

Sistema de análisis y búsqueda para sábanas telefónicas

Alurralde, Ramiro
ramo160689@gmail.com

Bigatti, Julián
jbigatti@gmail.com

Directora: Laura Alonso Alemany

Sistema de análisis y búsqueda para sábanas telefónicas

Resumen

En la actualidad el constante uso de la telefonía celular le da a la policía una herramienta más a la hora de afrontar una investigación criminal. Suponiendo la existencia de una comunicación durante la ejecución de un ilícito, el investigador logra obtener, a través de las empresas de telefonía, una determinada cantidad de información que sirve como prueba para demostrar la culpabilidad del sospechoso.

El trabajo de análisis y procesamiento de la información entregada por las empresas se realiza de forma manual consumiendo grandes cantidades de horas/hombre y arrastrando errores humanos propios de agotamiento intelectual que el análisis produce.

Astrea es un software para la unificación, búsqueda y análisis de datos en sábanas telefónicas de manera automatizada y sistemática. Proveyendo al investigador de una herramienta que maximiza el rendimiento y reduce los errores de búsqueda y análisis.

El software presenta una interfaz de usuario fácil e intuitiva, permitiendo que usuarios no expertos en informática y/o en el uso de sábanas telefónicas, se adapten a él pudiendo obtener los mismos resultados que un usuario experto.

Introducción y Motivación

En este trabajo presentamos un sistema para facilitar la exploración de datos sobre comunicaciones telefónicas en investigaciones policiales.

Actualmente, estos datos se trabajan de forma muy manual y poco sistemática, mediante hojas de cálculo o incluso hojas de papel. Cada empresa proporciona los datos en un formato propio e incompatible con el resto de empresas, con lo cual la integración de datos resulta prácticamente inviable. Por estas razones, el investigador debe prestar mucha atención a detalles de forma, lo cual le resta capacidad de concentración y tiempo para focalizarse en los conceptos a investigar.

Hemos desarrollado Astrea, un software para facilitar la exploración de los datos de comunicaciones telefónicas, con las siguientes funcionalidades:

- Integra los datos proporcionados por las diferentes empresas en una representación común que representa los conceptos cruciales de las comunicaciones telefónicas.
- Ofrece una interfaz gráfica para la exploración de los datos, con opciones de búsqueda y filtros organizados de forma intuitiva y bien documentados.
- Facilita la integración de los datos con otras aplicaciones, mediante una representación de los datos basada en una ontología según los estándares de la web semántica.

Con esta aplicación esperamos poder acercar a la gran mayoría de investigadores los métodos sistemáticos de exploración de los datos de comunicaciones telefónicas. Estos métodos han ofrecido muy buenos resultados en investigaciones policiales recientes, lo cual nos hace pensar que su masificación puede tener un impacto muy positivo en los resultados de las investigaciones. Al mismo tiempo, esta aplicación reduce significativamente el tiempo necesario para la exploración de los datos, lo cual también facilita su masificación, ya que en muchas ocasiones el tiempo / hombre es un recurso escaso en las investigaciones policiales.

Otro posible campo de aplicación es la minería de datos de comunicaciones telefónicas privadas, por ejemplo, para optimizar los patrones de uso y gasto en comunicaciones empresariales.

Este tipo de aproximación ha resultado muy útil en otras áreas de trabajo en las que también se manejan grandes cantidades de datos con orígenes diversos, como por ejemplo los estudios de mercado o la detección de fraude en tarjetas de crédito.

El resto del artículo se estructura como sigue. En la próxima sección presentamos las funcionalidades básicas y la arquitectura de Astrea. Desarrollamos la descripción de las partes más prominentes del sistema, como son la interfaz de usuario y la ontología subyacente. Finalmente, terminamos presentando algunas conclusiones y líneas de trabajo futuro.

Arquitectura del Sistema

Astrea es un software implementado en python con Qt, independiente de plataforma. Permite procesar datos de comunicaciones telefónicas en formato excel, csv o texto plano, las llamadas sábanas telefónicas. Podemos ver algunos ejemplos de sábanas telefónicas en el Anexo A. Estos datos se integran a una representación abstracta común para todas las empresas telefónicas, que se describe en la siguiente sección. Después, los usuarios pueden hacer búsquedas sobre los datos representados de esta forma, y los resultados de su búsqueda se ofrecen en diferentes formatos.

La arquitectura general del sistema puede verse en la Figura 1. Vemos que existen varios puntos de interacción con el usuario. Para la integración de datos, se solicita al usuario que resuelva los casos ambiguos, por ejemplo Fecha y Hora es la etiqueta utilizada por múltiples empresas para identificar la hora de inicio y/o la hora de finalización de una comunicación. Además, el usuario interactúa en el ingreso de datos, especificando los archivos con los datos y su origen, y también en la búsqueda.

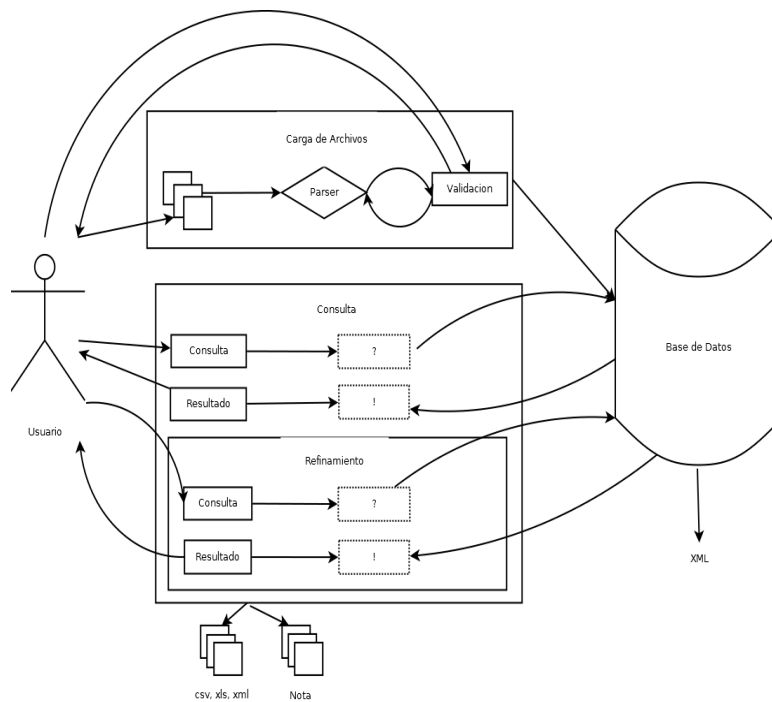


Figura 1

Interfaz de usuario

En la interfaz de usuario hemos tratado de mantener la mayor simplicidad de diseño posible, pero al mismo tiempo proporcionando información de ayuda sobre las diferentes funcionalidades de forma accesible, mediante mouse over, menús contextuales que se pueden acceder de diferentes formas. De esta manera, se garantiza que el usuario encuentre la información necesaria en todo momento. Estas decisiones de diseño han venido determinadas por la necesidad de interactuar con usuarios no expertos en informática, como es el caso de buena parte de los investigadores policiales.

La interfaz está compuesta de tres partes que se usan de forma secuencial:

- Carga de datos: se integran uno o más archivos de entrada a una representación común interna.
- Búsqueda: se seleccionan los filtros de búsqueda a aplicar sobre los datos cargados.

- Visualización de resultados: se le presentan los resultados al usuario de forma amigable para que pueda realizar diferentes acciones sobre los distintos tipos de datos.

Carga e integración de datos

El objetivo principal de la carga de datos es homogeneizar la información de diferentes empresas, que pueden tener distintos formatos y etiquetas identificatorias de los datos. Una vez homogeneizados los datos, resulta más sencillo realizar búsquedas, ya que el usuario se puede abstraer de las particularidades de cada documento. En esta interfaz, que se puede ver en la Figura 2, se observan los principales componentes de la interfaz: el listado de archivos a cargar, donde se especifica su origen (la empresa emisora), y las opciones para asignar un nombre canónico a los identificadores propios de cada empresa. De esta forma, identificadores como Fecha I., Fecha y Hora y Fecha serán unificados a un mismo nombre, Fecha, lo cual permite al usuario hacer búsquedas focalizándose en el concepto, sin tener que recordar los nombres específicos que usa cada empresa.

Cabe destacar que una vez ingresadas nuevas etiquetas, el sistema no volverá a preguntar por ellas a menos que presenten algún tipo de ambigüedad con otras existentes.

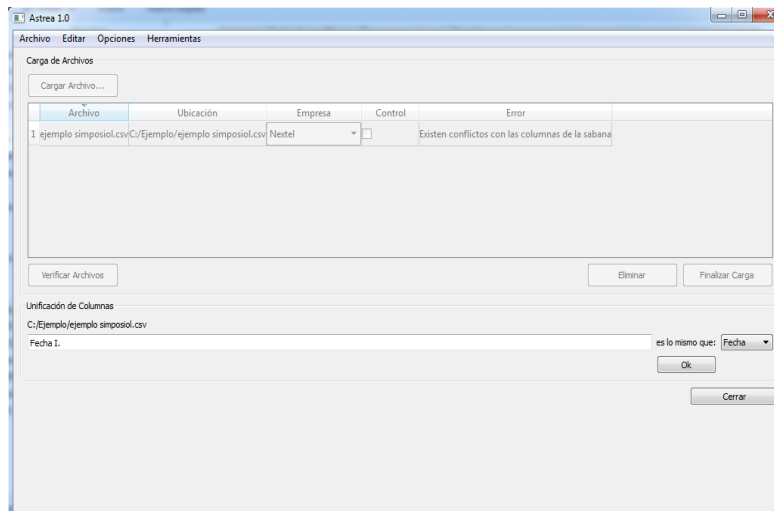


Figura 2

Búsqueda

En la Figura 3 se presentan las diferentes opciones de búsqueda sobre los datos previamente cargados. Se ofrecen las opciones más usadas de manera más accesible y rápida para realizar aquellas búsquedas más comunes. Se provee una búsqueda avanzada (Figura 4) para lograr resultados más específicos.

Además, el usuario puede decidir qué datos visualizar una vez realizada la búsqueda. Es decir, podrá ver la información que crea conveniente y de esta forma tener una visual más “limpia” de los datos.

The screenshot shows the 'Astrea 1.0' application window with a menu bar (Archivo, Editar, Opciones, Herramientas) and a search configuration panel. The panel is divided into several sections:

- Opciones de Búsqueda:** Contains input fields for 'Nro. de Origen' and 'Nro. de Destino'. It also has date and time range selectors: 'Fecha: Desde: 01/01/1800' to 'Hasta: 11/05/2013', 'Hora: Desde: 12:00:00 A.M.' to 'Hasta: 11:59:59 P.M.', and 'Duración: 0 seg.' to 'Hasta: más de 5 min.'. An 'Antena:' dropdown menu is set to 'Todos'.
- Código de Área:** Includes radio buttons for 'Igual a' and 'Distinto de', a text input '0', and a 'buscar en:' dropdown menu set to 'Ambos'.
- Opciones De Salida:** A grid of checkboxes for selecting data to display. Checked options include 'Nro. de Origen', 'Nro. de Destino', 'Fecha', 'Hora', and 'Dirección de Antena'. Other unchecked options include 'Localidad de Antena', 'Provincia de Antena', 'Id. de Celda', 'Dirección Nro. de Origen', 'Dirección Nro. Destino', 'Localidad Nro. Origen', 'Localidad Nro. Destino', 'Provincia Nro. Origen', 'Empresa Nro. Origen', 'Empresa Nro. Destino', 'Provincia Nro. Destino', 'Duración de Llamada', 'Nro. de Imei', 'Nro. de Sim', 'Titular Nro. Origen', 'Titular Nro. Destino', 'Contenido de Mensaje', and 'Tipo de Comunicación'.

At the bottom right, there are buttons for 'Limpiar Búsqueda', 'Búsqueda Avanzada', 'Buscar', 'Volver al Inicio', and 'Cerrar'.

Figura 3

The screenshot shows the 'Astrea 1.0' application window. The menu bar includes 'Archivo', 'Editar', 'Opciones', and 'Herramientas'. The main interface is divided into several sections:

- Opciones de Búsqueda:** Contains input fields for 'Nro. de Origen' and 'Nro. de Destino', date and time ranges ('Fecha: Desde: 01/01/1800' to 'Hasta: 13/05/2013' and 'Hora: Desde: 00:00:00' to 'Hasta: 23:59:59'), duration ('Duración: 0 seg.' to 'Hasta: más de 5 min.'), and antenna ('Antena: Todos').
- Código de Área:** Includes radio buttons for 'Igual a' and 'Distinto de', a numeric input field, and a 'buscar en:' dropdown menu.
- Búsqueda Avanzada:** A grid of fields for 'Nro. Imei', 'Nro. de Sim', 'Id. de Celdas', 'Tipo de Comunicación(Mej/Llamada)', 'Prov. de Antena', 'Loc. de Antena', 'Dirección de Nro. Origen', 'Dirección de Nro. Destino', 'Localidad de Nro. Origen', 'Localidad de Nro. Destino', 'Provincia de Nro. Origen', 'Provincia de Nro. Destino', 'Titular de Nro. Origen', 'Titular de Nro. Destino', 'Empresa de Nro. Origen', 'Empresa de Nro. Destino', 'Contenido de Mej.', and 'Estado de Mej.'.
- Opciones De Salida:** A grid of checkboxes for various search criteria: 'Seleccionar Todo', 'Nro. de Origen', 'Localidad de Antena', 'Localidad Nro. Origen', 'Provincia Nro. Destino', 'Titular Nro. Destino', 'Nro. de Destino', 'Provincia de Antena', 'Localidad Nro. Destino', 'Duración de Llamada', 'Contenido de Mensaje', 'Fecha', 'Id. de Celda', 'Provincia Nro. Origen', 'Nro. de Imei', 'Tipo de Comunicación', 'Hora', 'Dirección de Antena', 'Dirección Nro. de Origen', 'Empresa Nro. Origen', 'Nro. de Sim', and 'Titular Nro. Origen'.

At the bottom, there are buttons for 'Limpiar Búsqueda', 'Búsqueda Avanzada', 'Buscar', 'Volver al Inicio', and 'Cerrar'.

Figura 4

Visualización de Resultados

La Figura 5 muestra los resultados de la aplicación de los filtros seleccionados anteriormente por el usuario. Se presentan en forma de tabla donde las etiquetas de las columnas son las elegidas por el usuario.

Las distintas opciones que se ofrecen se dividen en 2 (dos) clases principales:

- Acciones sobre los resultados: los más destacados son Ampliar Búsqueda y Refinar Búsqueda. El primero concatena resultados de diferentes búsquedas realizadas, el segundo realiza la búsqueda sobre resultados obtenidos hasta el momento.
- Acciones sobre el tipo de datos de los resultados: dependen del dato seleccionado, de esta forma, si el usuario selecciona una dirección (Figura 6), se le ofrece la opción de visualizarla en google maps, según la dirección obtenida de los datos. Si lo seleccionado fuera un número telefónico, podrá identificarlo con un nombre más amigable, un alias, (Figura 7) o generar un

documento oficial para realizar la intervención de dicho número y/o pedir el arresto del titular de la línea si este dato existiera en el detalle. Astrea, además ofrece una herramienta para consultar las localidades abarcadas por un código de área determinado. Así como imprimir o exportar los resultados obtenidos.

Haciendo click en la celda deseada, se despliega el menú de opciones

Nro. de Origen	Nro. de Destino	Fecha	Hora	Duracion de LLamada	Nro. de Imei	Nro. de Sim	Id. de Celda	Direccion Nro. Orige
1 numero1	numero5	2009-09-21	18:33:11	00:00:03			NORTE	
2 numero2	numero5	2009-09-21	23:59:59	00:00:03			NORTE	
3 numero2	numero5	1800-01-01		00:00:03				
4 numero2	numero5	2009-09-21	18:36:57	00:00:03			NORTE	
5 numero2	numero5	2009-09-21	18:37:41	00:00:02			NORTE	

Buttons: Imprimir..., Refinar Búsqueda, Guardar..., Ampliar Búsqueda..., Nueva Búsqueda, Volver al Inicio, Cerrar

Figura 5

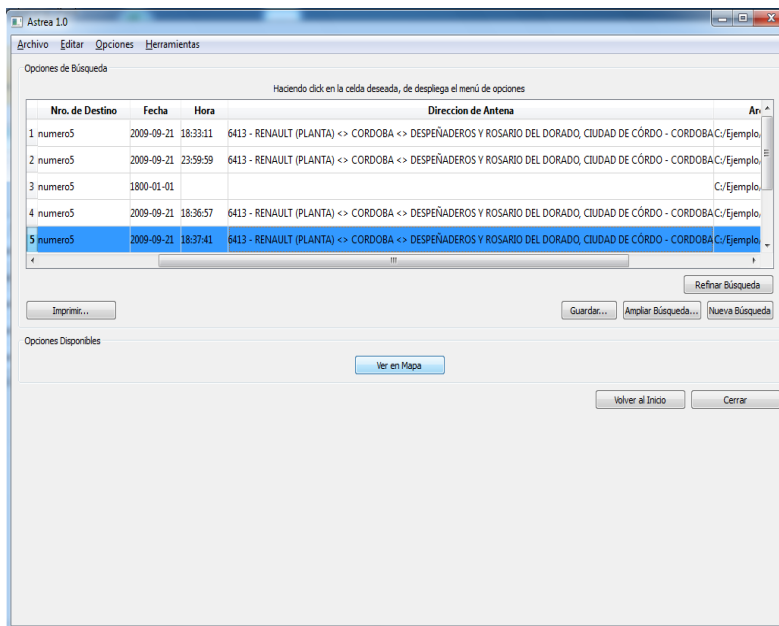


Figura 6

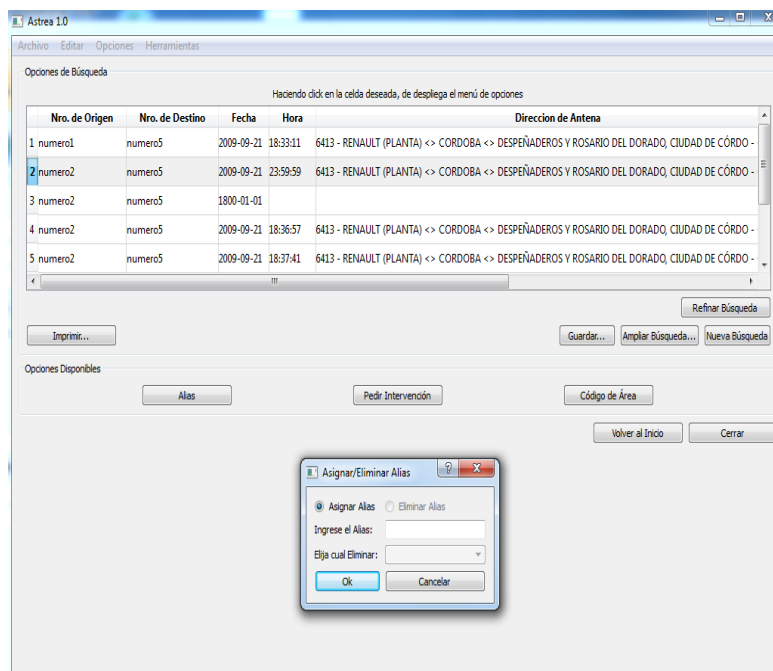


Figura 7

Ontología subyacente

Para facilitar la interoperabilidad de los resultados obtenidos de la búsqueda en los datos con otras bases de datos o aplicaciones, creamos una ontología subyacente para representar los datos de las sábanas telefónicas. Esta ontología modela los conceptos involucrados en la comunicación telefónica, de forma que resulte intuitivo trabajar con ellos y se puedan exportar para su fácil integración con diferentes herramientas.

En la versión actual se han implementado métodos de comunicación con google maps y se ha diseñado la interacción con weka. En versiones posteriores se desarrollarán los métodos para interactuar con google analytics y servicios en la nube, así como con herramientas de rastreo y surveillance.

La ontología se ha implementado siguiendo los estándares propuestos en web semántica. Se ha implementado en OWL (<http://www.w3.org/TR/owl-features/>), mediante la herramienta Protégé (<http://protege.stanford.edu/>), que facilita la visualización gráfica de la ontología y la comprobación de propiedades.

Un resumen de la ontología se puede ver en la Figura 8. Consta de 12 clases, con 13 relaciones entre ellas (Figura 9). Las clases principales son:

- Celda: Representa las celdas que se encuentran en las antenas.
- Teléfono: Representa al artefacto teléfono, dicho aparato, puede tener solo 1 chip (si bien en la actualidad existen aparatos que soportan hasta 2 (dos) chips, recordemos que solo modelamos una comunicación)
- Dirección: representa la dirección domiciliaria de determinadas clases como una antena, un titular. Tiene como atributos la provincia, localidad, código postal
- Antena: representa a las antenas físicas, que poseen en ellas una gran cantidad de celdas.
- Chip: representa al sim entregado por las empresas de telefonía, éste sólo puede tener a lo sumo 2 titulares, en cuyo caso uno de ellos es el usuario y el otro el titular propiamente dicho.
- Titular: representa a la persona que se supone adquirió el chip.

Las relaciones que se establecen entre las clases antes mencionadas modelan los eventos y relaciones propios de la comunicación telefónica.

- estaEn: es asimétrica, funcional e irreflexiva. Relaciona a las celdas con las antenas. Una celda esta solo en una antena.
- tieneCelda: es inversa funcional. Relaciona a las antenas con las celdas. Las antenas tienen enorme cantidad de celdas.
- usa: es asimétrica e irreflexiva. Relaciona a los teléfonos con las antenas, es decir, indica que antena está usando un teléfono determinado. Un teléfono usa solo una celda.
- hospeda: es asimétrica e irreflexiva. Relaciona a las antenas con los teléfonos, es decir, indica que teléfonos esta “hospedando” una antena determinada. Una antena hospeda gran cantidad de teléfonos.
- recibe: es asimétrica e irreflexiva. Relaciona un teléfono con otro indicando cuál de ellos es el receptor de la comunicación. Un teléfono recibe solo una llamada.
- Llama: es asimétrica e irreflexiva. Relaciona un teléfono con otro indicando cuál de ellos es el que inicia la comunicación. un teléfono solo llama a un teléfono.
- tieneTelefono: es asimétrica e irreflexiva. Relaciona un chip con un teléfono indicando que aparato está utilizando al momento de la comunicación. Un chip solo tiene un teléfono.
- tieneChip: es asimétrica e irreflexiva. Relaciona un teléfono con un chip indicando que chip se utilizó para la comunicación. Un teléfono tiene solo un chip.

- esTitularDe: es asimétrica e irreflexiva. Relaciona un usuario con un chip. Un usuario puede tener diferentes chips.
- tieneTitular: es asimétrica e irreflexiva. Relaciona un chip con un usuario. Un chip puede tener a lo sumo 2 (dos) titulares.
- tieneDireccion: es asimétrica, inversa funcional e irreflexiva. Relaciona un usuario o antena con una dirección domiciliaria. Solo se puede tener una dirección.
- direccionDe: es asimétrica, funcional e irreflexiva. Relaciona una dirección domiciliaria con un usuario o antena. Una dirección solo puede tener un usuario o antena.
- pedirNota: es asimétrica e irreflexiva. Relaciona una entidad “investigable” con la nota oficial correspondiente a la entidad.

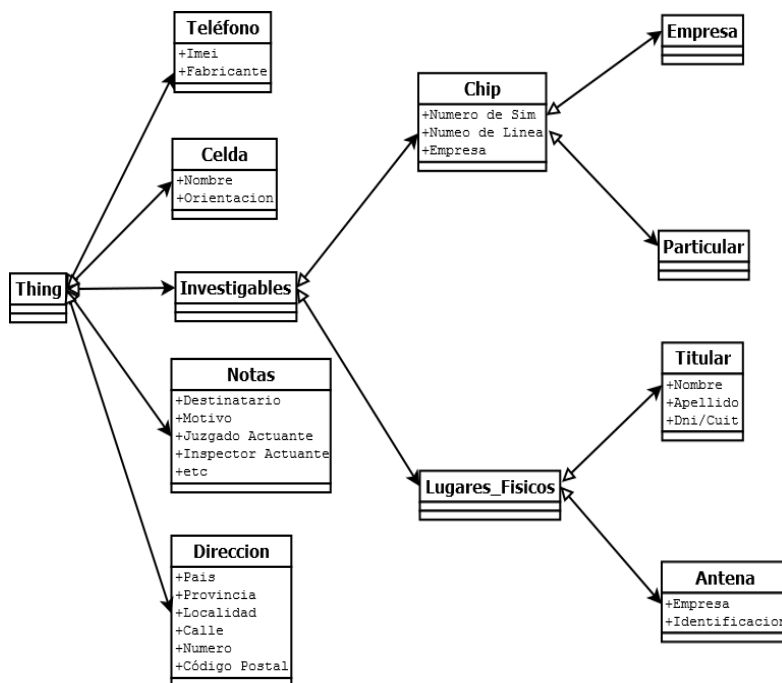


Figura 8

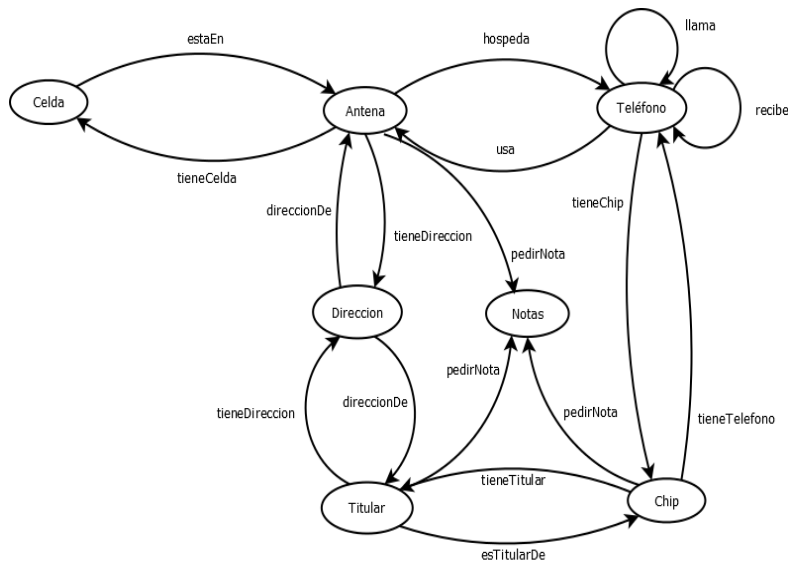


Figura 9

Conclusiones y Trabajo Futuro

La formalización del conocimiento experto y las búsquedas en el dominio restringido permitieron optimizar el rendimiento de los integrantes de las fuerzas policiales, pudiendo así contar con una herramienta de rápida aplicación en el campo.

Trabajo Futuro:


Trazados realizados por un teléfono móvil: Ofrecer la posibilidad de trazar en un mapa el recorrido realizado por los cambios de antenas de una comunicación. Uniendo las antenas que se fueron activando durante la misma.

La exportación de los datos de entrada en formatos compatibles con herramientas de minería de datos y agentes inteligentes para el análisis de patrones en las comunicaciones.

Anexo A

Personal

Ejemplo: Detalles de llamadas de un número en particular.



Línea	Fecha	Hora	Tipo	Otro	Durac	Serie	Celda Id	Celda direccion	Celda localidad	Celda provincia
351300000	20/04/05	00:05:16	S	351300000	320	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	11:01:23	S	351300000	12	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	11:04:41	S	351300000	32	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	11:05:35	S	351300000	1881	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	13:59:40	S	351300000	1959	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	14:25:23	S	351300000	335	6612522	KCZP05	0	Cordoba	Cordoba
351300000	20/04/05	14:35:36	S	351300000	933	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	15:43:32	S	351300000	435	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba
351300000	20/04/05	15:51:39	S	351300000	746	6612522	XCSANJAE	SanLuis 3007	Cordoba	Cordoba

Línea: Número telefónico del cual se pidió el detalle.

Fecha: Fecha de la comunicación

Hora: Hora de inicio de la comunicación

Tipo: Dirección de la comunicación. "E" significa entrante, "S" saliente.

Otro: Número con el cual se estableció la comunicación, ya sea entrante o saliente.

Durac: Duración en tiempo de aire.

Serie: Número de IMEI del teléfono perteneciente a la columna "Línea"

Celda Id: Celda utilizada para la comunicación.

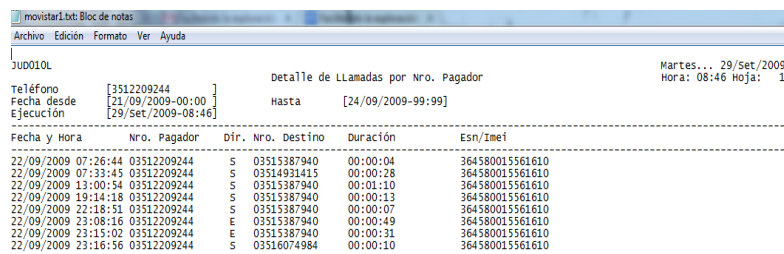
Celda direccion: Dirección de la antena donde está ubicada la celda utilizada para la comunicación.

Celda localidad: Localidad donde se encuentra la antena.

Celda provincia: Provincia donde se encuentra la antena.

Movistar:

Ejemplo: Detalle de llamadas de un número en particular.



movistar.txt: Bloc de notes

Archivo Edición Formato Ver Ayuda

JUD010L

Detalle de Llamadas por Nro. Pagador

Martes... 29/Set/2009
Hora: 08:46 Hoja: 1

Fecha y hora	Nro. Pagador	Dir.	Nro. Destino	Duración	Esn/Imei
22/09/2009 07:26:44	03512209244	S	03515387940	00:00:04	364580015561610
22/09/2009 07:33:45	03512209244	S	03514931415	00:00:28	364580015561610
22/09/2009 13:00:54	03512209244	S	03515387940	00:01:10	364580015561610
22/09/2009 19:14:18	03512209244	S	03515387940	00:00:13	364580015561610
22/09/2009 22:18:51	03512209244	S	03515387940	00:00:07	364580015561610
22/09/2009 23:08:16	03512209244	E	03515387940	00:00:19	364580015561610
22/09/2009 23:15:02	03512209244	E	03515387940	00:00:31	364580015561610
22/09/2009 23:16:56	03512209244	S	03516074984	00:00:10	364580015561610

Fecha y Hora: Fecha y hora del inicio de la comunicación

Nro. Pagador: Número del cual se pidió el detalle.

Dir.: Sentido de la comunicación, es decir, si fue entrante o saliente para el Nro. Pagador.

Nro. Destino: Segundo número que interviene en la comunicación.

Duración: Tiempo de llamada.

Esn/Imei: Número de Imei del número investigado.