Minería de grafos en el dominio de trayectos académicos

Smail Ana[1], Pompei Sabrina[1], Bendati Natalia[2], Russo Claudia[3], Ramón Hugo[3], Fanny Vizcaino[2], Emanuel Lazzari[4]

Instituto de Investigación y Transferencia de Tecnología (ITT)[5]
Comisión de Investigaciones Científicas (CIC)
Escuela de Tecnología (ET)
Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)

Sarmiento Nro 1119 3er Piso, Junín (B) – TE: (0236) 4477050 INT 11610

{ana.smail, sabrina.pompei, natalia.bendati, claudia.russo, hugo.ramon, emanuel.lazzari}@itt.unnoba.edu.ar; fannybv8@gmail.com

RESUMEN

Entre las problemáticas actuales en la educación superior argentina se encuentra la necesidad de disminuir la duración real de las carreras. Hoy, los estudiantes universitarios "en promedio, tardan media carrera más de lo estipulado", afirma Mónica Marquina, coordinadora general del Programa de Calidad Universitaria de la Secretaría de Políticas Universitarias. [1]

Distintos factores provocan esta sobre-duración ya sea por causas vinculadas a condiciones extrínsecas a los estudiantes como intrínsecas a los mismos: desigualdad social, débil formación en la escuela secundaria, rigidez de los currículos, entre otros.

El presente trabajo propone inicialmente generar una base de datos orientada a grafos con la trayectoria académica de los egresados de las carreras Licenciatura en Sistemas e Ingeniería Informática de la UNNOBA. A partir del análisis de los grafos resultantes de dichas trayectorias, mediante el cálculo del camino más transitado y su comparación con el óptimo, se espera detectar posibles estructuras que incidan en la sobre-duración de las carreras en estudio.

En cuanto al diseño operativo del proyecto se propone una triangulación metodológica a fin de validar la información obtenida contra datos de entrevistas y encuestas a los propios egresados.

Esperamos que este trabajo nos permita avanzar en la línea de investigación minería de grafos, tanto en su aplicación como también en aspectos teóricos.

Palabras clave: minería de grafos, base de datos orientada a grafos, trayectoria académica, duración real de carreras

[1] Docente Investigador - ITT
[2] Becario - ITT
[3] Docente Investigador ITT - Investigador Asociado Adjunto sin director CIC
[4] Becario CIN - ITT
[5] ITT - Centro Asociado CIC

CONTEXTO

La línea descripta está inserta en los proyectos de investigación “Informática y Tecnologías Emergentes" y “Tecnología y aplicaciones de Sistemas de Software: Innovación en procesos, productos y servicios”, presentados en la convocatoria “Subsidios de investigación bianuales” 2019 de la Universidad Nacional del Noroeste de Buenos Aires.

Dicho proyecto tiene lugar de trabajo en el Instituto de Tecnología y Transferencia (ITT) de la UNNOBA, Centro Asociado CIC. Dada la dependencia académica del mismo con la Escuela de Tecnología (ET) de la UNNOBA, se propone el estudio de duración de las carreras del área informática correspondiente a la oferta académica de la ET.

1.INTRODUCCIÓN

La creciente demanda de análisis basado en relaciones ha puesto en escena el uso de grafos para el descubrimiento de conocimiento. El modelo en grafo es útil particularmente cuando los datos a almacenar tienen multitud de interrelaciones entre sí, y cuando la importancia recae más en las interrelaciones que se establecen entre los datos, que en los propios datos. La minería de grafos es la encargada de encontrar patrones significativos, útiles y novedosos en una representación basada en grafos de los datos originales.

Un grafo G está compuesto por un conjunto de vértices V(G) y un conjunto de enlaces o arcos E(G). Un enlace es un par no ordenado de vértices que se denota {u, v}. Un grafo G’ = (V’, E’) es subgrafo de G si V’ □ V y E′□ E y se denota G’ □ G .

Instancia		Grafo
Elemento	⇔	vértice
Atributos Elemento	⇔	Etiqueta Vértice
Relaciones	⇔	Arcos
Tipo de Relaciones	⇔	Etiqueta Arcos

Una representación basada en grafos es lo suficientemente flexible para permitir tener más de una representación para un dominio dado. Esto permite al investigador experimentar para obtener la mejor representación para su dominio. [2] Las bases de datos pueden ser conjuntos de grafos de pequeño o mediano tamaño o un único grafo conexo de gran tamaño. En base a esto, las consultas más realizadas se consisten en buscar subgrafos similares a una estructura dada o, a partir de una frecuencia de ocurrencia determinada, buscar todos los subgrafos que se repitan con igual o mayor frecuencia entre todos los grafos de la base o dentro de un grafo de gran tamaño. [3]

La minería de datos basada en grafos no solo se enfoca en encontrar subestructuras repetitivas y comunes dentro de los datos de entrada, sino también en el proceso de identificar conceptos que describen a las subestructuras más importantes para una mejor interpretación de los datos. Una vez descubierta la misma podrá ser utilizada para simplificar el grafo original mediante el reemplazo de la subestructura por un vértice que represente a la recién descubierta subestructura.

En cuanto al espacio de búsqueda de un algoritmo basado en grafos, este está compuesto de todos los grafos derivados a partir del grafo de entrada. Es decir, la búsqueda de patrones discriminativos necesita realizar pruebas de isomorfismo de subgrafos en los grafos de entrada, el cual es un problema NP-Completo. [4]

A fin de reducir la complejidad algunos algoritmos dividen los datos de entrada en varios grupos y de esta manera el número de comparaciones entre los elementos se reduce. De hecho, existen varios algoritmos de minería de datos, entre los cuales se encuentran el FSG, FFSM, gSpan y GASTON. [5] Los mismos

varían en la estrategia que utilizan para recorrer los grafos, el tipo de entrada que utilizan y la información de salida que proveen. Para dar soluciones más eficientes a los problemas necesario compararlos y determinar cuál es el más apropiado para el dominio de datos con los que se cuenta.

Sobre el espacio de búsqueda es posible además observar distintos parámetros y/o métricas que proporcionan información sobre el dominio en estudio. Por ejemplo: la centralidad de grado, centralidad de cercanía, centralidad de intemediación; componentes fuertemente conectados, componentes débilmente conectados, tamaño de componente gigante; rutas más cortas; densidad del grafo.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Las líneas de investigación propuestas se enmarcan en la minería de grafos.

- Exploración de posibles representaciones para las trayectorias académicas mediante grafos.
- Generación de una base de datos orientada a grafos para almacenar trayectorias académicas.
- Estudio y evaluación de algoritmos para la detección de patrones en el dominio de los trayectos académicos.
- Estudio de herramientas de visualización y análisis de grafos.

3.RESULTADOS OBTENIDOS/ESPERADOS

Mediante la aplicación de minería de grafo al dominio de los trayectos académicos de los egresados de las carreras informáticas de la UNNOBA, se espera detectar posibles factores que inciden en la duración real de las carreras. Esta información puede ser clave para la toma de decisiones a fin de disminuir el índice de retraso en la carrera.

Se espera, además, alcanzar un modelo de grafo conceptual aplicable a otras carreras de la Universidad.

Para ello proponemos los siguientes objetivos específicos:

- Explorar patrones topológicos de grafos que contengan información relevante sobre los trayectos académicos de los egresados de las carreras informáticas de la UNNOBA.
- Analizar la existencia de patrones.
- Aplicar algoritmos que permitan obtener el camino más transitado.
- Comparar el camino más transitado con el camino óptimo.
- Inferir posibles causas de retraso en la carrera.
- Comparar la información obtenida con los datos obtenidos de entrevistas y encuestas de los egresados.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de trabajo está compuesto por docentes investigadores, becarios alumnos, y alumnos avanzados de las carreras del área de informática.

En el marco del proyecto se desarrollarán dos tesis de maestría, una en minería de datos y otra en metodologías de la investigación científica. Como así también una tesina de grado correspondiente a la Licenciatura en Sistemas.

5. BIBLIOGRAFÍA

[1] “Reconocimiento académico: pensar la formación con foco en el estudiante.”, Ministerio de Educación, Cultura, Ciencia y Tecnología, 2018. [En línea]. Disponible en: https://www.argentina.gob.ar/noticias/reconocimiento-academico-pensar-la-formacion-con-foco-en-el-estudiante [Accedido: 12-mar-2019]

[2] A. Cortez Vásquez y L. Pro Concepción, “Descubrimiento de conocimiento basado en

grafos", *Revista de Investigación de Sistemas e Informática*, vol 8, no. 2, pp 17-25, 2011 [En línea]. Disponible en: http://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/download/6321/5541 [Accedido: 26-feb-2019]

[3] S. Bianco, "Análisis comparativo de algoritmos de minería de subgrafos frecuentes", tesis de grado, Universidad Nacional de Lanus, 2016 [En línea]. Disponible en: http://sistemas.unla.edu.ar/sistemas/gisi/TFLS/Bianco-TFL.pdf. [Accedido: 1-mar-2019]

[4] R. Salomón Fonseca Delgado, "Diseño de un algoritmo de minería de datos basado en grafos para la tarea de aprendizaje de conceptos", tesis de maestría, Instituto Nacional de Astrofísica, Optica y Electrónica de Puebla, 2012 [En línea]. Disponible en: https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/745/1/FonsecaDR.pdf [Accedido: 26-feb-2019]

[5] S. Bianco, S. Martins y R. García Martinez, "Estudio compartivo de algoritmos de minería de subgrafos frecuentes" en 2016 XXII Congreso Argentino de Ciencias de la Computación, pp 748-457 [En línea]. Disponible en: http://hdl.handle.net/10915/56756 [Accedido: 3-mar-2019]

- P. Tan, M. Steinbach y V. Kumar, *Introduction to Data Mining*. New York, NY, USA: Pearson, 2005.
- T. Munzner, *Visualization Analysis & Design*. Boca Raton, USA: CRC Press, 2014.
- I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*. Sebastopol, USO: O'Reilly, 2015.
- L. Haberfeld, M. Marquina y S. Morresi, *El Sistema universitario argentino. Situación, problemas y políticas*. Centro de Estudios para el cambio estructural (CECE), 2018. [En línea] Disponible en: http://fcece.org.ar/el-sistema-universitario-argentino/ [Accedido: 14-nov-2018]