Data Loss Prevention System

Field of the Invention

This invention relates to a system for protecting data stored in a system connected to a network. More particularly, this invention relates to a system that detects a potential transmission of protected data via the network. Still more particularly, this invention relates to a system that detects a potential transmission of data even if the data is included with other data or has a different format.

Prior Art

In today's society, many businesses and private users rely on a network to connect multiple computers used by the company. In order to make business operations efficient and save space, most of the networks include a server or other memory system to store document. Often times, some of the documents include sensitive material that the business or user wishes to keep private. Some examples of sensitive material include customer lists, financial information, and proprietary technology.

To prevent dissemination of this material, files containing sensitive information have been restricted to specified users. However, restricted access does not prevent those users having access from intentionally or unintentionally copying and/or transmitting files and/or documents that include sensitive material to another employee not having access to the material over the network or to others outside the company via an Internet connection to the network. Administrators have noticed that it is often hard to prevent dissemination of material especially as the number of network connected devices used by employees increases.

In order to prevent the unauthorized transmission of sensitive material, administrators of networks have included data loss systems in network security devices such as firewalls, intrusion detection systems and intrusion protection systems. The systems monitor transmissions over the network to detect the transmission of sensitive material. Typically, these systems use the following search methods to detect sensitive material in a transmission: rule-based/regular expression, database fingerprinting, exact file matching, partial document matching, statistical analysis, conceptual/lexicon, and categorical. A rule-based/regular expression search analyses content of a transmission for specific rules to detect a match. For example, this type of search may look for 16 digit credit card checksum requirements. The database fingerprinting method searches transmission for matches to structured data from a reference database. The exact file matching method hashes a file to detect a match. The partial document matching method uses cyclical hashing on protected content to detect a match. Statistical matching uses machine learning and Bayesian analysis to determine a match. The conceptual or lexicon search method uses a combination of dictionaries, rules and other analysis to find a match. The categorical search method uses pre-built categories with rules and dictionaries for common types of sensitive data.

The use of the above described search methods has several shortcomings. The first shortcoming is that many of these types of searches are complex and require a large amount of device resources to be performed. Thus, systems using these types of searches often slow network traffic as a device, such as router, has to devote resources to perform the searches. This is often undesirable as the slowing of network traffic impedes communications across the network.

A second shortcoming is that most of these methods produce a great number of false positive matches between the documents under test and the sensitive material. Thus, a lot of transmissions that do not include sensitive material are mistakenly identified as including sensitive material. This causes unnecessary impedance of transmissions over network and requires an administrator to check an inordinate amount of work detecting actual transmissions of sensitive material. One solution to this problem is to use an exact match search method. However, exact match systems have the shortcoming that the search may be defeated by changing the transmitted document to not include the exact wording of the original document. Thus, some transmissions of sensitive material may not be detected.

Thus, those skilled in the art are constantly striving to provide a system that minimizes the amount of resources used to perform a search of transmitted documents while maximizing the rate of successful detections of transmissions of sensitive material.

Summary of the Invention

The above and other problems are solved and an advance in the art is made by a data loss prevention system in accordance with this invention. A first advantage of a system in accordance with this invention is that false positive detection rates are decreased. A second advantage in accordance with this invention is that searches may be more efficient. This reduces the time and resources needed to perform the search. A third advantage is that a threshold value for analysing the transmitted document for sensitive material may be adjustable to increase or reduce the number of reported matches from the searches.

The above and other features and advantages are provided by a data loss prevention system in accordance with this invention that detects a transmission of material in the following manner. The system begins by receiving a document to be transmitted. The system then generates a set of trigrams of the text of the document. Each trigram in the set of trigram includes three strings of characters including a middle string, a first string of characters immediately preceding the middle string of characters, and a last string of characters immediately following the middle string characters. The set of trigrams of the text of the document is then compared to a set of trigrams of information of interest, such as sensitive material. The system determines the number of trigrams from the set of the trigrams for the text of the received documents that match a trigram in the set of trigrams for the information of interest. In some embodiments, the number of matching trigrams is then compared to a threshold value. A possible transmission is indicated if the number of matches is greater than or equal to the threshold value. In other embodiments, a percentage of matches versus the number of trigrams in the set of trigrams of the information of interest may first be determined and compared to a threshold percentage value. Detection of a possible transmission of the information of interest is then indicated if the percentage value of matching trigrams is greater than or equal to the threshold percentage value.

In some embodiments, a notification of possible transmission message is generated in response to the detection of a possible transmission. This message may then be transmitted to an administrator via an e-mail or other means; or the message may be logged in a transmission detection file for future use. In other embodiment, the system may apply rules for handling the transmission in response to detecting the information of

interest. The rules may include preventing the transmission, delaying the transmission until an administrator can make a determination, transmitting a warning to the sender, or any other rule an administrator of the network may want to apply.

The system may be included in server or router that provides Internet service to devices connected to the network in some embodiments. In these embodiments, the system receives the packets being transmitted and assembles the text of the document from the data in the packets. The document may then be searched in the above described manner.

In some embodiments, the set of trigrams for the document is generated in the following manner. The system parses the text of the document to identify strings of characters in the text. The set of trigrams is then formed by grouping a middle string of characters with a first string of characters immediately preceding the middle string and a last string of characters immediately following the middle string. In some embodiments, the string of character may only include a single character and punctuation and spaces may or may not be included. In other embodiments, strings of characters of arbitrary lengths, such as words may be used. In some of these embodiments, the trigram may be formed only from words in an individual sentence. The generated trigrams may then be stored in a memory of the system as a test file or other data structure in some embodiments for performing the search.

In some embodiments, the document or file is first parsed and pre-processed prior to generating the trigrams. In these embodiments, pre-processing may include conversion of characters into a single case; removal of punctuation; and/or conversion of

the file into a common format, such as plain text. In some of these embodiments, character replacement may also be performed. For example, contracted words may be substituted for the contraction. In these embodiments, the substitution is performed in the following manner. First, the system searches the identified strings of characters for replaceable strings. The system then inserts replacement strings for each replaceable string found during the searching into the text of the document. The replacement strings are included in the trigrams generated for the document as alternatives to the replaceable string.

In accordance with some embodiments of the system, the received document may be compared to sets of trigrams for multiple documents of interest. The comparison is performed in the following manner. The set of trigrams for the received document is compared to the set of trigrams of the documents of interest. The system counts the number of matches between the set of trigrams for each document of interest and the set of trigrams of the received document. The system compares the number of matches for each document of interest with a threshold value. The received document is determined to be a possible transmission of a particular document of interest if the number of matches for the particular document is greater than or equal to the threshold value. In some embodiments, a percentage of matches versus the number trigrams in the set of trigrams of a document of interest may first be determined and compared to a percentage value.

In some of these embodiments, only a group of the documents of interest are compared to the received documents. In these embodiments, the system must determine the group of documents to compare to the received document. In some of these

embodiments, the trigrams of the received document are compared to an index set of trigrams. The index set of trigrams may be trigrams frequently found in a sub-set of the documents of interest or trigrams that indicate a specific subject matter.

In some embodiments that use an index set of trigrams to determine documents of interest to be compared to the received document, the system determines the documents of interest to compare in the following manner. The system first compares each trigram of the received document to each trigram in the index set. If a trigram from the received document matches a trigram in the index set, all documents of interest associated with the matched trigram are added to the group of documents of interest to compare to the received document.

In other embodiments using the index to determine documents of interest to compare to the received document, the system determines the documents of interest to compare in the following manner. First, the system compares the set of trigrams of the received document to an index set of trigrams. The number of matches for each trigram in the index is counted. The system then determines the documents of interest to compare from the counted number of matches of each trigram. The determination may be made by comparing the number of matches of each trigram in the received document to a stored number of occurrences of the trigram in each document of interest that is associated with the trigram in the index. In still other embodiments, documents of interest are selected for comparison when the number of matches for a trigram of the received document is equal to the number occurrences of the trigram in a particular document of interest. In other embodiments, a document of interest that has a number of occurrences of the trigram is selected for comparison if the number of occurrences of the

trigram in the document is within a predetermined deviation from the number of matches to the trigram in the set of trigrams of the received document.

Some embodiments add a new document to the set of documents of interest in the following manner. The system receives the new document. The system then parses the text of the new document to identify strings of characters in the text. The set of trigrams for the new document is then formed by grouping a middle string of characters with a first string of character immediately preceding the middle string and a last string of characters immediately following the middle string. In some embodiments, the string of characters may only include a single character and punctuation and spaces may or may not be included. In other embodiments, strings of characters of arbitrary lengths, such as words may be used. In some of these embodiments, the trigram may be formed only from words in an individual sentence. The generated trigrams may then be stored in a document trigram file or other data structure for use in the comparison process. The generated trigram will be consistent with the trigrams formed for the received document to allow comparisons of the trigrams.

In some of these embodiments, the new document or file is first parsed and pre-processed prior to generating the trigrams. The pre-processing may include conversion of characters into a single case; removal of punctuation; and conversion of the file into a common format, such as plain text. In some of these embodiments character replacement may also be performed. For example, contracted words may be substituted for the contraction. In these embodiments, the substitution is performed in the following manner. First, the system searches the identified strings of characters for replaceable strings. The system then inserts replacement strings for each replaceable string found

during the searching of the text of the document. The replacement strings are included in the trigrams generated for the document as alternatives to the replaceable strings.

In some of these embodiments, the index set of trigrams is maintained in the following manner. When a new document of interest is added to the system, each trigram in the set of trigrams for the new document is read from the set of trigrams of the new document and compared to the trigrams in the index. If the compared trigram is already in the index, the system determines whether a list of associated documents of interest includes the new document. If the new document is not in the list of associated documents of interest, the new document is added to the list.

In other of these embodiments where an occurrence count is used, the index set of trigrams is maintained in the following manner. When a new document of interest is added to the system, each trigram in the set of trigrams for the new document is read from the set of trigrams of the new document and compared to the trigrams in the index. If the compared trigram is already in the index, the system determines whether a list of associated documents of interest includes the new document. If the new document is included in the list of associated documents of interest, an occurrence counter is incremented. If the new document is not in the list of associated documents of interest, the new document is added to the list and a counter for the new document is set to one.

In the above embodiments, the system determines whether the trigram should be added to the index if the trigram was not found in the index. Preferably, the user provides some indication that the trigram is to be added. If the trigram is to be added, the trigram is added to the set of index of trigrams, and the new document is added to the

list of associated documents for the trigram. In embodiments using an occurrence count, an occurrence counter for the new document is set to one.

Brief Description of the Drawings

The above and other features and advantage of a data loss prevention system are described in the following Detailed Description and are shown in the following drawings:

Figure 1 illustrating an embodiment of a network including a data loss prevention system in accordance with one exemplary embodiment of this invention;

Figure 2 illustrating a processing device connected to the network that executes instructions to perform the processes of the data loss prevention system in accordance with an exemplary embodiment of the system;

Figure 3 illustrating a flow diagram of a process for comparing a received document with document of interest in accordance with one exemplary embodiment of this invention;

Figure 4 illustrating a flow diagram of a process for receiving a document to compare to the documents of interest in accordance with one exemplary embodiment of this invention;

Figure 5 illustrating a flow diagram of a process for generating a set of trigrams for a received document in accordance with one embodiment of this invention;

Figure 6 illustrating a flow diagram of a process for comparing the set of trigrams of the received documents to sets of trigrams of documents of interest in accordance with one exemplary embodiment of this invention;

Figure 7 illustrating a flow diagram of a process for adding a new document to the documents of interest in accordance with one exemplary embodiment of this invention;

Figure 8 illustrating a flow diagram of a process for generating an index set of trigrams for use in determining documents of interest to compare to the received document in accordance with one exemplary embodiment of this invention;

Figure 9 illustrating an example of a received document in accordance with one embodiment of this invention; and

Figure 10 illustrating an example of the set of trigrams formed for the received document shown in Figure 9 in accordance with one exemplary embodiment of this invention.

Detailed Description

This invention relates to a system for protecting data stored in a system connected to a network. More particularly, this invention relates to a system that detects a potential transmission of protected data via the network. Still more particularly, this invention relates to a system that detects a potential transmission of data even if the data is included with other data or data has a different format.

A data loss prevention system in accordance with this invention detects the possible transmission of information of interest over a network by generating a set of trigrams for the text in the transmission and comparing the set of trigrams to sets of trigrams for the information of interest stored in a library or other data structure for comparisons. For purposes of this discussion, information of interest, may be files, documents, or portions of text that an organization feels is proprietary, classified, or simply desires to know if the information is transmitted to unauthorized parties. For purposes of this discussion, unauthorized parties are either users of the network that do not have authorized access to the information or systems and/or parties outside the network.

Figure 1 illustrates an exemplary network 100 including a data loss prevention system in accordance with this invention. Servers 125 and 130 are connected to network 100. Servers 125 and 130 are processing systems that store and provide data and applications to other processing devices connected to the network 100. One skilled in the art will recognize that although two servers are shown, any number of servers may be connected to network 100 and the number of connected servers is left to the choice of a network administrator. Devices 135 and 140 are also connected to network 100. Devices 135 and 140 are processing systems that connect to the network for communications with other devices to receive data and applications to perform on the devices. These devices may be directly connected to the network via an Ethernet or other wire-based connection or may be connected using a wireless connection, such as a Radio Frequency (RF) using a known communication protocol. One skilled in the art will recognize that any number of devices may be connected to the network. The exact

number of connected devices is determined by network resource constraints and administrators of the network.

One or more routers 115 are also connected to network 100. A router 115 is a processing device that receives transmissions from the connected processing systems and transmits the transmission to the proper receiving devices. Router 115 may also connect network 100 to Internet 105 to allow devices connected to network 100 to communicate with devices connected to Internet 105.

Firewall 110 is included in the network which is shown conceptually as being between internet 105 and router 115. Firewall 110 prevents data for unauthorized material from being received by and/or transmitted from devices connected to network. One skilled in the art will recognize that the applications may be software, firmware, hardware, or any combination of the three. Furthermore, the firewall may be provided by applications performed by one or more servers 125, 130, router 115, and/or a combination of servers, routers and other connected devices.

A data loss prevention system in accordance with this invention may be included in the applications provided by firewall 110, an intrusion detection system, an intrusion protection system, and/or a stand alone security system. Furthermore, a data loss prevention system may be executed by a server, a router, a connected device, and/or any combination of the preceding systems.

In this embodiment of the invention, the firewall system including a data loss prevention system in accordance with this invention is implemented as hardware,

software, firmware, and/or a combination of any of the preceding three components of one or more processing system connected to the network. Figure 2 illustrates an exemplary processing system including the components needed to execute the applications from instructions stored in memory in accordance with this invention. One skilled in the art will recognize that the exact configuration of each processing system may be different and the exact configuration executing processes in accordance with this invention will vary and the figure is given by way of example only.

Processing system 200 includes Central Processing Unit (CPU) 205. CPU 205 is a processor, microprocessor, or any combination of processors and microprocessors that execute instructions to perform the processes in accordance with the present invention. CPU 205 connects to memory bus 210 and Input/Output (I/O) bus 215. Memory bus 210 connects CPU 205 to memories 220 and 225 to transmit data and instructions between the memories and CPU 205. I/O bus 215 connects CPU 205 to peripheral devices to transmit data between CPU 205 and the peripheral devices. One skilled in the art will recognize that I/O bus 215 and memory bus 210 may be combined into one bus or subdivided into many other buses and the exact configuration is left to those skilled in the art.

A non-volatile memory 220, such as a Read Only Memory (ROM), is connected to memory bus 210. Non-volatile memory 220 stores instructions and data needed to operate various sub-systems of processing system 200 and to boot the system at start-up. One skilled in the art will recognize that any number of types of memory may be used to perform this function.

A volatile memory 225, such as Random Access Memory (RAM), is also connected to memory bus 210. Volatile memory 225 stores the instructions and data needed by CPU 205 to perform software instructions for processes such as the processes for providing a system in accordance with this invention. One skilled in the art will recognize that any number of types of memory may be used to provide volatile memory and the exact type used is left as a design choice to those skilled in the art.

I/O device 230, keyboard 235, display 240, memory 245, network device 250 and any number of other peripheral devices connect to I/O bus 215 to exchange data with CPU 205 for use in applications being executed by CPU 205. I/O device 230 is any device that transmits and/or receives data from CPU 205. Keyboard 235 is a specific type of I/O that receives user input and transmits the input to CPU 205. Display 240 receives display data from CPU 205 and displays images on a screen for a user to see. Memory 245 is a device that transmits and receives data to and from CPU 205 for storing data to a media. Network device 250 connects CPU 205 to a network for transmission of data to and from other processing systems.

This invention relates to comparing a document to be transmitted to documents of interest to detect the possible transmission of sensitive material. To ensure accurate detections of the sensitive material, a data loss prevention system in accordance with this invention use comparisons of a set of trigrams generated from the document being transmitted and a set of trigrams of each of the documents including sensitive material.

A trigram is a grouping of three strings of characters. The strings of characters may be one character, or a string of characters of arbitrary length, for example words.

Each trigram has a middle string of characters, a first string of characters that immediately precedes the middle string of characters in the document, and a last string of characters that immediately follows the middle string of characters in the document. In some embodiments, strings of characters, such as words, are grouped in trigrams to reduce the number or trigrams generated and compared. However, it is envisioned that some embodiments may form trigrams of individual characters; and may include or exclude spaces and/or punctuation marks. In some other embodiments, the trigrams are bounded by sentences and thus trigrams are made for each sentence with none of the trigrams including words from two or more sentences. This is provided to reduce the number of trigrams compared. However, it is envisioned that other embodiments may use trigrams that cross sentences and include words from two or more sentences.

The following is an example to illustrate trigrams. Figure 9 illustrates file 900 storing phrase 905 that states "This document contains some sensitive information." Figure 10 is trigram file 1000 storing the trigrams 1005, 1010, 1015, and 1020 of phrase 905. Trigram 1005 includes the first three words of the document: "This document contains". Second trigram 1010 shifts the middle word trigram 1005 to the first word of the trigram and the fourth word of the document becomes the last word of the trigram forming the trigram: "document contains some". This process repeats until the end of the document producing trigram 1015: "contains some sensitive" and trigram 1020: "some sensitive information".

Process 300 is an embodiment of process performed by a data loss prevention system in accordance with this invention to detect a possible transmission of sensitive material using trigrams and is illustrated in Figure 3. Process 300 begins in step 305 by

receiving a document to test whether information of interest is included. This document may be received by intercepting packets or data being transmitted across a network or may be a document that a user wishes to transmit that is passed to process 300 being executed on the transmitting system. The exact embodiment is left to those skilled in the art designing and using a data loss prevention system in accordance with this invention.

In step 310, process 300 generates a set of trigrams for the received document. Typically, the generated trigrams are stored in a file or data structure for use in performing comparison search against multiple documents. An example of a process for generating the trigrams is illustrated in Figure 5 and is described below.

After the set of trigrams is generated for the received document, the set of trigrams for the received document is compared to each set of trigrams stored for each document of interest in step 315. These sets of trigrams for the documents of interest may be stored as files or other data structures in a library stored in the memory of a system performing the comparison or in the memory of another system accessible by the network.

In step 320, process 300 determines whether there is a match between the set of trigrams of the received document and the set of trigrams of one of the documents of interest. This may be performed by comparing the number of matching trigrams to a threshold value. However, a more accurate determination may be made by determining the percentage of trigrams of the document of interest matched by the trigrams of the received document and comparing the determined percentage to a threshold percentage value. The percentage of trigram matched may be calculated by taking the number of

matching trigrams detected and dividing by the total number of trigrams in the set of trigrams of the document of interest. The calculated percentage is then compared to a threshold percentage value, such as 80%, if the calculated percentage is greater than or equal to the threshold percentage value, a match is detected. The threshold percentage value may be determined by a network administrator and may be adjusted based upon the sensitivity of the material. For example, more sensitive material may have a lower threshold percentage value to ensure more documents are detected while less sensitive material may have a higher threshold percentage value to prevent too many false positive matches. Furthermore, the threshold percentage value should be set to allow detection of the documents while preventing a great number of false positive matches. This may be determined through statistical analysis of the comparisons.

If there is not a match between the trigrams of the received document and the sets of trigrams compared, process 300 ends. If a match is detected, a notification message is generated in step 325. The notification may include identification of the sender and/or receiver of the transmission; the time of transmission of the information of interest; and/or other information that an administrator deems relevant. The notification message is then either transmitted or written to a report file in step 330. The transmission may be an e-mail or other type of message sent to an administrator of the network to notify the administrator of the possible transmission of sensitive materials.

In optional step 335, a handling rule may then be applied for the transmission in response to the match. The rules may include: prevent transmission, delay transmission until an administrator reviews the transmission, notify the user that sensitive material is

possibly included in the message, and/or any other action that an administrator of the network may deem necessary.

Process 400 illustrates a process for receiving a document being transmitted over a network. Process 400 may be executed when process 300 is performed by a router, switch, server, or other device facilitating communication between devices over a network. Process 400 begins in step 405 with the packets being transmitted being received by the device executing process 300. One skilled in the art will recognize that packets is intended to cover cells, frames and other data structure used to transmit data over a network. In step 410, process 400 assembles the received document from the data in the received packets and process 400 ends.

Figure 5 illustrates a process 500 which is an embodiment of a process to generate the set of trigrams for a received document. Process 500 begins with pre-processing of the received document in step 502. Pre-processing may include conversion of characters into a single case; removal of punctuation; and/or conversion of the file into a common format, such as plain text. In step 505, the text of the received document is parsed to detect the strings of the documents. In some embodiments, optional step 510 may be performed to search for replacement strings. Replacement strings are strings that may represent other strings or have alternative strings. One example of replacement strings is contractions of words in documents. For example, in matching a document, it may be advantageous to match for occurrences of "don't" as well as "do not". Thus, it may be desirable to add "do not" as alternative when "don't" occurs in a file. This may be done by adding a marker indicating alternative words to allow trigrams with each

alternative to be formed. The replacement strings are then added to the document in step 515.

Process 500 forms the trigrams in step 520. This may be performed by an iterative process which begins with a counter, n, equal to 2. During each iteration, a trigram is formed that includes nth-1 string in the document, nth string in the document, and nth +1 string in the document. The counter is then incremented and trigrams are formed until n equals the total numbers of strings in the document minus 1. Each of the formed trigrams is then stored in a linked-list, file or other data structure for use in the comparisons in step 525 and process 500 ends.

Figure 6 illustrates process 600 which is an embodiment for a process of comparing the set of trigrams of the received document to sets of trigrams for documents including information of interest. Process 600 begins with steps for determining a group of documents including information of interest that is likely included in the received documents. The sets of trigrams for this group of documents are compared to the set of trigrams for the received document. This process is performed to attempt to reduce the documents needed to be compared to the received document in order to reduce the number of comparisons to be performed.

In process 600, the group of documents is determined by comparing the set of trigrams for the received document to an index set of trigrams. The index set of trigrams includes trigrams from documents of interest that are likely to indicate the information of interest in a specific document. The index has an associated list of documents for each trigram indicating the document of interest that include the trigram and in some

embodiment, an occurrence counter indicating the number of times the trigram occurs in the document. One skilled in the art will know that other means of determination may be used based upon the index.

Process 600 begins by comparing each trigram for the received document to the index set of trigrams in step 605. In some embodiments, all documents in the list of associated documents of interest for a trigram in the index are added to the group of documents of interest to compare to the received document when the trigram matches one of the trigrams from the set of trigrams for the received document. The process then proceeds to step 620.

In other embodiments, such as the embodiment shown in Figure 6, the number of matches of trigrams in the received document to each trigram in the index set is stored in step 610. From the number of matches between the set of trigrams of the received documents and each trigram in the index set, a group of documents likely to match the received document is generated in step 615. In this embodiment, the group of documents is generated by comparing the number of matches of trigrams for received document with a particular trigram and then selecting a document if the number of matches is equal or greater to a number of occurrences of the trigram in the document of interest. In other embodiments, a document may be selected if the number of matches is within a specified deviation or range of the number of occurrences in a specific document of interest.

After the group of documents to compare is generated or if all of the stored documents of interest are compared to the received document, process 600 does the comparisons of the documents of interest in the following manner. Process 600 begins

by selecting a set of trigrams for one of the documents of interest to be compared. The set of trigrams for the selected document of interest is then compared to the set of trigrams for the received document in step 620. In step 625, the number of matches between the set of trigrams for the selected document of interest and set of trigrams of the received document is counted.

The number of matches is then compared to a threshold value in step 630. This may be performed by comparing the number of matching trigrams to a threshold value. However, a more accurate determination may be made by determining the percentage of trigrams of the document of interest matched by the trigram of the received document and comparing the determined percentage to a threshold percentage value. The percentage of trigram matched may be calculated by taking the number of matching trigrams detected and dividing by the total number of trigrams in the document of interest. The calculated percentage is then compared to a threshold percentage value, such as 80%, if the percentage matched is greater than or equal to the threshold percentage value, a match is detected. The threshold percentage value may be determined by a network administrator and may be adjusted based upon the sensitivity of the material. For example, more sensitive material may have a lower threshold percentage value to ensure more documents are detected while less sensitive material may have a higher threshold percentage value to prevent too many false positive matches. Furthermore, the threshold percentage value should be set to allow detection of the documents while preventing a great number of false positive matches to the documents of interest. This may be determined through statistical analysis of the comparisons.

If it is not a possible match, process 600 proceeds to step 640. If number of matches is greater than or equal to the threshold value, a possible match is indicated in step 635. After step 635, process 600 may end or continue to step 640 to determine if another document of interest may match the received document. In step 640, process 600 determines if there is another document in the list of documents of interest to check. If there is another document of interest to compare, process 600 repeats from step 620 for another document. If there is not another document to compare, process 600 ends.

Figure 7 illustrates a process 700 which is an embodiment of a process to add a new document of interest to documents to monitor. Process 700 begins by receiving the new document to monitor in step 705. In step 707, an optional step of pre-processing of the new document is performed. Pre-processing may include conversion of characters into a single case; removal of punctuation; and/or conversion of the file into a common format, such as plain text. In step 710, the text of the new document is parsed to detect the strings of the documents. In some embodiments, optional step 715 may be performed to search for replacement strings. Replacement strings are strings that may represent other strings or have alternative strings as described with regard to step 510 of Figure 5. The replacement strings are then added to the new document in step 720.

Process 700 then forms the trigrams for the new document of interest in step 725. This may be performed by an iterative process which begins with a counter at 1 and forms trigrams having nth-1 string, nth string, and nth +1 string in the trigram. The counter is then incremented and new trigrams are formed until the last string in the document is reached. Each of the formed trigrams is then stored in a linked-list, file or

other data structure for use in the comparisons in step 730. The stored list or file is then stored in a library for use in comparisons in step 735 and process 700 ends.

Figure 8 illustrates process 800 that is an embodiment of a process for adding trigrams from a new document of interest to the index set of trigrams. Process 800 begins in step 805 by reading a trigram from the set of trigrams for the new document of interest. The set of trigrams may be in a linked list, file, or any other data structure used to store the trigrams for comparison. In step 810, the index set of trigrams is searched for the read trigram. If the trigram is not in the index, process 800 determines whether the trigram is to be added to the index in step 812. This determination is made by some kind of indication by the administrator or may be determined if the trigram occurs a certain number of times in the document. If the trigram is to be added, the trigram is inserted into the file or data structure for the index in step 814 and the document is placed in the list of associated document in step 820. If an occurrence counter is used to determine the documents of interest to add to the group to compare to the received document, an occurrence counter for the document of interest added to the list is also set to one in step 820.

If the trigram is determined to be in the index set in step 810, process 800 determines whether the new document is already in the list of associated documents for the trigram in step 815. If the document is in the list, the occurrence counter associated with the document is incremented in optional step 817 and process 800 proceeds to step 825. If the document is not in the list, the document is added to the list in step 820. If occurrence counters are used to determine the documents of interest to add to the group

to compare to the received document, an occurrence counter for the document of interest added to the list is also set to one in step 820.

After step 820, process 800 determines whether the set of trigrams of the new document includes more trigrams in step 825. If the set includes more trigrams, process 800 repeats from step 805. Otherwise, process 800 ends.

The above is a description of exemplary embodiments of a data loss prevention system in accordance with this invention. It is foreseeable that those skilled in the art can and will design alternative systems based on this disclosure that infringe on this invention as set forth in the claims below.

What is claimed is:

1. A method for detecting a possible transmission of information of interest over a computer network comprising:

receiving a document to be transmitted;

generating a set of trigrams of text of said document wherein each trigram in said set of trigrams comprises three strings of characters including a middle string of characters, a first string of characters immediately preceding said middle string of characters and, and a last string of characters immediately following said middle characters;

comparing said set of trigrams of said text of said document to a set of trigrams of said information of interest;

determining a number of matching trigrams from said comparing of said set of trigrams of said text to said set of trigrams of said information of interest;

comparing said number of matching trigrams to a threshold value; and

detecting a possible transmission of said information of interest in response to said number of matching trigrams being greater than or equal to said threshold value.

2. The method of claim 1 further comprising:

generating a notification of possible transmission message responsive to detecting said possible transmission.

3. The method of claim 2 further comprising:

transmitting said notification of possible transmission message to an administrator of said network.

4. The method of claim 1 further comprising:

applying a rule for handling said possible transmission in response to detecting a possible transmission of said information of interest.

5. The method of claim 1 wherein said step of receiving said document comprises:

receiving a plurality of packets to be transmitted; and

assembling text of said document from data in said plurality of packets.

6. The method of claim 1 wherein said step of generating said set of trigrams of said document comprises:

parsing text of said document to identify a plurality of strings of characters in said text of said document; and

generating said set of trigrams for said document from said plurality of strings of characters in said document.

7. The method of claim 6 wherein said step of generating said set of trigrams of said document further comprises:

pre-processing said document prior to parsing said text.

8. The method of claim 6 wherein said step of generating said set of trigrams of said received document further comprises:

searching said identified plurality of strings for replaceable strings; and

inserting replacement strings for each replaceable string found during said searching in said text of said document wherein said replacement strings are included in said trigrams generated for said document.

9. The method of claim 6 further comprising:

storing said set of trigrams of said text of said received document in a test file.

10. The method of claim 1 wherein each string of characters in each trigram in said set of trigrams comprises one character.

11. The method of claim 1 wherein each of said string of characters includes an arbitrary number of characters.

12. The method of claim 11 wherein said set of trigrams includes a group of trigrams for each sentence in said document.

13. The method of claim 12 wherein said set of trigrams excludes each trigram that includes string from a plurality of sentences.

14. The method of claim 1 further comprising:

comparing said set of trigrams of said text of said document to a plurality of sets of trigrams wherein each of said plurality of sets of trigrams is a set of trigrams for one of a plurality of documents of interest and wherein each of said plurality of documents of interest includes information of interest;

counting a number of matches between each of said plurality of sets of trigrams and said set of trigrams of said text of said document;

comparing said number of matches between each of said plurality of sets of trigrams and said set of trigrams of said text of said document to a threshold value; and

indicating a possible match between said document and one of said plurality of documents of interest in response to said number of matches between said one of said plurality of sets of trigrams of said one of said plurality of documents of interest and said set of trigrams of said text of said document.

15. The method of claim 14 further comprising:

determining a group of said plurality of documents of interest to compare to said text of said document.

16. The method of claim 15 wherein said step of determining said group of said plurality of documents of interest to compare comprises:

comparing each trigram in said set of trigrams of said text of said document to each trigram in an index set of trigrams;

determining said group of said plurality of documents to compare based upon matches between trigrams in said set of trigrams of said text of said document and said index set of trigrams.

17. The method of claim 16 wherein said step of determining said group of said plurality of documents of interest to compare further comprises:

counting a number of matches of each said trigram in said index set of trigrams with trigrams in said set of trigrams of said text of said document wherein said group of document of interest to compare is determined from said number of matches for each said trigram in said index set of trigrams.

18. The method of claim 17 wherein each said trigram in said index set of trigram has an associated number of occurrences in each of said plurality of sets of trigrams for said plurality of documents of interest.

19. The method of claim 18 further comprising:

selecting a one of said plurality of documents of interest for said group to compare to said text of said document responsive to said number of matches to an associated trigram in said index set of trigram that matches said number of occurrences of said trigram in said one of plurality of documents.

20. The method of claim 18 further comprising:

selecting a one of said plurality of documents of interest for said group of said plurality of documents to compare to said text of said document responsive to said number of matches between said set of trigrams of said text of said document and an associated trigram in said index set of trigrams being within a predetermined deviation from said number of occurrences of said trigram in a set of trigrams for said one of said plurality of documents.

21. The method of claim 16 further comprising:

generating said index set of trigrams.

22. The method of claim 21 wherein said step of generating said index set of trigrams comprises:

reading a trigram from one of said plurality of sets trigrams corresponding to one of said plurality of documents of interest;

determining whether said read trigram is included in said index set of trigrams;

determining whether said one of said plurality of documents of interest is included in a list of documents associated with said read trigram responsive to a determination said read trigram is included in said index set; and

adding said one of said plurality of documents of interest to a list of documents corresponding to said read trigram in said index set responsive to said one of said plurality of documents not being included in said list.

23. The method of claim 22 wherein said step of generating said index set of trigrams further comprises:

incrementing a counter associated with said one of said plurality of documents in said list responsive to a determination that said one of said plurality of documents is included in said list.

24. The method of claim 22 wherein said step of generating said index set of trigrams further comprises:

determining whether said read trigram is to be added to said index set of trigrams responsive to a determination that said read trigram is not included in said index set of trigrams; and

adding said read trigram to said index set of trigrams responsive to a determination that said read trigram is to be added.

25. The method of claim 24 wherein said step of generating said index set of trigrams further comprises:

adding said one of said plurality of documents of interest to a list of documents

corresponding to said read trigram in said index set responsive to adding said read trigram to said index set.

26. The method of claim 22 wherein said steps of generating said index set of trigrams is performed on each trigram of said set of trigrams for said one of said plurality of documents of interest.

27. The method of claim 12 further comprising:

adding a new document of interest to said plurality of documents of interest.

28. The method of claim 27 wherein said step of adding said new document of interest comprises:

generating a set of trigrams for said new document.

29. The method of claim 28 wherein said step of generating said set of trigrams for said new document comprises:

parsing text of said new document to identify a plurality of strings of characters in said text of said new document; and

generating said set of trigrams for said document from said plurality of strings of characters in said document.

30. The method of claim 29 wherein said step of generating said set of trigrams of said new document further comprises:

pre-processing said new document prior to parsing said text.

31. The method of claim 29 wherein said step of generating said trigrams of said new document further comprises:

searching said identified plurality of strings for replaceable strings; and

inserting replacement strings for each replaceable string found during said searching of said text of said document wherein said replacement strings are included in said trigrams generated for said document.

32. The method of claim 31 further comprising:

storing said set of trigrams of said new document in a trigram file.

33. The method of claim 32 further comprising:

storing said trigram file for said new document in a library of documents of interest storing said plurality of sets of trigrams of said plurality of documents of interest.

34. A product for detecting a possible transmission of information of interest over a computer network comprising:

instructions for a directing a processing system to:

receive a document to be transmitted,

generate a set of trigrams of text of said document wherein each trigram in said set of trigram comprises three strings of characters including a middle string of characters, a first string of characters immediately preceding said middle string of characters and, and a last string of characters immediately following said middle string of characters,

compare said set of trigrams of said text of said document to a set of trigrams of information of interest,

determine a number of matching trigrams from said comparing of said set of trigrams of said text of said document to said set of trigrams of said information of interest,

compare said number of matching trigrams to a threshold value, and

detect a possible transmission of said information of interest responsive to said number of matching trigrams being greater than or equal to said threshold value; and

a media readable by said processing unit for storing said instructions.

35. The product of claim 34 wherein said instructions further comprise:

instructions for directing said processing system to generate a notification of possible transmission message responsive to detecting said possible transmission.

36. The product of claim 35 wherein said instructions further comprise:

instructions for directing said processing system to transmit said notification of possible transmission message to an administrator of said network.

37. The product of claim 34 wherein said instructions further comprise:

instructions for directing said processing system to apply a rule for handling said possible transmission in response to detecting said possible transmission of information of interest.

38. The product of claim 34 wherein said instructions to receive said document comprise:

instructions for directing said processing system to:

receive a plurality of packets to be transmitted, and

assemble text of said document from data in said plurality of packets.

39. The product of claim 34 wherein said instructions of generate said set of trigrams of said received document comprises:

instructions for directing said processing system to:

parse text of said document to identify a plurality of strings of characters in said text of said document, and

generate said set of trigrams for said document from said plurality of strings of characters in said document.

40. The product of claim 39 wherein said instructions to generate said set of trigrams of said document further comprise:

instructions directing said processing system to pre-process said document prior to parsing said text.

41. The product of claim 39 wherein said instructions to generate said set of trigrams of said received document further comprise:

instructions for directing said processing system to:

search said identified plurality of strings for replaceable strings, and

insert replacement strings for each replaceable string found during said searching of said text of said document wherein said replacement strings are included in said trigrams generated for said document.

42. The product of claim 39 wherein said instructions further comprise:

instructions for directing said processing system to store said set of trigrams of said text of said received document in a test file.

43. The product of claim 34 wherein each string of characters in each trigram in said set of trigrams comprises one character.

44. The product of claim 34 wherein each of said string of characters includes an arbitrary number of characters.

45. The product of claim 44 wherein said set of trigrams includes a group of trigrams for each sentence in said document.

46. The product of claim 34 wherein said set of trigrams excludes each trigram that includes string from a plurality of sentences.

47. The product of claim 34 wherein said instructions further comprise:

instructions for directing said processing system to:

compare said set of trigrams of said text of said document to a plurality of sets of trigrams wherein each of said plurality of sets of trigrams is a set of trigrams for one of a plurality of documents of interest and wherein each of said plurality of documents of interest includes information of interest,

count a number of matches between each of said plurality of sets of trigrams and said set of trigrams of said text of said document,

compare said number of matches between each of said plurality of sets of

trigrams and said set of trigrams of said text of said document to a threshold value , and

indicate a possible match between said document and one of said plurality of documents of interest in response to said comparing of said threshold value to said number of matches between said one of said plurality of sets of trigrams of said one of said plurality of documents and said set of trigrams of said text of said document.

48. The product of claim 47 wherein said instructions further comprise:

instructions for directing said processing system to determine a group of said plurality of documents of interest to compare to said text of said document.

49. The product of claim 48 wherein said instructions to determine said group of said plurality of documents of interest to compare comprise:

instructions for directing said processing system to:

compare each trigram in said set of trigrams of said text of said document to each trigram in an index set of trigrams, and

determine said group of said plurality of documents to compare based upon matches between trigrams in said set of trigrams of said text of said document and said index set of trigrams.

50. The product of claim 49 wherein said instructions to determine said group of said plurality of documents of interest to compare further comprise:

instructions directing said processing system to count a number of matches of each said trigram in said index set of trigrams with trigrams in said set of trigrams of said text of said document wherein said group of document of interest to compare is

determined from said number of matches for each said trigram in said index set of trigrams.

51. The product of claim 50 wherein each said trigram in said index set of trigram has an associated number of occurrences of said trigram in each set of said plurality of sets of trigrams of each of said plurality of documents of interest that includes said trigram.

52. The product of claim 51 wherein said instructions further comprise:

instructions for directing said processing system to:

select a one of said plurality of documents of interest for said group to compare to said text of said document responsive to said number of matches to an associated trigram in said index set of trigram matches said number of occurrences of said trigram in said one of plurality of documents.

53. The product of claim 51 wherein said instructions further comprise:

instructions for directing said processing system to:

select a one of said plurality of documents of interest for said group of said plurality of documents to compare to said text of said document responsive to said number of matches between said set of trigrams of said text of said document and an associated trigram in said index set of trigrams being within in a predetermined deviation from said number of occurrences of said trigram in said one of plurality of documents.

54. The product of claim 49 wherein said instructions further comprise:

instructions for directing said processing system to generate said index set of trigrams.

55. The product of claim 54 wherein said instructions to generate said index set of trigrams comprise:

instructions for directing said processing system to:

read a trigram from one of said plurality of sets trigrams corresponding to one of said plurality of documents of interest,

determine whether said read trigram is included in said index set of trigrams,

determine whether said one of said plurality of documents of interest is included in a list of documents associated with said read trigram responsive to a determination, and

add said one of said plurality of documents of interest to a list of documents corresponding to said read trigram responsive to said one of said documents not being included in said list.

56. The product of claim 55 wherein said instructions to generate said index set of trigrams comprise:

instructions for directing said processing system to:

increment a counter associated with the said one of said plurality of documents in said list responsive to a determination that said one of said plurality of documents is included in said list.

57. The product of claim 55 wherein said instructions to generate said index set of trigrams further comprise:

instructions for directing said processing system to:

determine whether said read trigram is to be added to said index set of trigrams responsive to a determination that said read trigram is not included in said index set of trigrams, and

add said read trigram to said index set of trigrams responsive to a determination that said read trigram is to be added.

58. The product of claim 57 wherein said instructions to generate said index set of trigrams further comprise:

instructions for directing said processing unit to:

add said one of said plurality of documents of interest to a list of documents corresponding to said read trigram responsive adding said read trigram to said index set.

59. The product of claim 55 wherein said instructions to generate said index set of trigrams is performed on each trigram of said set of trigrams for said one of said plurality of documents of interest.

60. The product of claim 47 wherein said instructions further comprise:

instructions for directing said processing system to add a new document of interest to said plurality of documents of interest.

61. The product of claim 60 wherein said instructions to add said new document of interest comprise:

instructions for directing said processing system to generate a set of trigrams for said new document.

62. The product of claim 61 wherein said instructions to generate said set of trigrams for said new document comprise:

instructions for directing said processing system to:

parse text of said new document to identify a plurality of strings of characters in said text of said new document, and

generate said set of trigrams for said document from said plurality of strings of characters in said document.

63. The product of claim 62 wherein said instructions to generate said set of trigrams of said new document further comprise:

instructions directing said processing system to pre-process said document prior to parsing said text.

64. The product of claim 62 wherein said instructions to generate said trigrams of said new document further comprise:

instructions for directing said processing system to:

search said identified plurality of strings for replaceable strings, and

insert replacement strings for each replaceable string found during said searching into said text of said document wherein said replacement strings are included in said trigrams generated for said document.

65. The product of claim 64 wherein said instructions further comprise:

instructions for directing said processing system to store said set of trigrams of said new document in a trigram file.

66. The product of claim 65 wherein said instructions further comprise:

instructions for directing said processing system to store said trigram file for said new document in a library of documents of interest storing said plurality of sets of trigrams of said plurality of documents of interest.

67. A security system for a computer network that detects a possible transmission of information of interest over said computer network comprising:

circuitry configured to receive a document to be transmitted;

circuitry configured to generate a set of trigrams of text of said document wherein each trigram in said set of trigram comprises three strings of characters including a middle string of characters, a first string of characters immediately preceding said middle string of characters and, and a last string of characters immediately following said middle string of characters;

circuitry configured to compare said set of trigrams of said text of said document to a set of trigrams of said information of interest;

circuitry configured to determine a number of matching trigrams from said comparing of said set of trigrams of said text to said set of trigrams of said information interest;

circuitry configured to compare said number of matching trigrams to a threshold value, and

circuitry configured to detect a possible transmission of information of interest response to said number of matching trigrams being greater than or equal to said threshold value.

68. The system of claim 67 further comprising:

circuitry configured to generate a notification of possible transmission message responsive to detecting said possible transmission.

69. The system of claim 68 further comprising:

circuitry configured to transmit said notification of possible transmission message to an administrator of said network.

70. The system of claim 67 further comprising:

circuitry configured to apply a rule for handling said possible transmission in response to detecting a possible of transmission of information of interest.

71. The system of claim 67 wherein said circuitry configured to receive said document comprises:

circuitry configured to receive a plurality of packets to be transmitted; and

circuitry configured to assemble text of said document from data in said plurality of packets.

72. The system of claim 67 wherein said circuitry configured to generate said set of trigrams of said document comprises:

circuitry configured to parse text of said document to identify a plurality of strings of characters in said text of said document; and

circuitry configured to generate said set of trigrams for said text of said document from said plurality of strings of characters in said document.

73. The system of claim 72 wherein said circuitry configured to generate said set of trigrams of said document further comprises:

circuitry configured to pre-process said document prior to parsing said text.

74. The system of claim 72 wherein said circuitry configured to generate said set of trigrams of said document further comprises:

circuitry configured to search said identified plurality of strings for replaceable strings; and

circuitry configured to insert replacement strings for each replaceable string found during said searching of said text of said document wherein said replacement strings are included in said trigrams generated for said document.

75. The system of claim 72 further comprising:

circuitry configured to store said set of trigrams for said text of said document in a test file.

76. The system of claim 67 wherein each string of characters in each trigram in said set of trigrams comprises one character.

77. The system of claim 67 wherein each of said string of characters includes an arbitrary number of characters.

78. The system of claim 67 wherein said set of trigrams includes a group of trigrams for each sentence in said document.

79. The system of claim 78 wherein said set of trigrams excludes each trigram that includes string from a plurality of sentences.

80. The system of claim 78 further comprising:

circuitry configured to compare said set of trigrams of said text of said document to a plurality of sets of trigrams wherein each of said plurality of sets of trigrams is a set of trigrams for one of a plurality of documents of interest and wherein each of said plurality of documents of interest includes information of interest;

circuitry configured to count a number of matches between each of said plurality of sets of trigrams and said set of trigrams of said text of said document;

circuitry configured to compare said number of matches between each of said plurality of sets of trigrams and said set of trigrams of said text of said document to a threshold value; and

circuitry configured to indicate a possible match between said document and one of said plurality of documents of interest in response said number of matches between said one of said plurality of sets of trigrams of said one of said plurality of documents and said set of trigrams of said text of said document.

81. The system of claim 80 further comprises:

circuitry configured to determine a group of said plurality of documents of interest to compare to said text of said document.

82. The system of claim 81 wherein said circuitry configured to determine said of said plurality of documents of interest to compare comprises:

circuitry configured to compare each trigram in said set of trigrams of said text of

said document to each trigram in an index set of trigrams; and

circuitry configured to determine said group of said plurality of documents to compare based upon matches between trigrams in said set of trigrams of said text of said document and said index set of trigrams.

83. The system of claim 82 wherein said circuitry configured to determine said group of said plurality of documents of interest to compare further comprises:

circuitry configured to count a number of matches of each said trigram in said index set of trigrams with trigrams in said set of trigrams of said text of said document wherein said group of document of interest to compare is determined from said number of matches for each said trigram in said index set of trigrams.

84. The system of claim 83 wherein each said trigram in said index set of trigram has an associated number of occurrences of said trigram in each set of said plurality of sets of trigrams of each of said plurality of documents of interest that includes said trigram.

85. The system of claim 84 further comprising:

circuitry configured to select a one of said plurality of documents of interest for said to compare to said text of said document responsive to said number of matches to an associated trigram in said index set of trigram matches said number of occurrences of said trigram in said one of plurality of documents.

86. The system of claim 84 wherein said further comprises:

circuitry configured to select a one of said plurality of documents of interest for said group of said plurality of documents to compare to said text of said document

responsive to said number of matches between said set of trigrams of said text of said document and an associated trigram in said index set of trigrams being within in a predetermined deviation from said number of occurrences of said trigram in said one of plurality of documents.

87. The system of claim 82 further comprising:

circuitry configured to generate said index set of trigrams.

88. The system of claim 87 wherein said circuitry configured to generate said index set of trigrams comprises:

to:

circuitry configured to read a trigram from one of said plurality of sets trigrams corresponding to one of said plurality of documents of interest;

circuitry configured to determine whether said read trigram is included in said index set of trigrams;

circuitry configured to determine whether said one of said plurality of documents of interest is included in a list of documents associated with said read trigram responsive to a determination; and

circuitry configured to add said one of said plurality of documents of interest to a list of documents corresponding to said read trigram responsive to said one of said documents not being included in said list.

89. The system of claim 88 wherein said circuitry configured to generate said index set of trigrams comprises:

circuitry configured to increment a counter associated with the said one of said

plurality of documents in said list responsive to a determination that said one of said plurality of documents is included in said list.

90. The system of claim 88 wherein said circuitry configured to generate said index set of trigrams further comprises:

circuitry configured to determine whether said read trigram is to be added to said index set of trigrams responsive to a determination that said read trigram is not included in said index set of trigrams; and

circuitry configured to add said read trigram to said index set of trigrams responsive to a determination that said read trigram is to be added.

91. The system of claim 90 wherein said circuitry configured to generate said index set of trigrams further comprises:

circuitry configured to add said one of said plurality of documents of interest to a list of documents corresponding to said read trigram responsive adding said read trigram to said index set.

92. The system of claim 88 wherein said circuitry configured to generate said index set of trigrams is performed on each trigram of said set of trigrams for said one of said plurality of documents of interest.

93. The system of claim 91 further comprising:

circuitry configured to add a new document of interest to said plurality of documents of interest.

94. The system of claim 93 wherein said circuitry configured to add said new document of interest comprises:

circuitry configured to generate a set of trigrams for said new document.

95. The system of claim 94 wherein said circuitry configured to generate said set of trigrams for said new document comprises:

circuitry configured to parse text of said new document to identify a plurality of strings of characters in said text of said new document; and

circuitry configured to generate said set of trigrams for said document from said plurality of strings of characters in said document.

96. The system of claim 95 wherein said circuitry configured to generate said set of trigrams of said new document further comprises:

circuitry configured to pre-process said document prior to parsing said text.

97. The system of claim 95 wherein said circuitry configured to generate said trigrams of said new document further comprises:

circuitry configured to search said identified plurality of strings for replaceable strings; and

circuitry configured to insert replacement strings for each replaceable string found during said searching into said text of said document wherein said replacement strings are included in said trigrams generated for said document.

98. The system of claim 97 further comprising:

circuitry configured to store said set of trigrams of said new document in a trigram file.

99. The system of claim 98 further comprising:

circuitry configured to store said trigram file for said new document in a library of documents of interest storing said plurality of sets of trigrams of said plurality of documents of interest.