Exploración de grafos para el análisis de datos

Mario Alejandro García [1], Juan Bautista Cabral [1], María de la Paz Giménez Pecci [2], Rodrigo Liberal [1], Irma Graciela Laguna [2,3]

[1] Departamento de Sistemas, Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN FRC)
Maestro M. Lopez esq. Cruz Roja Argentina, Córdoba, Argentina

[2] Instituto de Patología Vegetal (IPAVE), Instituto Nacional de Tecnología Agropecuaria (INTA)
Camino 60 cuadras Km. 5 y ½, Córdoba, Argentina

[3] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

(mgarcia@sistemas.frc.utn.edu.ar)

Resumen

Se propone una técnica de análisis de datos a través de la exploración de redes que se forman con entidades existentes en la base de datos de estudio.

El principal aporte del método es la exploración de la red a través de ambientes definidos por los demás atributos de la base de datos.

Este método ha sido utilizado con éxito en el análisis de datos biológicos relacionados con el *Mal de Río Cuarto virus.*

Se presenta además a Yatel, la herramienta en desarrollo que da soporte al análisis propuesto.

Palabras clave:

Análisis de redes, Análisis de grafos, Yatel, Redes de haplotipos, KDD, Data Mining

Contexto

Este trabajo se realiza en el marco del proyecto UTN1685, "Análisis de datos de enfermedades infecciosas transmitidas por *Hemiptera Auchenorrhyncha*", el tercero de una serie de proyectos de cooperación UTN FRC - INTA IPAVE en los que se analizan fitopatologías mediante técnicas de Minería de Datos.

Introducción

Para estudiar la variabilidad genética del virus causante del Mal de Río Cuarto del Maíz (MRCV) [1] se desarrolló una técnica método que ahora se propone como método general de análisis de datos mediante redes.

El método propuesto consiste en representar entidades de la base de datos como nodos de una red cuyos arcos dependen de la diferencia que hay entre las entidades/nodos según el criterio elegido para medir distancias en el espacio multidimensional de los atributos de la base de datos.

Después de crear la red, se analizan sus propiedades y, lo más importante, se la explora dinámicamente de la misma forma que se hace con un cubo OLAP. Esta exploración permite al investigador

tener rápidamente un modelo mental del comportamiento de los datos representados en la red.

El método. Se compone de siete etapas, que al igual que las etapas del proceso de KDD (Knowledge Discovery in Database), se ejecutan cíclicamente. Estas son:

1- Identificación y representación de perfiles. Los perfiles son las entidades que serán los nodos en la red. El nombre "perfiles" se debe a que los datos analizados originalmente fueron perfiles electroforéticos, también llamados en las publicaciones haplotipos (genotipos haploides) [2]. Se mantiene el nombre porque se puede usar de forma general al analizar, por ejemplo, perfiles de redes sociales. Los perfiles pueden estar formados por una entidad única representada en la base de datos o por entidades creadas como combinaciones únicas de algunos atributos, como es el caso de los haplotipos del MRCV [3].

2- Definición de medidas de la distancia. Los arcos pueden ser conexiones naturales entre los nodos, como amistad en una red social o rutas entre ciudades, pero también pueden ser diferencias entre los perfiles, una medida de qué tan distintos son. En otras palabras, la distancia entre el perfil *x* y el perfil *y*, puede ser la cantidad de cambios o mutaciones se deberían realizar sobre *x* para convertirlo en *y*. En este último caso, pueden existir múltiples criterios y es necesario que un experto en el dominio del problema participe de la definición.

3- Cálculo de la distancia. En esta etapa se realiza el cálculo de las distancias según el criterio elegido en el paso anterior. Es importante definir correctamente, tanto la forma de ejecución del cálculo, como los detalles de almacenamiento de los resultados, porque la cantidad de valores calculados puede ser extremadamente grande.

4- Creación de la red. Cuando se crea la red de perfiles se debe decidir si todas las distancias calculadas generarán un arco o sí sólo lo harán aquellas que cumplan con determinado criterio, como por ejemplo, ser menor que algún valor. Como parte de esta etapa, se genera un gráfico de la red donde todas las entidades están presentes.

5- Visualización y análisis topológico. En analista junto con el experto en el dominio hacen un análisis visual del gráfico generado, calculan las propiedades topológicas (por ejemplo diámetro y cluster coefficient) y su interpretación en el problema estudiado.

6- Exploración. Esta etapa se analiza dinámicamente, en el sentido OLAP, la existencia de cada perfil de acuerdo a combinaciones de valores de varias dimensiones. Cada ciclo de exploración tiene las siguientes sub-etapas:

- Selección de dimensiones y niveles de detalles. La combinación de las dimensiones elegidas entre sí según el nivel de agregación de cada una define ambientes que contienen los hechos de la base de datos.

- Visualización de existencia de perfiles por ambiente. El analista navega a través de los ambientes creados (Figura 1). Para cada ambiente se muestra la red completa que se generó en la etapa 4 y se resaltan los nodos de los perfiles que existen para ese ambiente. Además, se muestra para cada ambiente el resultado de un set de cálculos, como cantidad de muestras, cantidad de perfiles distintos,

distancia promedio entre los perfiles existentes, o cualquier cálculo de complejidad arbitraria que el analista definiera.

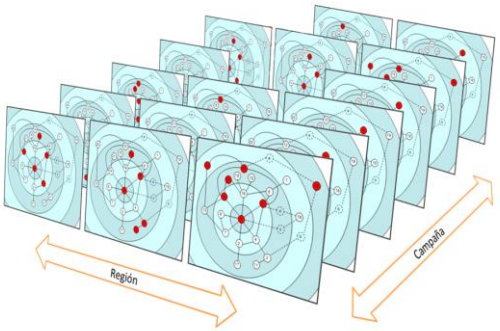

Figura 1. Ambientes generados para dos dimensiones. En todos los ambientes se puede ver la misma red y los perfiles existentes en el ambiente.

7- Generación de hipótesis y conclusiones. Las observaciones y cálculos realizados en las etapas 5 y 6 pueden dar lugar a hipótesis o conclusiones que se deben formalizar en esta etapa.

Yatel. Para dar soporte al método de análisis propuesto se desarrolló Yatel, una base de datos que brinda la infraestructura al proceso antes descrito de una manera genérica, robusta y escalable. Yatel representa la información en una estructura de datos que se decidió llamar NWOLAP (del inglés "Network OnLine Analytical Processing"), claramente inspirando en las infraestructuras de la Inteligencia de Negocios [4].

Desde el punto de vista técnico se siguió una implementación similar a la proporcionada por cualquier herramienta ROLAP (del inglés: "Relational OnLine Analytical Processing") [5] la cual se brinda una capa de abstracción sobre una base de datos relacional para representar la información como nodos, arcos y hechos. Toda la implementación fue realizada en el lenguaje de programación Python.

La arquitectura de Yatel está dividida en los siguientes módulos (Figura 2):

db.YatelNetwork es el componente de más bajo nivel de toda la infraestructura, genera (utilizando sqlAlchemy[1]) la abstracción sobre el almacén de datos para representar red de perfiles.

etl.ETL Es un micro-framework incorporado para la realización de herramientas de extracción transformación y carga de datos en el almacén de redes de perfiles.

El paquete de **data mining**, contiene algoritmos que operan sobre las instancias de las redes de perfiles para descubrir relaciones existentes entre los datos almacenados.

stats contiene un conjunto de funciones estadísticas para medir la variabilidad en las redes.

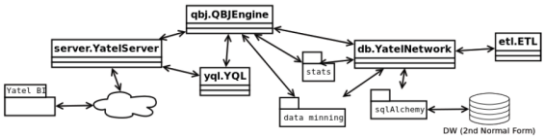

Figura 2. Módulos de Yatel.

Los objetos **qbj.QBJEngine** son envoltorios sobre la red que permiten realizar búsquedas sobre los datos utilizando un lenguaje agnóstico que se decidió llamar "QBJ" (del inglés: Query By Json). Este lenguaje es considerado de bajo nivel, con lo cual es verboso y difícil de escribir por una persona; pero permite acceder a toda la funcionalidad de Yatel de manera declarativa.

[1] http://www.sqlalchemy.org/

El objeto **yql.YatelQueryLanguage** (YatelQL) es el lenguaje de alto nivel implementado sobre QBJ que aún esta en etapas tempranas de diseño, pero persigue el objetivo de ser sencillo para realizar las consultas sobre las redes de perfiles.

server.YatelServer es en objeto encargado de servir la red utilizando el protocolo HTTP. Integra en su diseño la posibilidad de realizar consultas en QBJ y YatelQL sobre la red servida. En el futuro contará con las funcionalidades de los, todavía en diseño, módulos de seguridad y cache.

Yatel BI: Es un proyecto en implementaciones tempranas separado de Yatel que se encargará de brindar al analista una interfaz amigable para operar los datos de la red en busca de conocimiento.

Si bien quedan fuera de esta descripción existen otros módulos, como los encargados de la utilización de Yatel por línea de comandos y la exportación e importación de datos de manera agnóstica de la base subyacente en formatos basados en JSON y XML.

Líneas de Investigación, Desarrollo e Innovación

Las principales líneas de I/D/I de este trabajo son:

- Aplicaciones del análisis de redes
- Visualización de datos
- Almacenamiento de datos orientado a redes
- Optimización de consultas orientadas a redes

Resultados y Objetivos

Los resultados obtenidos son positivos.

Utilizando este método se logró evidenciar que la variabilidad del *Mal de Río Cuarto virus* ha disminuido después de la gran epidemia de 1996/97 [6]. Para este caso se definieron perfiles (haplotipos) con las bandas electroforéticas del virus, se crearon medidas de distancia basadas en la distancia de Hamming [7] más modificaciones fundadas en el conocimiento biológico del virus [8], se exploraron las redes generadas, donde se pudo ver que en las primeras campañas muestreadas los perfiles existentes eran más y con mayores distancias entre ellos, para luego confirmar la observación mediante la creación de una prueba basada en el indicador SDH (Suma de distancia entre haplotipos) y su valor esperado E(SDH) [9].

Nuestra conclusión es que, en un proceso centrado en la persona (*human-centered*), donde la creatividad y experiencia del analista juega un rol fundamental [10], la herramienta propuesta es capaz de ofrecer una perspectiva novedosa y complementaria con las demás técnicas de KDD.

Entre los objetivos del proyecto se encuentran:

- Con respecto al método planteado, su difusión y aplicación a distintos dominios.
- Con respecto a Yatel, su difusión como proyecto *open source*[2] y la extensión/optimización de las funcionalidades.

[2] http://getyatel.org/

Formación de Recursos Humanos

La estructura del grupo según la afiliación de los integrantes es la siguiente:

- UTN FRC
  - (1) Director
  - (1) Investigador graduado
  - (2) Investigador alumno
- INTA IPAVE
  - (3) Investigador
- CONICET
  - (1) Investigador

Referencias

[1] Laguna I.G., Remes Lenicov A.M.M., Virla E., Avila A., Giménez Pecci, M.P., Herrera P., Garay J., Ploper L.D., Mariani R.: *Difusión del "Mal de Río Cuarto" (MRCV) del maíz, su vector, delfácidos asociados y hospedantes alternativos en Argentina*. Revista de la Sociedad Entomológica Argentina 61, 87--97 (2002)

[2] Gimenez Pecci M.P., Bruno C., Balzarini M., Laguna I.G.: *Aplicación del análisis de la varianza molecular en datos de perfiles electroforéticos de segmentos genómicos del Mal de Río Cuarto virus (MRCV) del maíz (Zea mays L.) en Argentina*. Actas de la Academia Nacional de Ciencias 13, 141--152 (2007)

[3] Gimenez Pecci M.P., Carpane P., Murua L., Bruno C., Balzarini M., Laguna I.G.: *Variabilidad del Mal de Río Cuarto virus (MRCV) del maíz según frecuencia de haplotipos obtenidos desde perfiles electroforéticos de los segmentos genómicos*. Actas de la Academia Nacional de Ciencias 14, 99--107 (2008)

[4] Surajit C., Umeshwar D.: An overview of data warehousing and OLAP technology SIGMOD Rec. 26, 1, 65-74 (1997)

[5] Bach Pedersen T.; S. Jensen C.: *Multidimensional Database Technology*. Distributed Systems Online (IEEE): 40–46 (2001)

[6] García, M.A., Maurino M.F., Cucco N,, Laguna I.G., Giménez Pecci M.P., Nieto A., Cabral J.B.: *Aplicación de redes de haplotipos al análisis espaciotemporal de la variabilidad del virus del Mal de Río Cuarto (MRCV)*. 4ta Escuela Argentina de Matemática y Biología (2010).

[7] Hamming R.W. *Error Detecting and Error Correcting* Codes. Bell System Tech Journal, Vol. 9, pp.147-160. (1950)

[8] García, M.A., Gimenez Pecci M.P., Cabral J.B., Laguna I.G., Maurino F., Vera C.H.: *Analysis of variability of Mal de Río Cuarto virus (MRCV) through haplotype networks*. Memorias del 3er Congreso Argentino de Bioinformática y Biología Computacional, 30 (2012)

[9] García, M.A., Giménez Pecci M.P., Cabral J.B.,Nieto A., Laguna I.G.: *Interactive network exploration in the KDD process, Contributions in the study of population variability of a Corn Fijivirus*. Journal of Data Mining in Genomics & Proteomics ISSN: 2153-0602 (2012)

[10] Brachman R.J., Anand T.: *The Process of Knowledge Discovery in Databases: A Human-Centered Approach*. Advances in Knowledge Discovery and Data Mining, MIT Press, pp.37-58. (1996)