Investigación y Tecnología Vol. 3(1-2): 46-55. Diciembre 2015. ISSN 2306-0522

Data-Text Mining y construcción de *datawarehouse* para la imputación formal en delitos de acción pública

Data-Text Mining and Construction Datawarehouse for formal charges incrimes against Public order

Celia Elena Tarquino Peralta
Instituto de Investigaciones en Informática
Carrera de Informática
Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés
La Paz - Bolivia
Autor de correspondencia:celiaetp@hotmail.com

Resumen

Todos los días el Ministerio Publico recibe denuncias de la comisión de delitos y es en la etapa preparatoria que se realiza la investigación a cargo del Fiscal conjuntamente la Policía. En esta fase se unifican los datos de diferentes órganos, fuentes de datos heterogéneas como textos, hojas electrónicas, bases de datos (de diferentes gestores) en un repositorio denominado *datawarehouse* sobre el cual se aplica la *minería de datos*, además, el Fiscal para calificar el hecho debe considerar, analizar la normativa nacional e internacional, jurisprudencia, doctrina, diccionarios jurídicos y bibliografía adicional, sobre esta última información no estructurada se realiza la *minería de texto*. Se inicia por investigar el proceso penal, la calificación del delito, el análisis del expediente y de las instituciones que coadyuvan a la imputación formal de la comisión de un delito a una persona, siguiendo con la investigación sobre las metodologías, técnicas, herramientas para el *datawarehouse* y *datamining* y *textmining*, con este análisis se ingresa al desarrollo multidimensional concretizado en las tablas de hechos y dimensiones en el cual se aplica la minería de datos, con lo que se logró aminorar los tiempos de investigación que realizan el Fiscal con la ayuda de los policías y realizar la calificación teniendo la información en forma instantánea.

Palabras clave: *Datawarehouse, Datamining, Pentaho, Weka, e-juicio, Tableau, Ethnograph*.

Abstract

Every day the Public Prosecutor receives allegations of criminal offenses and in the preparatory stage is that the research is conducted jointly by the Fiscal Police. At this stage data from different organs, heterogeneous data sources such as text, spreadsheets, databases (different managers) in a repository called the data warehouse on which data mining applies also unify the Prosecutor to qualify the fact should consider, analyze national and international regulations, case law, doctrine, legal dictionaries and additional literature on the latter unstructured text mining is done. It begins to investigate the criminal case, the classification of the crime, analysis of the file and institutions that contribute to the formal accusation of a crime to a person, according to research on methodologies, techniques, tools for datawarehouse and datamining and textmining with this analysis is entered into the multidimensional development materialized in the fact tables and dimensions in which data mining is applied, so that it was possible to reduce the time of research carried out by the Prosecutor's help policemen and perform qualifying having information instantly.

Keywords*: Datawarehouse, Datamining, Pentaho, Weka, e-juicio, Tableau, Ethnograph*.

Introducción

Según la estadística realizada por el Consejo de la Magistratura en la gestión 2013, se reportó un total de 801.523 causas ordinarias en todas las materias, civil, penal, familiar, administrativa, laboral, niñez y adolescencia, etc., que se encuentran en trámite o en movimiento en el sistema judicial boliviano.

De este total, más del 50% son causas nuevas que ingresaron durante la gestión y el resto fueron procesos pendientes de resolverse de gestiones anteriores. A ello se suma el volumen de la demanda superior a la capacidad razonable de respuesta del Órgano Judicial.

Entre algunos factores que dan lugar al congestionamiento están: Suspensión de audiencia por inconcurrencia de las partes,- programación de audiencias a la misma hora, prolongación de las audiencias, declaratoria de recesos que duran días y hasta semanas, indebida ampliación de investigaciones, inefectividad en los controles del juez cautelar, desnaturalización del principio de oralidad, falta de uniformidad interpretativa y aplicación de normas y otros.

Con la tecnología informática, *datawarehouse*, *datamining y textmining* se pretende aminorar este tiempo del proceso penal apoyando a la investigación y calificación del hecho delictivo a través de la creación de un *datawarehouse* y la minería de datos y texto[1].

Se crea un perfil del imputado en base a los datos obtenidos de diferentes fuentes. Como estrategia, entre las fuentes se consideró las bases de datos de: la Etapa Preparatoria, del Proceso Penal Ordinario, del Sistema de Medidas Cautelares, del Sistema Penitenciario, y datos en Excel de Antecedentes Policiales y de Antecedentes Judiciales lo que da un historial casi completo del imputado, pudiendo incluirse el acceso a los datos del Instituto de Investigaciones Forenses, de donde se podrían extraer los dictámenes de los peritos en el tipo de delito, el Sistema de otros Procesos Judiciales para determinar la litispendencia, el Sistema de SERECI (Servicio de Registro Cívico) para certificar directamente la identidad de la persona objetivamente y en forma inmediata y otros.

Estado de Arte

Proceso Penal Ordinario:se tienen las fases: etapa preparatoria, juicio oral, recursos y ejecución penal.

La etapa preparatoria consiste en los actos iniciales como la intervención policial, denuncia, querella.

En el desarrollose realiza actos de investigación, de prueba, medidas cautelares.

En los actos conclusivos se tiene la acusación ya sea formal y particular con lo que se inicia el juicio oral y en otro caso se concluye el proceso con sobreseimiento o salidas alternativas.

¿Qué actores intervienen en los actos iníciales?	Policía (investigador).-Investiga el hecho delictivo. Ministerio Público (Fiscal).-Tiene una función de dirección y control sobre los actos investigativos. Poder Judicial (Juez Instructor o Cautelar).-Ejerce control jurisdiccional durante los actos investigativos. Defensor.-Defiende al procesado.

[1] Consejo de la Magistratura. Descongestión de la carga procesal - Una medida para agilizar los procesos penales y superar la retardación de justicia

Los involucrados en la investigación en etapa preliminar están el policía, el Ministerio Público y el Poder Judicial a través del Juez Instructor.

El delito es definido como una conducta típica, antijurídica e imputable, sometida a una sanción penal (Harb M. 1998)

La imputación, en derecho procesal penal, es el acto mediante el cual se le acusa formalmente a una persona de un delito concreto.

Órganos Estatales de Investigación de Delitos

Según normativa:

- Constitución Política del Estado. Artículo 225. I. El Ministerio Público defenderá la legalidad y los intereses generales de la sociedad, y ejercerá la acción penal pública.
- Código Procesal Penal. Artículo 70º.(Funciones del Ministerio Público).Corresponderá al Ministerio Público dirigir la investigación de los delitos y promover la acción penal pública ante los órganos jurisdiccionales. Con este propósito realizará todos los gastos necesarios para preparar la acusación y participar en el proceso, conforme a las disposiciones previstas en este Código y en su Ley Orgánica.
- Ley del Ministerio Público. Artículo 74º.- (Policía Nacional).La Policía Nacional en la investigación de los delitos, se encargará de la identificación y aprehensión de los presuntos responsables, de la identificación y auxilio a las víctimas, de la acumulación y aseguramiento de las pruebas y de toda actuación dispuesta por el fiscal que dirige la investigación; diligencias que serán remitidas a los órganos competentes.
- Artículo 75º.- (Instituto de Investigaciones Forenses).El Instituto de Investigaciones Forenses es un órgano dependiente administrativa y financieramente de la Fiscalía general de la República. Estará encargado de realizar, con autonomía funcional, todos los estudios científico – técnicos requeridos para la investigación de los delitos o la comprobación de otros hechos mediante orden judicial

Datawarehouse (DWH).

Según Bill Inmon "Un *Data Warehouse* es una colección de datos orientados a un tema, integrados, no volátiles, e historiados, organizados para dar soporte al proceso de ayuda a la toma de decisiones"

Sus características son:

Integrado	Recoge los datos de diferentes sistemas operacionales.
Orientado a temas	Excluye la información que no será utilizada por el soporte de decisiones, y la clasifica en base a los aspectos que son de interés para la empresa.
De tiempo invariante	La información del DWH es requerida en cualquier momento
No volátil	Nos proporciona una base de datos multidimensional estable donde se cargan datos y se acceden.

Como se observa en la figura No. 1 la arquitectura de un datawarehouse compren de:

a) Fuentes de extracción de datos: estas fuentes, tales como los sistemas OLTP (*On-Line Transaction Processing*), fuentes externas de datos, etc., son las que proveen toda la información que será almacenada en el DWH, sin antes haberla transformado e integrado.

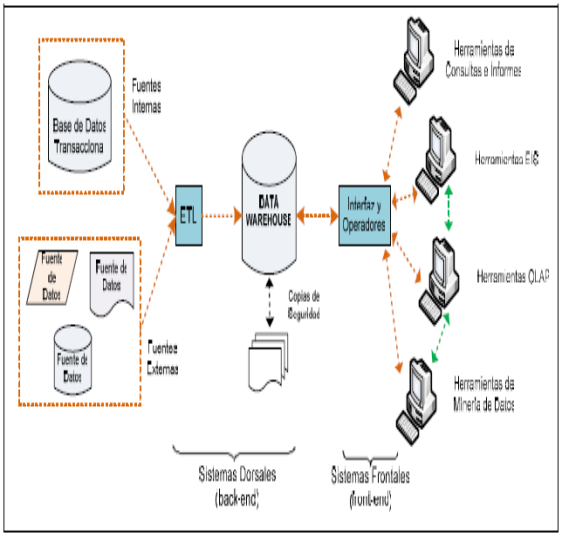

Figura No. 1 Arquitectura Real del Data Warehouse
Fuente: Sepúlveda, J & Urrutia, L

b) Sistemas dorsales (*back-end*): son herramientas que se encargan de la extracción, limpieza, transformación, integración, carga y refresco de los datos de las fuentes externas. Dentro de la clasificación de sistemas dorsales, podemos mencionar los sistemas ETL y el mismo DWH, que se describen a continuación.

c) Sistemas ETL (*Extract, Transform and Load*): realizan funciones como:

- Extracción de los datos desde las distintas fuentes de datos.
- Transformación o filtrado de los datos: limpieza, consolidación, etc.
- Carga inicial del *Datawarehouse*: ordenación, agregaciones, etc.
- Refresco del *Datawarehouse*: es una operación periódica que propaga los cambios de las fuentes externas al *Data Warehouse*.

d) *Datawarehouse*: posee información relevante de los datos o metadatos, estos metadatos, contienen información relativa a los datos y comúnmente, permiten mantener información sobre:

- La procedencia y frecuencia de refresco de los datos.
- La fiabilidad y forma de cálculo de los datos
- Semántica de los datos y su localización en el DW, etc.

e) Sistemas frontales (*front-end*) Herramientas *front-end*: ofrecen al usuario final mecanismos de acceso a la información simples y potentes, permitiendo obtener una correcta y eficaz conexión con el DWH. Con estos sistemas se pueden realizar consultas, informes, análisis y extracción del conocimiento (*data mining*). Dentro de esta clasificación podemos mencionar las Interfaces y Gestores de Consultas, y los Sistemas de Integridad y Seguridad.

DataMarts

A partir del depósito de datos integrado del DWH, se pueden crear los *datamarts*. Los *datamarts* son repositorios de datos o pequeños DWH centrados en un tema o un área de negocio específico. En muchos casos, los *DataWarehouse* comienzan siendo *DataMarts* con el objetivo de minimizar los riesgos para luego ir ampliando su espectro gradualmente. (Ver Figura No. 2).

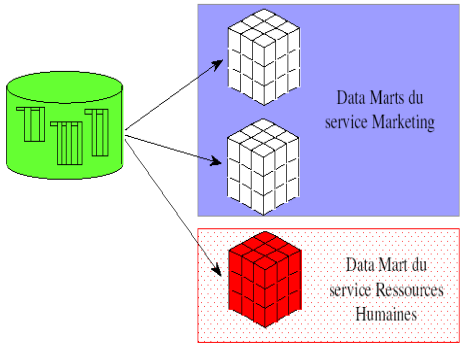

Figura No. 2 Ejemplo de *Datamarts* de Marketing y Recursos Humanos
Fuente: Donsez, D. (2012)

Modelo Multidimensional

Es una técnica que les permite visualizar a los usuarios finales las relaciones que existen en el modelo de una manera más simple y entendible. Este modelo permite el empleo de cualquier base de datos como MOLAP (matrices multidimensionales), base de datos relacionales**,** ROLAP(transforma datos multidimensionales en operaciones relacionales en SQL), etc. Un ejemplo es el que muestra la figura No. 3. Dimensiones de mercado, tiempo y producto.

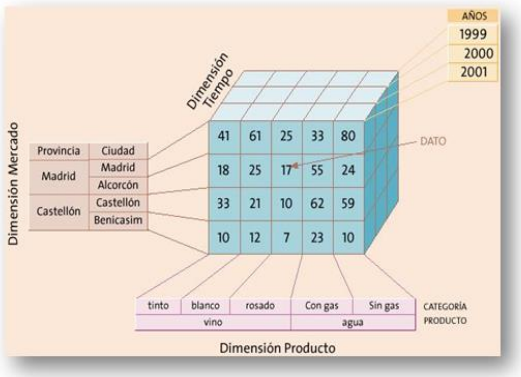

Figura No. 3 Las Dimensiones de Ventas
Fuente: Herrera, C. (2007)

Elementos

- Tabla de hechos o síntesis: contienen hechos. Un hecho es un dato sensible al tiempo que es funcionalmente dependiente de las dimensiones que lo definen.
- Tabla de dimensiones: poseen atributos conocidos.

Ejemplo: Una matrícula es un hecho en el que un estudiante, en una determinada fecha, a través del banco o el departamento de finanzas, paga el arancel de una determinada carrera. Las dimensiones son estudiante, fecha, medio de pago y carrera. Otro ejemplo de modelo de ventas pero con cuatro dimensiones se muestra en la figura No. 4.

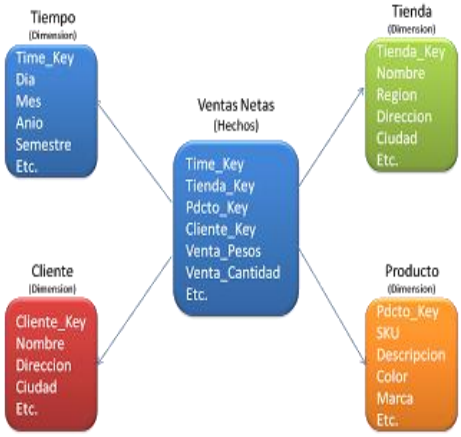

Figura No. 4 Tablas de Dimensiones del Hecho Ventas

Representación del Sistema Multidimensional.

Existen tres esquemas de representar que son: estrella, copo de nieve y constelación.

Métodos

Las metodologías que posibilitan un buen *datawarehouse* están las propuestas por Ralph Kimbal, Bill Inmon, Ricardo Bernabeu y otros.

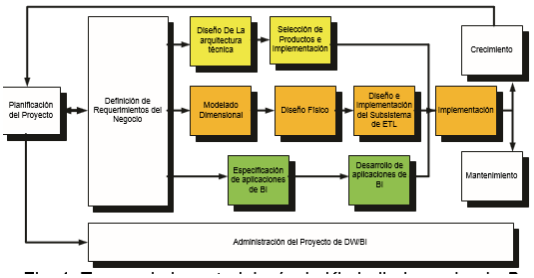

Figura No. 5 Metodología de Kimball
Fuente: Kimball et al *(1998).*

La figura No. 5 muestra las fases de la metodología de Kimball. Inicia con la definición de requerimientos, modelados dimensional, diseño físico, implementación del ETL y la implementación del *datawarehouse*.

Herramientas para creación de *datawarehouse*

- *Tableau:* software con licencia, se fundó sobre la idea de que el análisis de datos y los informes subsiguientes no deben ser actividades aisladas, sino que deben integrarse en un proceso único de análisis visual: uno que les permita a los usuarios ver rápidamente patrones en sus datos y cambiar las vistas al instante para seguir su línea de pensamiento
- *Pentaho,* software libre, consiste en una *Suite* Completa de Inteligencia de negocio permite implementar soluciones de *Business Intelligent* , tales como: Informes , Dashboards, Cubos OLAP, Procesos ETL , *Data integration*, Subscripciones, *Data Mining*, Alertas

Minería de datos

Es la extracción no trivial de información implícita, desconocida previamente, y potencialmente útil desde los datos".

- Técnicas de visualización: son buenas para crear patrones en un conjunto de datos que puede ser utilizado al comienzo de un proceso de *Data Mining.*
- Redes neuronales artificiales, se trata de modelos predecibles -no lineales que aprenden a través del entrenamiento y semejan la estructura de una red neuronal biológica.
- Árboles de decisión, son estructuras en forma de árbol que representan conjuntos de decisiones, las cuales generan reglas para la clasificación de un conjunto de datos.
- Reglas de asociación,establecen asociaciones en base a los perfiles de los usuarios sobre los cuales se está realizando el *Data Mining*.
- Algoritmos genéticos,se trata de técnicas de optimización que usan procesos tales como combinaciones genéticas, mutaciones y selección natural en un diseño basado en conceptos relacionados con la evolución.
- Método del vecino más cercano, es una técnica que clasifica cada registro en un conjunto de datos basado en una combinación de clases de los K registros más similares a él en un conjunto de datos históricos
- Regla de inducción, es una extracción de reglas Si-entonces de datos basados en significados estadísticos.
- Redes Bayesianas, buscan determinar relaciones causales que expliquen un fenómeno en base a los datos contenidos en una base de datos.
- OLAP (Procesamiento Analítico en Línea) es en realidad un conjunto de herramientas que gozan de un mayor poder para revisar, graficar y visualizar información multidimensional.

Así el *DataMining* surge como una tecnología que intenta ayudar a comprender el contenido de las bases de datos.

La minería de datos es parte de KDD(*Knowledge Discovery in Databases*) es el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos. KDD se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles".

Herramientas para minería de datos

Algunos de ellos son:*XLMiner, Matlab, IBM SPSSModeler, SAS ENTERPRISEMiner, SalfordSystems Data*

Mining, OracleDataMining, K-NimeOrange ,weka.

- *Weka* es acrónimo de *Waikato environment for Knowledge Analysis*, es un conjunto de librerías java para la extracción de conocimientos desde la base de datos. Está constituido por una serie de paquetes de código abierto con diferentes técnicas, de procesado, de clasificación, agrupamiento, asociación y visualización.

Minería de texto

Según Dan Sullivan es el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos

Herramienta para minería de texto

- *Ethnograph*: es un programa para el análisis descriptivo-interpretativo de textos.Puede analizar datos iconográficos, mapas, croquis, diagramas, fotografías, videos audio.Es una herramienta para la investigación cualitativa.

Resultados

- Arquitectura de *datawarehouse* para el Ministerio Publico: el modelo comprende las fuentes de datos son los datos obtenidos de los sistemas de la misma Fiscalía, como: de la Etapa Preparatoria, de las Medidas Cautelares, del mismo Proceso Penal Ordinario, del Sistema de Proceso Abreviado, del Sistema Procesos de Daños y Perjuicios, el IDIF (Instituto de Informaciones Forenses) con dictámenes de los distintos peritos de huellas, de médicos, etc. Del Consejo de la Magistratura,antecedentes judicialesy de la FELCCFuerza Especial de Lucha Contra el Crimeny FELCN Fuerza Especial de Lucha Contra el Narcotráfico así también se considera la fuente del sistema de los Procesos Civiles. (Este modelo es ampliable a otras fuentes necesarias para completar un análisis completo de la información).

- Las bases de datos, de los cuales se tiene el script de creación, de inserción de datos (el script permite crear en cualquier gestor sea comercial o libre estas bases de datos) y las consultas que permiten probar la consistencia de los resultados son:

- Base de datos de la etapa preparatoria del proceso penal**.**se registran: fiscales, abogados, peritos y policías; vehículo de policía; victimas; intervención Policial; testigos, objetos Secuestrados, Requerimiento de intervención; Declaraciones; Estudio Forense; Muestrario Fotográfico; Indicios materiales; Cadena de custodia
- Base de datos de procesos penales ordinarios. Se registran:Los datos de querellante(s), querellado(s), testigo(s), etc. Se recuperan unos datos de la base de datos de la etapa preparatoria. El fiscal, el juez, el o los abogados. Las audiencias del juicio oral público y contradictorio. Las pruebas.Las excepciones. Las resoluciones judiciales. Los recursos de apelación y casación. La ejecución de sentencia.
- Base de datos del sistema penitenciario.Se registran:Las características somáticas del detenido, la razón de su privación de libertad; fecha de ingreso, tiempo que debe

permanecer recluido. Seguimiento a su estadía en la penitenciaria. Su situación jurídica. Las partes del proceso.El estado en que se encuentra el proceso.Los abogados defensores.La conducta dentro de la penitenciaria. La sanción impuesta y fecha de libertad.Datos de sus familiares. Como se puede observar en la figura No. 6.

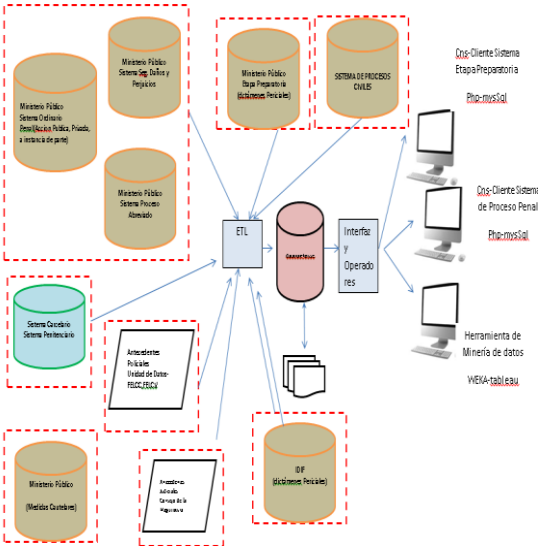

Figura No. 6 Arquitectura Datawarehouse para la Etapa Preparatoria en La imputación Formal del Ministerio Público.

- Base de datos de medidas cautelares. Se registran:Las audiencias, el tribunal, el abogado defensor, patrocinante, otros representante de protección al menor (si es el caso) en defensa de la víctima, las razones de las medidas a tomar, que articulados son subsumidos para determinar la detención del posible responsable
- Unificación y ETL en el Pentaho-*Creación de Datawarehouse***.** Se hicieron las conexiones a los diferentes gestores la etapa preparatoria en Oracle, la de procesal ordinario en *postgres*, la de Medidas Cautelares en *SqlServer*, el Sistema Penitenciario en *MySql* y los Antecedentes Judiciales y Policiales en Excel. Se realiza la extracción cargado y la transformación además de la unificación en el depósito-*datawarehouse* creado en el mismo servidor del Oracle. Como se aprecia en la figura No. 7.

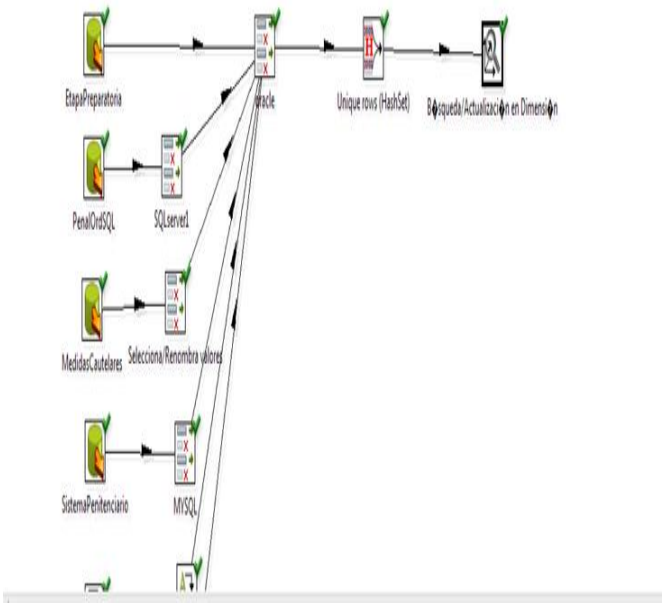

Figura No. 7 ETL y Unificación en Pentaho.

- Diseño Multidimensional: en lo siguiente se muestran el diseño de la tabla de hechos y dimensiones en estrella para las distintas fuentes de los datos.
- Diseño Físico: por ejemplo para el diseño dimensional nro. 1 (perfil del imputado) se crea la tabla de hechos y dimensiones a través del mapeo de cubos Las fuentes fueron: Sistema de la Etapa Preparatoria, Sistema del Proceso Penal Ordinario, Sistema de Medidas Cautelares, Sistema Penitenciario, Antecedentes Policiales y Judiciales en una sola tabla.
- Minería de Datos; se aplica a las tablas multidimensionales los algoritmos de clasificación para búsquedas de datos por delitos, por fiscales, por ocupación de imputados, de agrupación por

juzgados, y de asociación por los tipos de delitos, por los lugares de comisión del delito.

- Modelo de Minería de Texto.-Como se muestra en la figura No. 8. Las fuentes de datos para su procesamiento con el e*thnograph* de datos consistentes en documentos digitales necesarios para la calificación del hecho a cargo del Fiscal de Materia asignado al caso.

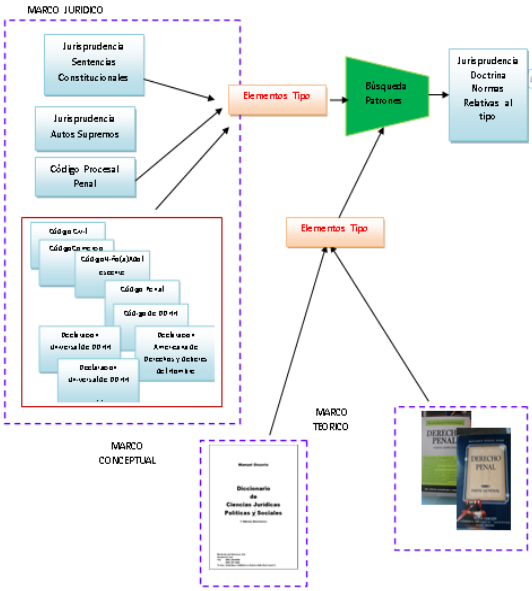

Figura No. 8 Modelo de Fuentes de Datos no Estructurados para la Minería de Texto

- Minería de Texto: al *etnograph* se alimenta con la normativa actual como código civil, penal, procesal penal, código de niño niña adolescente, normativa de la violencia contra la mujer y otros, también normativa internacional como derechos humanos, así también de la jurisprudencia de sentencias constitucionales, autos supremos, diccionarios jurídicos doctrina y demás textos digitales concernientes para la calificación del hecho. Esta herramienta en base a categorías definidas encuentra los fragmentos de texto de la ocurrencia del patrón.

Discusión

- Realizar el perfil del posible responsable de un delito con información proveniente de distintas instituciones permite tomar decisiones en el momento de la imputación formal con mayor certeza.
- Contar con información de normativa vigente incluyendo jurisprudencia logra una calificación subsumida no discutible.
- La experimentación sobre datos hipotéticos, y diseños de bases de datos no refleja la real dimensión del impacto del proyecto ni sus limitaciones.
- Con la presente investigación la minería de datos y texto tiene directrices para su sencillo y claro desarrollo, además de incluir tutoriales sobre el uso de las herramientas necesarias efectivizar el diseño en la implementación.

Conclusiones

- Se minimiza el tiempo de investigación de la Fiscalía, en uno de sus factores, respecto a la búsqueda de información respecto a información de la policía, de los procesos judiciales y del IDIF sobre el presunto responsable.
- Para el diseño de las distintas bases de datos en diferentes gestores se realizó el estudio del proceso penal y el análisis de datos que son requeridos de las instituciones públicas que emiten información de interés para procesos de acción penal para la fase preparatoria y en otros contextos de donde provengan los datos.
- Se desarrolló un diseño e implementación del *datamart* del etapa preparatoria y sistema penal ordinario, con posibilidad de expandir a un *Datawarehouse* que ayudara a la

Fiscalía y se aplica minería de datos y texto sobre este repositorio.
- Se creó la primera versión de la plataforma de investigación de fiscalía para la investigación preliminar.

Agradecimientos

A los Abogados José María Rivera Ibáñez e Iván Córdova por su valiosa enseñanzas que inspiraron la creación del proyecto.

Referencias

Coops S, (2012). *The judicial datawarehouse.*

Donsez, D. (2012). *Systémesd`informationdecisionnels*. Disponible en: http://fr.slideshare.net/demahomdidjoub/bd dwdm

Gil, G., (2009). *Datamining*, Editorial Megabyte.

Gil, Harjinder S., (1996).*Data Warehousing*, Prentice Hall,

Harb, Miguel, (1998). "*Derecho Penal*". Editorial Juventud.

Herrera, C. (2007). "*Todo lo que querias saber sobre DatawareHouse (IV)*". Disponible en: http://www.adictosaltrabajo.com/tutoriales/ datawarehouse-4/#2.10.DataMart|outline

Kimball et al.,(1998)*The DatawarehouseLifecicle Toolki*t. New York, Wiley.

Lobos R, (2010). *El Uso de Nuevas Tecnologías en el Sistema Judicial: Experiencias y precauciones*.

Moine, J.et al, (2011). *Estudio comparativo de metodologías para minería de datos*.

Molina J. M., & Garcia J. (2006).*Técnicas de Análisis de datos. Aplicaciones Prácticas utilizando Microsoft Excel y Weka*.

Perez, A *etal ,(*2010). *Minería de texto para la categorización automática de documentos*

Sepúlveda J. Urrutia, L. (2004). *Diseño e Implementación de un DataWarehouse para la gestión de ventas de la empresa vitivinícola*. Miguel Torres Chile.

Código Penal Boliviano
Código Procesal Boliviano
Constitución Política del Estado Plurinacional de Bolivia
Ley del Ministerio Público

Presentado: La Paz, 9 de octubre de 2015

Aceptado: La Paz, 27 de noviembre de 2015