Población de ontologías con datos no estructurados utilizando herramientas de minería de datos

Beatriz P. de Gallo[a], Marcela Vegetti[b], Horacio Leone[b]
[a]*I.Es.I.Ing. /Facultad de Ingeniería, Universidad Católica de Salta*
Campo Castañares S/N, Salta, Argentina
[b]*INGAR/ Facultad Regional Santa Fe UTN*
Avellaneda 3657, Santa Fe, Argentina
[a]*bgallo@ucasal.net,* [b]*{mvegetti, hleone}@santafe-conicet.gov.ar*

Abstract

Este trabajo tiene como objetivo abordar el estudio de las tecnologías sobre búsqueda y recuperación de información disponibles para la preparación de datos no estructurados, provenientes de correos electrónicos, como paso previo para la preparación del reservorio que se utilizará para poblar una ontología para el análisis forense de correo electrónico. Se formula un diagrama de procesos que representa la extracción semiautomática de datos, que posteriormente nutrirán ese reservorio, utilizando herramientas de minería de datos. En este trabajo se realiza una aplicación con una pequeña base de 1200 correos para realizar una validación preliminar de la estrategia propuesta.

1. Introducción

El análisis forense de correos electrónicos requiere acceder a los datos de cabecera y cuerpo que se encuentran residentes en archivos propietarios de los clientes de correo que gestionan la cuenta de mail. Estos datos se almacenan según la estructura propia del software de gestión de la cuenta y no es sencillo extraerlos para su posterior análisis. Resulta conveniente, entonces contar con un marco de referencia basado en la conceptualización formal de la materia de discusión. Y en particular, las ontologías sirven como herramienta universal o pluridisciplinar para facilitar el análisis de la prueba documental, por parte de todo los actores involucrados en el análisis forense (abogados, jueces, investigadores y peritos).

La construcción de una ontología contempla como última etapa, la validación del modelo instanciado desde un contexto real y acorde a los objetivos que se persiguen con la misma. Cuando las instancias provienen de repositorio con datos mayormente no estructurados (como los correos electrónicos), es importante recurrir a los procedimientos y herramientas de búsqueda y recuperación de información o *Information Search and Retrieval (ISR)* [1]. En base a la problemática cada vez más compleja para poblar las ontologías con información cierta y pertinente, surge un nuevo campo de estudio que se aboca a la *Instanciación Automática de Ontologías* (IAO) a partir de textos o datos no estructurados.

Este trabajo tiene como objetivo abordar estas tecnologías con el fin de identificar las herramientas de ISR e IAO disponibles para la preparación de datos provenientes de correos electrónicos, como paso previo para la preparación del conjunto de datos que se utiliza para poblar una ontología, también descripta brevemente en este artículo, que será utilizada para realizar análisis forense de correos electrónicos. Atendiendo a esto, se propone un proceso semiautomático, que utiliza la herramienta *RapidMiner*[2], para la extracción de datos .y se muestra su aplicación en un caso de estudio.

La organización de este trabajo es la siguiente: la sección 2 describe la Minería de Datos o *Dataminig (DM)*, sus características y enfoque. En la sección 3 se exponen los antecedentes sobre el tema, en cuanto a la aplicación de tecnologías de DM referidas al pre-procesamiento de datos y las metodologías y técnicas disponibles para un conjunto de datos documentales, detallan el caso particular del corpus de archivos proveniente de una casilla de correo electrónico. La sección 4 describe de manera general el proceso de población de ontologías. El apartado 5 introduce la ontología propuesta para el análisis forense de correo electrónico, así como el procedimiento para la población de ontologías con datos no estructurados utilizando herramientas de minería de datos y su aplicación en un caso práctico. Finalmente, en la sección 6 se presentan las conclusiones y trabajos futuros.

---

[1] ISR: encontrar material (generalmente documentos) de naturaleza no estructurada (generalmente de texto) que satisface una necesidad de información desde dentro de grandes colecciones (generalmente almacenada en las computadoras).
[2] https://rapidminer.com/

2. Minería de Datos como tecnología para el análisis de los datos

Existen varias definiciones formales de Minería de Datos, la más difundida es la de Fayyad et al. [1]: "*Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos*"

A lo largo del tiempo, la minería de datos se ha dotado de diferentes herramientas que ayudan al tratamiento de los datos –no ya desde el enfoque transaccional- sino más bien desde la necesidad de generar un contexto de trabajo analítico. Con el tiempo, la minería de datos ha abordado diferentes espacios, según sea el contexto tecnológico de aplicación. Se puede hablar ahora de minería de textos, minería web, minería de redes sociales, minería de datos complejos (como series temporales por ejemplo).

Sin abordar la diferencia puntual entre cada uno de estos espacios, herramientas, y tecnologías que los investigadores del tema definen para esos componentes, la característica que las unifica a todas ellas es la necesidad de *tomar la mejor decisión en un contexto de incertidumbre*. En base a ese objetivo, la DM se ha provisto de importantes herramientas no sólo para el análisis sino también para la búsqueda y recuperación de información de interés, que se encuentra almacenada en corpus documentales de datos no estructurados.

Wen-Yang Lin [2] sintetiza en un esquema muy claro la relación de DM y las ontologías durante el proceso de gestión del conocimiento, vinculando ambos temas a partir de la multidimensionalidad y de la especialización de los reservorios de datos. El autor destaca el rol de las ontologías como marco de referencia para la representación de los datos en todas las etapas del proceso de KDD[3].

3. Aplicación de minería de datos en el análisis de correos electrónicos

Mohapatra et al. [3] describe ampliamente los diferentes ámbitos y tipos de actividades en que se utiliza la minería de datos: está a disposición de expertos en los diferentes dominios del conocimiento (como la educación o la salud) y produce resultados interesantes en relación a la mejora en los métodos de producciones industriales o de fabricación, en las estrategias de juegos deportivos, y en los estamentos de control estatales como los servicios de inteligencia antiterroristas o el control de la evasión fiscal.

En cuanto a la aplicación de la minería de datos en relación a los correos electrónicos, existen varios antecedentes que se pueden tomar como referencia.

Abbasi [4] señala la importancia de la minería en e-mail para resolver cuestiones de robo de identidad y de plagio en las investigaciones de forensia digital. Otros autores [5] han utilizados técnicas de clasificación de minería de textos en e-mails con dos propósitos: para clasificar nuevos mensajes de correo electrónico sobre la base del tema; y, para identificar la identidad del verdadero autor de un correo electrónico anónimo.

Por su parte, Iqbal et al. [6] aborda ampliamente el problema del análisis de autoría, cuando se trata de documentos en línea (como correos electrónicos o mensajes instantáneos) que –en cuanto a párrafos de escritura- usualmente son cortos, mal estructurados sintácticamente y con errores ortográficos y de sintaxis. Basando el trabajo en el concepto de *writeprint*, una analogía de una huella dactilar, sugieren que las personas a menudo dejan rastros de su personalidad en su trabajo escrito. Tomando estas características estilográficas, los autores proponen la identificación de patrones para casos en los que se cuenta con una gran colección de mensajes, con una escasa cantidad de mensajes o bien cuando es necesario caracterizar al autor anónimo de los mensajes electrónicos.

En el 2004, el trabajo de Airoldi et al. [7] muestra la ineficiencia de los algoritmos anti-spam de correos electrónicos para filtrar el ingreso de mails fraudulentos o que promueven las estafas por internet, presentando un sistema *ScamSlam* que básicamente es una herramienta para el análisis de textos.

En [8] se presenta el trabajo de análisis de correos electrónicos para la detección de amenazas terroristas basado en el algoritmo de árboles de decisión de inducción llamado *Ad Infinitum* que actúa en bases de datos multidimensionales de grandes volúmenes.

Uno de los principales retos de seguridad para la comunidad en línea está en la identificación de páginas web de *phishing*[4] debido principalmente a la gran cantidad de transacciones on-line que se realizan diariamente. El trabajo de Abdelhamid [9] trata el tema con máquinas de vectores soporte.

Incluso se observa la preocupación por establecer criterios restrictivos en el uso de la minería de datos sobre correos electrónicos. Armknecht et al. [10] proponen un esquema criptográfico para cifrar previamente los mensajes y buscar por cierta cantidad de palabras claves (cifradas también), de manera que el algoritmo de búsqueda solo muestra el mensaje descifrado en caso de que supere la cantidad de palabras claves.

Otro estudio destacado es el de Stuit [11] que trabajó sobre el análisis de correos electrónicos en el ambiente

[3] KDD: (Knowledge Discovery from Databases) Descubrimiento de Conocimiento en Bases de Datos

[4] Sitio web de confianza para obtener información sensible de los usuarios en línea, tales como nombres de usuario y contraseñas

corporativo de las empresas, proponiendo un modelo centrado en las interacciones temporales y de similitud de temas de las "cadenas de mensajes" y buscar mejorar los procesos de negocios.

Resulta de especial interés la propuesta de Tang [12] sobre la limpieza en los correos electrónicos, como paso previo para el procesamiento con técnicas analíticas, a fin de extraer solo la información relevante al caso. Los autores proponen realizar esta depuración utilizando máquinas de vectores soporte, con un trabajo "en cascada", abordando primero la limpieza de los datos no textuales en el cuerpo del mail y a continuación identificados párrafos y palabras hasta lograr una normalización.

3.1. Pre-procesamiento de datos mediante DM

Existe abundante bibliografía acerca de las metodologías propuestas para trabajar en minería de datos. Si bien existen otras y muy variadas metodologías específicas, se presenta *CRISP-DM*[5], que bajo el principio de iteración creciente, propone un modelo de tareas sencillo, fácil de implementar y que no requiere experticia o dominio en el proceso técnico de la minería de datos. CRISP-DM propone las siguientes fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

De todas estas actividades, nos interesa particularmente la fase dedicada a la *preparación de los datos*, ya que es una de las actividades que mayores inconvenientes presenta al momento de hacer minería de datos, y que es altamente estudiada por los investigadores del tema, pues se pretende que esa fase concluya con un conjunto de datos sólidos y consolidados sobre el cual avanzar en las siguientes fases.

Este objetivo, es compartido en el ámbito de las ontologías semánticas, en donde también se reconoce la importancia y necesidad de lograr un conjunto de datos pre-procesados a conveniencia previo a la etapa de implementación del modelo de análisis semántico formulado.

La fase de preparación de datos, que normalmente insume el mayor esfuerzo del tiempo total de construcción del producto, comprende todas las actividades que se debe realizar sobre los datos en bruto para conformar un conjunto de datos confiable. Shearer [13] propone realizar cinco actividades en esta fase: selección, limpieza de los datos, construcción, integración y formateo de los datos.

La *selección de los datos* debe centrarse en el objetivo principal que se persigue, sin descuidar criterios de calidad ni restricciones técnicas como el volumen o el tipo de datos y marcando además los criterios de inclusión (o exclusión) de los datos y el grado de importancia de unos sobre otros. La etapa de *limpieza de los datos* tiene por objetivo trabajar en la calidad de los datos, si no se consideran criterios definidos en el objetivo del trabajo y rigurosos en su metodología, es altamente probable que los resultados no sean confiables, pues cuestiones como atributos faltantes o con valores no plausibles impactan de manera exponencial en las conclusiones a obtener con la minería de datos. Durante el paso de *construcción de los datos*, puede resultar conveniente generar nuevos registros o atributos derivados para los datos que se han depurado, o bien deben discretizarse o normalizarse para ajustarse a las herramientas de minería de datos que se utilizará. Por *integración de los datos*, se entiende al proceso de combinación de datos provenientes de diversos registros o fuentes, unificándolos en una única entidad que contenga todos los atributos necesarios para el estudio, incluyendo instancias de agregación/desagregación de los datos. El proceso culmina con el paso de *formateo de los datos*, que tiene por objetivo realizar ajustes sobre las propiedades de los atributos (longitud, tipo de dato, etc.)

3.2. Pre-procesamiento de datos para el caso particular de correos electrónicos

Para el presente trabajo es de especial interés abordar las técnicas de trabajo en datos documentales, y en particular, los orientados a archivos de salida de los clientes de correo. Estos archivos se generan en los más variados formatos, desde *.mbx* que es un archivo de texto plano fácilmente legible con un editor de notas, hasta estructuras complejas como *.pst* que son legibles solo desde "dentro" del propio cliente de correo (como es el caso de Microsoft Outlook).

En caso de necesidad de convertir el archivo de un formato a otro para obtener un archivo plano, puede tomarse el trabajo de Berg et al. [14] que propone una secuencia de tres pasos para migrar correos electrónicos desde clientes que los almacenan en archivos *.pst* a *.mbox*. A través de un método iterativo plantea realizar un pre-procesamiento que consiste en la identificación de las características del archivo, y la selección de la infraestructura de hardware y software necesaria para un análisis de detección de virus (aunque también aconsejan realizar ese análisis en la última etapa de la migración de los datos), luego realizar la migración de los correos al soporte solicitado y por último realizar una rutina de post-procesamiento orientada particularmente a trabajar los elementos complementario de la cuenta que no se hubieran migrado (calendario, contactos, etc.).

[5] CRISP-DM: (Cross Industry Standard Process for Data Mining) Modelo de procesos de Minería de Datos.

Un aporte interesante es el realizado por Squire [15] que presenta un sistema llamado *Apachemail,* integrado en la plataforma *FlossMole*[6], en el que se propone fijar como estándar una infraestructura reproducible y compartible, para el procesamiento de mensajes de correo electrónico por parte de los grupos de investigación que toman esta estructura de datos como insumo para realizar sus investigaciones. El sistema funciona en cualquier colección de correo electrónico en formato *mbox* e incluye comandos y técnicas para el procesamiento de listas de correos, mejorando la replicabilidad al proporcionar un lenguaje común para la comunicación acerca de cómo fueron recogidos y almacenados los mensajes. Partiendo de un enfoque orientado al documento (y no a la estructura relacional del archivo *pst*), el sistema Apachemail se sustenta en tres premisas: a) el sistema debe permitir adiciones incrementales en el formato nativo como resultado del uso regular y continuo del correo electrónico que se está analizando; b) el sistema debe permitir el análisis de las partes estructuradas (cabecera) y no estructuradas (cuerpo) del correo electrónico; y c) el sistema debe estar disponible también en modo distribuido y accesible desde internet.

Por su parte, Bacchelli [16] presentó un conjunto de herramientas para importar, procesar, almacenar y analizar tanto correo electrónico y código fuente de datos, incluyendo métricas sobre el proceso de exportación con indicadores como número de autores de correo electrónico, o las líneas de texto en los contenidos.

El trabajo de Kota [17] aborda detalladamente esta cuestión, destacando la necesidad de contar con algoritmos de clasificación eficientes acompañados de técnicas de visualización que faciliten la tarea al investigador forense. Recurriendo a LUCENE[7] se extrae cada correo electrónico del corpus documental y se interpreta, analiza, tokeniza y almacena en el repositorio resultante. Así, el correo electrónico se clasifica en función de su relevancia y se presenta al usuario, siendo posible además, poblar la ontología a partir de esta última etapa.

La literatura especializada [18] sobre *ISR* menciona como objetivo de las metodologías y herramientas la necesidad de interpretar el contenido de los elementos de información o documentos de la colección y ordenarlos de acuerdo con el grado de relevancia necesario.

En el ámbito propio de la forensia digital, existen diversas herramientas de búsqueda y extracción de datos que brindan la posibilidad de obtener directamente un archivo plano con la información de un correo electrónico (metadatos y contenido)[8].

Podría suponerse que utilizando cualquiera de estas técnicas es suficiente para obtener el conjunto de datos que luego se utilizarán para la instanciación de la ontología, pero no es tan simple. Si bien las herramientas enunciadas permiten trabajar sobre los correos electrónicos con cierta confianza en cuanto a la solidez de los procesos de migración y de detección de virus, no resultan totalmente adecuadas para el trabajo que nos ocupa, ya que no cuentan con herramientas o componentes destinados a la depuración semántica inicial que se requiere para poblar la ontología posteriormente.

Y en este punto puede realizar un aporte muy importante la minería de textos.

4. Proceso de poblar ontologías

El proceso de instanciar o poblar manualmente una ontología requiere mucho tiempo y es propenso a errores [19]. Es por ello que la instanciación automática de ontologías (IAO) a partir de textos ha surgido como un nuevo campo de aplicación para las técnicas de adquisición de conocimiento. El objetivo de IAO es la construcción de una base de conocimientos ontológica, un repositorio de datos acerca de los objetos de la vida real considerados como instancias de conceptos de una ontología.

A lo largo de todo el proceso de construcción de una ontología es de importancia la confiabilidad de la información contenida en el dominio que se está estudiando, según criterios de pertinencia, coherencia y consistencia sobre el objeto de estudio.

Según sea el tipo de datos que almacena el reservorio que representa el dominio (datos estructurados o no estructurados), se presentará en mayor o menor grado la necesidad de utilizar técnicas automáticas o semiautomáticas de extracción de datos para obtener un conjunto que luego permita la instanciación de la ontología.

Son varios los autores que mencionan la necesidad de recurrir a tecnologías de recuperación de la información para poblar las ontologías [20], [21], [22], [23] marcando la necesidad de desarrollar herramientas inteligentes y métodos de extracción de datos para el conocimiento (procesamiento de metadatos, identificación y categorización de conceptos claves, detección primaria de inconsistencias, técnicas de búsquedas automáticas o semiautomáticas, etc.).

Es de interés en este punto el trabajo de Hacherouf et al. [24] en el que se discuten las ventajas y desventajas de las técnicas de conversión de documentos XML en

[6] FLOSSmole: repositorio colaborativo para la investigación y análisis de software libre

[7] Librerías JAVA para la recuperación de información. Es posible hacer búsqueda de documentos almacenados en un árbol de directorios y sobre cualquier documento convertido en texto plano.

[8] Varias de ellas se citan en [30]

ontologías OWL, en referencia la riqueza de la transformación de los datos, la complejidad de los mismos y los riesgos de pérdida de información, sumado a ello el costo del pre-procesamiento de los datos.

Por su parte Fiel Cortes [25] propone una herramienta semiautomática para importar información de propiedades desde archivos RDF (que contienen las URI de origen) y asignarla a un individuo de la ontología creada. En [26] Ruiz Martínez propone metodologías para poblar ontologías mediante el análisis lingüístico tradicional y tecnologías para la extracción de conocimiento textual.

Santosh [27] hace un aporte interesante desde el punto de vista de la minería de opiniones al proponer la conversión de textos en datos no estructurados mediante RDF[9] y algoritmos de machine learning para poblar una ontología OWL sobre como los clientes toman decisiones sobre sus compras.

En [28] los autores desarrollan ALLRIGHT una ontología basada en el análisis de escenarios en los que no se pueden aplicar de manera directa las técnicas usuales de extracción de información. Se destaca en este trabajo el estudio sobre el uso de técnicas de *crawling* (minería web) y de extracción de información para la población de una ontología.

En [29] se marca la importancia de la tarea de población de la ontología tanto durante el proceso de construcción como de mantenimiento, en particular, cuando el dominio se nutre de datos no estructurados (texto). Señala los procesos *OBIE* (Ontology Based Information Extraction) habituales para poblar una ontología: técnicas de procesamiento de lenguaje natural para extracción de términos o relaciones semánticas, técnicas de reconocimiento de nombres de entidades, entre otras.

La etapa de población de una ontología está orientada a enriquecer una ontología pre-existente, mediante la incorporación de instancias de clases y/o relaciones. En esta etapa tiene especial importancia las técnicas de búsqueda y recuperación de la información, ya sea por razones de actualización periódica de la ontología o cuando se abordan diferentes y numerosos dominios específicos que requieren una tarea cuidadosa al momento de instanciar.

El proceso de adquisición del conocimiento para la población de una ontología comprende tres etapas: recuperación de la información, extracción de la información necesaria y carga de la ontología. La primera etapa consiste en encontrar los documentos que contengan la información requerida para la ontología. Aquí es de mucha utilidad la *categorización de textos*, que identifica subconjuntos de textos relacionados por un criterio predefinido, aunque también puede recurrirse a otras herramientas de minería de textos como las redes bayesianas, las redes neuronales o las máquinas de soporte vector. La etapa de extracción de la información necesaria se ocupa de encontrar los objetos, clases, instancias, restricciones y propiedades que conforman el corpus de la ontología. En esta etapa usualmente hay un experto que valida los elementos extraídos antes de pasar la última etapa de población de la ontología.

La etapa de extracción de la información necesaria, es la que se aborda en esta propuesta, la cual se introduce en la próxima sección.

5. Procedimiento de extracción y limpieza de datos relevantes de correos electrónicos utilizando DM.

Este trabajo tiene como objetivo abordar el estudio de las tecnologías sobre búsqueda y recuperación de información disponibles para la preparación de datos no estructurados, provenientes de correos electrónicos, como paso previo para la preparación del reservorio que se utilizará para poblar una ontología para el análisis de correos electrónicos. En esta sección se formula un proceso semiautomático para la extracción de datos que posteriormente nutrirán ese reservorio, utilizando para ello herramientas aportadas por la minería de datos. En primer lugar se introduce brevemente la ontología propuesta y, luego se presenta el procedimiento para la extracción y limpieza de datos relevantes de correos electrónicos, mediante el uso de herramientas de DM.

5.1. Ontología propuesta

En [30] se presentó una ontología que permite definir la semántica de los conceptos relacionados con el análisis forense de un correo electrónico a fin de formalizar los mismos, así como sus relaciones y restricciones, llegando a una primera conceptualización que sirvió de insumo para el proceso iterativo que marca el proceso de construcción del modelo semántico.

La taxonomía desarrollada –que se muestra en la Figura 1- señala la relación de subclase de entre los usuarios más destacados (el emisor y el receptor del correo), que se distinguen a la vez del resto de los gestores del servicio porque son aquellos sobre quienes finalmente impacta el carácter probatorio del correo electrónico.

[9] RDF (Resource Description Framework) es un marco para la representación de recursos de la web recomendado por la W3C

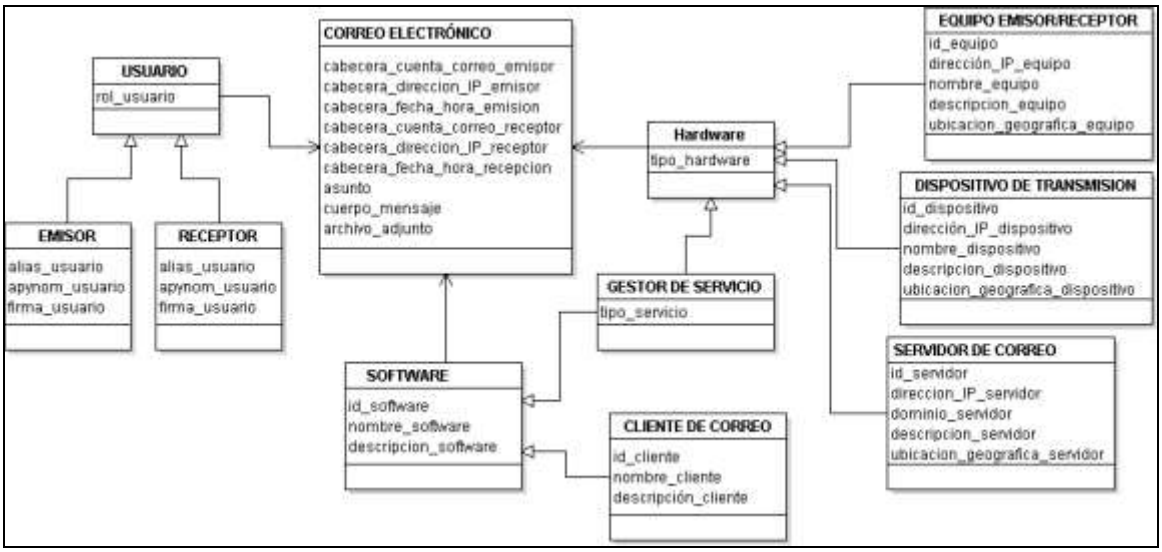

Figura 1: Taxonomía de Conceptos

Partiendo de la narrativa escrita en lenguaje natural –los puntos de pericia- se establecen las ***relaciones ad hoc*** existentes entre los conceptos definidos en la taxonomía. Las relaciones deben establecer con exactitud y precisión, indicando el origen y destino de cada una, evitando imprecisiones o sobreespecificación de esos puntos.

La Figura 2 muestra la *relación ad-hoc* para la emisión/recepción del correo electrónico, proceso principal en el análisis forense, señalando los conceptos que participan con sus atributos y la vinculación que existe entre ellos.

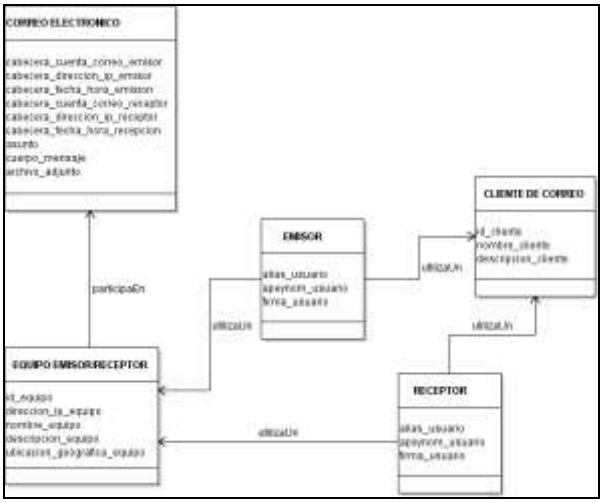

Figura 2: Relaciones ad-hoc para emisión/recepción del correo electrónico

Los axiomas destacados, y que en cierta forma, conforman la base del análisis forense de un correo electrónico son los siguientes:

- ***Axioma 1: sobre la autenticidad de un correo electrónico***

Un correo electrónico es auténtico cuando se identifican: los datos del remitente (nombre de usuario, cuenta de correo y dirección IP), la trazabilidad del mismo (diferentes servicios o agentes que intervienen en la transmisión) y los datos del destinatario (nombre de usuario, cuenta de correo y dirección IP).

- ***Axioma 2: sobre la existencia de un correo electrónico***

Un correo electrónico existe cuando se comprueba la presencia del archivo digital del mismo tanto en el dispositivo emisor (ó en el servidor del ISP del emisor) como en el dispositivo receptor del correo (ó en el servidor ISP del receptor); y ambos archivos digitales son idénticos.

El proceso iterativo en la construcción de la ontología derivó en la revisión del modelo conceptual inicialmente propuesto, en función de la necesidad de ajustar las clases y propiedades de los objetos de la vida real, en cuanto a instancias de los conceptos de la ontología.

Del proceso de contrastación entre el modelo conceptual y los datos propios del objeto de estudio (los correos electrónicos) surgió la necesidad de estudiar las técnicas de pre-procesamiento de estos datos, a fin de instanciarlos con márgenes adecuados de certeza y consistencia de los datos.

En el análisis forense de correos electrónicos pueden intervenir conjuntos de datos provenientes de orígenes diferentes (como en el caso en que se deba analizar la casilla del emisor y del receptor), de manera que es posible que se requieran diversas fuentes de información para extraer las instancias de las relaciones ad-hoc presentadas.

5.2. Procedimiento propuesto para la extracción y depuración de datos

De por sí la forensia digital recurre a las técnicas de recuperación de información, con el objetivo de reunir la evidencia necesaria en base a una búsqueda selectiva de datos particulares en un gran conjunto de información digitalizada.

Según Baeza et al. [31], el modelo de Recuperación de Información es el resultado de una asociación funcional entre cuatro elementos: la colección de documentos, la consulta del usuario, el entorno de trabajo y la relación existente entre la consulta y la representación de la colección de documentos.

Desde esta perspectiva, es esencial realizar adecuadamente la selección, limpieza, enriquecimiento, reducción y transformación de datos, como paso inicial de validación de la información entrante.

Así, se debe trabajar identificando:

- *datos incompletos* (atributos faltantes o incompletos, datos fuera de rango):
- *datos con ruido* (fallas en las rutinas de validación de los datos de ingreso, limitaciones tecnológicas)
- *datos inconsistentes* (ajuste de códigos o conversiones de tipo al provenir de diferentes fuentes, tuplas duplicadas).

La minería de datos propone diversas técnicas para esta etapa, como ser, la utilización de valores estadísticos para resolver la ausencia de valores en un atributo, clustering para valores anómalos o fuera de rango o la agregación y normalización para la transformación de los datos.

Además de las citadas, existen otras herramientas provenientes de la minería de textos, que se pueden aprovechar para esta etapa de extracción y limpieza de datos, con atención en los objetivos de generación del conocimiento derivado del modelo ontológico.

Tal es el caso de *RapidMiner* una herramienta ágil para el análisis predictivo, catalogada como una de las más difundidas en el área de la minería de textos. Bajo el concepto de software abierto, RapidMiner cuenta con un enfoque por *procesos* en los cuales se insertan *operadores*, que realizan el procesamiento analítico de los datos.

A continuación se muestra el diagrama de procesos semiautomáticos para la extracción y limpieza de datos de una casilla de correos electrónicos, realizado a partir de la utilización de la herramienta *RapidMiner*, tomando como ejemplo una vista parcial de la ontología que se está trabajando.

En particular, se utilizarán los conceptos y asociaciones involucrados en la relación ad-hoc para la emisión/recepción del correo electrónico que se muestran en la Fig. 2.

La primera actividad consiste en realizar la extracción de todos los correos electrónicos obrantes en la casilla de interés. Esta ***Extracción Inicial*** debe permitir la obtención de todos los datos necesarios para la instanciar la ontología.

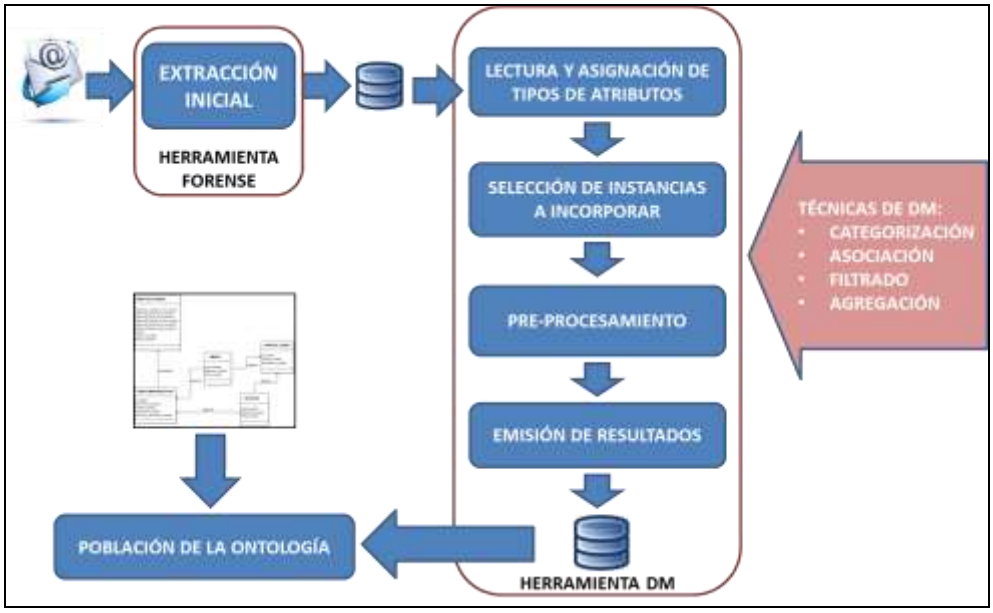

Figura 3: Diagrama de Proceso para Poblar la Ontología

Al abrir una casilla de correo se observan a simple vista los datos de identificación del correo electrónico (emisor, receptor, asunto, fecha y contenido), pero los *metadatos* (direcciones IP, message ID, entre otros) se encuentran en la *cabecera* del correo electrónico que no es visible de manera directa (ver Figura 4).

Así que el proceso de extracción debe prever el acceso a ambos espacios para obtener los datos que sustentan técnicamente el trabajo pericial.

Según sea el producto, los clientes de correo electrónico que almacenan los datos de cabecera y cuerpo de los correos electrónicos utilizan diferentes estructuras de almacenamiento para los correos electrónicos, no obstante, las herramientas forenses identificadas en [30] permiten el acceso a todos estos datos de manera directa, sin necesidad de realizar procesos de conversión previos.

Figura 4: Cabecera de un correo electrónico

Una vez obtenido el corpus de correos electrónicos, se inicia el trabajo de procesamiento con herramientas propias de DM. Por empezar, de manera automática es posible asignar el tipo de dato y otras características necesarias (identificación del campo ID, identificación de campo etiqueta para procesos de predicción, etc.) en el mismo proceso de lectura del archivo. Esta etapa es la denominada ***Lectura y asignación de tipos de atributos*** en el proceso señalado en la Figura 3.

Para la ***Selección de Instancias a Incorporar*** se recurre a las técnicas de DM que sean necesarias para conformar el conjunto de instancias que se utilizarán posteriormente para poblar la ontología.

En esta etapa es cuando se agrega valor al proceso de instanciación, ya que –usualmente- esta identificación de información relevante se realiza *manualmente* y con la presencia del experto.

En la etapa de ***Pre-procesamiento*** se recurre a los componentes para depuración y limpieza de datos que habitualmente están incorporadas en las herramientas de DM, y que en el caso de los procesos de actualización de las ontologías se trabajan desde las técnicas de ISR.

Según sea el caso, se puede trabajar con técnicas de Categorización, Asociación, Filtrado, Agregación, entre otras.

Por último, en la etapa de ***Emisión de Resultados*** se obtiene el conjunto de datos seleccionados y depurados que poblarán la ontología, en formato de archivo tabular.

Tomando como entradas el archivo tabular resultante y un archivo de mapeo, un parser generará las instancias correspondientes a los conceptos y relaciones de la ontología propuesta.

Existen herramientas semi-automáticas para realizar esta actividad, se deberá definir cuál es la más pertinente y que más se adecue al modelo ontológico y al dominio de estudio.

5.3. Caso práctico

A modo de ejemplo se trabajará con un conjunto de 1162 correos electrónicos[10] alojados en un archivo *.pst*.

Se puede suponer un caso judicial en el cual se pretende probar que existe una asociación ilícita entre un grupo de empleados, y se presume que en la casilla de mail del supuesto cabecilla hay información para establecer la cadena de relación con el resto de los imputados de quienes no se sabe su identidad.

Se solicita al perito el análisis forense de la cuenta del actor, según el siguiente punto de pericia: "*identificar todos las usuarios con quienes el actor se comunicó por correo electrónico durante el período considerado, señalando para cada caso en cuantas oportunidades remitió correos a cada usuario identificado, y a quienes lo hizo en forma conjunta".*

A continuación se describirán los pasos del procedimiento propuesto.

5.3.1. Extracción Inicial

De la narrativa sobre el punto de pericia se obtienen los datos requeridos para instanciar la ontología: email del actor (*emisor*), email de los usuarios con los que se comunicó (*receptores*), período considerado (*fecha_de_emisión)*[11].

En el caso que nos ocupa, el conjunto de correos electrónicos de una cuenta están encapsulados en un archivo *.pst* según se muestra en la Figura 5.

Identificado el equipo en el cual el actor gestiona habitualmente su correo electrónico, se recurre a la herramienta *Autopsy*[12], para realizar la extracción de los datos necesarios que permitirán responder a los puntos de pericia, generando un archivo de texto plano con los

[10] El conjunto de correos que se toman como banco de pruebas del presente trabajo fue extraído del corpus de correos electrónicos de la empresa ENRON, que forman parte del proyecto *The Enron Data Set Cleansed of PII by Nuix and EDRM.* Ver: http://info.nuix.com/EnronDownload2013.html?mkt_tok=3RkMMJWWfF9wsRokvK7BZKXonjHpfsX56OgvWKa%2FlMI%2F0ER3fOvrPUfGjI4EScVlI%2BSLDwEYGJlv6SgFQ7XCMap637gOUhg%3D

[11] Cuando el caso amerite, deberán trabajarse el resto de las propiedades que se deban instanciar.

[12] Herramienta libre para la extracción de metadatos de correos electrónicos (Versión 3.1.1) la http://www.sleuthkit.org/autopsy/

datos de cabecera de los 1162 correos electrónicos que contiene la casilla.

Figura 5**:** Vista de correos con MS-OUTLOOK

La Figura 6 muestra la salida de datos correspondiente a esta extracción, obtenido como archivo en formato de planilla de cálculo.

Figura 6**:** Datos Entrada a RapidMiner

La Figura 7 muestra una *vista minable*, con la secuencia de operadores de *RapidMiner*, que intervienen en el resto de los pasos del procedimiento señalado.

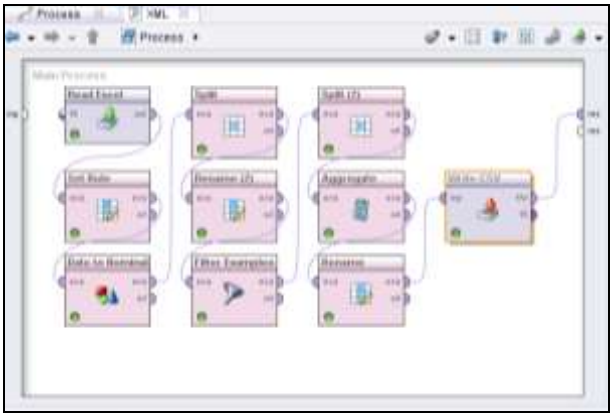

Figura 7**:** Vista Minable

5.3.2. Lectura y asignación de tipos de atributos

Ya desde *RapidMiner,* se carga el archivo generado en el paso anterior, utilizando los siguientes operadores:

- El operador *Read_XLS* permite obtener los datos directamente desde el archivo exportado por la herramienta forense,
- Con *Set_Role* se asigna al atributo Message_ID como campo ID para asegurar la unicidad de los registros.

5.3.3. Selección de Instancias a Incorporar

Una vez leidos y cargados los datos, se procede a seleccionar las instancias teniendo presente las restricciones señaladas en los puntos de pericia: email del actor (*emisor*), email de los usuarios con los que se comunicó (*receptores*), período considerado (*fecha_de_emisión)*

Las instancias que se deben seleccionar son aquellas cuya *fecha_de_emisión* se incluya en el período considerado, que servirán de párametros de ingreso del operador *Filter_Examples.* Para ello, se aplican previamente los operadores *Date_to_nominal, Split y Rename* que ajustan el formato de *fecha_de_emisión* a lo requerido por el operador *Filter_Examples*.

5.3.4. Pre-procesamiento

En esta etapa, se utiliza los operadores necesarios para ajustar los datos:

- Con el operador *Splits* se desagregan los atributos *E-mail CC* y *E-mail To*, separando el texto original del atributo en columnas, tantas como partes del texto se encuentre ante el símbolo ";" que usualmente separa los destinatarios en una lista de correo. Para el ejemplo en cuestión, estas operaciones generan 3 columnas por cada atributo *E-Mail CC y E-Mail To.*
- *Aggregate* realiza una selección de los datos de agrupados para identificar a los destinatarios que se les remitió correos de manera simultánea.

5.3.5. Emisión de Resultados

La emisión de resultados incluye dos operadores: *Rename* para prolijar los nombres de campos del archivo de salida y *Write CSV* para generar un archivo plano con los datos procesados.

En la Figura 8 se muestra parte de los datos resultantes.

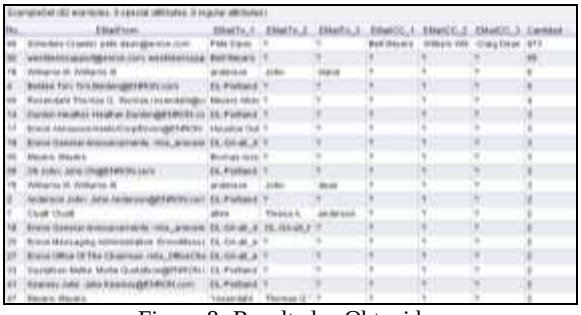

Figura 8**:** Resultados Obtenidos

5.3.6. Población de la ontología

Los datos generados con el pre-procesamiento realizado en *RapidMiner* serán luego instanciados en la ontología propuesta, según el siguiente mapeo: *EMaiFrom* se instanciará como *emisor*, mientras que *EmailTo_n* como *EMailCC_n* representan los correos que se instanciarían con *receptor* al poblar la ontología.

6. Conclusiones

En este trabajo se incursionó en el estudio de técnicas de ISR y DM como herramientas que permitan construir procesos semiautomáticos destinados a seleccionar y depurar la información que servirá luego para poblar una ontología.

Se describió en un diagrama de procesos tal propuesta, y se aproximó el estudio con un caso práctico que permitiera mostrar la factibilidad de ello, partiendo de un corpus de datos documentales (como son los correos electrónicos obrantes en una casilla).

El proceso semiautomático señalado deberá validarse a futuro para generalizar la propuesta a partir de la prueba de aplicación en otros casos prácticos y estableciendo métricas de rendimiento (precisión de la extracción, exhaustividad, velocidad de procesamiento, etc).

Se entiende que se ha cumplido con el objetivo de establecer las bases para formular una herramienta que permita poblar la ontología, restando todavía la validación de la generalización de la propuesta a partir de pruebas experimentales que incluyan las variantes necesarias, como ser:

- a) la comprobación con más casos periciales como el citado, e incluso de diferente contenido al del ejemplo;
- b) la cobertura plena del modelo de la ontología, i.e., a todas las clases, propiedades y relaciones definidas en la misma; y
- c) la inclusión de otras técnicas de DM que también pueden ser de utilidad para la depuración previa de los datos (la categorización de textos, por ejemplo).

7. Agradecimientos

Este trabajo ha sido financiado en forma conjunta por CONICET, la UTN (PID 25-O156) y el Consejo de Investigaciones de la Universidad Católica de Salta. Se agradece el apoyo brindado por estas instituciones, y en particular por el Grupo de Investigación sobre Minería de Datos de la Universidad Católica de Salta. Se considera importante destacar la colaboración técnica brindada por el Ing. Esteban Rivetti en la ejecución de las pruebas del caso de estudio con las herramientas forenses utilizadas.

8. Referencias

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

[2] Wen-Yang Lin, Ontology-Based Data Mining A Case in Multidimensional Association, Dept. of Computer Science and Information Engineering , National University of Kaohsiung, 2006

[3] Mohapatra, M. S., Ramesh, D., Dash, M. C., & Behera, M. P. C. *DATA MINING-A DOMAIN SPECIFIC ANALYTICAL TOOL FOR DECISION MAKING*.

[4] Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 7.

[5] Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., & Benredjem, D. (2009). Towards an integrated e-mail forensic analysis framework. *digital investigation*, *5*(3), 124-137[6] Iqbal, F., Binsalleeh, H., Fung, B. C., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications.*Information Sciences*, *231*, 98-112.

[7] Airoldi, E., & Malin, B. (2004, November). Data mining challenges for electronic safety: the case of fraudulent intent detection in e-mails. In *Proceedings of the workshop on privacy and security aspects of data mining* (pp. 57-66).

[8] Appavu, S., Rajaram, R., Muthupandian, M., Athiappan, G., & Kashmeera, K. S. (2009). Data mining based intelligent analysis of threatening e-mail. *Knowledge-Based Systems*, *22*(5), 392-393.

[9] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications*,*41*(13), 5948-5959

[10] Armknecht, F., & Dewald, A. (2015). Privacy-preserving email forensics. *Digital Investigation*, *14*, S127-S136.

[11] Stuit, M., & Wortmann, H. (2012). Discovery and analysis of e-mail-driven business processes. *Information Systems*, *37*(2), 142-168.

[12] Tang, J., Li, H., Cao, Y., & Tang, Z. (2005, August). Email data cleaning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 489-498). ACM.

[13] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing* 5(4):2000

[14] Berg, R., Dickson, E., & Choi, C. (2013). Warren Spector Email Migration: Final Presentation, University of Texas at Austin• School of Information•INF392K• Spring 2013

[15] Squire, M. (2013, October). A replicable infrastructure for empirical studies of email archives. In *Replication in Empirical Software Engineering Research (RESER), 2013 3rd International Workshop on* (pp. 43-49). IEEE.

[16] Bacchelli, A., Lanza, M., & D'Ambros, M. (2011, May). Miler: A toolset for exploring email data. In *Software Engineering (ICSE), 2011 33rd International Conference on* (pp. 1025-1027). IEEE.

[17] Kota, V. K. (2012). An ontological approach for digital evidence search.*International Journal of Scientific and Research Publications*, *2*(12).

[18] Bender, C., Deco J., González S., J., Hallo, M., Ponce G., J. (2014) Tópicos Avanzados de Bases de Datos 1a ed. - *Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn), 2014*.

[19] Shchekotykhin, K., Jannach,D., Friedrich, G. & Kozeruk, O. (2005). ALLRIGHT: Automatic Ontology Instantiation from Tabular Web Documents. *Lecture Notes in Computer Science 4825, 466–479*

[20] Fierros, J. D. G. TESIS DE MAESTRÍA EN CIENCIAS, Poblado Automático de Ontologías Espaciales a Partir de Texto no Estructurado, *Centro Nacional de Investigación y Desarrollo Tecnológico, Departamento de Ciencias Computacionales, Cuernava, Mexico, 2012*.

[21] Paredes Moreno, A. (2007). Técnicas de depuración e integración de ontologías en el ámbito empresarial.

[22] Cala A., Schorlemmer M, Noriega P., PROTOTIPO DE UN MODULO DE BUSQUEDA SEMANTICA PARA LA PLATAFORMA GreenIDI. TR--IIIA--2013--01, *IIIA-CSIC Barcelona, 2013*

[23] Daly, M., Grow, F., Peterson, M., Rhodes, J., & Nagel, R. L. (2015, April). Development of an automated ontology generator for analyzing customer concerns. In *Systems and Information Engineering Design Symposium (SIEDS), 2015* (pp. 85-90). IEEE.

[24] Hacherouf, M., Bahloul, S. N., & Cruz, C. (2015). Transforming XML documents to OWL ontologies: A survey. *Journal of Information Science*,*41*(2), 242-259

[25] Fiel Cortes, D. (2012). XML y Web Semántica. Diseño y población semiautomática de ontologías.

[26] Ruiz Martínez, J. M. (2012). Metodología para la población automática de ontologías: aplicación en los dominios de medicina y turismo.

[27], Santosh D. T., & Vardhan, B. V. OBTAINING FEATURE-AND SENTIMENT-BASED LINKED INSTANCE RDF DATA FROM UNSTRUCTURED REVIEWS USING ONTOLOGY-BASED MACHINE LEARNING, *International Journal of Technology (2015) 2: 198-206*

[28] Shchekotykhin, K., Jannach, D., Friedrich, G., & Kozeruk, O. (2007). *AllRight: Automatic ontology instantiation from* tabular Web documents (pp. 466-479). *Springer Berlin Heidelberg.*

[29] Buitelaar, P., & Cimiano, P. (2008). Ontology learning and population: bridging the gap between text and knowledge (Vol. 167). *Ios Press.*

[30] Beatriz P. de Gallo, Marcela Vegetti, Horacio Leone, "Ontología para el Análisis Forense de Correo Electrónico", *CoNaIISI 2014* - ISSN: 2346-9927 - Página 1008-1018

[31] Baeza-Yates, R. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*. Addison-Wesley. ISBN 0-201-39829-X