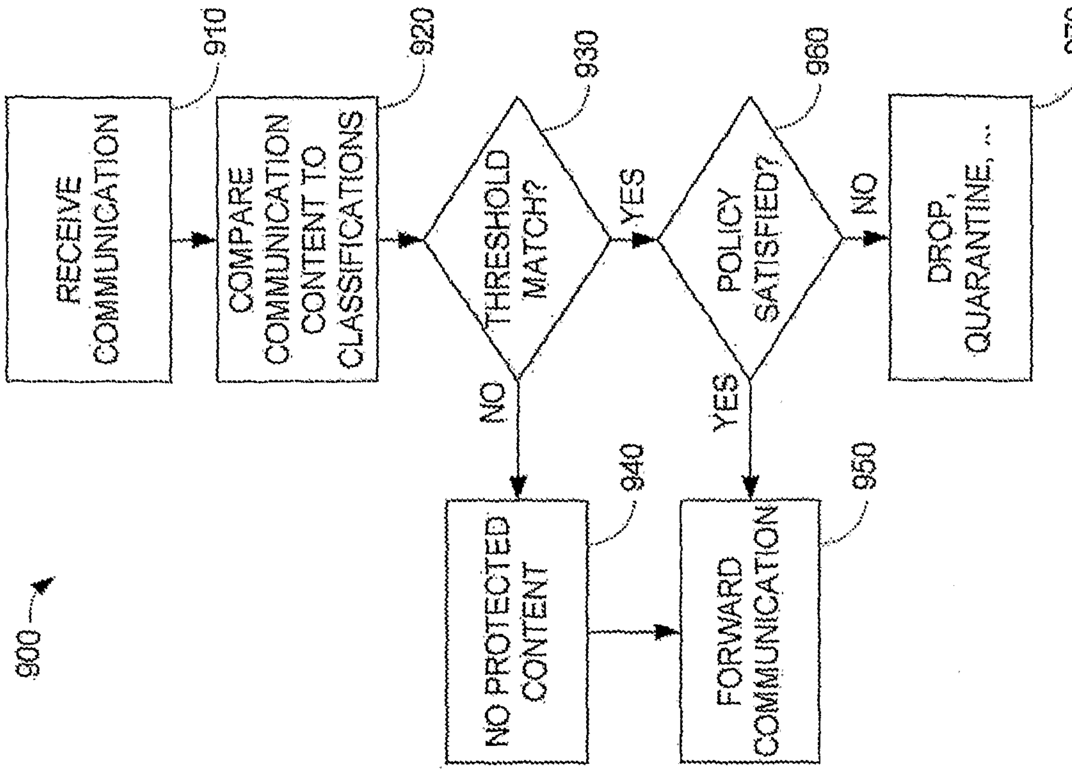

900
RECEIVE COMMUNICATION
910
COMPARE COMMUNICATION CONTENT TO CLASSIFICATIONS
920
THRESHOLD MATCH?
930
YES
NO
NO PROTECTED CONTENT
940
POLICY SATISFIED?
960
YES
NO
FORWARD COMMUNICATION
950
DROP, QUARANTINE, ...
970

FIG. 10

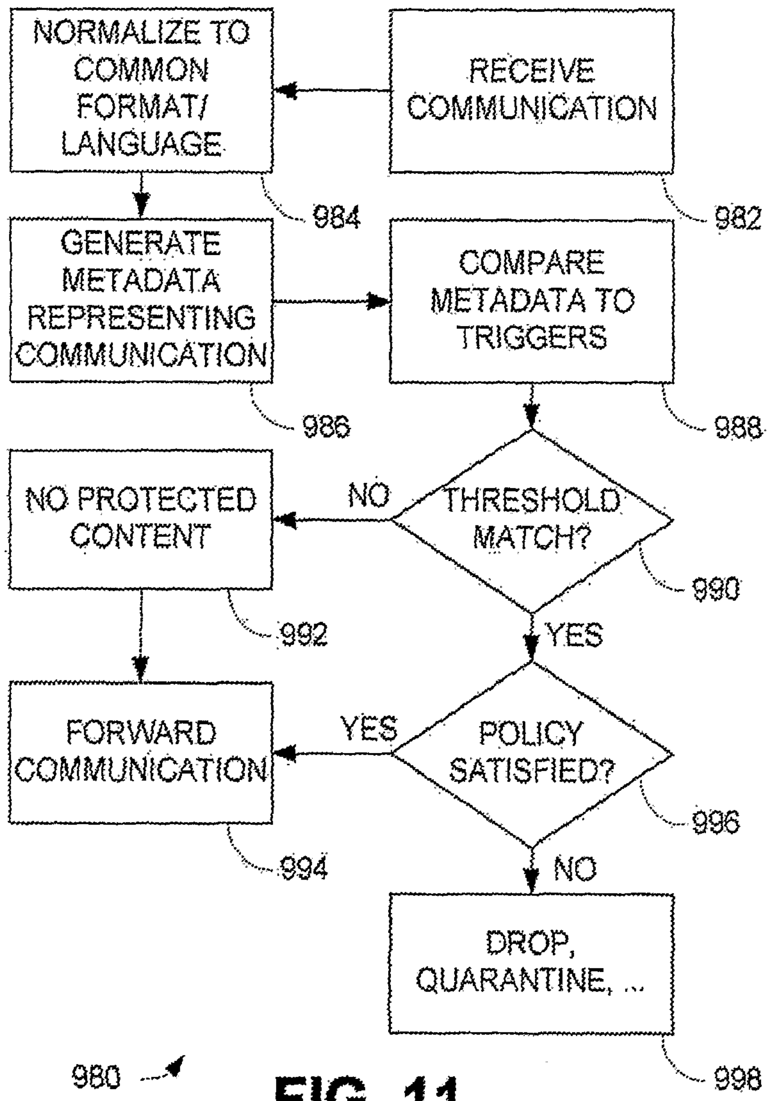

RECEIVE COMMUNICATION
982
NORMALIZE TO COMMON FORMAT/ LANGUAGE
984
GENERATE METADATA REPRESENTING COMMUNICATION
986
COMPARE METADATA TO TRIGGERS
988
THRESHOLD MATCH?
990
NO
NO PROTECTED CONTENT
992
YES
POLICY SATISFIED?
996
YES
FORWARD COMMUNICATION
994
NO
DROP, QUARANTINE, ...
998
980

FIG. 11

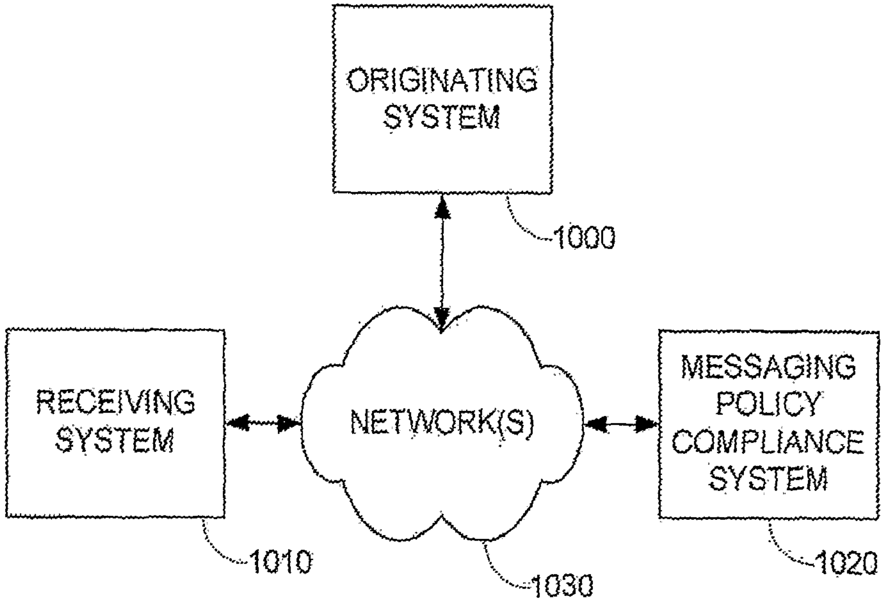

ORIGINATING SYSTEM
1000
RECEIVING SYSTEM
1010
NETWORK(S)
1030
MESSAGING POLICY COMPLIANCE SYSTEM
1020

FIG. 12

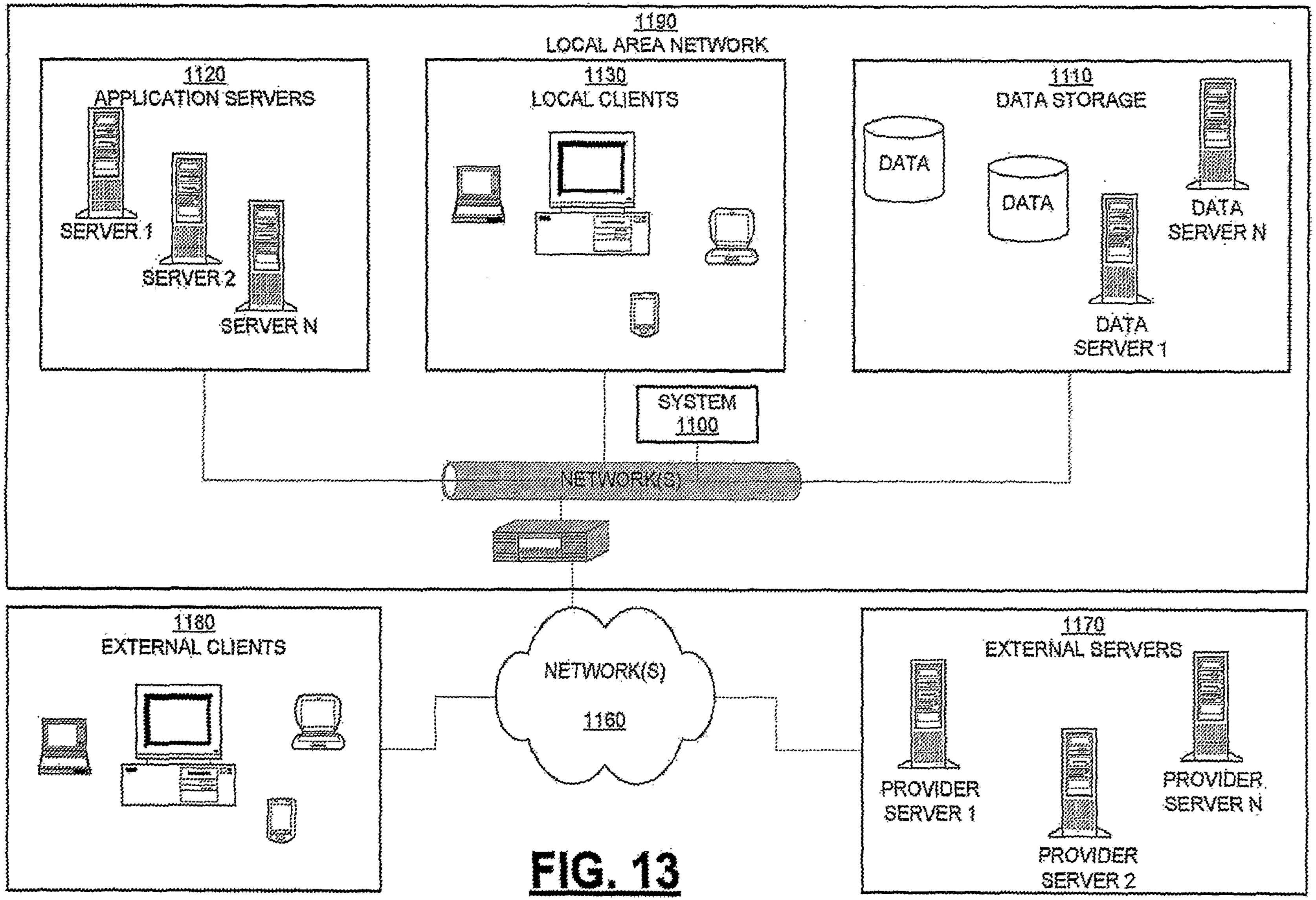

1190
LOCAL AREA NETWORK
1120
APPLICATION SERVERS
SERVER 1
SERVER 2
SERVER N
1130
LOCAL CLIENTS
1110
DATA STORAGE
DATA
DATA
DATA SERVER N
DATA SERVER 1
SYSTEM 1100
NETWORK(S)
1180
EXTERNAL CLIENTS
NETWORK(S)
1160
1170
EXTERNAL SERVERS
PROVIDER SERVER 1
PROVIDER SERVER 2
PROVIDER SERVER N

FIG. 13