Optimización de data warehouse mediante desarrollos colaborativos de modelos de cubos semánticos

Vargas Navarro Rocío [#1]

[#] *Facultad de Tecnología Informática, Universidad Abierta Interamericana*
Chacabuco 90, 1°piso, CABA, Buenos Aires, Argentina

[1] rocio.vargasn@gmail.com

***Abstract*— Most data warehouses have limitations in terms of efficiency, flexibility and scalability. One difficulty is the maintenance and the analysis due to data growth and overload. This paper presents the methodology for collaborative working between users through a tool which allows the design of cube structures based on the Object-Oriented Data Model or Ontological maps. The users will provide their knowledge and experience for the development of these structures, which have the aim of improving the integrity of information queries, reduce duplicate data and prevent the loss of semantic data.**

***Keywords:* semantic cube, Ontology, natural language processing, linguistic, collaborative knowledge**

***Resumen*— La mayoría de los data warehouses poseen limitaciones en cuanto a eficiencia, flexibilidad y escalabilidad. Una dificultad es el mantenimiento y el análisis dado por el crecimiento y sobrecarga de datos. Este artículo presenta la metodología de trabajo colaborativo entre usuarios a través de una herramienta que permite diseñar estructuras de cubos basándose en el modelado Orientado Objetos o en mapas Ontológicos. Los usuarios aportarán su conocimiento y experiencia para el desarrollo de dichas estructuras, las cuales tienen como objetivo mejorar la integridad en las consultas de la información, reducir datos duplicados y prevenir la perdida de datos semánticos.**

***Palabras claves*: cubo semántico, Ontología, procesamiento del lenguaje natural, lingüística, conocimiento colaborativo.**

I. INTRODUCCIÓN

Múltiples recursos son consolidados en el diseño de un data warehouse, los desarrolladores deben analizar la estructura y el contenido de los recursos y luego definir las reglas para combinarlos.

Los servicios On-Line Analytical Processing ofrecen herramientas de Bussines Intelligent que proveen mejoras a la arquitectura de un data warehouse. Los cubos OLAP de los cuales se emplean estrategias de uso multidimensional permiten realizar diversas combinaciones entre sus elementos para poder visualizar desde distintas perspectivas lo resultados variando su nivel de detalle.

La relación semántica definida en los cubos es menor entre cada cubo. Los usuarios aportan el conocimiento desde su perspectiva de trabajo y no desde una mirada global incluyendo la relación con otras áreas de negocio, a raíz de esto los problemas que se desencadenan son los de duplicación de datos, inconsistencias, incremento de datos independientes y problemas de integridad en las consultas. La integración de cubos semánticos dentro de un data warehouse podría resolver dichos problemas.

Actualmente hay muchas investigaciones que discuten el mantenimiento de los cubos de datos en un data warehouse (e.g. [1], [2]). El objetivo de este artículo es dar pautas iniciales para establecer relaciones semánticas entre los cubos integrando la metodología Orientada a Objetos y/o un modelo Ontológico integrando una herramienta colaborativa de la cual la interacción de los usuarios es a través de un aplicativo que permite compartir conocimientos basados de experiencias. Debido a que la mayor parte de las base de datos instaladas comerciales son relacionales, este trabajo se enfoca sólo en este tipo. A su vez el análisis que se realiza contempla métricas básicas ya que lo que se pretende mostrar es el modelo y sus alternativas.

En un modelo de cubos semánticos que utiliza tecnología orientada a objetos para data warehouse, proporciona a los usuarios diseñar mediante la generalización de relaciones entre diferentes cubos donde sus objetivos son mejorar la performance en la integridad de las consultas y reducir la duplicación de datos [3]. En un diseño Ontológico estas relaciones se basan en interpretar esta generalización en forma de jerarquías relacionando conceptos conocidos.

Hay estudios relacionados con la mejora de la eficiencia en la estructura de los data warehouses a través de la identificación de las relaciones entre cubos.

La gran cantidad de investigaciones que discuten en la implementación de los cubos en un data warehouse (e.g. [4], [5], [6]) tienen en común que se definen sobre una base de gran cantidad de atributos y tablas de hechos y en una pequeña cantidad de dimensiones con respecto a los datos analizados [4].

En investigaciones (e.g. [4], [7], [8]) sugieren que el modelo orientado a objetos es apropiado para el proporcionar grandes beneficios en el diseño de data warehouses facilitando mejoras sustanciales de flexibilidad y superioridad en lo que respecta a las búsquedas y a la organización de bases de datos como a las consultas que provienen de la búsqueda relacional de datos.

Cuando se describe a los cubos semánticos se los asocia con la acción de generar estructuras con dimensiones que mantienen una cierta semántica. Las herramientas OLAP poseen la capacidad de organizar datos semánticos dentro de estadísticas o de información concisa que incrementa la eficiencia en el análisis y en la visualización, tal como menciona [9]. A su vez en [9] describe el caso en el que una especificación semántica puede aportar performance en las operaciones de las herramientas OLAP sobre los datos, construir formatos comunes y hasta combinar cierta

información para construir respuestas o para materializarlas en nuevas bases de datos.

Las tecnologías semánticas en el ámbito de Business Intelligent (BI) proporcionan acotar la brecha entre el desarrollo de sistemas y el analista de negocios (e.g. [10], [11]). El proyecto Semantic Cockpit [10] describe la semántica de las dimensiones y métricas dentro de una estructura de ontológica multimedial con el objetivo de facilitar la formulación de consultas OLAP y la interpretación de resultados asistiendo y guiando al analista de negocios a la tarea de razonar sobre varios tipos de conocimientos, capaces de entender el dominio interno y externo de la organización, la semántica de medidas y alcances, el conocimiento sobre análisis previamente obtenidos y cómo actuar ante inusuales comparaciones de alcance. Existe un Modelo de Inteligencia de Negocios (BIM) (e.g. [11], [12]) que tiene como fin presentar los datos en términos amigables y familiares para el analista de negocios. Nebot et al. [13] describe en su investigación la organización multidimensional de ontologías y la extracción de cubos OLAP semi-automáticos. En [14] describe la utilización de ontologías de dominio como semántica para generar las dimensiones en cubos OLAP. Estos enfoques son complementarios para el uso o el modelado de métricas con valores ontológicos.

Es probable que se presenten dificultades a los usuarios en cuanto a la tarea de descubrir las relaciones semánticas entre cubos, pero la colaboración entre ellos es el aporte que proporcionan para el análisis de la integración de estas estructuras.

En el desarrollo de las siguientes secciones se conforman de la siguiente manera. En la sección 2 se describen diversas investigaciones vinculadas con el diseño de data warehouse, donde enfocará atención en el diseño de cubos Orientados a Objetos y en la semántica entre ellos.

Para la sección 3 se explicará con ejemplos el modelo de cubo semántico basado en mapas ontológicos y las conceptualizaciones de las relaciones abstractas.

La sección 4 describe el rol de los usuarios para interpretar las relaciones entre cubos mediante la interacción entre ellos con una herramienta colaborativa sin determinación de roles y basándose en su conocimiento. En la sección 5 se mencionan las conclusiones y trabajos futuros en vistas de adicionar ciertos prototipos que concluyan en pruebas para ampliar la investigación y de mejorar la performance de la propuesta.

II. MODELO DE CUBO SEMÁNTICO ORIENTADO A OBJETOS

Un objeto dentro de él posee ciertas características de valores y propiedades que podrían modificarse con determinadas acciones. La estructura que posee el modelo OO agrupa a los objetos con las mismas características en un contenedor denominado clase. A su vez la clase se la puede caracterizar como la estructura que define el tipo de objeto.

Las investigaciones de modelos Orientados a Objetos permiten definir una arquitectura semántica de cubos con la interacción de los usuarios [3] otorgando la posibilidad de proveer ciertas especificaciones sobre la base de la estructura del cubo raíz.

En [3] presenta la técnica de agregación como un cubo combinado con otros cubos. Esta combinación podría utilizar esos cubos de datos para hacer otros cubos que representen el objeto o el sujeto en un data warehouse.

El método de categorización permite al usuario identificar nuevos cubos de datos, esto en parte resuelve un problema pero surgen nuevos inconvenientes devenidos por la multi-herencia dado que a través de la definición de atributos podrían existir atributos con el mismo nombre, por este motivo los usuarios necesitan para identificar las reglas de herencia que la redundancia de eventos sean presentadas.

Existen artículos en los cuales se basan en el modelo Orientado a Objetos, dando la posibilidad al usuario de crear entre los diversos cubos, un número de relaciones abstractas tales como la generalización, agregación y categorización con el objetivo de definir varios significados atribuibles a dichos cubos [3].

El modelado Orientado a Objetos se vincula directamente al desarrollo de sistemas, en el cual su concepto primario de representación es mediante una colección de objetos.

A. Semántica de generalización en un modelo de cubo semántico

La generalización es la relación entre clases en las cuales una subclase hereda de otras clases llamadas superclases, sus propiedades. Si los cubos se interpretarían como clases determinaríamos que un cubo podría heredar ciertas características de otro, para demostrarlo se debe crear un cubo (subcubo) que herede las características de otro cubo (supercubo). La Fig. 1 muestra la relación de herencia.

La generalización es posible que sea una de las características más importante para evitar la duplicación de datos y ayude a la reutilización de objetos.

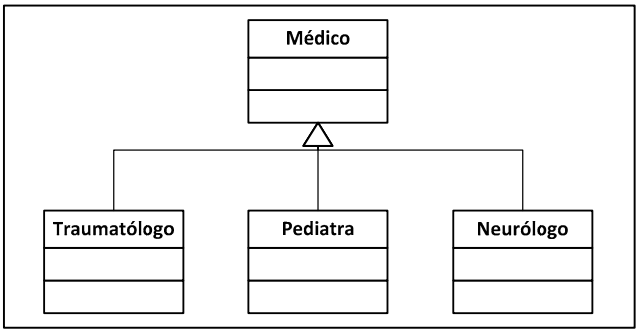

Figura 1. Relación de generalización

B. Semántica de agregación en un modelo de cubo semántico

La agregación es una forma de vincular relaciones entre objetos. La generalización representaría la relación vertical de los cubos y la agregación tendría lugar a la representación de una relación en forma horizontal. En la Fig. 2 se muestran los dos tipos de asociaciones existentes, con y sin dependencia.

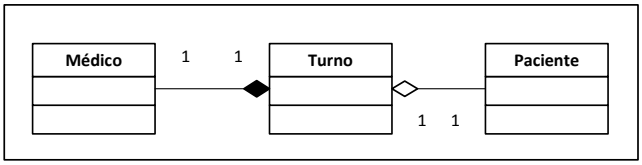

Figura 2. Relación de asociación con y sin dependencia.

C. Semántica de categorización en un modelo de cubo semántico

La categorización es otra de las características de modelo Orientado a Objetos. Esta especificación se da mediante la herencia múltiple como se representa en la Fig. 3.

En este modelo se permiten las múltiples herencias y permite a los objetos tener multi-referencias para construir relaciones con semántica.

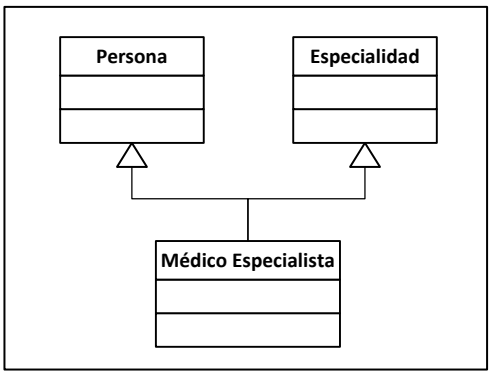

Figura 3. Relación de categorización

III. MODELO DE CUBO SEMÁNTICO BASADO EN MAPAS ONTOLÓGICOS

Como define [14], el dominio ontológico consiste de conceptos y roles, junto con axiomas que definen su significado. Cada concepto representa un conjunto de objetos, estos son miembros o instancias. La definición de un concepto viene dada mediante criterios de objeto de los cuales deben ser considerados miembros del concepto, estos son organizados asumiendo jerarquías de las cuales pueden ser acertadas para el diseño de la ontología o ser derivadas por un razonamiento automático basado en definiciones de concepto.

Un concepto puede ser un subconcepto de otro o de múltiples otros conceptos, la relación comparativa con el modelo Orientado a Objetos estaría reflejándose en la relación de asociación (es-un).

El dominio de ontologías definido como dimensiones semánticas en un data warehouse según [14] se pueden definir mediante la identificación de partes relevantes incluidas en un dominio ontológico a través de la recolección de conceptos (i.e., el dominio del atributo semántico), generar sub-secciones jerárquicas unidas mediante conceptos iniciales (roll-up) y por último por medio de la agregación de cada concepto de un nodo este asociado a cada cubo base.

Las dimensiones semánticas difieren de las dimensiones convencionales dado que no tienen niveles predefinidos ni descripciones en sus nodos en forma de definición de conceptos.

En [10] propone el acceso a recursos de conocimiento antes conocido por el analista de negocios el cual tendrá asistencia interactuando con la herramienta propuesta de Cockpit. Un dominio ontológico comprende entidades individuales de un data warehouse y enlaza estos conceptos de ontologías, métricas multidimensionales y los alcances ontológicos describen la semántica de estas las cuales están basadas en conceptos definidos previos a la descripción del dominio ontológico, de un juicio de reglas ontológicas que capturan conocimiento sobre antecedentes de análisis y relacionada a la historia de acciones.

En el dominio ontológico que describe la fig. 4 abstrae desde datos específicos de un data warehous, las instancias de dimensión (entidades) y las jerarquías de estas (roll-up) son incluidas en la ontología, las tablas de hecho solo se incorporan en el data warehouse. Inicialmente el dominio ontológico se enriquece de conceptos adicionales que constituyen conjuntos de entidades relevantes para el negocio con un nivel de dimensión y por enlaces a ontologías externas de las cuales el nivel de detalle incluye al anterior. Las medidas ontológicas muestran la representación de ontologías multidimensionales con medidas de esquemas definidas por el analista. Una medida de esquema consiste en el alcance que tiene un punto de calificación multidimensional, opcionalmente con una o más calificaciones que son conceptos o expresiones ontológicas sobre otros conceptos del mismo nivel de dimensión debajo del punto multidimensional y la instrucción de medida. La medida es definida por cada punto multidimensional incluido por el alcance del dominio.

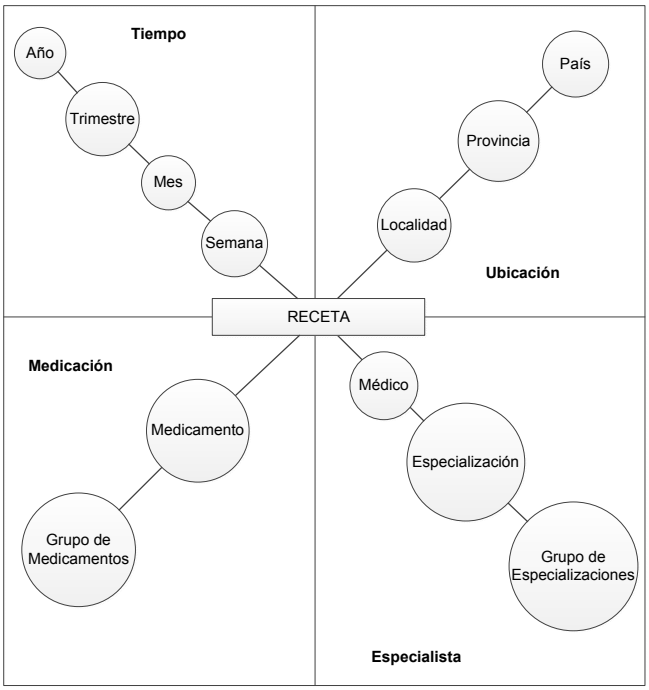

Figura 4. Dominio ontológico.

IV. HERRAMIENTA PARA EL MODELO PROPUESTO.

Para la definición de la propuesta se pretende utilizar un software diseñado e implementado en previas investigaciones de autoría [15] que ha sido probado en ámbitos educativos para generar mapas conceptuales colaborativos con el objetivo de generar conexiones de

saberes provenientes por usuarios que poseen conocimientos aislados pero conceptualmente relacionados con los de otros colaboradores.

El objetivo de utilizar este software proporcionará una interfaz amigable para el modelo de diseño de cubos semánticos en ambas alternativas de diseño, el modelado de cubos Orientados a Objetos o el modelado de mapas Ontológicos, en el cual diversos usuarios que interactúan en el contexto de negocios va a proporcionar el conocimiento de su área o especialidad compartiendo la estructura hacia el resto de los participantes con el objetivo de facilitar la relación de los objetos o nodos (que componen las dimensiones de un cubo) que agreguen al mapa o diagrama.

La herramienta se basa en una plataforma con tecnología Visual Basic .net de Microsoft desarrollada con mediante módulos integrados que permiten la escalabilidad y mantenibilidad futura y una funcionalidad reutilizable y escalable por su estructura en capas. Los módulos correspondientes son tres y cada uno es responsable de una tarea específica.

El primer módulo es el de Comunicación que va a ser el responsable de permitir la sincronización y comunicación entre las múltiples maquinas. Esto se lleva a cabo mediante un socket que permite intercambiar flujos de datos manejando un modelo de documento con formato XML para dicho pasaje.

En el caso de los Clientes, el socket de conexión provee el Puerto y la IP del Host por el cual va a realizar la conexión. En el Servidor se define el Puerto de Escucha esperando las solicitudes de conexión que se generen de los Clientes. El puerto definido es el 6000/TCP. Para lograr dicho objetivo se utiliza el espacio de nombres System.Net.Sockets que ofrece la clase TcpListener que se encarga de proporcionar las propiedades y métodos para escuchar y aceptar solicitudes de conexión. Por otro lado es el cliente quien utiliza TcpClient para conectarse y así poder comenzar con el envío y recepción de datos.

Al contar con un Servidor Sincronizador y un Cliente se obtiene la posibilidad de dividir en grupos de trabajo, en el cual siempre se dispondrá de un Sincronizador y varios Clientes.

El XML definido para la comunicación realiza el pasaje del conjunto de datos que se van a transmitir en entre el Cliente y el Servidor. Cuando se crea un elemento, se mueve o edita, se guarda en formato XML las coordenadas, color y texto (si es que posee), para luego enviarlas a los demás.

El segundo módulo integrado es el de Componentes Gráficos el cual va a proporcionar los elementos utilizados para el dibujo del mapa. Tiene como característica principal generar los nodos y conectores del esquema mediante código en forma dinámica. A su vez realiza la clonación del componente gráfico, mediante un método Clone de la interface ICloneable, provocando la apariencia de visualizar al control moviéndose en el área de trabajo con el objetivo de ser colocado en el lugar correspondiente dentro del mapa conceptual.

El tercer módulo es el de Interfaz de Usuario (UI) el cual incluye todas las interfaces que se requieren para la ejecución de la herramienta. Como interfaz principal contamos con el formulario MDI en el cual se despliega el escenario para la interacción de los elementos del menú y la interconexión de las computadoras. Otro de los formularios que se despliega en la interacción del uso de la herramienta es el que provee el área de trabajo y es allí donde se concretará el desarrollo de dibujo del mapa. El formulario Toolbox se expone las herramientas necesarias para ser incluidas en la elaboración de nodos y conectores que conforman el diseño del mapa. Por último, el formulario que contiene las propiedades del diseño el cual permite al usuario modificar el formato a los nodos con la posibilidad de cambiar el alto, ancho, color de relleno y hasta adicionarle un texto con la opción de variar el tamaño del mismo.

A continuación en la fig. 5 se despliega la imagen del escenario de trabajo de la herramienta colaborativa.

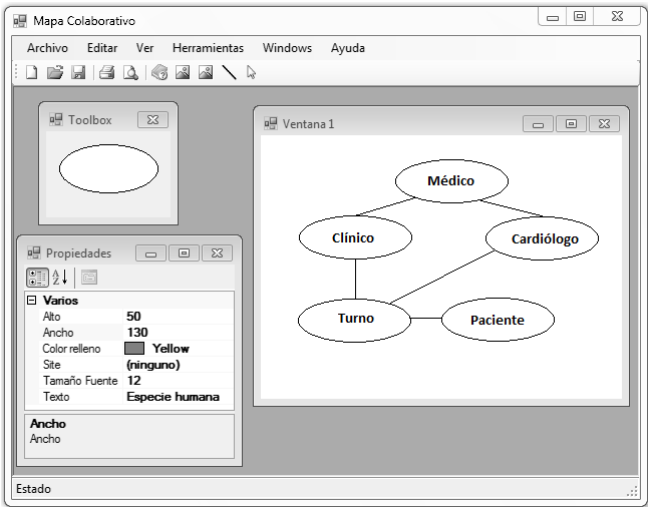

Figura 5. Entorno de desarrollo de la herramienta colaborativa para el diseño de mapas o estructuras conceptuales.

En cuanto al desempeño global que se obtienen de los resultados de la herramienta, el objetivo final es obtener estas estructuras de pensamiento de parte de los usuarios con la finalidad de ser integrados mediante la arquitectura física descripta en la Fig. 6.

No solo podremos interpretar la información desde diversas perspectivas del negocio sino que también nos ayudará a disminuir la cantidad de información duplicada dentro de un data warehouse y comprobar performance en la integridad de las consultas.

Veamos el siguiente ejemplo, un cliente A requiere de cierta información de un producto en un tiempo determinado pero el cliente B necesita esta información pero a un nivel aún más detallado. En el caso de B se requerirá duplicar la información implementando otro cubo de datos diferente al que consulta el cliente A dado a que se requiere un nivel de detalle más dentro de la estructura del cubo. Debido a esto lo que se quiere demostrar en esta investigación es que el modelo de cubo semántico creado a través del conocimiento de los usuarios en forma colaborativa resolvería esta problemática mejorando la performance y reduciendo la cantidad de datos almacenados.

El usuario es el que selecciona el sujeto de análisis y el tipo de semántica para las dimensiones y medidas.

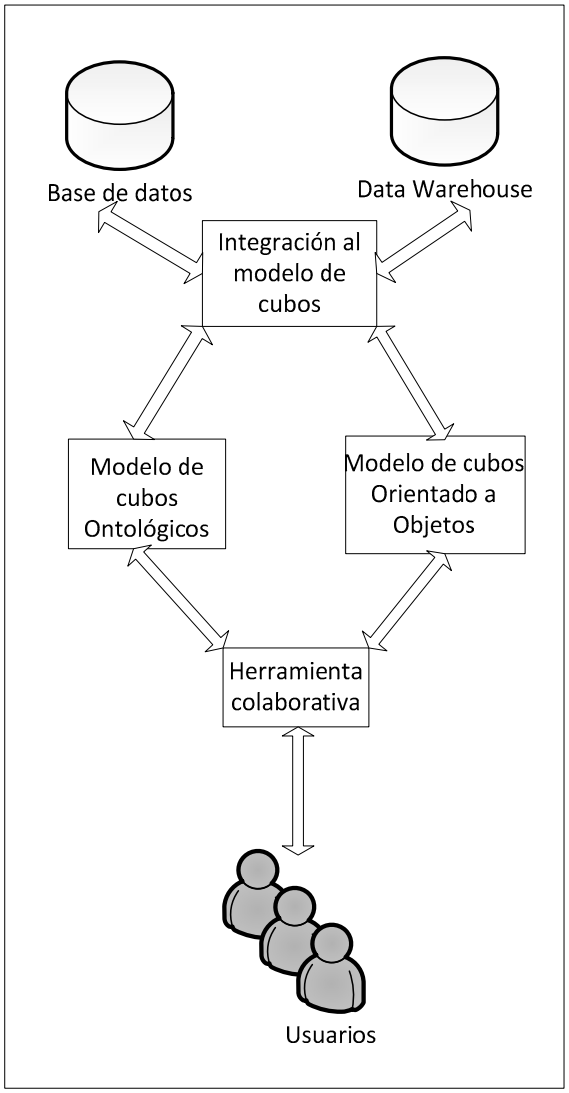

Figura 6. Modelo de arquitectura de data warehouse para la generación de cubos semánticos.

V. CONCLUSIÓN

En esta investigación define teóricamente como diseñar colaborativamente una estructura semántica de cubos OLAP Orientados a Objetos y mediante un modelo Ontológico los cuales proporcionan beneficios para el diseño de data warehouse.

El objetivo de los cubos semánticos es el de generar un flexible y ágil sistema donde la interacción de los usuarios cumplan un papel importante para la creación de los mismos. El modelo de agregación, generalización y categorización vinculado a la estructura Orientada a Objetos adiciona una metodología de semántica entre los cubos que proporciona un aval factible y válido propio de la arquitectura. Los modelos Ontológicos para la generación del diseño en una estructura colaborativa proporciona la posibilidad de que diversos usuarios con distintos roles aporten su conocimiento y experiencia para generar las relaciones necesarias entre diversos conceptos. La semántica incorporada mediante la interacción del usuario incrementa la potencialidad en la generación de los cubos.

En trabajos futuros se continuará con la implementación del prototipo en un ámbito real para validar lo desarrollado en forma teórica y encontrar las complicaciones que podría demandar el uso de este tipo de herramientas.

Sería interesante hacerlo extensivo a otro tipo de bases de datos y adicionar en el estudio un conjunto de métricas que amplíe la visión propuesta en este trabajo. También se evaluarán otros tipos de consultas y mayores testeos para poder evaluar de manera más precisa a nivel estadístico las propuestas. Por último se pretende incorporar cubos sofisticados para el análisis.

Las investigaciones relacionadas a la generación e integración de cubos mencionan diversas alternativas para incorporar ontologías y modelos multidimensionales. Por ejemplo en Netbot et al. [13] propone Ontologías multidimensionales integradas como base para el análisis de datos provenientes de la Web Semántica. Existen investigaciones como la propuesta de Romero y Albelló [16] quienes introducen a un enfoque que deriva en modelos de conceptos multidimensionales partiendo desde dominios de conocimientos representados en ontologías. Otras investigaciones [17] discuten la manera en que las ontologías pueden ser utilizadas para razonar sobre la lógica de sumarización en OLAP. Sell et al. [18] focaliza el uso de ontologías como un significado para representar datos en términos de negocios con el objetivo de simplificar el análisis de datos y customizar aplicaciones BI.

REFERENCIAS

[1] H. Gupta, V. Harinarayan, A. Rajaraman, J. Ullman, "Index selection for OLAP", in: *Proceedings of the International Conference on Data Engineering*. Birmingham, UK, pp. 208–219, 1997.

[2] V. Harinarayan, A. Rajaraman, J.D. Ullman, "Implementing data cubes efficiently", in: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, pp. 205–216, 1996.

[3] H. Shi-Ming; C. Tung-Hsiang; S. Jia-Lang. "Data warehouse enhancement: A semantic cube model approach". *Information Sciences*, vol. 177, no 11, p. 2238-2254, 2007.

[4] J. Han, S. Nishio, H. Kawano, W. Wang, "Generalization-based data mining in object-oriented databases using an object cube model", *Data and Knowledge Engineering*, vol. 25, no 1, p. 55-97, 1998.

[5] L.D. Chen, T. Sakaguchi, M.N. Frolick, "Data mining methods, applications, and tools", *Information Systems Management*, vol. 17, no 1, pp. 65–70, 2000.

[6] S. M. Huang, C.H. Su, "The development of an XML-based data warehouse system", in: *Intelligent Data Engineering and Automated* Learning - *IDEAL 2002*, Springer Berlin Heidelberg, pp. 206–212, 2002.

[7] C.M. Chao, "Incremental maintenance of object-oriented data warehouses", *Information Sciences, vol.* 160, no 1, pp 91–110, 2004.

[8] W.A. Giovinazzo, *Object-Oriented Data Warehouse Design: Building a Star Schema*, Prentice-Hall, New Jersey, 2000.

[9] A. Matei, K. M. Chao, N. Godwin. "OLAP for Multidimensional Semantic Web Databases". *En Enabling Real-Time Business Intelligence*. Springer Berlin Heidelberg, p. 81-96, 2015.

[10] B. Neumayr, M. Schrefl, & K. Linner, "Semantic cockpit: an ontology-driven, interactive business intelligence tool for comparative data analysis". In *Advances in Conceptual Modeling. Recent Developments and New Directions,* Springer Berlin Heidelberg, pp. 55-64, 2011.

[11] L. Jiang, D. Barone, D. Amyot, & J. Mylopoulos, "Strategic models for business intelligence. In *Conceptual Modeling–ER",* Springer Berlin Heidelberg, pp. 429-439, 2011.

[12] D. Barone, T. Topaloglou, & J. Mylopoulos, "Business intelligence modeling in action: a hospital case study". In *Advanced Information Systems Engineering,* Springer Berlin Heidelberg, pp. 502-517, Junio 2012.

[13] V. Nebot, & R. Berlanga, "Building data warehouses with semantic web data". *Decision Support Systems*, vol. *52*, no 4, 853-868, 2012.

[14] S. Anderlik, B. Neumayr, & M. Schrefl, “Using domain ontologies as semantic dimensions in data warehouses”. In *Conceptual Modeling,* Springer Berlin Heidelberg, pp. 88-101, 2012.

[15] P. Vilaboa, M. Paris & R. Vargas N. “Aplicando mapas conceptuales como estrategia para generar un espacio de aprendizaje colaborativo.” *VIII Congreso de Tecnología en Educación y Educación en Tecnología*, Córdoba, 2013.

[16] O. Romero, A. Abelló, “A framework for multidimensional design of data warehouses from ontologies”. *Data & Knowledge Engineering.* vol. 69, no. 11, 1138–1157, 2010.

[17] T. Niemi, M. Niinimaki, “Ontologies and summarizability in olap”. In*: Proceedings of the 2010 ACM Symposium on Applied Computing,* ACM , pp.1349–1353, Marzo 2010.

[18] D. Sell, D. C. da Silva, F. D. Beppler, M. Napoli, F.B. Ghisi, R.C. dos Santos Pacheco, J.L. Todesco*, “*Sbi: a semantic framework to support business intelligence*”.* In*: In Proceedings of the first international workshop on Ontology-supported business intelligence,* ACM, vol. 308, p. 11, New York, 2008.