BASES DE DATOS DE GRAFOS COMO MECANISMO DE REPRESENTACIÓN DE INFORMACIÓN HETEROGÉNEA PARA LA DETECCIÓN DE PATRONES

Martín Gustavo Casatti

Plan de Tesis para el Título de Magister en Ingeniería de Sistemas

Diciembre de 2016

El presente trabajo se realiza como parte de los requerimientos para la obtención del título de Magíster en Ingeniería de Sistemas, tal como dicta la normativa de la Universidad Tecnológica Nacional. *22 de diciembre de 2016*

Índice

1. Justificación

Actualmente la mayor dificultad que pueden encontrar las personas no es el acceso a la información en si sino la selección de información relevante frente a un océano de datos que crece constantemente.

Con una cantidad de datos en constante aumento y la multiplicación de las diversas fuentes que proveen esta información, el volumen de contenido a analizar aumenta exponencialmente y son cada vez más necesarios métodos automatizados para su proceso y categorización.

Por otra parte, tan importantes como los datos en sí, resultan ser las relaciones entre los mismos, lo que establece un conjunto completamente nuevo de temas a considerar, tales como la cardinalidad y la direccionalidad de las relaciones, la distancia entre datos y medidas tales como la centralidad, el contorno y el peso de los componentes, entre otras.

Si a esto sumamos el avance cada vez más acentuado de IoT (Internet of Things, Internet de las Cosas), el cual busca interconectar cada dispositivo imaginable y recolectar la información que cada uno de ellos suministra, estamos frente a un escenario inmanejable desde el punto de vista humano del acceso a la información.

La elaboración de un modelo para la representación de datos heterogéneos de forma consistente y el desarrollo de técnicas de consulta tendientes a obtener información valiosa, no de uno o más datos puntuales, sino desde el punto de vista de estructuras de información completas, reviste fundamental importancia y tiene el potencial de brindar la base para el desarrollo de herramientas nuevas y poderosas para obtener información valiosa a partir de una cantidad enorme de datos relacionados, con independencia de la fuente que los origina.

El presente trabajo propone un plan de tesis que buscar abordar el problema de representación de información a partir de fuentes heterogéneas, de una manera completa y consistente que posibilite usar técnicas de consulta para la detección de patrones entre los datos.

2. Fundamentación

GIRÓ:Bien pero ampliar. No hay referencias a antecedentes o motivación del tesista por el tema abordado

El problema de la representación del conocimiento de tal forma que sea procesable por técnicas automáticas no es nuevo y abarca prácticamente toda la historia de la informática como disciplina. Todos los mecanismos de almacenamiento de información tienen como finalidad última el permitir que un sistema de cómputo pueda acceder y procesar dicha información sin la intervención humana.

Las bases de datos relacionales son quizá el mecanismo más establecido y popular a la hora de almacenar información que debe ser procesada por una computadora. Pero existe un problema inherente a la representación y consulta de la información almacenada y es que para que las consultas puedan realizarse

en lenguaje natural[1] el usuario no tiene por qué conocer la estructura en la cual se hayan almacenados los datos.

Como vemos, las bases de datos relacionales no cumplen con este requisito ya que su estructura, modelada como tablas, compuestas cada una por filas y columnas, debe ser perfectamente conocida para permitir la consulta y el acceso a la información representada.

En 1976, un trabajo de John Sowa, un investigador de IBM (International Bussiness Machines), define las bases de lo que se conoce como grafos conceptuales, con el objetivo de poder realizar consultas a una base de datos sin conocer la estructura subyacente de la misma.[**Sowa1976**]

La representación del conocimiento es una rama de la Inteligencia Artificial (IA) que se dedica a estudiar la representación del conocimiento del mundo de forma tal que una computadora pueda interpretarla y utilizarla para resolver problemas complejos. Y es dentro de este ámbito que los grafos conceptuales se han probado de gran valor.

En palabras de Marko Rodriguez, Doctor por la Universidad de California en Santa Cruz:

> "Una base de datos de grafos, y su ecosistema asociado, puede llevarnos a una solución elegante y eficiente a diversos problemas en el ámbito de la representación del conocimiento y el razonamiento."
>
> (**markorodriguez2011 markorodriguez2011**)

En el año 1983, John Sowa ampliaría y compaginaría todos los trabajos previos realizados sobre sus ideas de los grafos conceptuales en el libro **Sowa1983** en donde tocaría temas tan diversos como la filosofía, la psicología, la lingüística y la inteligencia artificial.

Actualmente se deben tener en cuenta al menos cuatro factores fundamentales a la hora de diseñar un sistema de representación del conocimiento en cualquier dominio dado:

Adecuación Representacional: Habilidad para representar todas las clases de conocimiento que son necesarias en el dominio.[**van2008handbook**]

Adecuación Inferencial: Habilidad de manipular estructuras de representación de tal manera que devengan o generen nuevas estructuras que correspondan a nuevos conocimientos inferidos de los anteriores.[**van2008handbook**]

Eficiencia Inferencial: Capacidad del sistema para incorporar información adicional a la estructura de representación, llamada metaconocimiento, que puede emplearse para focalizar la atención de los mecanismos de inferencia con el fin de optimizar los cómputos.[**van2008handbook**]

[1] En la filosofía del lenguaje, el lenguaje natural es la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación. Se diferencia de los lenguajes artificiales, que han sido diseñados con finalidades específicas.

Eficiencia en la Adquisición: Capacidad de incorporar fácilmente nueva información. Idealmente el sistema por sí mismo deberá ser capaz de controlar la adquisición de nueva información y su posterior representación.[**van2008handbook**]

3. Objetivos

GIRÓ: Deben agruparse en principal y secundarios. No es conveniente que haya llamadas a pié de página. Cualquier aclaración de terminología debe hacer sido hecha con anterioridad

- Definir las estructuras que permitan registrar y relacionar información no homogénea[2] que pueda ser utilizada para inferir patrones de comportamiento aplicables al tráfico de mercaderías a traves de una frontera.
  - Definir los atributos generales y particulares que modelaran el tipo de contenidos y sus relaciones.
  - Diseñar mecanismos de recolección automatizada de información que permitan la actualización permanente y automática de la base de conocimientos.
  - Diseñar algoritmos que permitan la detección de patrones conocidos y la inferencia de patrones desconocidos en las relaciones entre los datos.
  - Implementar, en la forma de un proyecto piloto, las herramientas diseñadas de forma tal que se demuestre la validez de las técnicas mencionadas.

4. Metodología

4.1. Hipótesis

"El uso de grafos conceptuales para la representación de información heterogénea, permitirá la detección de patrones de comportamiento asociados al tráfico de mercadería en una frontera geográfica."

En el presente trabajo se iniciará evaluando las distintas alternativas disponibles para el almacenamiento de información en forma de grafos dirigidos, para soportar el modelo de grafos conceptuales de **Sowa1976**

A continuación se realizará un análisis detallado de las distintas fuentes de información actualmente disponibles para poder determinar un conjunto de atributos que permitan representar cada una de ellas. Estos atributos o descriptores se consolidarán luego de forma tal que se definan un conjunto de descriptores comunes a todas las fuentes de información y un conjunto propio de cada una de las fuentes.[**findler2014associative**]

[2] Se considera información "no homogénea" a toda aquella información que difiera de otra tanto en la estructura de representación como en el tipo de datos que contiene.

Una vez determinados los descriptores generales y particulares se modelarán los distintos tipos de relaciones que pueden unir dos conceptos cualquiera y se definirán los atributos asociados a las mismas.

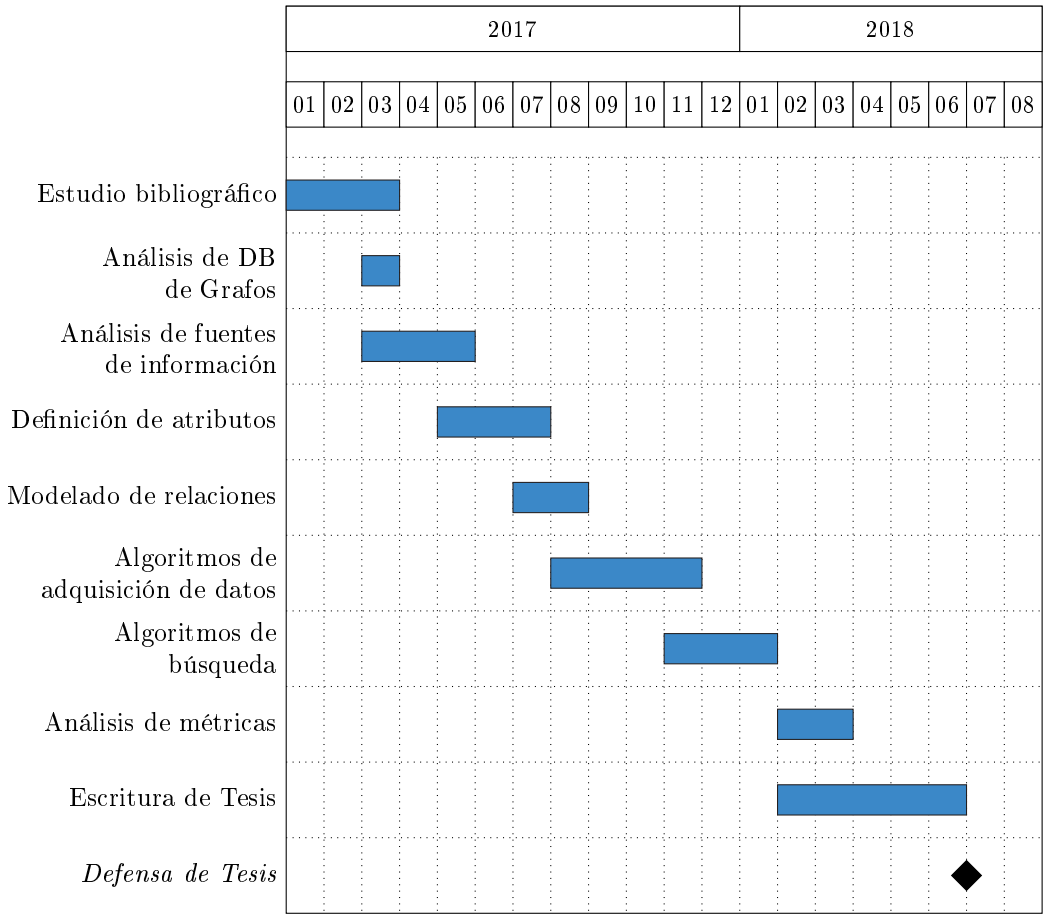
Una vez que los modelos conceptuales estén plasmados en la base de datos elegida, se diseñaran los algoritmos necesarios que permitirán adquirir la información de las distintas fuentes y que, luego de realizar los mapeos y adecuaciones necesarios, registraran la misma en la base de conocimientos, con el mayor nivel de detalle posible, y establecerán las relaciones necesarias con la información asociada, ya existente en la base de conocimentos.

Habiendo concluido con el modelado conceptual de la base de conocimiento y teniendo implementados los mecanismos de carga inicial y mantenimiento, se desarrollarán algoritmos de búsqueda que tengan en cuenta el modelo conceptual existente en la base de datos. Hay que mencionar que no se realizará una consulta relacional sino conceptual, de más alto nivel, pero sin llegar a implementar una consulta en lenguaje natural.

Se implementarán, posteriormente, el análisis de métricas del grafo conceptual constituído, lo cual dará lugar a la detección de patrones que no son evidentes o detectables mediante la simple consulta de los datos. Se buscarán medidas de centralidad, dispersión, cercanía, camino más corto, más largo, perímetro y otras que sirvan para caracteriza la información del grafo o un subconjunto de la misma.

Finalmente se elaborarán las conclusiones, necesarias para determinar si la utilización de un grafo conceptual es un mecanismo válido y eficiente para la representación de información heterogénea para detección de patrones, y se enunciarán los trabajos futuros que surjan a partir de dicho análisis.

5. Cronograma



6. Condiciones institucionales

Gran parte del trabajo mencionado en el presente documento se puede hacer sin el auxilio de ningún tipo de equipamiento especializado o de software propietario. Aún así, es probable que el acceso a algunas fuentes de datos, de caracter gubernamental, puede requerir de ciertas autorizaciones que será necesario gestionar por parte de la Universidad.

El resumen de las necesidades de hardware, software y acceso se encuentra en la tabla 1:

Etapa	Necesidades	Acceso
Estudio bibliográfico	Internet, WebSites de búsqueda de papers y material bibliográfico	Libre
Modelado de DB	Base de datos de grafos[3]	Libre (Mandatorio)
Acceso a información	Internet, Bases de datos abiertas, Websites de noticias, Repositorios de datos de gobierno	Libre[4]
Redacción de tesis	Notebook, Herramientas de redacción (LaTeX, paquetes necesarios), Herramientas de corrección, Graficadores	Propiedad del tesista
Prototipo experimental	Herramientas de desarrollo (Java, Netbeans, Eclipse o similar)	Libre
Todo el desarrollo del trabajo	Ubicación física, aula u otro con acceso a pizarra y archivo	UTN LIS[5]

Cuadro 1: Condiciones institucionales para el desarrollo de la tesis

7. Referencias y Bibliografía

[3] Las bases de datos deben incluir sus correspondientes herramientas de consulta y administración

[4] El acceso a datos de gobierno puede requerir permisos especiales o firma de convenios entre la Universidad Tecnológica Nacional y el organismo que corresponda

[5] Laboratorio de Investigación de Software