ANALES DE LA
Academia de Ciencias de Cuba

Algoritmos para la minería de patrones frecuentes aproximados en colecciones de grafos

ENTIDAD EJECUTORA PRINCIPAL: Centro de Aplicaciones de Tecnologías de Avanzada. División de Investigaciones, CENATAV, Datys Soluciones Tecnológica)

Otras entidades participantes: Instituta Nacional de Astrofísica, Óptica y Electrónica (INAOE), México

AUTOR PRINCIPAL: Dr. C. Niusvel Acosta Mendoza[1]

Otros autores: Dr. C. Andrés Gago Alonso[1], Dr. C. José Eladio Medina Pagola[1], Dr. C. Jesús Ariel Carrasco Ochoa[2], Dr. C. José Francisco Martínez Trinidad[2]

Filiación: [1]CENATAV, Datys Soluciones Tecnológicas; [2]Instituta Nacional de Astrofísica, Óptica y Electrónica (INAOE), México

Palabras clave

minería; subgrafos frecuentes aproximados; patrones frecuentes; multigrafos

RESUMEN

En numerosas aplicaciones reportadas, los grafos han permitido naturalmente representar las entidades y sus relaciones. La búsqueda de subgrafos frecuentes aproximados (SFA) en colecciones de grafos ha sido de gran ayuda en diferentes tareas donde se permiten variaciones en la correspondencia entre grafos. Sin embargo, no existen algoritmos para la minería de este tipo de patrones donde se permitan variaciones en las etiquetas de los vértices y aristas manteniendo la estructura de los grafos. Por este motivo, en esta investigación se propone el primer algoritmo de este tipo para la minería de SFA en colecciones de grafos. Además, con el objetivo de mostrar la utilidad de dichos patrones, se propone un método para la clasificación de imágenes representadas mediante SFA y otro para el agrupamiento de imágenes. Con el objetivo de mejorar la eficiencia del algoritmo propuesto, se proponen dos podas que inciden en la disminución del espacio de búsqueda, del número de candidatos y de las pruebas de formas canónicas. Con fines similares, pero para reducir la representación de la colección de imágenes basada en SFA, se explora el uso de algoritmos de selección de características, se propone el uso de patrones emergentes y se extiende el algoritmo propuesto para identificar solo los SFA maximales durante el proceso de minería. Consecuentemente, se realiza un estudio del impacto de la selección de patrones en la clasificación de imágenes basada en este tipo de minería. Los algoritmos y métodos propuestos trabajan sobre colecciones de grafos simples; sin embargo, en algunas aplicaciones los multigrafos han sido utilizados para modelar los datos, porque en la realidad existe comúnmente más de una relación (arista) entre las entidades representadas como vértices. Sin embargo, los algoritmos reportados para la minería de SFA han sido diseñados para trabajar con grafos simples. Por tanto, con el objetivo de solucionar el problema de minar SFA en colecciones

de multigrafos, en esta investigación se proponen dos algoritmos —el primero basado en matrices de adyacencia, y el segundo basado en búsqueda en profundidad— para la minería de SFA directamente sobre las colecciones de multigrafos. Para la confección de estos dos últimos algoritmos fue necesario extender las formas canónicas basadas en matrices de adyacencia y en búsqueda en profundidad para que estas pudieran representar multigrafos, ya que están diseñadas solo para representar grafos simples. Además, con el objetivo de reducir el conjunto de SFA que se minan, se propone la extensión de un algoritmo para la minería tanto de los SFA cerrados, como de los maximales, así como de otros patrones con pequeñas diferencias en sus soportes. La novedad de esta investigación está avalada por la comunidad científica internacional a través de 15 publicaciones y una tesis doctoral.

En minería de datos, la minería de patrones frecuentes se ha convertido en un importante tópico con aplicaciones en varios dominios de la ciencia, donde las técnicas para la minería de subgrafos frecuentes han sobresalido. El objetivo de estas técnicas es encontrar subgrafos que se repiten frecuentemente en un conjunto de grafos. Los grafos son utilizados para modelar los datos en aplicaciones reales, donde las entidades u objetos se representan como vértices y las relaciones entre ellos como aristas.

En la literatura se han reportado varios algoritmos para la minería de subgrafos frecuentes, basados en correspondencias aproximadas entre grafos, en colecciones de grafos que permiten variaciones en los datos. Estos algoritmos surgen porque existen varios problemas en la vida cotidiana donde algunas variaciones en los datos son permitidas y se ve afectada la correspondencia exacta entre grafos. Sin embargo, ninguno de los algoritmos reportados permite variaciones en las etiquetas de los vértices y las aristas manteniendo la estructura de los grafos. Este tipo de minería puede ser de gran utilidad en la clasificación de imágenes y redes sociales donde los objetos puedan ser sustituidos por otros con cierto grado de semejanza manteniendo la estructura espacial de la representación. Debido a las variaciones que se tienen en cuenta en este tipo de minería, es común que se obtenga un gran número de patrones cuando es aplicada en una colección de grafos. Por este motivo, surge la necesidad de identificar patrones de interés dentro de la colección total de patrones, pero no se conocen algoritmos, en este contexto aproximado, que identifiquen patrones interesantes.

Por otro lado, a pesar de la aplicabilidad que han tenido los algoritmos para la minería de subgrafos frecuentes aproximados (SFA) en colecciones de grafos simples, en algunas aplicaciones puede existir más de una relación entre dos vértices, lo que da como resultado una representación en forma de multigrafo. Un ejemplo de esta situación puede observarse en el análisis de una red social, donde las entidades (personas, videos, objetos, etc.) pueden ser modelados como vértices y las multiaristas pueden representar diferentes interacciones entre ellos. Otras redes como las de rutas de transporte, vías férreas y de viajes pueden ser modeladas con multigrafos para determinar el costo mínimo de entregas mediante la predicción de contactos entre estaciones de ómnibus, o encontrar caminos más baratos para viajar en avión, entre otros.

De igual manera, varios trabajos utilizan multigrafos para representar imágenes en diferentes aplicaciones. Los autores de estos trabajos han afirmado que mediante el uso de multigrafos se modela mejor el problema que utilizando los grafos simples. Sin embargo, este tipo de representación no ha sido explotada apropiadamente debido al déficit de algoritmos para la minería de SFA en colecciones de multigrafos.

Sobre la base de lo explicado antes y los problemas detectados en la literatura relacionada, en el marco de la investigación, se obtuvieron los siguientes resultados, los cuales constituyen aportes en el área de estudio. En primer lugar, se desarrolló un nuevo algoritmo (VEAM) para la minería se SFA en colecciones de grafos simples, el cual permite aproximaciones en las etiquetas de los vértices y aristas manteniendo la estructura de los grafos. Este algoritmo se basa en un enfoque de crecimiento de patrones para identificar subestructuras (subgrafos) que se repiten con una cierta frecuencia en colecciones de grafos simples. Para representar los patrones y acelerar el proceso de minería se utiliza la forma canónica basada en matrices de adyacencia. Además, se propusieron dos podas para reducir el espacio de búsqueda, el número de candidatos y el número de pruebas de formas canónicas, con el objetivo de mejorar la eficiencia del algoritmo propuesto.

Se desarrollaron un método para la clasificación de imágenes representadas mediante SFA y otro método para el agrupamiento de imágenes con esa representación. Estos métodos permitieron mostrar la utilidad de los SFA identificados por nuestro algoritmo VEAM en tareas de clasificación y agrupamiento de imágenes.

Se exploró el uso de algoritmos de selección de características y de los patrones emergentes para la reducción del conjunto de SFA. Se realizó un estudio del impacto de la selección de patrones en la clasificación de imágenes basada en la minería de SFA. Se presentó una extensión del método para la clasificación de imágenes, donde se le incluyó la reducción del conjunto de SFA identificados por los algoritmos de minería. De esta manera (mediante el uso de los selectores de características, así como los patrones emergentes) se elimina información redundante y se mejoran significativamente los resultados obtenidos en tareas de clasificación de imágenes.

Se extendió el VEAM para la minería de los SFA maximales en colecciones de grafos simples. De esta manera se atacó el problema de la redundancia entre los patrones identificados en el proceso de la minería. El algoritmo propuesto (*maxVEAM)* mantiene el proceso de minería de VEAM almacenando aquellos patrones que son identificados como maximales sin adicionar costo computacional en el proceso de minería.

Se desarrollaron dos algoritmos (*MgVEAM y AMgMiner)* para la minería de SFA en colecciones de multigrafos sin necesidad del uso de las transformaciones de grafos: MgVEAM está basado en matrices de adyacencia y AMgMiner está basado en búsqueda en profundidad. Para desarrollar estos algoritmos fue necesario extender las formas canónicas basadas en matrices de adyacencia y en búsqueda en profundidad para que trabajaran con multigrafos, ya que originalmente están diseñadas para tratar solo con grafos simples.

Se desarrolló una extensión del algoritmo MgVEAM (llamado *GenCloMgVEAM*) para la minería de SFA generalizados cerrados en colecciones de multigrafos. Con el uso de un umbral, en GenCloMgVEAM se suavizan las definiciones de maximales y cerrados para encontrar, además, algunos SFA con ligeras diferencias en sus soportes.

El impacto económico y social de los algoritmos y métodos obtenidos no podrá ser apreciado en su totalidad hasta que se encuentren desplegados algunos de los productos de DATYS en los que se proyecta la implantación de los aportes científicos de este trabajo. Es importante señalar que el método de clasificación de imágenes basado en subgrafos frecuentes aproximados fue extendido para ser aplicado en la tarea de detección de autoría. Esto permitió la aplicación de los algoritmos y métodos propuestos en el área de procesamiento de textos, aumentando el espectro de las aplicaciones de los resultados científicos obtenidos.

El impacto científico de los resultados obtenidos en esta investigación está en que constituyen una fuente nueva de conocimiento propio del que dispone nuestro país para apoyar al desarrollo y la seguridad. Además, los aportes de esta investigación han sido avalados por investigadores de la comunidad científica internacional, que finalmente aceptaron la publicación de los resultados en revistas y eventos de impacto internacional. De esta manera, se aumentó la visibilidad a Cuba en el área de la ciencia.

El conjunto de resultados presentados tiene alcance nacional e internacional, ya que serán introducidos en productos de la empresa DATYS que serán utilizados tanto dentro como fuera de Cuba. Estos resultados tributan al desarrollo y la seguridad interna del Ministerio del Interior (MININT) y del país. Los resultados de esta investigación tienen un gran impacto en la independencia científico/tecnológica de Cuba.

El principal aspecto para garantizar la sostenibilidad de los resultados alcanzados es el desarrollo de algoritmos con conocimiento propio que permitan el estudio y la solución de problemas en diferentes áreas de la ciencia. La obtención de estos algoritmos es un importante aporte de la Empresa DATYS para con el desarrollo científico/técnico del país, manteniendo la seguridad interna del MININT. A partir del conocimiento adquirido en esta investigación puede ser utilizado para mejorar la eficacia de los métodos de identificación de objetos e imágenes, así como la detección de autoría en tareas de procesamiento de imágenes y textos, respectivamente. Además, se identificarán nuevos problemas y aplicaciones para garantizar la continuidad del impacto de los resultados.

AUTOR PARA LA CORRESPONDENCIA

Dr. C. Niusvel Acosta Mendoza. *7ma #21406 e/ 214 y 216, Reparto Siboney, Playa. La Habana, CP 12200. Correo electrónico: nacosta@cenatav.co.cu*