RNPS No. 2143
ISSN 2072-6260
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Minería de Datos

SERIE GRIS

Métodos para la determinación de la sensibilidad de los documentos: un estado del arte

Saturnino Job Morales Escobar,
Osvaldo Andrés Pérez García y
José Ruiz-Shulcloper

RT_036 **octubre 2016**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2143
ISSN 2072-6260
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Minería de Datos

SERIE GRIS

Métodos para la determinación de la sensibilidad de los documentos: un estado del artes

Saturnino Job Morales Escobar,
Osvaldo Andrés Pérez-García y
José Ruiz-Shulcroper

RT_036 **octubre 2016**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Métodos para la determinación de la sensibilidad de los documentos: un estado del arte

Saturnino Job Morales Escobar[1], Osvaldo Andrés Pérez García[2] y José Ruiz-Shulcloper[3]

[1]Centro Universitario UAEM Valle de México
Universidad Autónoma del Estado de México (UAEM)
Estado de México, México
sjmoralese@uaemex,mx
[2] Equipo de Investigaciones de Minería de Datos, CENATAV - DATYS, La Habana, Cuba
osvaldo.perez@cenatav.co.cu
[3]Equipo de Investigaciones de Reconocimiento de Patrones, CENATAV - DATYS, La Habana, Cuba
jshulcloper@cenatav.co.cu

RT_036, Serie Gris, CENATAV - DATYS
Aceptado: 12 de septiembre de 2016

Resumen. En este reporte técnico se presenta un estado del arte acerca del problema de la determinación de la sensibilidad en documentos para enfrentar el problema de la fuga de información. Motivados por la relevancia, actualidad y complejidad que presenta esta problemática se hace un análisis enfocado a la determinación y protección de información sensible en documentos, los principales enfoques y métodos para la solución de este problema. Partiendo del análisis realizado y teniendo en cuenta las necesidades prácticas planteadas por los expertos de las áreas de posible aplicación, se bosquejan las principales metas de un proyecto de investigación que aborde esta temática. Además, se plantean las posibles extensiones que estos estudios pueden tener en áreas similares de aplicación partiendo de otros portadores de la información.

Palabras clave: fuga de información, sensibilidad de documentos, sistemas de protección de información, clasificación supervisada.

Abstract. In this technical report, we present a state of the art about the problem of determining the sensitivity of documents to address the problem of information leakage. Motivated by the relevance, actuality and complexity that has this problem, we analyze focused on the identification and protection of sensitive information in documents, the principal approaches and methods for solving this problem. Based on this analysis and taking into account the practical needs raised by consulted experts in the areas of possible applications, the main goals of a research project that addresses this issue are outlined. Furthermore, we formulate the possible extensions that these studies may have to similar areas of application from other carriers of information.

Keywords: information leakage, sensitive documents, information protection systems, supervised classification.

1 Introducción

Sin duda, uno de los recursos más valiosos para cualquier organización es la información y los datos que utilizan en todos sus procesos, sin embargo, debido a la propia naturaleza en la que ambos se presentan, procesan, envían o almacenan, se deben tratar de diferentes maneras. Es de particular interés la información y datos que se pueden considerar sensibles, o de uso restringido por su confidencialidad, y que son utilizados en las actividades que involucran el uso de computadoras y dispositivos móviles. Estos dispositivos, son capaces de almacenar y/o procesar *objetos de información* (datos, documentos, imágenes, videos, audio, etc.), en diversos formatos, y en diferentes plataformas. Ante esta situación, las organizaciones, entre las que podemos incluir, empresas, institutos de investigación, industrias, o instancias de gobierno, deberían tomar en cuenta que la fuga de información y datos sensibles es un riego que se ha incrementado durante los últimos años y que su impacto puede incluso facilitar la comisión de delitos informáticos como el fraude. Por ejemplo, tan solo en la industria, considerando los fraudes de telecomunicaciones y de acuerdo con la Communications Fraud Control Association (ACCP), en 2011, las pérdidas globales por fraude se estimó que le costaron más de $40 mil millones (USD) al año, considerando solamente los realizados mediante sistemas de correo de voz, robo de identidad, beneficios internacionales por participación, por bypass[1] y por tarjetas de crédito [1], amenazas posiblemente materializadas después de ocurrida una fuga de datos conteniendo información sensible.

Por otra parte, en los procesos de generación de datos e información sensible, manejo o almacenamiento, es difícil tener la certeza que el personal sigue las políticas de seguridad o que al utilizar alguna aplicación de ayuda al aseguramiento de las mismas, los usuarios las cumplan. En el caso de la fuga de información, se tiene gran cantidad de reportes de incidentes y de acuerdo a la organización Nonprofit Consumer Privacy Rights Clearinghouse, un total de 227052199 registros conteniendo información personal sensible estuvieron involucrados en violaciones de seguridad en los Estados Unidos entre enero de 2005 y mayo de 2008 [2]. En ese mismo rubro se puede consultar WikiLeaks, sitio en el que se expone información clasificada como secreta, de gobiernos y organizaciones, lo que ha causado gran impacto para algunos gobiernos y empresas privadas [3].

De acuerdo con [4] la fuga de información o datos contenidos en objetos de información puede ser el resultado de acciones deliberadas o errores espontáneos, la cual se puede incrementar por su transmisión interna o externa vía correo electrónico, mensajes instantáneos, formularios de páginas web, entre otros medios y aún más, el riesgo se acrecienta cuando los datos e información sensible son compartidos por clientes, socios comerciales, empleados externos, etc.

De esta manera, la fuga de datos e información es considerada como un problema emergente de amenaza a la seguridad de las organizaciones, toda vez que el número de incidentes y el costo que implican continúan creciendo.

En [5] definen la fuga de datos como la distribución accidental o involuntaria de datos sensibles a una entidad no autorizada. Los datos sensibles para una organización incluyen la propiedad intelectual, información financiera, información de pacientes, datos personales, entre otros. Así, es de vital importancia continuar el desarrollo de herramientas que protejan la información.

Hasta el momento, con la intención de resolver este problema, se han desarrollado sistemas para la Prevención de Fuga de Datos (Data Leakage Prevention, DLP) también conocidos como sistemas para la prevención de pérdida de datos, prevención de pérdida/fuga de información, prevención de extrusión, monitoreo y filtrado/protección de contenido, los cuales están diseñados para identificar, monitorear y proteger datos confidenciales y detectar su mal uso basados en reglas predefinidas. De esta manera, los sistemas DLPs, se agregan a las medidas de seguridad tradicionales como los Sistemas de Detección de Intrusos (IDSs, por su sigla en inglés); Sistemas de Prevención de Intrusión (IPSs, por su sigla en inglés), redes privadas virtuales (VPNs, por su sigla en inglés), entre otras, las cuales funcionan adecuadamente para datos bien definidos, estructurados y constantes [4].

[1] Desvío para evitar las medidas de seguridad implementadas y penetra un sistema.

Nuestro país se ha planteado como una necesidad para el desarrollo, la informatización de la sociedad, lo cual implica ampliar el uso de Internet como un medio de acceso público, fortalecer su empleo en instituciones del estado y del gobierno, así como en los medios de comunicación y centros de interés social y económico, tales como universidades, escuelas, empresas, entidades gubernamentales, entre otros. Esto trae acompañado nuevos desafíos en materia de seguridad.

Una de las principales estrategias para la puesta en marcha de esta idea, consiste en integrar las redes institucionales del país, actualmente desconectadas entre sí. Para ello se hace imprescindible la creación de una infraestructura que facilite el intercambio de tráfico de nacional e internacional. Esto traerá consigo un incremento notable de los riesgos de fuga o pérdida de información sensible, y la necesidad de incorporar medidas de protección, como los DLPs, que posibiliten minimizar tales riesgos.

Continuando con la búsqueda de solución al problema de DLP, en este trabajo se presenta un análisis de estos sistemas, las técnicas para la solución de la fuga de datos y documentos sensibles, se profundiza en las propuestas para la determinación de la sensibilidad de datos y documentos y se formulan las líneas en las que los autores consideran se deben desarrollar nuevos métodos para la determinación de la sensibilidad de la información en documentos.

2 Sistemas para la prevención de pérdida/fuga de información (DLP)

En [5] se presenta un estudio sobre las propuestas de solución al problema de DLP, visto como un problema de la seguridad de la información, en donde la confidencialidad, la integridad y la disponibilidad de los datos son conceptos básicos de seguridad. En el contexto de la fuga y mal uso de datos, identifican como actores físicos principales a la organización, los atacantes y los bienes de la organización, y como actores lógicos a los mecanismos de seguridad, vulnerabilidades, amenazas y riesgos de seguridad. Clasifican las soluciones DLP con base en la fuente de la fuga, el estado de los datos, canal de fuga, esquema de despliegue, enfoques de prevención y detección, y por las acciones tomadas para remediar los problemas.

Por otra parte, en [5] realizan un estudio sobre los medios tecnológicos para tratar el problema de DLP y los clasifican en las siguientes categorías: medidas de seguridad estándar, medidas inteligentes o avanzadas de seguridad, control de acceso y cifrado, y sistemas DLPs designados.

Las medidas de seguridad estándar incluyen firewalls, IDS y software de antivirus, las que ofrecen protección en alguna medida contra ataques externos e internos. Por otra parte, en las medidas de seguridad inteligentes se incluyen, el aprendizaje automático y algoritmos de razonamiento temporal para detectar accesos anormales a los datos, verificación basada en la actividad (pulsaciones de teclado o movimiento del ratón), la detección de patrones de intercambio de correo electrónico anormales, y la aplicación del concepto honeypot[2] para detectar accesos maliciosos. En la categoría de control de dispositivos, control de acceso, y cifrado, los métodos son utilizados para evitar el acceso de usuarios no autorizados. Finalmente los sistemas DLPs designados, están destinados a detectar y prevenir intentos de copiar o enviar datos sensibles sin autorización, con o sin intención, principalmente por personal que está autorizado a acceder a la información sensible. Entre los métodos utilizados se mencionan las *huellas de objeto de información* (fingerprinting), aprendizaje automático y uso de reglas y expresiones regulares.

Con todo este panorama, desde el punto de vista comercial Brian y Neil [6] comentan que las empresas que ofrecen sistemas DLPs continúan evolucionando para soportar capacidades tanto de análisis de contenido como de contexto para la detección y prevención, además de mantener la compatibilidad con los líderes de ese mercado, de tal manera que cubran casos de uso más allá del cumplimiento de las normativas y la protección de la propiedad intelectual. En su planeación

[2] Pote de miel: Sistema o parte de un sistema desarrollado y desplegado intencionalmente para tentar y atraer a un intruso y aprender, por ejemplo, de sus técnicas de ataques.

estratégica, consideran que para el año 2018, el 90% de las organizaciones implantará dentro de sus sistemas de seguridad al menos una forma de DLP, contra el 50% actual, y de ese 90%, menos del 10% tendrá una política de seguridad de datos bien definida, de casi cero en la actualidad.

Ante tantas alternativas y propuestas, en [4], para realizar un análisis comparativo de técnicas implantadas en los DLPs se toman las siguientes consideraciones:

- La división entre las propuestas académicas y las que ofrece la industria.
- La clasificación de los datos en tres posibles estados en los que se pueden encontrar: en uso, en tránsito o en reposo.
- El tipo de análisis lo distinguen entre los que realizan el análisis de contexto y los de contenido.
- Las acciones remediales que realiza.
- Canales por los cuales se puede presentar la fuga de datos.
- El factor humano.
- Derechos de acceso.
- Cifrado y Esteganografía[3].
- Modificación de datos.
- Escalabilidad e integración.
- Clasificación de datos en tráfico normal y confidencial.

En [6] con el objetivo de clasificar proveedores de DLPs en el cuadrante mágico de Gartner[4], definen el mercado de la prevención de pérdida de datos como aquellas tecnologías que, como una función esencial, realizan tanto inspección de contenido como análisis contextual de datos, ya sea en reposo (Data at Rest), ubicados en las instalaciones, en aplicaciones o almacenados en la nube, en movimiento sobre la red (Data in motion), o en uso en un dispositivo de punto final (Data at End Points). Las soluciones de DLPs pueden ejecutar respuestas basadas en las políticas y reglas definidas -que van desde la notificación hasta al bloqueo activo- para hacer frente al riesgo de fugas fortuitas o accidentales, o a la exposición de datos sensibles fuera de los canales autorizados.

Así, en una primera instancia, clasifican a las tecnologías de prevención de pérdida de datos en dos categorías: soluciones DLPs empresariales y soluciones DLPs integradas, dependiendo de las soluciones, forma y facilidades que ofrecen.

Mientras, las DLPs empresariales, funcionan como una solución completa para descubrir datos sensibles dentro de una organización y mitigar el riesgo de su pérdida en el punto final, en el almacenamiento y por la red; las soluciones integradas, por lo general se centran en implementar un conjunto limitado de características de los DLPs enfocándose en un conjunto reducido de casos de uso relacionados con el cumplimiento de normativas y de la propiedad intelectual básica, donde los datos a proteger son fácilmente identificables y la política de remediación es sencilla.

Los criterios que los proveedores deben cumplir para ser incluidos en el cuadrante mágico de Gartner son:

- 8 millones de dólares anuales de ventas de su producto DLP empresarial;
- capacidad para detectar contenido sensible en tráfico de red sin necesidad de un agente de punto final;
- capacidad para detectar y descubrir contenido sensible tanto en datos en reposo como en datos en uso;
    - las soluciones más completas serán aquellas que logren resolver el problema para los tres escenarios: de red, de punto final y el descubrimiento de datos;

[3] Ciencia que estudia cómo esconder información de forma tal que sea muy difícil detectar el canal de comunicaciones. De acuerdo al contexto, la información podrá ocultarse en contenedores físicos o digitales. En este último caso los contenedores de información son objetos de información.

[4] El cuadrante mágico de Gartner, es una representación gráfica de la clasificación de proveedores de Sistemas de Prevención de Fuga de Datos y es elaborado por la empresa Gartner, Inc.
http://www.gartner.com/technology/about.jsp

- tener una política centralizada y una consola de gestión de eventos;
- capacidad para detectar contenido sensible usando al menos tres de las siguientes técnicas: coincidencia parcial o exacta de documentos, huella de objetos de información en datos estructurados, análisis estadístico, coincidencia en expresiones regulares extendidas y análisis léxico y conceptual;
- soportar la detección de contenido de datos sensibles en datos estructurados y no estructurados usando registros o definiciones de datos descritos;
- capacidad para bloquear como mínimo las violaciones de las políticas que se producen a través de las comunicaciones por correo electrónico.

Por otra parte, para excluir un proveedor del cuadrante mágico consideran:

- que la solución dependa de la integración en otro producto, por ejemplo, servidor de correos;
- que las soluciones no tengan una interfaz única y centralizada y un repositorio para la gestión de flujo de eventos de trabajo, tanto para el descubrimiento de datos como para los DLPs de punto final y de red;
- que las soluciones solo utilicen mecanismos de detección de datos simples (por ejemplo, concordancia de palabras clave, vocabulario o expresiones regulares simples);
- soluciones con funciones basadas en redes que soportan menos de cuatro protocolos (por ejemplo, solo el correo electrónico SMTP, FTP y HTTP);
- soluciones que soportan políticas DLP por medio de la asignación de etiquetas de contenido a los objetos;
- soluciones que no pueden detectar accesos a contenido sensible en la red sin necesidad de instalar software DLP de punto final, y en particular, que no detectan datos en movimiento a sistemas o dispositivos no administrados.

El cuadrante tiene como eje horizontal la integridad de visión y como eje vertical la capacidad de ejecución, y los criterios de inclusión o exclusión son ajustados dependiendo del cambio en los mercados.

La integridad de la visión es evaluada en términos de las estrategias para el futuro y los proveedores se clasifican en función de la capacidad para mostrar un compromiso con la evolución de la tecnología DLP empresarial y la comprensión de las necesidades de negocio de los clientes de DLP. Así, los proveedores se deben centrar en la regulación impulsada por las organizaciones para identificar, localizar y controlar los datos sensibles almacenados en sus redes o al cruzar sus fronteras.

La ponderación de la integridad de la visión es influenciada por cuatro categorías básicas de capacidad: rendimiento en la red, rendimiento en punto final, desempeño en el descubrimiento de datos y de sus consolas de administración.

Con respecto a los criterios para evaluar la habilidad para ejecutar, se ubica al proveedor dependiendo de qué tanto cumple los requisitos del cliente en cuanto a características de capacidad/función, así como su capacidad para entregar y ejecutar el producto con un alto nivel de garantías de servicio y soporte al cliente.

Las calificaciones de los proveedores están influenciadas por la comprensión que el proveedor tiene del mercado, sus procesos para solicitar retroalimentación del cliente y la experiencia del cliente.

En estas condiciones, presentan diez empresas como las principales proveedoras de soluciones de DLP, las cuales quedan ubicadas en el cuadrante mágico de la siguiente forma: Symantec, Forcepoint, Digital Guardian e Intel Security como los líderes, seguidos por Fidelis Cybersecurity y GTB Technologies y al final a InfoWatch, Zecurion, Clearswift, Somansa. En el reporte resaltan las fortalezas y debilidades de cada producto en términos de los criterios señalados, sin embargo, sobre los métodos para la determinación de los datos sensibles, nosotros consideramos que se requiere de un estudio más profundo que lo presentado en el mencionado reporte.

Por todo lo anterior, es claro que los sistemas DLPs están dentro del conjunto de tecnologías de seguridad empleadas para atender los problemas relacionados con las amenazas internas, las cuales

son diseñadas con el propósito de prevenir de forma automática la fuga o pérdida de datos sensibles en cualquiera de los tres estados ya mencionados.

La arquitectura básica de los sistemas DLPs está constituida por tres módulos (ver Figura 1). El primero detecta si se está enviando, creando o accediendo a un documento[5] (para su impresión, copia, edición, envío por la red, etc.) sin importar su contenido. El segundo módulo analiza el documento detectado en el filtro, lo revisa y envía hacia el tercer módulo una valoración en correspondencia con la política establecida. Este último módulo responde permitiendo o bloqueando, si es necesario, las acciones sobre la información a proteger, emitiendo la alerta correspondiente.

Las características más importantes de cada módulo están definidas por la política de la entidad que aplica la solución DLP para proteger su información. Por ejemplo, el filtraje de los documentos tendrá en cuenta el Modelo de Amenazas[6] y los Vectores de Ataque[7] identificados en este, información que es dependiente del dominio. El análisis de los documentos detectados se realiza a nivel de contenido o de metadatos, ambos aspectos se relacionan estrechamente con el dueño de la información y de los ficheros, y por último, las respuestas que debe emitir el sistema estarán matizadas por el nivel de seguridad que se desee obtener con la aplicación del DLP, lo cual se expresa en la política de seguridad definida para el sistema.

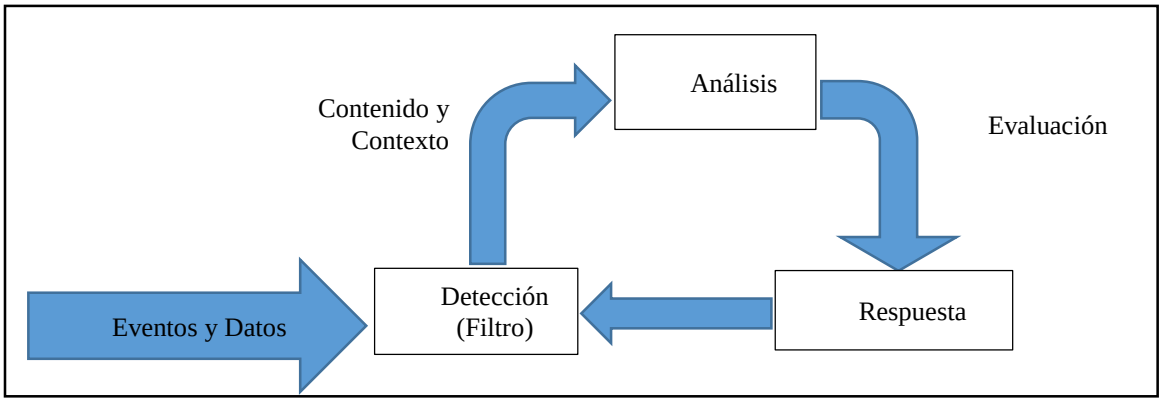

Fig. 1. Arquitectura básica de los DLP.

Desde el punto de vista técnico, las complejidades asociadas a los módulos de Detección y Respuesta están identificadas [7, 8], y se vinculan estrechamente con el soporte tecnológico del sistema informático sobre el que se implementa el DLP. Es en el módulo de Análisis donde se localizan los problemas teóricos relacionados con la determinación de la sensibilidad de la información a proteger.

El análisis en los sistemas DLPs estudiados se puede agrupar de acuerdo al enfoque que asumen. En la Tabla 1 se presenta un resumen a modo de taxonomía. Estos métodos, en su mayoría, se han desarrollado con el objetivo de evaluar el nivel de sensibilidad de la información, independientemente del momento en que se produce el evento que vincula al documento con una amenaza de seguridad. Están orientados a procesar todo el contenido, presuponiendo que la clasificación será booleana: Si o No.

[5] La elección del objeto de información “documento” es a los efectos de la explicación.
[6] Es en esencia una representación estructurada de todo lo que afecta a la seguridad de un sistema.
[7] Elementos imprescindibles para conformar el modelo de amenaza. Explican las posibles vías de entradas a los sistemas que pueden ser aprovechadas por los atacantes.

Tabla 1. Módulos de análisis de los DLPs.

Enfoque	Métodos y Descripción	Referencias
Contextual	Uso de metadatos asociados con los datos confidenciales, por ejemplo, en envío de datos, la fuente, el destino, tiempo, tamaño, formato, frecuencia, registro de temas. Los metadatos pueden ser usados en procesos o en patrones de transacciones y están basados en las políticas definidas. En [9] se propone un algoritmo para obtener características para detectar correos mal dirigidos.	[9, 10]
Contenido	**Expresiones regulares:** Conjunto de términos o caracteres usados para formar patrones de detección, típicamente son usados para detección parcial o exacta en números de seguridad social, tarjetas de crédito, registros personales y corporativos. Técnicas basadas en diccionarios específicos pueden acelerar y mejorar la detección significativamente.	[11-13]
	Clasificadores: Dependen considerablemente de una adecuada clasificación de los datos, de otra manera los DLPs no serán capaces de distinguir entre el tráfico confidencial y normal. La práctica usual es que el propietario de los datos es el responsable de la determinación de la sensibilidad de los datos y si deben o no ser protegidos. La mayoría de las soluciones se han basado en *etiquetas*, en *lista de palabras sucias* (*dirty list*) y las limitaciones que éstas tienen. También se asume que para permitir el acceso a los datos sensibles o moverlos entre diferentes dominios, todos los datos deben estar bien etiquetados con la correspondiente clasificación.	[14-16]
	Huellas de objetos de información: Son utilizadas especialmente en datos no estructurados para detectar coincidencia parcial o exacta. Es la técnica más común usada para detectar fuga de información, DLPs con funciones de hash como MD5, y SHA1 pueden alcanzar hasta el 100% de exactitud si los archivos no son alterados. Se han realizado propuestas para superar las alteraciones de datos y mantener la detección de datos sensibles en tránsito. En [Shu y Yao, 2013] se usa un algoritmo de marcas difusas para detectar fuga de datos en tráfico de redes.	[17]
	N-grama[8]: Utilizado ampliamente en procesamiento del lenguaje natural, en aprendizaje automático y en recuperación de información por términos pesados. Depende principalmente del análisis de frecuencia de términos y n-gramas en los documentos. Los primeros en usarlo en DLPs fueron Hart y Johnson para clasificar documentos empresariales en sensibles y no sensibles; utilizan Máquinas de Soporte Vectorial (Support Vector Machines, SVM por sus siglas en inglés) para clasificar tres tipos de datos: empresariales privados, públicos y no empresariales.	[18], [12], [17]
	Ponderación o pesado de términos: El pesado de términos es un método estadístico que indica la importancia de un término en un documento, usado en clasificación de textos y modelos de espacios vectoriales en donde los documentos son tratados como vectores y se utilizan funciones para determinar las frecuencias de los términos.	[19-21]

[8] Un n-grama es una subsecuencia de n elementos de una secuencia dada.

Varios son los problemas aun no resueltos, un grupo de ellos enmarcados en los aspectos técnicos y de seguridad [22] y otros agrupados de acuerdo al tipo de análisis que se realiza y el contexto de los documentos, ver Tabla 2. De la valoración realizada de la literatura académica sobresale un detalle que marca el actual estado de cosas en el campo de los sistemas DLPs, y es que existe distanciamiento entre la manera de abordar la Detección (ver Figura 1) de los eventos relacionados con los datos y los métodos o algoritmos de análisis que se proponen para evaluar la sensibilidad de esa información, desaprovechándose, la posibilidad de potenciar la detección de intentos reales e ilegales de extraer información sensible. La clave del problema radica en que para crear el contexto (imagen, por ejemplo) con información sensible, o no, embebida, no se realiza análisis alguno, perdiéndose la oportunidad para detectar una posible fuga de información.

Un aporte en ese sentido pudiera estar dirigido a proporcionar una solución a la detección de información sensible para su aplicación en el momento que está siendo creada esa información y antes de su inclusión en contextos asociados a los problemas descritos, por ejemplo, en la Tabla 2, lo cual permitiría disminuir el riesgo de fuga de información por esos conceptos.

Tabla 2. Problemas no resueltos de acuerdo al contexto.

Problema	Descripción
Escaneo de imágenes	Detección y extracción de textos introducidos en imágenes. El problema radica en ubicar el texto en la imagen y extraerlo.
Formatos de ficheros y protocolos de comunicaciones	Nuevos formatos y protocolos desarrollados para el intercambio de información y el almacenamiento que complejizan la detección de la información sensible.
Cifrado de documentos	La aplicación de técnicas de cifrado de información impide analizar su contenido.
Esteganografía	En imágenes y textos. El primer problema es detectar que hay información esteganografiada, y después debe intentar extraerla para su análisis.

3 Análisis de métodos para la determinación de la sensibilidad de documentos: ventajas y limitaciones

Como se ha visto, los sistemas DLPs incorporan un conjunto de capacidades para la detección y/o prevención de la fuga de datos sensibles tratando de cubrir la mayor cantidad de variantes, desde la aplicación de políticas hasta monitorear los medios por los cuales se puede presentar la fuga. Es de particular interés desde el punto de vista de la arquitectura de los sistemas DLPs el módulo de análisis, toda vez que en ese módulo se pueden ofrecer nuevos métodos y técnicas que fortalezcan la determinación del nivel de sensibilidad de los objetos de información, en este punto coincidimos con [4] y [16].

Como resultado de la evolución de los métodos, se espera que se pueda realizar la determinación de la sensibilidad de los documentos de manera automática y que la tarea de etiquetar los documentos deje de ser responsabilidad exclusiva del dueño de los datos quien en la mayoría de las ocasiones la realiza de manera manual con todos las desventajas que esto implica, entre otras, el tiempo que consume en esta tarea, el que puede dar como resultado una clasificación inconsistente al ser un proceso subjetivo que depende del conocimiento y entrenamiento del responsable de asignar la sensibilidad del documento.

Comentan estos autores sobre la tendencia de errores de sobre clasificación de documentos, en particular, sobre un testimonio de experto que reportó ante el congreso de los Estados Unidos que se estimaba que el 50% de los documentos militares relacionados con el gobierno fueron sobre clasificados [23]. El Mayor General Flynn, militar de inteligencia en Afghanistan, reportó que la sobre

clasificación obstaculizó los esfuerzos de inteligencia en Afghanistan [24]. Ambos problemas, la inconsistencia y la sobre clasificación, perjudican las actividades de los analistas de información, pues se impide o limita el acceso a información que no es clasificada pero que está en un documento clasificado, siendo necesario compartirla. Tanto el uso de los metadatos, como el uso de la identificación de los que acceden a la información, están basados en lo que se denomina Política de Seguridad.

En cuanto a los principales métodos mencionados, de contexto y de contenido, a continuación se presenta un análisis con la intención de detectar las ventajas y limitaciones de cada uno de ellos.

3.1 Análisis de contexto

Según el diccionario Oxford, "El contexto se puede definir como las circunstancias que forman el escenario de un evento, declaración o idea en términos de los cuales se puede entender plenamente"[9], mientras que en el diccionario de la Real Academia Española, "Contexto es el entorno lingüístico del que depende el sentido de una palabra, frase o fragmento determinados"[10]. Si se generaliza el alcance de esta definición, se puede decir que el contexto, es una extensión que proporcionan los metadatos sobre la generación y manipulación de los datos sensibles que contiene un documento.

Por ejemplo, en envío de datos, los metadatos que se podrían considerar incluirían: el origen de los datos, destino, tiempo de transmisión, tamaño del archivo, medio utilizado, formato, frecuencia, registro de temas, datos del usuario emisor, entre otros. Estos metadatos pueden ser usados en el desarrollo de procesos o en la generación de patrones de transacciones y normalmente son tomados con base en las políticas definidas por la organización.

En [13] especifican que el desarrollo de los esquemas de análisis basados en el contexto ha conducido al empleo de factores como: el propietario del fichero y permisos que tiene asignados, cuáles protocolos de red o formatos de ficheros cifrados emplea, cuál es el rol de ese usuario dentro de la empresa, qué servicios web utiliza, direcciones web, información asociadas a los dispositivos de tipo USB empleados (ejemplo: fabricante, número del modelo) o la aplicación de escritorio empleada para editar, leer o enviar la información. A partir de poseer este tipo de conocimiento puede orientarse el trabajo de descubrimiento de posibles canales de fuga de información aplicando la detección de anomalías.

Los sistemas DLPs que utilizan el análisis de contexto, comparten características con sistemas IDSs basados en anomalías, motivo por el cual emplean un conjunto de técnicas similares en la solución de ambos problemas. En [25] presentan una revisión de las técnicas de detección de intrusos aplicadas al cómputo en la nube. Esto sugiere que se deben identificar eventos que parecen ser anómalos en relación con el comportamiento normal del sistema. Entre las técnicas utilizadas para la detección de anomalías mencionan la minería de datos, modelos estadísticos y modelos ocultos de Markov.

De manera general, en el enfoque basado en anomalías, durante un período de tiempo, se recopilan datos del comportamiento de los usuarios legítimos, y luego se aplican pruebas estadísticas para compararlo con el comportamiento observado. En base a esta comparación, se determina si éste es legítimo o no. El elemento principal de este enfoque, es la generación de reglas de tal manera que se pueda reducir la proporción de falsas alarmas, tanto en la detección de nuevos ataques como de ataques ya conocidos.

[9] Oxford Dictionaries, "http://www.oxforddictionaries.com/definition/english/context", consultado: 7 de agosto de 2016.
[10] Diccionario de la Real Academia Española, "http://dle.rae.es/?id=AVBbFZW", consultado: 7 de agosto de 2016.

3.2 Análisis de contenido

Como se ha visto en el presente trabajo, la detección de objetos de información sensible, se puede abordar desde diferentes enfoques y utilizando distintos métodos. En este epígrafe, se abordará el problema desde el punto de vista del contenido de los objetos de información. La sensibilidad del contenido, normalmente está ligada al significado que puedan tener los datos. Es claro, que en sí mismo, cada dato, puede contener gran cantidad de información, sin embargo, ésta, se puede incrementar o decrementar si se relaciona con otros datos. En este sentido, y con base en el Diccionario de la Real Academia Española, "contenido"[11], tiene dos acepciones relacionadas con lo que se aborda en este trabajo: la primera "cosa que se contiene dentro de otra", y la segunda "en una obra literaria, tema o idea tratados, distintos de la elaboración formal". Los autores de este trabajo consideramos, que el contenido de un objeto de información, va a estar muy relacionado al contexto en el cual se origina dicho objeto, por ejemplo, en una institución bancaria, una cadena formada por combinaciones de 8 caracteres, en sí misma es no sensible, pero relacionada con las cadenas "clave" y "acceso" cambiaría a dato sensible. En las siguientes secciones se presentan métodos para tratar este reto.

3.2.1 Expresiones regulares

Este método es tomado de la Teoría de la Computación en donde se emplea para representar lenguajes regulares y en términos generales se toma un conjunto finito de símbolos para formar cadenas, las cuales son utilizadas para crear patrones de detección. Estos patrones son llamados expresiones regulares (ER) y son usados por las máquinas de búsqueda y en el procesamiento de textos para validar, generar, extraer o remplazar datos. Típicamente son usados en detección parcial o exacta en números de seguridad social, tarjetas de crédito, registros personales y corporativos. Técnicas basadas en diccionarios sobre campos específicos pueden acelerar y mejorar la detección de datos sensibles de manera significativa.

Una aplicación de las expresiones regulares la podemos encontrar en [11] donde son utilizadas con el fin de detectar la aparición de patrones críticos en las cargas útiles de los paquetes de red. Para esta tarea, deben realizar comparaciones de expresiones regulares en tiempo real, y aun cuando los autómatas finitos determinísticos (DFA) realizan la operación en un tiempo de orden lineal, los requerimientos de memoria pueden hacer prohibitivo su uso. En [26] proponen retardar la entrada para el DFA lo que proporciona un equilibrio entre los requisitos de memoria del DFA y el número de estados visitados por cada carácter procesado, que corresponde directamente con el ancho de banda de memoria necesaria para evaluar las expresiones regulares.

En [11] se presenta una técnica de compresión general que se traduce en recorridos de a lo más 2N estados del autómata al procesar una cadena de longitud N y un esquema de reducción del alfabeto para estructuras basadas en el DFA que pueden producir reducciones importantes en tamaño de estructura de datos. La técnica se aplicó en los sistemas de detección de intrusión basados en contexto, obteniendo resultados satisfactorios.

Como comentan los autores de ese trabajo, una desventaja grande de los DFA para que reconozcan el lenguaje generado por la ER puede ser la complejidad en espacio.

Desde nuestro punto de vista otra limitación importante es la complejidad de expresar los requerimientos por medio de una ER y otra es la limitación intrínseca de las ERs para representar procesos deterministas.

3.2.2 Clasificadores

En esta sección se revisan algunas propuestas para la evaluación de la sensibilidad de objetos de información, vista como un problema de clasificación supervisada. En [16] los autores hacen énfasis en la aplicación de una herramienta y un enfoque en particular, los métodos del procesamiento

[11] Diccionario de la Real Academia Española "http://dle.rae.es/?id=AUhqW5L", consultado: 10 de agosto de 2016.

estadístico del lenguaje natural y el uso de algoritmos de aprendizaje por computadora, en particular utilizan los métodos de los k-vecinos más cercanos, Naive Bayes y SVM.

Estos autores comentan que hay pocas investigaciones publicadas sobre la forma de evaluar automáticamente la sensibilidad de un documento y su trabajo es el primer reporte en la literatura científica en dar a conocer los resultados de la utilización del aprendizaje por computadora para determinar la clasificación de seguridad de documentos, a pesar de que, la categorización automática de documentos, de acuerdo con el tema, ha sido ampliamente estudiada. Sobre este particular la situación se mantiene, en la actualidad este tema no cuenta con una divulgación adecuada, es decir, los resultados que pueden haberse alcanzado se mantienen bajo la confidencialidad de las empresas propietarias de dichos resultados.

La propuesta presentada en [16] es el resultado de la aplicación de métodos estadísticos de categorización de textos al problema de la clasificación automática desde el punto de vista de la seguridad de textos no estructurados. Los métodos tradicionales de aprendizaje automático fueron probados y optimizados para determinar cuáles eran los más apropiados para categorizar documentos con base en la sensibilidad, y se realizaron investigaciones sobre el entrenamiento para adaptarse a documentos de diferentes temas, lo que intenta reflejar el cambio de políticas de seguridad.

En ese trabajo siguieren la aproximación estándar de SNLP (por sus siglas en inglés Statistical Natural Language Processing) para la categorización de textos. El documento es considerado como una bolsa de palabras siguiendo la propuesta de [27]; así, solo las palabras y su frecuencia relativa son de interés en la caracterización del documento, en oposición al orden de las palabras, partes del discurso y contexto. Se asume que existe una colección de documentos llamada corpus la cual es adecuadamente categorizada y se puede utilizar como conjunto de entrenamiento para los algoritmos de clasificación automática supervisada empleados en el clasificador.

Otra manera en la que se ha abordado el problema de fuga de datos, es desde el campo de la seguridad [28]. En ese trabajo, presentan a los protectores o guardias de información, los cuales son similares a los filtros de correo no deseado que examinan el contenido de los mensajes intercambiados. Un protector se define como dispositivo de alta seguridad utilizado para controlar el flujo de información, típicamente de un dominio con "alto" nivel de confidencialidad, a un dominio con un nivel "bajo".

El control es realizado utilizando listas simples de clasificación (listas de la palabra sucias) para que, de forma automática se pueda evaluar la clasificación de seguridad (por ejemplo, "Público" o "Confidencial") del objeto de información (documentos o mensajes de texto). Los objetos se liberan al dominio "bajo", sólo si la política le permite ese nivel de clasificación. En caso contrario, será bloqueado y posiblemente puesto en cuarentena hasta la inspección humana. Las listas de clasificación son generalmente simples y configuradas manualmente.

En [28] los autores muestran el uso de una máquina de aprendizaje para crear una lista de clasificación más avanzada de forma automática. Mencionan que un obstáculo importante para utilizar la máquina de aprendizaje es que los usuarios suponen que la misma podría crear largas listas de clasificación difíciles de inspeccionar, analizar y controlar por seres humanos. Además, algunas de las técnicas de aprendizaje automático más eficientes, en particular SVM y Redes Neuronales, son clasificadas como "cajas negras", lo que significa que no poseen un carácter explicativo. En ese trabajo, exploran el uso de una reducción de dimensionalidad masiva/estricta con el fin de crear una solución dispersa que resulte en una lista de clasificación corta que sea más fácil de analizar por seres humanos.

Según los autores, una lista de clasificación avanzada debe ser similar a una lista de sentimientos, lo que complica esta opción, porque se debe tener una lista considerablemente grande de palabras, *darle a cada palabra un peso* para cada nivel de la clasificación, asumen que la suma de esos pesos da una probabilidad de la pertenencia del objeto de información a un cierto nivel de clasificación, para lo cual es necesario la determinación de umbrales que permitan decidir a cuál de las clases pertenece. Los autores confiesan que para un humano es complejo crear una lista con estas

características que tenga una buena eficacia, en particular por la dificultad de determinar los pesos y umbrales de modo tal que el resultado sea eficaz y consistente.

Otra opción que proponen para crear una lista de palabras de manera automática es basada en la frecuencia de aparición de las mismas en las diferentes clases y consideran que la variante más avanzada, es haciendo uso de las técnicas de *aprendizaje automático*. Basados en esta idea, los autores descartan las redes neuronales por considerar que crean ¨soluciones opacas que son difíciles de analizar¨. Se enfocan en el análisis de las técnicas: SVM, K- Vecinos más Cercanos (k-Nearest Neighbor k-NN por sus siglas en inglés) y Naive Bayes (NB) [29].

Así, la contribución principal del trabajo **[**28], es que realizaron una exploración profunda de cómo usar el aprendizaje automático para crear listas de clasificación más avanzadas, de manera automática, que las listas simples de palabras sucias que tengan buen desempeño y al mismo tiempo sean cortas y fácilmente analizables e inspeccionables por las personas.

Adicionalmente, muestran que el método Lasso (por las siglas en inglés Least Absolute Shrinkage and Selection Operator) tiene el potencial para crear listas de clasificación muy cortas sin sacrificar el desempeño.

Las listas que tomaron en cuenta en su trabajo para la clasificación de los objetos de información fueron:

1) Lista negra de palabras prohibidas/Lista de palabras sucias.
2) Listas de sentimiento genéricas.
3) Listas de clasificación avanzadas.
4) Listas de clasificación opacas.

- **1)** Lista negra de palabras prohibidas /Lista de palabras sucias: tienen típicamente salidas booleanas. Si un objeto de información contiene una palabra o expresión de la lista, el objeto de información es bloqueado.
- **2)** Listas de sentimiento genéricas: generalmente asignan un número (peso) a cada término de la lista. Las palabras que contiene el objeto de información son comparadas con la lista y se suman los pesos y sobre esa base se clasifica el objeto de información. Esta herramienta se usa también en minería de opiniones, ver en detalles en [30, 31].
- **3)** Listas de clasificación avanzadas: si el problema tiene solo dos clases, la lista funciona como en el caso de la lista anterior con la diferencia que contiene un término de sesgo. En el caso que hayan más de dos clases, se separan las listas por clase y se opera de manera análoga. En esta dirección los autores consideran que las listas más avanzadas son aquellas que pueden convertir la puntuación recibida por el objeto de información en una probabilidad de la pertenencia del mismo a la clase.
- **4)** Listas de clasificación opacas: según los autores, estas listas son generadas mediante aprendizaje automático y los resultados son difíciles de entender por los usuarios. Esto hace que en los casos que se haga necesaria la intervención humana para la inspección y control del funcionamiento del sistema de seguridad estas listas no son recomendables.

En ese trabajo hacen referencia a [28] del que dicen que trata de llevar el tema de la clasificación de seguridad automática al dominio abierto de las obras publicadas, para explicar los experimentos y los parámetros relacionados y los métodos en detalle. Además de explorar las técnicas de aprendizaje automático, tales como SVM, k-NN o Naïve Bayes, también se exploran el uso de Lasso. En el documento se amplió la técnica para un problema de clasificación de dos clases a una solución para la clasificación multiclase, y se discutió la importancia de considerar falsos positivos y falsos negativos. Finalmente, el documento identifica la idea de utilizar Lasso para generar listas de clasificación cortas.

La conclusión es que el procedimiento Lasso es el más adecuado para la creación automática de listas avanzadas de clasificación que pueden parcialmente reemplazar a las listas de palabras sucias y que son más cortas y fáciles de interpretar.

Desventajas

- Una de las limitaciones de usar lista de palabras, es que por lo general se hacen de manera manual.
- La comparación entre las listas de palabras sucias que se usan en los filtros de spam es la misma, desempeñan el mismo papel y tienen las mismas limitaciones.
- Deben tener una lista considerablemente grande de palabras, *darle a cada palabra un peso* para cada nivel de la clasificación, asumen que la suma de esos pesos da una probabilidad de la pertenencia del objeto de información a un cierto nivel de clasificación, para lo cual es necesario la determinación de umbrales que permitan decidir a cuál de las clases pertenece. Los autores confiesan que para un humano es complejo crear una lista con estas características que tenga una buena eficacia, en particular por la dificultad de determinar los pesos y umbrales de modo tal que el resultado sea eficaz y consistente.
- Los autores de ese trabajo descartan las redes neuronales por considerar que crean ¨soluciones opacas que son difíciles de analizar¨.
- Se enfocan pues en el análisis de las técnicas: SVM, k-NN y Naive Bayes (NB) que solamente utilizan valores de tipo numérico.
- El procedimiento está basado en regresión logística, norma L1 para crear predictores lineales, etc., conceptos que no son tan sencillas de entender y sobre todo de fundamentar en cualquier aplicación de este tipo.
- No se fundamenta el uso de los algoritmos mencionados.

Estas consideraciones de los autores ratifican la posición nuestra de que la semántica de la información sobre la base de la cual se tomen las decisiones tiene que ser tenida en cuenta, no se pueden tomar decisiones de seguridad sobre la base de información que no podemos entender su significado y no podemos dejarnos engañar con límites y procedimientos que no tienen una fundamentación adecuada.

Algo en lo que coincidimos con los autores de ese trabajo es que hace falta continuar las investigaciones en este campo, ya que es sorprendente la falta de publicaciones realizadas hasta el 2016.

Como la mayoría de las soluciones actuales se basan en *etiquetas y* en *lista de palabras* sucias con las limitaciones que éstas tienen, entre las más importantes se pueden mencionar que se ignora el orden en que aparecen las palabras y la semántica asociada, sin omitir lo que comentan los autores de [16], que para gestionar el acceso a los datos de diferente nivel sensibilidad y para moverlos entre diferentes dominios de seguridad, todos los datos deben estar debidamente etiquetados con su nivel de seguridad.

En [32] abordan el problema de encontrar la similitud semántica entre pares de textos desde la perspectiva del Procesamiento de Lenguaje Natural (PLN), donde determinar la similitud entre documentos se aplica en varias tareas, como en máquinas de traducción, construcción automática de resúmenes, atribución de autoría, pruebas de lectura comprensivas, recuperación de información, y recientemente sensibilidad de documentos, entre otras.

Presentan un modelo para resolver el problema de similitud semántica entre textos de diferente longitud, utilizando características léxicas, características basadas en conocimiento y características basadas en corpus, con el objetivo de desarrollar un modelo de aprendizaje supervisado. Entre las características léxicas, tomaron: la frecuencia de ocurrencia de los n-gramas de caracteres; k-skip- n-gramas (salto-*k* en *n*-gramas); palabras y relaciones léxicas como sinónimos; el coeficiente de similitud de Jaccard, expandiendo cada término con un conjunto de sinónimos; y la similitud coseno entre los dos textos representados por la bolsa de n-gramas y skip-gramas de caracteres.

Para las características basadas en conocimiento, determinan la similitud semántica entre dos textos como el máximo valor de similaridad obtenido entre los pares de palabras y para medidas basadas en corpus, Información Mutua para el cálculo de la similitud entre pares de palabras y análisis semántico latente.

Con estas características, con los datos de entrenamiento construyen un vector para cada par de textos y lo introducen en Weka para construir un modelo de clasificación basado en regresión logística.

Entre sus conclusiones mencionan que su propuesta tiene buen comportamiento con la detección del grado de similitud semántica entre párrafo-sentencia y sentencia-frase, pero la metodología de expansión para detectar el grado de similitud semántica entre los pares frase a palabra y palabra a sentido no fue correcta, con resultados extremadamente bajos. En trabajos futuros consideran construir vectores incorporando 5 términos a la derecha y 5 a la izquierda de la palabra para detectar relaciones.

Desventajas:

En el análisis de sensibilidad de objetos de información, para evitar la fuga de información sensible, se tienen que homologar conceptos con los del corpus de entrenamiento.

El manejo de vectores numéricos para la representación de documentos y los ajustes que se tienen que realizar cuando tienen diferente longitud, es también una limitación.

3.2.3 Análisis estadístico

Huella de objetos de información en datos no estructurados

Uno de los métodos utilizados para la detección de fuga de datos basado en el contenido es el de huella de objetos de información, el cual ha sido usado en problemas de detección de plagio [33], duplicación de archivos [34] y detección de autoría [35]. En [17] se aplica este método por primera vez, con modificaciones, al problema de la detección de fuga de datos.

El método consiste en calcular un conjunto de valores de funciones de hash que corresponden a combinaciones de palabras (o n-gramas) tomadas del contenido del documento, estas combinaciones son utilizadas como las características para describirlo. Al conjunto de valores obtenido se le conoce como la *firma del contenido del documento*. Posteriormente, esta firma será igualada con la firma del contenido de un documento de salida con el fin de detectar si se está presentando o no la fuga de datos sensibles.

En [17] presentan un análisis de los métodos de huella de objetos de información existentes, e identifican dos limitaciones principales. La primera, que se puede evitar la detección de la fuga reescribiendo el contenido sensible y la segunda, debido a que por lo general todo el contenido del documento es procesado, incluyendo partes no sensibles, se producen falsas alarmas. Para subsanar lo anterior, proponen una extensión del método de coincidencias de *n*-gramas llamado salto-*k* en *n*-gramas-ordenados. A manera de descripción de la idea principal del método, se inicia con el método *n*-gramas, en el cual se van tomando *n* cadenas (o porciones) de una larga secuencia de cadenas de texto, cada selección de *n* cadenas es utilizada para calcular su función de hash. En el método salto-*k* en *n*-gramas se permite saltar o ignorar hasta *k* elementos de los *n*-gramas. Esta posibilidad, al considerar relaciones entre cadenas que no son adyacentes, opinan los autores, que permite agregar información contextual que no se logra con el método de los *n*-gramas. Finalmente en la propuesta de salto-*k* en *n*-gramas-ordenados, se ofrece la ventaja de que puede saltar k en las *n* cadenas, pero ahora están ordenadas en forma alfabética.

Con las mejoras realizadas, el método es capaz de producir huellas de objetos de información de lo que llaman el núcleo del contenido sensible omitiendo las secciones no sensibles. De acuerdo con los autores, es más robusto ante la reescritura y puede detectar documentos sensibles no vistos en el entrenamiento, lo anterior sugiere una mejor detección de los incidentes de fugas intencionales.

Con base en las definiciones dadas en [17], el método debe ser capaz de: detectar texto duplicado totalmente (copia exacta); la duplicación cercana (donde el número de diferencias es pequeño); el duplicado parcial (en él, partes del texto sensible son embebidas en un texto largo no sensible); detectar segmentos de texto reescrito (aquí, partes del texto sensible son modificadas y embebidas en texto no sensible); texto sensible no visto (texto desconocido), se determina que un texto es

desconocido, si se verifica que el tema del texto no está incluido en el conjunto de temas de los documentos revisados; y finalmente manejar contenido estándar (ignora texto no sensible). Con base en estas capacidades, los autores realizan una tabla de comparación de varios métodos de análisis de contenido, entre los que se encuentran métodos de filtros globales, huella de objetos de información, métodos basados en tokens y métodos de aprendizaje automático. Ver Tabla 3, tomada de [17].

Tabla 3. Comparación de métodos para la detección de la fuga de datos basada en contenidos.

	Filtros globales	Basado en tokens	Aprendizaje automático	Huella de objetos de información				
				Huella de objetos de información clásico	Basado en LHS	Basado en Colección estadísitca	Basado en ancla	Método propuesto
Duplicación completa	**SE**	**SP**	**SE**	**SE**	**SE**	**SE**	**SE**	**SE**
Duplicación cercana	**NA**	**SP**	**SE**	**SE**	**SE**	**SE**	**SE**	**SE**
Duplicación parcial	**NA**	**SP**	**NA**	**SE**	**NA**	**SP**	**SE**	**SE**
Segmento de texto reescrito	**NA**	**SP**	**NA**	**SP**	**NA**	**SP**	**SP**	**SE**
Texto desconocido	**NA**	**SP**	**SE**	**NA**	**NA**	**NA**	**NA**	**SP**
Manejo de contenido estándar	**SE**	**SE**	**SP**	**NA**	**SP**	**NA**	**SP**	**SE**
SE: Solución efectiva **SP**: Solución parcial **NA**: no apropiada								

El método que proponen tiene dos fases: la de indexación y la de detección.

El esquema del método en la fase de indexación para las muestras de documentos sensibles y no sensibles es el siguiente:

- Primero: se realiza la extracción de características. Se obtienen palabras y opcionalmente se pueden eliminar artículos, preposiciones, etc.
- Segundo: La separación de características. Aquí se aplica el salto-k en n-gramas-ordenados, lo que significa que se pueden saltar hasta k gramas en la cadena de n gramas y cada conjunto de cadenas se ordena de manera alfabética.
- Tercero: Se aplica la función de hash.
- Cuarto: Selección de hashes. Esto se aplica solamente a los documentos sensibles.

Los hashes que aparecen en menos de *m* documentos no sensibles, son considerados como la huella (fingerprint) del documento, el núcleo del documento. En su evaluación preliminar *m* fue igual a 1.

Los valores de los hashes de documentos sensibles y no sensibles son almacenados en la base de datos.

En la fase de detección, se requiere como entrada el documento *d* y la salida es la "puntuación de sensibilidad" de *d*. Si la puntuación de sensibilidad del documento está por encima de cierto umbral, se considera sensible y se detecta como una fuga.

Al documento se le aplica el esquema usado en la fase de indexación, resultando en una lista de valores hash que representan al documento. A continuación, se obtiene una lista de los documentos de la base de datos que contienen cada uno de los hashes del documento *d*. Los documentos que comparten un número de hashes con *d*, por encima de un umbral, son considerados similares. Esto

hace que el tiempo del proceso sea lineal con respecto a la longitud de *d* y que el método sea de gran escalabilidad.

Por otro lado, con el fin de reducir el tamaño de la base de datos de los valores de hash, la mayoría de los métodos seleccionan solo un pequeño subconjunto de los n-gramas o términos del documento que no afecten significativamente la eficacia del método, sin embargo, se reportan casos en los cuales aunque se incluyan todos los términos de un documento, la eficacia no mejora [33]. Por lo tanto, los términos para la formación de las huellas de los objetos de información deben ser cuidadosamente seleccionados para una óptima representación de un documento.

En [17] se describe el esquema de los métodos de huella de objetos de información considerando los siguientes pasos:

1. Extracción de términos: el documento de entrada se divide en tokens (componentes léxicos con significado, desde símbolos hasta sentencias). Luego se aplican técnicas de pre-procesamiento donde se eliminan artículos, preposiciones, etc.
2. Separación de características: se extraen los n-gramas con solapamiento de la secuencia de tokens.
3. Función de hash: se aplica la función de hash (MD5, SHA Rabin, entre otras) a los n-gramas.
4. Selección de Hash: el objetivo de este paso es reducir la cantidad de huellas del documento, si este paso no se realiza, se conoce como huella de objetos de información completo.

En su propuesta, hacen un ajuste en el paso de separación de características, donde aplican la extracción de n-gramas de salto-k, de las secuencias de palabras obtenidas en el paso de extracción de características. Los n-gramas de salto-k, no son continuos, en contraste con los n-gramas, y se permiten saltos a lo más de k-gramas en los n-gramas. Otra diferencia con el método de los n-gramas es que aquí éstos están ordenadas de manera alfabética. Otro ajuste es realizado en el paso de selección de hash, donde solamente se aplica a los documentos sensibles y solo hashes que aparecen en menos de m documentos no confidenciales son considerados como huella del documento.

Su método permite determinar los siguientes parámetros: si se eliminan palabras o no; si se utiliza o no el ordenamiento alfabético de las palabras de los n-gramas; n es el número de palabras en el n-grama; k es número de saltos permitidos en los n-gramas de salto-k y m es el número mínimo de documentos no sensibles que contienen un n-grama de salto-k específico.

Concluyen que su método es más robusto en la detección de contenido sensible reescrito y puede filtrar partes no relevantes o sin contenido sensible. Además, permite ordenar los n-gramas de salto-k para lograr una mayor precisión en la detección con respecto a los basados en n-gramas. Mejora la eficacia cuando se considera el espacio completo de huellas de objetos de información. Sin embargo, utiliza demasiado espacio de almacenamiento, que es un aspecto crítico cuando la detección se realiza en los puntos finales. En situaciones en donde se tienen problemas de conexión a la red de la organización, la sincronización de la base de datos de las huellas de los objetos de información grandes entre el servidor y el sistema final puede llegar a ser inviable.

Las desventajas de los métodos de huella de objetos de información tomando como referencia lo reportado por los mismos autores en [17] son:

Los métodos basados en funciones de hash sensibles localmente (LSH por sus siglas en inglés) se desarrollaron para la detección de duplicado de documentos, lo que justifica que funcionan muy bien cuando se aplican a documentos similares, pero tiene deficiencias con documentos con pequeñas partes en común. Estos métodos, además, son sensibles al ruido en los documentos de entrada.

Desde nuestro punto de vista el considerar solamente dos clases, sensibles o no sensibles, limita su aplicación a problemas en los cuales se requieren diferentes niveles de sensibilidad e incluso utilizar grados de sensibilidad.

Al eliminar artículos, preposiciones, frases comunes, se elimina información importante sobre el contexto en el cual se genera el objeto de información.

En la propuesta de [17] no se ofrece un criterio claro para asignar los valores a *m* y *k*. El primero utilizado como el valor de referencia para elegir los hashes a conformar la huella del documento y el segundo como la cantidad de gramas a saltar.

Por otra parte, están los basados en colecciones estadísticas, los cuales son utilizados para la selección de términos, asumiendo que entre menos frecuentes sean en la colección completa y más frecuentes en el documento analizado, serán los que mejor describan la idea principal del documento. Este enfoque es utilizado en la recuperación de información [36]. Es de señalar que los autores de ese trabajo, en ningún momento dicen quiénes son los parámetros λ, σ_j con j=1,…,m, o por qué utilizar la exponencial y ponderar por 0.5 la similitud. Solamente mencionan que se pueden determinar por ajustes en las pruebas. Otra limitación es el uso de vectores para la descripción de documentos y métricas para determinar la similitud entre pares de documentos.

Pero la más significativa de las desventajas es que en la representación de los documentos se está pasando por alto la semántica de los términos, lo cual consideramos es una deficiencia.

Para concluir con los métodos de huella de objetos de información, en [17] hacen referencia a métodos basados en *anclas*, en los cuales se selecciona cuidadosamente secuencias de términos que serán las palabras de anclaje. Estas palabras deben ser comunes en el contenido principal del documento y poco frecuente en documentos a comparar, lo que implica que deben ser seleccionadas en cada dominio específico. Así, estos términos utilizados como anclas, serán los términos sensibles.

Mencionan otros métodos que utilizan la transformada rápida de Fourier o agregan filtros para comprobar la similitud de forma rápida, e incluso métodos para la detección de documentos derivados de plantillas, sin embargo, todos ellos son inapropiados para la detección de contenido sensible reescrito.

Sobre el método propuesto en [17], los autores comentan que al eliminar términos correspondientes a artículos, preposiciones, entre otros, disminuye la eficiencia en la detección de sensibilidad de objetos de información y que en trabajos futuros, se debe mejorar el espacio de huellas de objetos de información sin perder precisión en la detección y manteniendo un rendimiento aceptable.

N-gramas

Utilizado ampliamente en lingüística, en aprendizaje automático y en recuperación de información por términos pesados. Depende principalmente del análisis de frecuencia de términos y n-gramas en documentos. Los primeros en usarlo en DLPs fueron Hart y Johnson [18] para clasificar documentos empresariales en sensibles y no sensibles, utilizando SVM para clasificar tres tipos de datos: empresariales privados, públicos y no empresariales. Reportan 97% de eficiencia y solo 3% de falsos negativos. Entre sus desventajas están el que solo pueden tratar con 2 clases de documentos lo cual puede hacer difícil la tarea de aplicar políticas de seguridad.

En su trabajo desarrollan un algoritmo de aprendizaje automático para aprender lo que es sensible y clasificar documentos empresariales estructurados y no estructurados como públicos o privados. El algoritmo es entrenado con muestras de documentos públicos y privados. No utiliza lista de palabras claves y afirman que puede reconocer información sensible en documentos que no tienen solapamiento sustancial con documentos usados en el entrenamiento.

Inicialmente trabajaron con unigramas, es decir, palabras simples con pesos binarios, eliminaron las palabras comunes como artículos, preposiciones, etc. Además, limitaron el número total de palabras a 20000. Si un corpus contenía más de 20000, elegían las 20000 palabras más frecuentes como las características.

Utilizaron Máquinas de Soporte Vectorial y la técnica de regresión propuesta por Cortes y Vapnik [37].

Entre los resultados que mencionan está la creación del primer corpus para la tarea de DLP, siendo la primera aproximación para entrenar un clasificador solamente con documentos empresariales. Sin embargo, en pruebas con documentos reales, tuvo un pobre desempeño.

4 Conclusiones

En este estado del arte, que sin duda puede no ser exhaustivo en el análisis de todos los métodos y alternativas de solución al problema de la determinación de la sensibilidad de los objetos de información, en específico de los documentos, se ha mostrado que estas soluciones resuelven parcialmente el problema y que aún existe un largo camino por recorrer para llegar a ofrecer una herramienta para la determinación automática del grado de sensibilidad de cualquier objeto de información, en particular de los documentos con información no estructurada.

En la atención y formulación de propuestas de métodos de solución a este problema, determinar si un objeto de información es o no sensible, es la aproximación más simple que se puede hacer, puesto que en general se puede hablar de niveles o grados de sensibilidad del mismo objeto de información.

Por lo expuesto en este documento, el problema tiene la suficiente complejidad como para considerar que aún la aproximación más simple, es una solución aceptable dada la importancia que para cualquier organización tiene la protección de su información sensible, como se ha analizado en las primeras páginas de este reporte. Sin embargo, nuestro objetivo es buscar desde diferentes enfoques, la aplicación o desarrollo de métodos que tomen en cuenta todos los factores presentes en el problema de manera integral, y que sin duda incrementan la complejidad del problema.

Por lo anteriormente expuesto, nuestro proyecto es desarrollar una herramienta que sea independiente del contexto de modo tal que dicho contexto se genere a partir de un proceso de entrenamiento en el que el contexto y contenido son tenidos en cuenta. El entrenamiento se realizaría con documentos sensibles y no sensibles determinados por el especialista del área de la aplicación. Debe permitirse incluso la posibilidad que estas clases no sean disjuntas. En dependencia del problema, la determinación de la sensibilidad se podría realizar por documento, párrafo, oración o palabra. La idea central es aprender a diferenciar, con base en el contenido, objetos de información sensibles de los no sensibles. Una de las interrogantes que debemos abordar es qué valor semántico tiene el orden de las palabras y en qué medida la combinación de palabras, oraciones o párrafos que en sí no son sensibles pueden aportar sensibilidad al documento.

La herramienta que se desarrolle, teniendo en cuenta los problemas planteados por los especialistas del área de aplicación consultados, debe ser incremental a partir del entrenamiento; debe permitir la desclasificación (con base en subcadenas) cuando la información deja de ser sensible; si se agrega un documento nuevo, se deben agregar nuevas subcadenas sensibles. Dicha herramienta debe además ser capaz de trabajar con grados de sensibilidad, es decir, no solo en modo booleano sino permitir niveles (modo k-valente) e incluso grados de sensibilidad (modo difuso), y esto debe realizarse con cualquiera de los elementos sujetos a análisis (documento, párrafo, oración, palabra).

En principio la herramienta debe trabajar de forma automática aunque no se debe descartar la posibilidad que, para algunas aplicaciones, puede ser necesario el trabajar de forma interactiva. La herramienta debe estar preparada también para en el caso que se desclasifiquen documentos, detectar en los mismos, párrafos, oraciones y/o palabras sensibles que deben ser protegidas. Con este objetivo las subcadenas sensibles (párrafos, oraciones o palabras) deben ser obtenidas de la muestra de entrenamiento y debe guardarse la referencia de los documentos en donde aparecen y en el reentrenamiento se deben emitir alertas sobre la desclasificación de documentos en caso que éstos contengan elementos sensibles.

Es importante subrayar que aunque en este estado del arte hemos hecho énfasis en la automatización de la clasificación de la sensibilidad de documentos textuales, la problemática de la automatización de la clasificación de la sensibilidad de los objetos de información es más amplia y compleja y se hace necesario un estudio detallado de cada uno de los posibles objetos de información como son los casos de las imágenes, las grabaciones y otras formas de objeto de información, cada una de las cuales conlleva niveles de complejidad que ameritan estudios análogos a los que hemos iniciado sobre la automatización de la clasificación de textos sensibles.

Referencias bibliográficas

1. Neustar: WHAT THE FRAUD? A Look at Telecommunications Fraud and Its Impacts. Resources & Tools, (2016)
2. Lopez, G., Richardson, N., Carvajal, J.: Methodology for Data Loss Prevention Technology Evaluation for Protecting Sensitive Information. Revista Politécnica 36, (2015)
3. https://www.wikileaks.org/
4. Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V.: A survey on data leakage prevention systems. Journal of Network and Computer Applications 62, 137-152 (2016)
5. Shabtai, A., Elovici, Y., Rokach, L.: A survey of data leakage detection and prevention solutions. Springer Science & Business Media (2012)
6. Reed, B., Wynne, N.: Magic Quadrant for Enterprise Data Loss Prevention. Gartner Institution (2016)
7. Wadkar, H., Mishra, A., Dixit, A.: Prevention of information leakages in a web browser by monitoring system calls. In: Advance Computing Conference (IACC), 2014 IEEE International, pp. 199-204. IEEE, (2016)
8. Torsteinbo, T.: Data Loss Prevention Systems and Their Weaknesses. Kristiamand and Grimstad, Norway: University of Agder (2012)
9. Liu, T., Pu, Y., Shi, J., Li, Q., Chen, X.: Towards misdirected email detection for preventing information leakage. In: Computers and Communication (ISCC), 2014 IEEE Symposium on, pp. 1-6. IEEE, (Year)
10. Zilberman, P., Dolev, S., Katz, G., Elovici, Y., Shabtai, A.: Analyzing group communication for preventing data leakage via email. In: Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on, pp. 37-41. IEEE, (2011)
11. Becchi, M., Crowley, P.: An improved algorithm to accelerate regular expression evaluation. In: Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems, pp. 145-154. ACM, (2007)
12. Sokolova, M., El Emam, K., Rose, S., Chowdhury, S., Neri, E., Jonker, E., Peyton, L.: Personal health information leak prevention in heterogeneous texts. In: Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains, pp. 58-69. Association for Computational Linguistics, (2009)
13. Mogull, R.: Understanding and Selecting a Data Loss Prevention Solution. pp. 1-26. Securosis (2010)
14. Chen, K., Liu, L.: Privacy preserving data classification with rotation perturbation. In: Fifth IEEE International Conference on Data Mining (ICDM'05), pp. 4 pp. IEEE, (2005)
15. Aggarwal, C.C., Philip, S.Y.: A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining, pp. 11-52. Springer (2008)
16. Brown, J.D., Charlebois, D.: Security Classification Using Automated Learning (SCALE): Optimizing Statistical Natural Language Processing Techniques to Assign Security Labels to Unstructured Text. DTIC Document (2010)
17. Shapira, Y., Shapira, B., Shabtai, A.: Content-based data leakage detection using extended fingerprinting. arXiv preprint arXiv:1302.2028 (2013)
18. Hart, M., Manadhata, P., Johnson, R.: Text classification for data loss prevention. In: Privacy Enhancing Technologies, pp. 18-37. Springer, (2011)
19. Salton, G., Wong, A., Yang, C.-S.: A vector space model for automatic indexing. Communications of the ACM 18, 613-620 (1975)
20. Manning CD, Raghavan P, H., S.: "Introduction to information retrieval". Cambridge University (2008)
21. Carvalho, V.R., Balasubramanyan, R., Cohen, W.W.: Information leaks and suggestions: A case study using mozilla thunderbird. In: CEAS 2009-Sixth Conference on Email and Anti-Spam. (2009)
22. Pshehotskaya, E., Nikitinsky, N., Sokolova, T.: DLP Technologies: Challenges and Future Directions. In: The International Conference on Cyber-Crime Investigation and Cyber Security (ICCICS2014), pp. 31-36. The Society of Digital Information and Wireless Communication, (2014)
23. U.S. Government Printing Office, "Too Many Secrets: Over classification as a Barrier to Critical Information Sharing." Hearing Before the Committee on Government Reforms, US House of Representatives, p. 82. 2004.
24. Carlstrom, G. "Connecting agencies: Can Obama end mistrust in intel community?" (online), ttp://www.federaltimes.com/article/20100110/ AGENCY04/1100308/-1/RSS (Access Date: July 8, 2010), 2010.

25.Modi, C., P, D., Patel, H., Borisaniya, B., Patel, A. & Rajarajan, M., “A survey of intrusion detection techniques in Cloud”. Journal of Network and Computer Applications, 36(1), pp. 42-57. doi: 10.1016/j.jnca.2012.05.003, 2013.
26.Kumar S., et alt., “Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection,” in ACM SIGCOMM, Sept 2006
27.Sebastiani, Fabrizio (2002), Machine learning in automated text categorization, ACM Computing Surveys, 34(1), 1–47
28. Engelstad, P. E., Hammer, H., Yazidi, A., & Bai, A. Advanced classification lists (dirty word lists) for automatic security classification. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference on (pp. 44-53). IEEE. (2015, September)
29. Entezari –Maleki, C., A. Rezaei, and B. Minaei-Bidgoli, “Comparison of classification methods based on the type of attributes and sample size,” Journal of Convergence Information Technology, vol. 4, no. 3, pp. 94–102, 2009.
30. Bing, L. Sentiment Analysis and Opinion Mining. Claypool Publisher, 2012
31. Hammer, H. L., A. Bai, A. Yazidi, and P. E. Engelstad, “Building sentiment lexicons applying graph theory on information from three norwegian thesauruses,” Norwegian Informatics Conference (NIK 2014).
32. Vilarino, D., Tovar, M., Beltrán, B., & León, S. (2014). Un modelo para detectar la similitud semántica entre textos de diferentes longitudes. Avances en la Ingeniería del Lenguaje y del Conocimiento, 57.
33. Hoad T.C. and Zobel J., "Methods for identifying versioned and plagiarized documents," Journal of the American Society for Information Science and Technology, vol. 54, pp. 203-215, 2003.
34. Bernstein Y. and Zobel J., "Accurate discovery of co-derivative documents via duplicate text detection," Information Systems, vol. 31, pp. 595-609, 2006.
35. Kolcz A., Chowdhury A., and Alspector J., "Improved robustness of signature-based near-replica detection via lexicon randomization," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, p. 610.
36. Kumar, B. V., & Basha, M. S. K. Optimal Similarity Measure to Ensure Robustness in Text Classification and Clustering. International Journal of Electronics Communication and Computer Engineering Volume 6, Issue (5) Sept., NCRTCST-2015, ISSN 2249–071X 3rd National
37. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20, 273-297 (1995)

RT_036, octubre 2016
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2143
ISSN 2072-6260
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba