Knowledge Discovery Process in the Open Government Colombian Model

Johnny Alexander Salazar Cardona, Carlos Hernan Gomez G, Marcelo López Trujillo
Departamento de Sistemas e Informática, Universidad de Caldas

Manizales, Colombia

alexander9052@gmail.com
ch@ieee.org
mlopez@ucaldas.edu.co

Abstract -- Currently Colombia is pushing its "gobierno en linea" project, which corresponds to a consolidation of government processes with different technological tools to have an o-Gov maturity level (open government), seeking liberation and publication of the information generated in the public sector, allowing citizens to have access to this based on the paradigm of Open Data. With this type of paradigm must implement knowledge discovery process (KDP), for the purpose of enhance the data have been released and to facilitate their understanding by stakeholders, but this is not being taken into consideration in the Colombian model. Therefore, this article aims to explain the different elements that involve the use of KDP in national governance model, and the approach that should be given to this when working on public sector data.

Keywords -- Data warehause, data mining, knowledge management, Open government, o-Gov, open data, knowledge discovery process, KDP

I. INTRODUCCIÓN

En este mundo altamente tecnológico donde la velocidad en la que se genera, transmite y se almacena informacion es cada dia mayor, los seres humanos aspiran a darle mejor utilidad a los datos generados, para administrar, controlar y optimizar el proceso de toma de decisiones en cualquier campo. El sector gubernamental global no ha sido indiferente a este fenómeno, hasta el punto que se convirtió en un paradigma llamado gobierno abierto (o-Gov) permitiendo una relación con los ciudadanos.

Este paradigma, consiste en que la informacion debe ser compartida al ciudadano, para tener una mayor cercanía y fortaleciendo la confianza y participación mutua [1].

Colombia no es ajeno a este paradigma gubernamental, debido a que en su plan de gobierno en línea ya está dando los primeros pasos para establecer un modelo de gobierno abierto, el cual está liberando algunos conjuntos de datos para el acceso público por parte de los ciudadanos y terceros interesados. Además Colombia ha establecido un modelo para la liberación de datos abiertos del sector público establecido por el MinTIC [2] . Cuando se habla de acceso público y liberación de informacion, se requiere un nuevo paradigma llamado Open Data entendido como los datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona o entidad, siempre y cuando se le de reconocimiento al autor que definió [3]. Este tipo de acciones cuando se aplican a un gobierno es que permite que sea llamado gobierno abierto (e-Gov + Open data), debido a que los datos del sector público son publicados permitiendo el acceso a todos los ciudadanos para su consulta, comprensión y posterior análisis.

Si se aborda el paradigma de Open data, se debe tratar la informacion que se libera por medio de KDP (procesos de descubrimiento de conocimiento), siendo esto facilitador de la comprensión de los datos

Johnny Alexander Salazar Cardona Universidad de Caldas alexander9052@gmail.com, Marcelo López Trujillo Universidad de Caldas/Universidad Nacional de Colombia mlopez@ucaldas.edu.co, Carlos Hernán Gómez Gómez Universidad de Caldas/Universidad Nacional de Colombia, ch@ieee.org

978-1-4799-6717-9/14/$31.00 ©2014 IEEE

y guía para la toma de decisiones. El punto de conexión entre gobierno abierto, Open Data y procesos de descubrimiento de conocimiento se da cuando los datos que se publican deben ser tratados para facilitar la comprensión por parte de los ciudadanos o terceros interesados quienes son realmente las personas que tendrán acceso a dicha informacion, y con esto brindar formas y mecanismos para aplicar procesos de mineria de datos para encontrar informacion oculta en los datos, con el fin de explotar todo el potencial de los estos para encontrar conocimiento que sea realmente util y que pueda ser utilizado para un bien común, pero Colombia en su modelo nacional no lo está teniendo en cuenta.

El presente artículo tiene como objetivo exponer los diferentes elementos que involucran el uso de KDP en el modelo nacional de gobierno abierto y el enfoque que se debe dar a este cuando se trabajan con datos del sector público. Inicialmente en el marco de este documento se dará una explicación de la liberación de datos públicos en el sector público colombiano, luego se contextualizaran los diferentes modelos existentes de procesos de descubrimiento de conocimiento y, de estos cual es el que mejor se adapta a el enfoque de un gobierno abierto, después se explicara la integración del modelo KDP elegido con el modelo Colombiano de gobierno abierto y finalmente se darán algunas conclusiones respecto a toda la temática tratada.

II. PUBLICACIÓN DE DATOS

Según el modelo nacional de apertura de datos del sector público colombiano, el proceso de publicación de datos depende del nivel de madurez en el que se encuentre el modelo de o-Gov. Según el nivel de este el formato de los archivos publicados varía desde archivos TXT, hojas de cálculo CSV, archivos estructurados XML e integración de este con web semántica, pero cualquier dataset debe estar acompañado de un conjunto de metadatos que describirán el dataset publicado para facilitar su comprensión. El modelo colombiano de gobierno abierto también define una serie de roles o equipo de trabajo que se encargan del proceso de publicación de datos, entre los que se encuentran un rol funcional, técnico, seguridad y jurídico [2]. Cada uno de estos desempeña una función específica en el proceso de apertura de datos del sector público pero ninguno está enfocado al proceso de descubrimiento de conocimiento sobre los datos publicados. De los roles nombrados anteriormente es relevante el rol funcional, debido a que se encarga de la selección y priorización de la informacion que va a ser publicada y por lo tanto, debe gestionar los datos que se están publicando en cuanto a confianza, seguridad e integridad, además de la dinámica del entorno que estos tienen. Adicionalmente en el marco de este artículo se definirá un nuevo rol el cual se llamara "analista de datos", que se encargara del proceso de descubrimiento de conocimiento sobre los datos.

III. PROCESOS DE DESCUBRIMIENTO DE CONOCIMIENTO (KDP)

Unos de las metas de un gobierno abierto, aparte de publicar y liberar la informacion a sus ciudadanos, es brindar mecanismos que agilicen la interpretación de los datos liberados, y proporcionar elementos que faciliten los procesos de descubrimiento de conocimiento sobre estos, para que así los usuarios puedan realizar este proceso de manera autónoma sin tener vastos conocimientos ni experiencia. Esto es logrado desde múltiples enfoques: Documentación de los datasets con base a los metadatos requeridos, conexión entre fuentes de datos con web semántica y la aplicación de procesos de descubrimiento de conocimiento publicando los resultados obtenidos por parte de la entidad dueña de un conjunto de datos, brindando el dataset tratado computacionalmente para que cualquier ciudadano que lo desee pueda aplicar su propio KDP que constaten los resultados publicados o para que pueda realizar sus propios descubrimientos de nuevo conocimiento que se haya pasado por alto, sin que tengan que manipular los datos, el cual es un proceso bastante arduo y que puede generar una barrera para el ciudadano sin experiencia.

Luego de que una entidad pública o privada con funciones públicas, libera o publica un determinado conjunto de datos o dataset en su página web o portal nacional designado para la publicación de datos abiertos, se puede proceder a aplicar procesos de descubrimiento de conocimiento o KDP (Knowledge discovey process) sobre estos, con el fin de encontrar informacion o conocimiento sobre los datos que aún se desconozcan o para tener una descripción detallada. Actualmente existen múltiples modelos y metodologias de KDP, como lo es el modelo KDD (descubrimiento de conocimiento en bases de datos), SEMMA, las metodologias CATALYST o P3QT, CRISP-DM

(Cross Industry Standard Process for Data Mining), entre otras. Inicialmente fue definido el modelo KDD en 1996 por Fayyad, et al. [4] como un modelo de descubrimiento de conocimiento para el sector académico. Con el transcurso del tiempo se han generado variaciones y propuestas alternas con otros enfoques, como lo son los otros modelos y metodologias mencionados anteriormente. Entre estas destaca la metodología CRISP-DM que a diferencia del modelo KDD tiene un enfoque netamente industrial y es mundialmente utilizado y reconocido. De todas estas metodologías y modelos disponibles se debe utilizar el modelo KDD para descubrimiento de conocimiento en el sector público, debido a su enfoque que facilita la implementación de variaciones con base al caso de estudio, ya que no especifica actividades puntuales en lo que se debe realizar en cada etapa, permitiendo ajustes con base a criterios establecidos por el equipo de trabajo [5].

El proceso de KDD fue definido como el proceso no trivial de descubrir conocimiento e informacion portencialmente util dentro de los datos contenidos en algun repositorio de informacion [6]. Este repositorio puede ser estructurado como una base de datos relacional, archivos XML - CSV o no estructurada como texto, documentos e imagines. Este no es un proceso automático, es iterativo (la salida de alguna de las fases puede retroceder a pasos anteriores y a menudo se necesitan varias iteraciones para extraer conocimiento de calidad) e interactivo (se necesita alguien que domine el entorno de los datos para apoyar el proceso) y exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones [5].

Se puede decir que KDD permite la comprensión de un dominio con base a un conjunto de datos, por lo que a partir de un dataset publicado de un determinado sector público se busca la comprensión de este, incluso prediciendo que está pasando en ese sector a partir de descubrimiento de conocimiento. Esta comprensión es lograda con base a la búsqueda de patrones comprensibles que pueden ser interpretados como conocimiento util e interesante, por ejemplo: el uso de redes neuronales son una poderosa herramienta de modelado pero son más difíciles de entender en comparación a un árbol de decisión [7], por lo que puede no ser una muy buena opción para un modelo de gobierno abierto, ya que los resultados deben ser muy claros para la ciudadanía en general.

IV. INTEGRACION O-GOV COLOMBIANO Y KDP

Hasta este momento se ha nombrado en múltiples ocasiones el proceso de mineria de datos, pero este no se debe confundir con el proceso de descubrimiento de conocimiento en bases de datos, debido a que KDD se refiere al proceso general de descubrimiento de conocimiento sobre los datos, y la mineria de datos se refiere a la aplicación de algoritmos para la extracción de patrones, es decir la mineria de datos es una etapa definida dentro de KDD. El proceso de descubrimiento de conocimiento está compuesta de nueve etapas que definen el "que" se debe realizar sobre los datos para poder descubrir conocimiento sobre esto. En el marco de este artículo estas se ajustaron al modelo de gobierno abierto para facilitar la aplicación de este en datos del sector público.

A. Comprensión del Dominio de los Datos

Cuando el analista de los datos va a realizar un proceso de descubrimiento de conocimiento sobre un conjunto de datos, debe comprender el entorno que los rodea, de que se tratan y que representan, con el fin de que cuando se obtengan los resultados durante este proceso, él pueda brindar una evaluación más profunda sobre resultados obtenidos, y así brindar una mejor interpretación y análisis de los resultados. Es difícil que el analista de datos comprenda el dominio de todas las áreas que involucren los diferentes dataset que deban ser tratados, para esto se recomienda trabajar en conjunto con el rol funcional el cual tendrá una compresión del dominio de los datos, y adicionalmente se puede apoyar en los metadatos publicados en el portal frente al dataset de estudio, ya que estos brindan una descripción detallada este.

La comprensión del dominio por parte del rol funcional permite definir el objetivo o la meta a alcanzar sobre la aplicación de procesos de mineria de datos en los datasets disponibles, para esto el analista y el rol funcional deben tener claro el proceso detrás de la generación de los datos, para formular un objetivo correcto, además podrán seleccionar las variables relevantes del conjunto de datos para el alcance del objetivo, y así el analista podrá interpretar de una manera clara los resultados y realizara una difusión clara de estos. Cuando se indica que se debe entender el dominio de los datos también se refiere a que se debe comprender los

diferentes campos del dataset, cuáles son sus posibles valores, los diferentes tipos de datos, etc. Adicionalmente el analista debe definir el objetivo de la aplicación del proceso de KDD, pero esto no llega a ser complejo, debido a que generalmente lo que se busca en el entorno de un gobierno abierto es describir los datos para que la ciudadanía en general pueda entender más claramente los datos es decir, dar una visión general del comportamiento de los datos que han sido publicados a la ciudadanía y que ellos los puedan entender con facilidad

B. Creación de la Base de Datos de Trabajo o Selección de Datos

Luego de tener comprendido el dominio de los datos por parte del analista de datos, ayudado por el rol funcional con los metadatos disponibles se debe pasar a la recolección de los datos. Puede suceder que las fuentes de los datos sean diferentes si se está utilizando web semántica si el nivel de o-Gov es muy alto, o que los datos se encuentren separados en archivos por fechas o se encuentren versionados. Cuando esto sucede el analista de datos debe buscar la centralización de los estos en un solo punto, ya sea en un único archivo manipulable o en la creación de un data warehause con el fin de acceder a los datos para su tratamiento de una manera más sencilla y eficaz, estructurando y unificando los datos en un formato especializado y tratable por la herramienta de mineria de datos que se vaya a utilizar. La unificación o centralización de los datasets en un único archivo manipulable, o la centralización de las fuentes en un data warehause son igualmente validas, ninguna es obligatoria o tiene mayor peso una sobre otra, solo se necesita un formato valido para el tratamiento idóneo para una herramienta de mineria de datos [8].

Cuando la informacion que se desea tratar se encuentra descentralizada, es decir tiene múltiples fuentes de datos, el analista debe determinar cuáles son los atributos compartidos entre estos para realizar una integración de los datos de una manera adecuada. Además sin importar que los datos estén descentralizados o no, el analista debe definir cuáles son los atributos relevantes para cumplir el objetivo de la aplicación de KDD, y cuáles serán las instancias de muestra de los cuales se extraerá conocimiento, con el fin de eliminar atributos redundantes e inconsistencias en los datos. Con esto se trata todo como una sola fuente de datos consistente y lista para ser procesada.

C. Limpieza y Pre-Procesamiento de los Datos

Cuando el analista tiene los datos centralizados en un formato tratable por una herramienta de mineria de datos se deben limpiar y pre-procesar los datos. Ningún conjunto de datos del sector público está libre de que tenga valores de atributos faltantes o vacíos por instancia, datos que se encuentren fuera de rango con valores inapropiados para el atributo que representa como por ejemplo: un atributo que represente la edad de una persona tenga un valor de “998”, lo cual claramente es un error o que los valores de los atributos no se encuentren en las mismas unidades de medida o formatos. Adicionalmente según el objetivo del proceso el analista puede aplicar discretizacion o binning, que consiste en transformar las variables numéricas de tipo real o entero sean a categorías o a variables de tipo nominal, como por ejemplo: las edades que normalmente son representadas con valores enteros pueden representarse con rangos, como una persona que tenga 18 años puede estar en el rango de personas de 18 a 25 años. Este proceso de discretizacion se recomienda cuando se desea aplicar asociaciones o procesos de clasificación entre los datos. Igualmente existe el proceso inverso llamado numerizacion, el cual consiste en convertir las variables nominales a numéricas pero este proceso es menos común. Para los datos o atributos faltantes se deben generar estrategias para su tratamiento [9].

Con la estructura de los datos que se tengan en este punto se debe generar los metadatos que documentaran todo la estructura del dataset, y que se publicara junto a los resultados de este proceso para que cualquier ciudadano o tercero que desee realizar un KDP sobre los datos, tenga una documentación clara de los datos que está analizando. Esta estructura de los metadatos debe contener el tipo de datos para el proceso de descubrimiento de conocimiento, posibles valores por atributo y total de instancia entre otros, con el fin de facilitar el proceso de tratamiento de los datos y futuros accesos a este por parte de ciudadanos o terceros.

D. Reducción de Datos y Proyección o Transformación de los Datos

Esta etapa se enfatiza en el tratamiento adicional se los datos por parte del analista de los datos, buscando disminuir la dimensionalidad a nivel de columnas con base a la generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada, se consolidan los datos con operaciones de agregación y normalización buscando tener unos resultados más precisos. De ser necesario el conjunto de datos de estudio puede ser divido en un conjunto de entrenamiento, uno de prueba y otro de evaluación para disminuir la dimensionalidad a nivel de filas. En algunos casos cuando el dataset se encuentra en bases de datos relacionales se debe realizar agrupación, cuando se tienen relaciones 1 – N entre tablas, agregando un campo que contabilice la cantidad de registros por relación. Los cambios que se hagan en esta etapa sobre los datos deben ser actualizados en los metadatos generados en la etapa anterior, con el fin de tener los datos lo más procesadamente posible y así facilitar el proceso a terceros que deseen corroborar o descubrir su propio conocimiento sobre los datos, incluyendo los datos derivados, reformateados y creados.

E. Elegir la Función o el Método de Algoritmo y Mineria de Datos

Los objetivos planteados por el analista de datos en el proceso de KDD pueden estar enfocadas a 2 grandes paradigmas: La verificación y el descubrimiento. La verificación se centra en el hecho de comprobar una hipótesis existente por parte del analista de datos, y con el descubrimiento se busca de manera autónoma patrones desconocidos que es lo que realmente importa en el proceso que se aplica a los datos en un gobierno abierto, debido a que se busca una descripción que no se tiene sobre los datos para tener un mejor entendimiento sobre estos. El descubrimiento a su vez se divide en dos métodos según el objetivo que se haya establecido en etapas anteriores: Métodos predictivos o supervisados y no supervisados o descriptivos [4, 10].

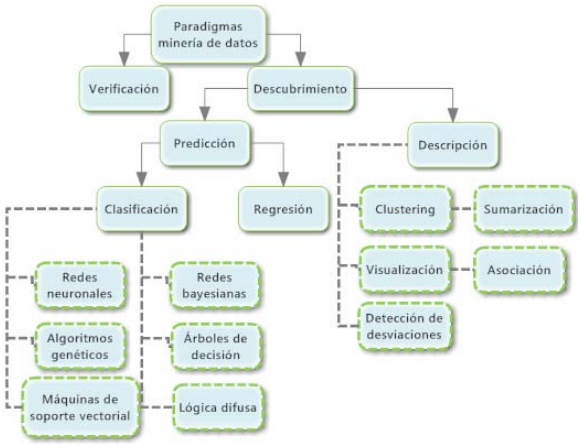

Fig. 1 Paradigmas de la mineria de datos

F. Elegir el Algoritmo de Mineria de Datos

Luego de que el analista seleccione el método a utilizar (supervisado o no supervisado) con base a los resultados que se quieren alcanzar (predicción o descripción) se debe seleccionar la tarea o tareas a utilizar que se encuentran disponibles en el método elegido. Con esto el analista de datos refina el alcance de la tarea anterior utilizando el algoritmo más adecuado que ayude a alcanzar el objetivo propuesto [11]. Además el analista debe definir los parámetros que pueden ser apropiados para el conjunto de datos analizados para optimizar los resultados obtenidos por la tarea aplicada.

Aunque el alcance entre la predicción y la descripción no es clara, algunos modelos predictivos pueden ser descriptivos y entendibles fácilmente por un ser humano y algunos modelos descriptivos pueden dar índices predictivos [7]. La existencia de múltiples algoritmos para resolver algunas tareas de mineria de datos dentro de un KDP no se debe al hecho de que una técnica sea mejor que otra: Algunas tareas son numéricas, otras expresan reglas, unas son rápidas y precisas, otros son lentas y aún más precisas [12].

G. Mineria de Datos

Luego de que el analista de datos seleccione el método o métodos y la tarea o tareas a utilizar de mineria de datos, solo falta aplicarlos con una herramienta que facilite el proceso, teniendo en cuenta el proceso de parametrizacion de los algoritmos utilizados para obtener el mejor resultado posible; esto se convierte simplemente en una tarea de optimización con base a los parámetros elegidos. Se debe tener en cuenta que los datos de aprendizaje

no deben ser utilizados para probar el modelo, por lo tanto existen varias técnicas para hacer esto de la mejor manera: validación cruzada, bootstrapping, entre otros. Ya es tarea del analista de datos determinar cuál es la herramienta que puede ofrecer mejores resultados, y que estos sean comprensibles, para que facilite el proceso de interpretación por parte de la ciudadanía en general.

H. Interpretación y Evaluación

El analista de los datos debe entender los resultados del análisis y sus implicaciones, que son dados por el software utilizado, identificando los patrones obtenidos, evaluándolos y documentándolos, utilizando técnicas de visualización en los patrones y modelos extraídos que pueden ser útiles para facilitar el entendimiento de los resultados. Cuando los resultados dados por el software no son óptimos, el analista debe repetir el proceso de KDD, buscando la mejora en los procesos y así repetir nuevamente el análisis en la herramienta elegida para comparar resultados hasta obtener unos que sean aceptables. Los resultados obtenidos deben ser comprensibles, validos (información real), útiles y novedosos (información desconocida) [13].

I. Uso del Conocimiento Obtenido

Como el objetivo del proceso de KDD en el entorno actual es poder ofrecer informacion adicional a la ciudadanía para explicar el comportamiento del conjunto de datos que se tengan publicados, se deben compartir los resultados de estos procesos de una manera sencilla de entender (ya sean simbólicos o visuales) y transparentes. Por eso se recomienda apoyar la interpretación de los resultados con tecnicas de visualización, para así brindar medios de interpretación de los resultados, para que los ciudadanía en general pueda interpretar con facilitad los datos y los resultados obtenidos por el proceso de KDD. Estos resultados deben ser publicados junto con los metadatos generados en el proceso en el portal o sitio destinado para la publicación de datos.

V. CONCLUSIONES

El impacto de la tecnología en el sector gubernamental está permitiendo que se tenga una mayor cercanía con el ciudadano, debido a que se comparten libremente la documentación generada en el sector público. Es necesario que la informacion sea procesada y aprovechada, para que la ciudadanía en general realmente pueda entender esa informacion a la que tiene acceso. Por eso se de implementar procesos de descubrimiento de conocimiento con un enfoque descriptivo y no predictivo, debido a que a los ciudadanos les interesa realmente es qué muestran los datos que el sector público que están accediendo, mas no en base a ellos predecir comportamientos futuros. En el modelo Colombiano definido por el MinTIC se debe definir un rol o encargado adicional que trabaje en conjunto con el rol funcional, para aplicar procesos de KDD. Adicional al proceso, se deben generar futuras investigaciones para integrar procesos de visualización de resultados como los ofrecidos por el campo de BI (Inteligencia de negocios) para fortalecer el proceso de interpretación de resultados y así ofrecer resultados a la ciudadanía de mayor calidad.

REFERENCIAS

[1] A. Sourouni, G. Kourlimpinis, S. Mouzakitis, and D. Askounis, "Towards the government transformation: An ontology-based government knowledge repository," *Computer Standards & Interfaces,* vol. 32, No 2., p. 44, ene 2010.

[2] MinTIC, "Lineamientos para la implementación de datos abiertos en Colombia," vol. 1-13, Ministerio de tecnologias de la informacion y las comunicaciones, Ed., 1 ed. Bogota (Colombia), 2011, pp. 1-13.

[3] Open Knowledge Foundation, *Open Data Handbook Documentation: Release 1.0.0 [online]* vol. 1. Cambridge (UK), 2012.

[4] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, "The KDD process for extracting useful knoledge from volumes of data," *Communications of the ACM,* vol. 39, pp. 27-34, 1996.

[5] J. M. Moine, "Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo " Magister en Ingenieria de Software, Facultad de informática, Universidad Nacional de la Plata, Argentina, 2013.

[6] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, *Advances in knowledge discovery and data mining*. Masachussetts: MIT Press, 1996.

[7] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *IA Magazine,* vol. 17, No 3, pp. 37-54, 1996.

[8] J. Carreño, "Descubrimiento de conocimiento en los negocios," *Panorama,* vol. 4, 2008.

[9] C. L. Hernandez and J. E. Rodrigez, "Preprocesamiento de datos estructurados," *Revista Vinculos,* vol. 8, pp. 27-48, 2008.

[10] H. Howard. (2012). *Knowledge Discovery in Databases*. Available: http://www2.cs.uregina.ca/~dbd/cs831/index.html

[11] J. Reyes and R. García, "El proceso de descubrimiento de conocimiento en bases de datos," *Ingenierias,* vol. 8, 2005.

[12] J. Hernandez Orallo, *Encycpledia of database technologies and applications* vol. 54. Valencia, España: Technical University of Valencia, Spain, 2005.

[13] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*, 2 ed. vol. 1, 2010.