Análisis y detección de patrones en un grafo conceptual construido a partir de respuestas escritas en forma textual a preguntas sobre un tema específico.

María Alejandra Paz Menvielle, Cynthia Lorena Corso, Karina Ligorria, Analía Guzmán, Martín Casatti

Departamento de Ingeniería en Sistemas de Información
Facultad Regional Córdoba – Universidad Tecnológica Nacional
Maestro Marcelo López esq. Cruz Roja Argentina – Córdoba
0351 – 4686385

pazmalejandra@gmail.com, cynthia.corso@gmail.com, karinaligorria@hotmail.com, aguzman@sistemas.frc.utn.edu.ar, mcasatti@gmail.com,

Resumen

En este trabajo se presenta una línea de Investigación que busca descubrir patrones asociados a la evolución de una base de conocimiento representada en una base de datos orientada a grafos, la misma contiene respuestas de exámenes en formato de texto de redacción libre relacionadas a un dominio específico, utilizada para realizar el análisis de texto en respuestas a preguntas de exámenes en la cátedra de Paradigmas de Programación, con el objetivo de detectar el grado de acierto de las respuestas de los alumnos. Dentro de los patrones que se pretenden descubrir se encuentran aquellos asociados a las respuestas de los alumnos, a la forma de representación de las preguntas de los docentes, entre otras. Es por ello que el presente proyecto busca avanzar en la línea de investigación relacionada a la detección de patrones a partir de grafos dirigidos, tanto en sus aspectos teóricos como prácticos y en sus aplicaciones.

Palabras clave: grafos – patrones – rutas – comunidades.

Contexto

El presente trabajo forma parte del proyecto de investigación y desarrollo que ha sido homologado por la Secretaría de Investigación, Desarrollo y Posgrado de la Universidad Tecnológica Nacional, reconocido con el código PIDEIUTNCO0004812, el mismo forma parte del Centro de Investigaciones, Desarrollo y Transferencia de Sistemas de Información – CIDS.

Para su desarrollo se utilizará como caso testigo a la cátedra de Paradigmas

de Programación, perteneciente a la carrera Ingeniería en Sistemas de Información, dictada en la Facultad Regional Córdoba, de la Universidad Tecnológica Nacional. Se respetarán los contenidos mínimos fijados para esta asignatura tal cual figuran en la ordenanza 1150 de la carrera, los cuales pertenecen al bloque de tecnologías básicas dentro del área programación, que están principalmente referidos a los paradigmas lógicos, funcional y de orientación a objetos. Además se cumplirá con los descriptores y criterios de intensidad de formación práctica de la Resolución Ministerial 786/09, los que se encuentran definidos en el área de tecnologías básicas, sub-área programación que incluyen a los paradigmas y lenguajes de programación.

El trabajo que aquí se presenta es la continuación de los trabajos realizados durante el desarrollo del PID EIUTNCO0003592 "Metodología para determinar la exactitud de una respuesta, escrita en forma textual, a un interrogatorio sobre un tema específico". Durante el transcurso de dicho proyecto se generó una base de datos de grafo cuyo objetivo principal es la registración y almacenamiento de todos los conceptos contenidos en la curricula de la materia Paradigmas de Programación. En el presente proyecto se propone complementar la funcionalidad del proyecto anterior, mediante la búsqueda, el análisis y la propuesta de patrones topológicos frecuentes en un grafo conceptual construido para determinar la exactitud de las respuestas, escritas en forma textual sobre un tema específico.

Introducción

Un patrón es una entidad a la que se le puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas (*vector de características*) [1].

En el campo del Reconocimiento de Patrones un enfoque que está ganando popularidad es el de la aplicación de grafos como herramienta para la representación de entidades con estructuras complejas [2].

En una representación de este tipo los vértices y sus atributos representan objetos (o partes de ellos) mientras que los arcos representan relaciones entre estos objetos. Este enfoque explota la generalidad inherente de las representaciones basadas en grafos y gracias a las mejoras en la capacidad de procesamiento de las computadoras, estas representaciones estructurales y los algoritmos que sobre ellas se aplican se han vuelto más eficientes en su aplicación [2].

Caracterización

A la hora de determinar patrones dentro de un grafo dirigido existen un conjunto de medidas que caracterizan el grafo y que determinan qué tipos de técnicas se pueden utilizar de manera más o menos eficiente para obtener información estructural del mismo [3]. Estas medidas se denominan "características" o "métricas" y son inherentes al grafo en su conjunto, son dinámicas y cambian a medida que nuevos nodos y arcos se van incorporando a la estructura. Algunas de las

características más importantes de los grafos son:

Tamaño: El tamaño de un grafo se determina por la cantidad de nodos que lo componen.

Grado de un vértice: Es la cantidad de arcos que convergen en el mismo. Esta característica es particular de cada vértice pero se utilizan algunas medidas generales, como el grado máximo, mínimo o promedio, para caracterizar el grafo de manera general.

Densidad: La densidad es la relación existente entre la cantidad total de arcos del grafo con respecto a los nodos que lo conforman. Un grafo denso, en este contexto, es un grafo con una gran cantidad de interconexiones entre nodos.

Isomorfismo en grafos

Consideramos un grafo *G* definido como *G = (V,E,L)* siendo *V* un conjunto finito de vértices, *E* un conjunto finito de arcos entre los mismos y *L* un conjunto finito de etiquetas que se pueden aplicar tanto a vértices como a arcos.

Un grafo *G' = (V',E',L')* es isomórfico de *G* si: Existe un mapeo biyectivo entre los vértices *V* y *V'*. Si existe un arco *E* entre dos vértices de *G* existe también un arco *E'* entre los dos vértices correspondientes en *G'*. Las etiquetas utilizadas en *G* son preservadas al realizar el mapeo entre *G* y *G'* [4].

Especificación de patrones

Una forma simple de detección de patrones en grafos es el problema de encontrar un subgrafo (el "resultado") de un grafo de entrada dado (el "objetivo") tal que ese subgrafo sea isomórfico de otro grafo de entrada (el "patrón") [5].

Un enfoque más general busca encontrar todos los subgrafos ("resultados") dentro del grafo objetivo y no solamente uno de ellos.

Análisis de patrones comunes

Mientras que la minería de datos se enfoca principalmente en los valores de los datos que se están buscando, en los esquemas semi-estructurados y en los grafos, el enfoque se encuentra en etiquetas frecuentes y topologías comunes [6]. En estos la estructura de los datos es tan importante como su contenido.

Originalmente se plantearon soluciones para el hallazgo de estructuras representadas por una ruta simple (single-path) y para estructuras de tipo árbol, pero actualmente muchas de las estructuras que se encuentran en la Web, así como en redes sociales o comunidades online tienen la forma de grafos más complejos, tanto cíclicos como acíclicos.

Es por eso que ésta disciplina ha experimentado un resurgimiento debido a que el descubrimiento de los patrones subyacentes posibilita un mejor diseño de las bases de datos que gestionan estas estructuras y un mejor indexamiento en la aplicación de algoritmos que tienen en cuenta las preferencias de los usuarios (recomendaciones online, compras, sugerencias de grupos afines, etc.), en las predicciones de comportamiento, entre otras.

Si bien surgen y se mejoran algoritmos de uso específico para detección de patrones conocidos, un problema inherente a éstas técnicas, radica en el

planteo de algoritmos de uso general (que presenten un rendimiento razonable) y en el descubrimiento de patrones, hasta el momento desconocidos mediante técnicas generales [9].

Líneas de Investigación, Desarrollo e Innovación

- Estudio de los diferentes patrones topológicos de grafos que puedan ser relevantes en la búsqueda de información en el dominio elegido, analizando si dichos patrones tienen comportamientos recurrentes o subyacentes.
- Estudio de algoritmos que permitan detectar patrones conocidos en la teoría de grafos como son las “comunidades”, “pares”, “rutas principales” y otros patrones comunes por medio del análisis de las métricas de la base de conocimientos.
- Detección de patrones que, aún no siendo comunes en otras áreas de la teoría de grafos, si lo son recurrentes en el dominio bajo estudio.
- Aplicación de los patrones encontrados sobre las respuestas elaboradas por los alumnos en un examen posibilitará descubrir algunas características importantes que se relacionan con el aprendizaje [7].
- Estudio de herramientas de visualización y análisis de grafos (como Gephi o GUESS, entre otras), para realizar los análisis preliminares y la determinación de los parámetros y métricas de la base de datos [8].
- Automatización de algunos de estos análisis para incluirlos en una herramienta ad-hoc.

Resultados y Objetivos

El objetivo del presente estudio es analizar, detectar y evaluar patrones topológicos frecuentes en un grafo conceptual construido para determinar la exactitud de las respuestas, escritas en forma textual sobre un tema específico, utilizando una base de conocimientos diseñada como un grafo dirigido. Para ello hemos identificado los siguientes objetivos particulares:

1. Explorar patrones topológicos de grafos que contengan información relevante para la identificación de estructuras dentro de la base de conocimientos de la materia Paradigmas de Programación.
2. Analizar la existencia de patrones recurrentes o subyacentes en los grafos generados a partir de las respuestas base de los docentes y los obtenidos de las respuestas dadas por los alumnos.
3. Proponer algoritmos que permitan detectar patrones conocidos en la teoría de grafos como son las "comunidades", "pares", "rutas principales" y otros patrones comunes, por medios del análisis de las métricas sobre la base de conocimiento.
4. Identificar características como exactitud, coherencia y consistencia, entre otras, de las respuestas escritas en forma textual, en la base de conocimiento diseñada como un grafo dirigido.

Formación de Recursos Humanos

Dentro del desarrollo de este proyecto de investigación se está desarrollando el trabajo de Tesis de Maestría de dos integrantes docentes del presente proyecto. Se incorporan al equipo de trabajo docentes-investigadores de la carrera Ingeniería en Sistemas de Información como investigadores de apoyo con la finalidad de que inicie su formación en investigación científica y tecnológica, se incorpora un becario graduado BINID y becarios alumnos, quienes colaborarán en la recolección, manipulación y desarrollo de este marco metodológico. En el marco del proyecto los estudiantes tendrán la posibilidad de hacer la Práctica Supervisada de quinto año. Los avances, propuestas y herramientas construidas, estarán disponibles para su transferencia y aplicación en el Centro de Investigaciones, Desarrollo y Transferencia de Sistemas de Información - CIDS. Del mismo modo la detección de patrones sobre el dominio de conocimiento de la materia Paradigmas de Programación continuará beneficiando a los integrantes de la cátedra y a los estudiantes.

Referencias

[1] Watanabe, Satosi. Pattern recognition: human and mechanical. John Wiley & Sons, Inc., 1985.

[2] “Graph for pattern recognition” (Doctoral dissertation, Ph. D. dissertation, Universite Francois-Rabelais). Raveaux, R., & Abu, Z. (2013).

[3] Joyner, David, Minh Van Nguyen, and David Phillips. "Algorithmic graph theory and sage." *Version 0.8-r1991* (2013)

[4] VAN STEEN, Maarten. An Introduction to Graph Theory and Complex Networks. *Copyrighted material*, 2010.

[5] “An algorithm for subgraph isomorphism”. J. R. Ullmann. Journal ACM, 23(1):31–42, 1976.

[6] “Computing frequent graph patterns from semistructured data. En Data Mining”. VANETIK, Natalia; GUDES, Ehud; SHIMONY, Solomon Eyal. 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002. p. 458-465.

[7] “Introduction to information retrieval”. C. D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Cambridge University Press. Julio 2008. También disponible on line: www.safaribooksonline.com

[8] "Classification of graph metrics". Hernández, Javier Martín, and Piet Van Mieghem. Delft University of Technology, Tech. Rep (2011).

[9] FOGGIA, Pasquale; SANSONE, Carlo; VENTO, Mario. A performance comparison of five algorithms for graph isomorphism. En *Proceedings of the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition*. 2001. p. 188-199.