LA MINERÍA DE SUBGRAFOS FRECUENTES APROXIMADOS
Frequent Approximate Subgraph Mining

Niusvel Acosta-Mendoza, Andrés Gago-Alonso, José E. Medina-Pagola

Centro de Aplicación de Tecnologías de Avanzada, Cuba, La Habana
(Minería de Datos y Textos)
nacosta@cenatav.co.cu

Resumen: En los últimos años se ha podido observar un incremento en el uso de la minería de subgrafos frecuentes aproximados (MSFA). Este tipo de minería, poco a poco, se ha convertido en una importante tarea con un amplio espectro de aplicaciones en varios dominios de la ciencia. Sin embargo, existen muchas áreas donde pudieran obtenerse buenos resultados utilizando estas técnicas y aún no se han aplicado. En este trabajo se realiza un estudio de la MSFA, específicamente de los métodos que tratan las aproximaciones semánticas entre etiquetas manteniendo la topología de los grafos. En este estudio, no solo se hace un resumen los resultados obtenidos al aplicar este tipo de minería sino que además se presentan varias áreas de la ciencia como fuertes candidatas para su aplicación. Por otro lado, en varios de estos trabajos previamente analizados se han enfocado los esfuerzos en mejorar la eficiencia de estas técnicas. Por lo que en este trabajo también se muestra un resumen de las podas propuestas para lograr una mejor eficiencia en la MSFA.

Palabras claves: Minería de grafos aproximados, cotejo de grafos aproximados, subgrafos frecuentes aproximados, grafos etiquetados.

Abstract: In the last years we can seen an increment in the use of frequent approximate subgraph mining (FASM). This kinds of mining, bit by bit, have become important task with wide applications in several domains of the science. However, there are many areas where relevant results can be obtained using these techniques and they have not been yet applied. In this paper, a study of FASM is performed, specifically the models that treat the semantics approximations between labels, preserving the topology of graphs. In this study, a summary of the results obtained by the application of this type of mining is performed, and several areas of the science are presented as strong candidates where this mining process can be applied. On the other hand, in several of the previously analyzed works have been focused on improving the efficiency of these techniques. For this reason, a summary of proposed prunes to get better efficiency in FASM is is also shown in this paper.

Keywords: Approximate graph mining, approximate graph matching, frequent approximate subgraphs, labeled graphs.

1 Introducción

Los datos en múltiples dominios pueden ser naturalmente modelados como grafos [21] ya que estos son estructuras de datos generales y poderosas que pueden ser usadas para representar diversos tipos de objetos [5]. Varios autores han desarrollado técnicas basadas en grafos y métodos para satisfacer la necesidad de convertir grandes volúmenes de datos en información útil [11]. La MSFA es un ejemplo de estas técnicas [3, 8, 9]. Esta técnica se ha convertido en un tema importante en las tareas de minería donde los patrones son detectados teniendo en cuenta distorsiones en los datos. Estos subgrafos brindan información más cercana a la realidad ya que no es común la existencia de dos objetos reales exactamente iguales en el mundo. Por esta razón, se ha convertido en una necesidad el evaluar la similitud entre grafos permitiendo diferencias estructurales, es decir, técnicas de cotejo aproximado. El cotejo aproximado consiste en encontrar una correspondencia entre vértices o aristas de dos grafos para determinar su similitud permitiendo diferencias entre su estructura o etiquetas.

Teniendo en cuenta este hecho, varios algoritmos han sido desarrollados para la MSFA en diferentes dominios de la ciencia, tales como: clasificación de imágenes [3]; análisis de redes genéticas [22], de estructuras bioquímicas [9]; análisis de circuitos, citas, redes sociales y vínculos [8]. Sin embargo, solo *VEAM* [3] y *APGM* [9] identifican los subgrafos frecuentes aproximados (SFA) en colecciones de grafos, donde la aproximación consiste en considerar variaciones en los datos mediante probabilidades de sustitución, manteniendo la topología de los grafos. Estos algoritmos especifican cuáles vértices, aristas o etiquetas pueden reemplazar a otras. De esta forma se defiende la idea de que no siempre una etiqueta de un vértice o de una arista pueda ser reemplazada por cualquier otra. APGM solo trata las variaciones entre el conjunto de etiquetas de los vértices mientras que VEAM realiza el proceso de minería utilizando los conjuntos de etiquetas de los vértices y las aristas.

En este trabajo se realiza un estudio del enfoque utilizado en VEAM ya que es el único trabajo de MSFA que trata las aproximaciones semánticas entre el conjunto de etiquetas de vértices y aristas, manteniendo la topología de los grafos. Mediante este estudio se mostrarán y analizarán los resultados y aportes alcanzados con este enfoque. Por otro lado, se mencionarán algunas posibles aplicaciones donde aún no se ha utilizado la MSFA, en los cuales pudieran obtenerse buenos resultados.

Este trabajo está organizado de la siguiente manera. En la sección 2 se presentan algunos conceptos básicos; en este también se encuentra la descripción de un método aproximado, se define el problema de la MSFA y se presenta el diseño de un algoritmo para la MSFA. Los resultados alcanzados utilizando la MSFA se detallan en la sección 3. Luego, en la sección 4 se presentan las posibles aplicaciones del proceso de MSFA en diferentes dominios de la ciencia. Finalmente, las conclusiones de esta investigación y algunas ideas de trabajo futuros son expuestas en la sección 5.

2 Marco teórico

En esta sección se comenzará con la explicación de los conceptos básicos y las notaciones utilizadas a lo largo de este trabajo. Se presenta el algoritmo más relevante del estado del arte y se describe de manera general su enfoque. Luego, se muestra la función de similitud del algoritmo para la MSFA utilizado como base en este trabajo. Finalmente, se plantea el problema de la MSFA y el diseño de dicho algoritmo.

2.1 Conceptos básicos

Este trabajo es enfocado en grafos etiquetados simples y no dirigidos. En lo adelante cuando se hable de grafo se suponen todas estas características y en otro caso se especificará explícitamente. Antes de presentar su definición formalmente, se define el dominio de etiquetas.

Sean L_V y L_E conjuntos de etiquetas, donde L_V es un conjunto de etiquetas de vértices y L_E es un conjunto de etiquetas de aristas, el dominio de todas las posibles etiquetas es denotado por $L = L_V \cup L_E$.

Un *grafo etiquetado* en L es una 4-tupla, $G = (V, E, I, J)$, donde V es un conjunto en el que sus elementos son conocidos como *vértices*, $E \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$ es un conjunto en el que sus elementos son conocidos como *aristas* (la arista $\{u, v\}$ conecta el vértice u con el vértice v), $I : V \to L_V$ es una *función etiquetadora* que asigna etiquetas a los vértices y $J : E \to L_E$ es una *función etiquetadora* que asigna etiquetas a las aristas.

Sean $G_1 = (V_1, E_1, I_1, J_1)$ y $G_2 = (V_2, E_2, I_2, J_2)$ dos grafos etiquetados en L, se dice que G_1 es un *subgrafo* de G_2 si $V_1 \subseteq V_2$, $E_1 \subseteq E_2$, $\forall u \in V_1, I_1(u) = I_2(u)$ y $\forall e \in E_1, J_1(e) = J_2(e)$. En este caso, se usa la notación $G_1 \subseteq G_2$ y se dice que G_2 es un *supergrafo* de G_1.

Dados dos grafos $G_1 = (V_1, E_1, I_1, J_1)$ y $G_2 = (V_2, E_2, I_2, J_2)$ etiquetados en L, donde $G_1 \subseteq G_2$, se dice que $e = \{u, v\} \in E_2$ es una *extensión* de G_1 si: $V_2 = V_1 \cup \{v\}$ y $E_1 = E_2 \setminus \{e\}$. Este hecho puede ser denotado por $G_2 = G_1 \diamond e$. Se dice que e es una *extensión hacia atrás* si $v \in V_1$, en otro caso se dice que es una *extensión hacia delante* (este extiende el conjunto de vértices de G_1).

Dados G_1 y G_2, se dice que f es un *isomorfismo* entre esos grafos si: $f : V_1 \to V_2$ es una función biyectiva, donde $\forall u \in V_1, f(u) \in V_2 \wedge I_1(u) = I_2(f(u))$ y $\forall \{u, v\} \in E_1, \{f(u), f(v)\} \in E_2 \wedge J_1(\{u, v\}) = J_2(\{f(u), f(v)\})$. Cuando existe un isomorfismo entre G_1 y G_2, se dice que G_1 y G_2 son *isomorfos*. Una manera de enfocar las pruebas de isomorfismo es utilizando formas canónicas (FC) para representar los grafos.

Sea Ω el conjunto de todos los posibles grafos etiquetados en L, la *similitud* entre dos elementos $G_1, G_2 \in \Omega$ es definida como una función $sim : \Omega \times \Omega \to [0, 1]$. Se dice que los elementos son muy diferentes si $sim(G_1, G_2) = 0$, con el aumento del valor de $sim(G_1, G_2)$ más similares son los elementos y si $sim(G_1, G_2) = 1$ entonces existe un isomorfismo entre estos elementos.

Sean $G_1 = (V_1, E_1, I_1, J_1)$, $G_2 = (V_2, E_2, I_2, J_2)$ y $T = (V_T, E_T, I_T, J_T)$ tres grafos etiquetados en L, donde $T \subseteq G_2$. Utilizando un umbral de isomorfismo τ,

se dice que T es una *ocurrencia* de G_1 en G_2 si $sim(G_1, T) \geq \tau$. El *conjunto de ocurrencias* de G_1 en G_2 es denotado por $O(G_1, G_2)$.

Sea T una ocurrencia de G_1 en $G_2 = (V_2, E_2, I_2, J_2)$, utilizando un umbral de isomorfismo τ. El *conjunto de extensiones* de T es denotado por $ExtSet(T) = \{e \in E_2 \mid e \text{ es una extensión de } T\}$.

Sea $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos y G un grafo etiquetado en L, el valor de *soporte* de G en D se obtiene mediante la siguiente ecuación:

$$supp(G, D) = \sum_{G_i \in D} sim(G, G_i)/|D| \tag{1}$$

Cuando $supp(G, D) \geq \delta$, entonces el grafo G ocurre frecuentemente en la colección D, siendo G un *SFA* en D. Nótese que cuando nos referimos a una colección de grafos estamos asumiendo que es una representación construida a partir de la colección de grafos real. El valor del umbral de soporte δ está en $[0, 1]$ asumiendo que la similitud se normaliza a 1. La *MSFA* consiste en encontrar todos los SFAs en una colección de grafos D, utilizando una función de similitud sim y un umbral de soporte δ.

2.2 Un método aproximado

Antes de presentar el método aproximado del algoritmo VEAM [3], donde el cotejo aproximado se basa en los conjuntos de etiquetas de los vértices y aristas, se muestra la definición de matriz de sustitución. Esta matriz puede tener una interpretación probabilística, con la cual se ofrece un esquema probabilístico para esta tarea de minería de subgrafos frecuentes.

Una *matriz de sustitución* $M = (m_{i,j})$ es una $|L| \times |L|$ matriz indizada por el conjunto de etiquetas L. Una celda $m_{i,j}$ $(0 \leq m_{i,j} \leq 1, \sum_j m_{i,j} = 1)$ en M es la probabilidad de que la etiqueta i sea reemplazada por la etiqueta j. Cuando M es diagonal dominante (i.e. $M_{i,i} > M_{i,j}$, $\forall j \neq i$) entonces M se dice que es una *matriz estable* [9]. En lo adelante, cuando se hable de matriz de sustitución se asume este tipo de matriz.

Sean $G_1 = (V_1, E_1, I_1, J_1)$ y $G_2 = (V_2, E_2, I_2, J_2)$ dos grafos etiquetados en L, MV una matriz de sustitución indizada por L_V, ME una matriz de sustitución indizada por L_E, y τ el umbral de isomorfismo. Se dice que G_1 es *isomorfo aproximado* a G_2, denotado por $G_1 =_A G_2$, si existe una función biyectiva $f : V_1 \to V_2$ tal que:

1. $\forall \{u, v\} \in E_1, \{f(u), f(v)\} \in E_2$,
2. $S_f(G_1, G_2) = \prod_{u \in V_1} \frac{MV_{I_1(u), I_2(f(u))}}{MV_{I_1(u), I_1(u)}} * \prod_{e=\{u,v\} \in E_1} \frac{ME_{J_1(e), J_2(\{f(u), f(v)\})}}{ME_{J_1(e), J_1(e)}} \geq \tau$.

La función f es un isomorfismo aproximado entre G_1 y G_2, y $S_f(G_1, G_2)$ es el producto de las probabilidades normalizadas conocido como *grado del isomorfismo aproximado* de f. Cuando G_1 es isomorfo aproximado a un subgrafo de G_2, se dice que G_1 es *sub-isomorfo aproximado* a G_2.

Análogamente, el *grado del cotejo aproximado* entre dos grafos, denotado por $S_{max}(G_1, G_2)$, es el mayor de los grados de isomorfismos aproximados:

$$S_{max}(G_1, G_2) = max_f\{S_f(G_1, G_2)\} \tag{2}$$

Dada una colección D y un umbral de isomorfismo τ, el *soporte aproximado* de un grafo G, denotado por $supp(G, D)$, es el promedio de los grados de cotejos aproximados del grafo en la colección, donde G es isomorfo aproximado a un subgrafo de los grafos de la colección:

$$supp(G, D) = \sum_{G_i \in D} S_{max}(G, G_i)/|D| \tag{3}$$

Cuando $supp(G, D) \geq \delta$, entonces el grafo G es frecuente en la colección D, siendo G un *SFA* en D, con δ como umbral de soporte. Nótese que el valor del producto normalizado de probabilidades $S_f(G_1, G_2)$ está en el intervalo $(0, 1]$. El valor del umbral de soporte δ está en $[0, 1]$ asumiendo que $S_{max}(G, G_i)$ es normalizado. La tarea de *MSFA* usada en este trabajo consiste en encontrar todos los subgrafos conexos frecuentes aproximados en una colección de grafos D, utilizando (3), δ como umbral de soporte y τ como umbral de isomorfismo.

3 Resultados alcanzados utilizando la MSFA

En esta sección se exponen los aportes y resultados obtenidos haciendo uso de la MSFA, en específico el enfoque de VEAM [3]. Primero se describen las bases de datos utilizadas para probar la utilidad de la MSFA en tareas de clasificación. Luego se muestran los resultados alcanzados en la clasificación de imágenes sobre varias bases de datos reales y sintéticas de imágenes representadas en forma de grafos. Además, se mencionan varias podas para la MSFA que permiten acelerar el proceso de la minería, las cuales fueron aplicadas en VEAM.

3.1 Colecciones de grafos utilizadas

Se utilizan varias bases de datos bien conocidas para probar el enfoque de VEAM como son: el conjunto de imágenes reales COIL-100 [18] y ETH-80 [14], los cuales contienen imágenes de objetos reales tomados desde diferentes puntos de vistas. Otra colección usada es la "test" de la colección de imágenes, conocida como GREC [21], la cual representa símbolos de los planos arquitectónicos o electrónicos. La última colección de imágenes sobre la que se trabaja está compuesta por 700 imágenes obtenidas mediante el Generador aleatorio de imágenes de Coenen[1], la cual fue dividida en seis sub-colecciones con diferentes cantidades de imágenes (desde 200 hasta 700 con un incremento de 100).

En todas estas colecciones cada imagen se representa como un grafo no dirigido y etiquetado. Las características específicas de cada colección se muestran en la tabla 1. Nótese que en el caso de COIL se utilizan 25 objetos aleatorios de 100 que posee la colección y en ETH-80 se usan 6 categorías de las 8 existentes.

[1] www.csc.liv.ac.uk/∼frans/KDD/Software/ImageGenerator/imageGenerator.html

Table 1. Colecciones de imágenes utilizadas.

Colección	COIL-100	ETH-80	GREC (test)	Coenen images
Cantidad grafos	1800	2460	528	200–700
Cantidad etiquetas de vértices	150	150	4	18
Cantidad etiquetas de aristas	27	27	30	24
Tamaño promedio de los grafos	80	69	12	43–47
Cantidad clases	25	6	22	2

3.2 Clasificación de imágenes

Muchas han sido las técnicas desarrolladas para representar imágenes en forma de grafos [15, 21]. Estos tipos de representaciones han sido de gran ayuda para el procesamiento de imágenes ya que los grafos pueden describir la información estructural y topológica de las imágenes. La idea principal de esta representación es que las regiones de la imagen con propiedades similares se denotan mediante los vértices y las relaciones entre las diferentes regiones se denotan mediante las aristas del grafo. Los atributos de los vértices y aristas usualmente describen las características de las regiones y sus relaciones respectivamente. Mediante el modelado de las imágenes en forma de grafos, la tarea de la clasificación de imágenes se convierte en una de clasificación de grafos.

Con la representación en forma de grafos de un conjunto de imágenes pre-etiquetadas se inicia el proceso de clasificación de imágenes. El segundo paso es la confección de las matrices de sustitución utilizadas por el modelo aproximado. Luego se identifican los SFA haciendo uso del algoritmo para la MSFA (VEAM), los cuales son detectados según la definición de dos argumentos que reducen el espacio de búsqueda en el proceso de minería: (1) umbral de soporte ($0 < \delta \leq 1$), (2) umbral de isomorfismo ($0 < \tau \leq 1$). A partir de estos patrones se confeccionan los vectores de características necesarios para los algoritmos de clasificación. De esta manera queda descrito el proceso general de clasificación de imágenes mostrado en la figura 1.

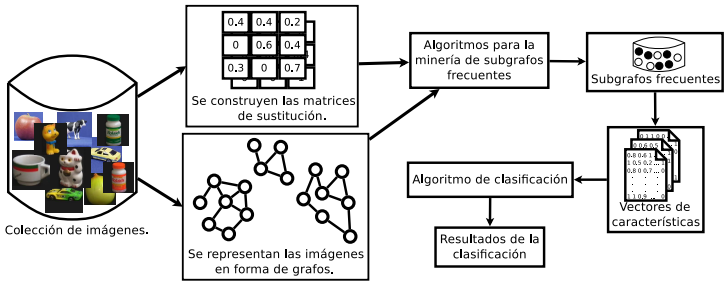

Fig. 1. Esquema de clasificación de imágenes basadas en grafos.

El esquema de la figura 1 fue utilizado por diferentes trabajos reportados [2, 3, 5], donde difiere la representación en forma de grafos de las imágenes. Por otro

lado, solamente uno de estos trabajos construye de forma automática las matrices de sustitución utilizando técnicas de agrupamiento sobre las regiones de las imágenes con características similares [5]. De este modo se le dio respuesta a la interrogante sin solución de los trabajos anteriores: ¿Cómo se obtienen las matrices de sustitución en caso de la ausencia de un especialista?, ya que las matrices tenían que ser construidas por un especialista en el conjunto de datos. En estos trabajos se utilizan dos funciones núcleo: Radial Basis Function (RBF) kernel y linear kernel; en el algoritmo de clasificación SVM (de sus siglas en inglés, Support Vector Machine) y las precisiones de la clasificación son probadas utilizando la validación cruzada con una ventana de tamaño 10 (10 cross-validation).

Los resultados obtenidos, en los trabajos antes mencionados, en tareas de clasificación de imágenes se resumen en la tabla 2, donde se muestran los mejores resultados alcanzados en cada colección de imágenes utilizada. Estos resultados han mostrado la utilidad de los patrones que se identifican mediante la MSFA, específicamente los patrones detectados con VEAM. Dichos resultados han sido comparados con los obtenidos al utilizar algoritmos convencionales que no se basan en el cotejo aproximado entre grafos demostrando gran mejoría en la mayoría de los casos. Además, se comparan con las precisiones alcanzadas utilizando APGM, el cual tiene un enfoque cercano al de VEAM.

La primera columna de la tabla 2 indican la colección usada. Desde la segunda hasta la quinta columna se muestran las precisiones obtenidas utilizando la función núcleo RBF y de la sexta en adelante se muestran las precisiones obtenidas utilizando la función núcleo lineal, donde la segunda y la sexta columna muestran el valor del umbral de soporte utilizado en la MSFA por los diferentes algoritmos especificados.

Table 2. Mejores precisiones alcanzadas utilizando SVM en diferentes conjuntos de grafos (imágenes).

Colección	Utilizando Linear kernel				Utilizando RBF kernel			
	δ	**Exactos**	**APGM**	**VEAM**	δ	**Exactos**	**APGM**	**VEAM**
Coenen–700	25%	90.58%	90.29%	**95.56**%	20%	95.29%	93.71%	**95.86**%
Coenen–600	20%	90.67%	90.50%	**95.67**%	20%	94.83%	93.00%	**95.83**%
Coenen–500	25%	91.20%	91.20%	**95.40**%	25%	92.80%	94.00%	**97.20**%
Coenen–400	30%	92.00%	91.25%	**96.75**%	25%	92.75%	93.75%	**97.75**%
Coenen–300	25%	91.00%	91.00%	**95.33**%	20%	94.33%	94.33%	**97.33**%
Coenen–200	25%	91.50%	91.00%	**93.50**%	20%	93.00%	92.00%	**97.50**%
GREC	3%	77.27%		**84.66**%	2%	86.55%		**92.80**%
COIL	40%	-	68.11%	**69.91**%	40%	-	91.39%	**92.18**%
ETH	30%	29.92%	**76.34**%	**76.34**%	30%	28.70%	**82.03**%	**82.03**%

Los resultados reportados sobre las colecciones sintéticas de imágenes obtenidas mediante el Generador aleatorio de imágenes (ver la primera fila de la tabla 2) fueron los primeros pasos y la motivación de la aplicación de la MSFA [3] en tareas de clasificación de imágenes. Como se puede observar en esa tabla, la diferencia entre las precisiones alcanzadas son bastante ilustrativas. Estas difer-

encias, que benefician a VEAM, muestran que tener en cuenta las distorsiones en las aristas juega un papel importante en la clasificación de imágenes, donde los objetos pueden tener variaciones espaciales en imágenes de una misma clase. En dicho trabajo se utiliza $\tau = 0.4$ como valor del umbral de isomorfismo con diferentes valores para el umbral de soporte.

De manera similar se muestran los resultados reportados sobre la colección de imágenes reales GREC [2], específicamente la sub-colección test (ver la segunda fila de la tabla 2). En esta ocasión el algoritmo APGM identifica los mismos patrones que los algoritmos exactos debido a que la colección GREC no contiene semejanzas semánticas en los vértices. Por esta razón, solo se tratan las aproximaciones en las aristas y APGM no utiliza este tipo de variaciones a diferencia de VEAM. En dicho trabajo se utiliza $\tau = 0.2$ como valor del umbral de isomorfismo con diferentes valores para el umbral de soporte. Por otro lado, los resultados de los promedios de reconocimiento reportados en la Segunda edición de reconocimiento de símbolos [6] no sobrepasan el 83.33% sobre esta colección utilizando otro esquema de clasificación y sin el uso de la minería. Sin embargo, los resultados en GREC alcanzan el 92.80% utilizando la minería, específicamente la MSFA, en el proceso de la clasificación.

Los resultados más recientes reportados con el uso de VEAM son alcanzados sobre las colecciones COIL-100 y ETH-80 [5]. En ese trabajo se utiliza $\tau = 0.4$ como valor del umbral de isomorfismo con diferentes valores para el umbral de soporte. Lo primero a notar en estos resultados (ver tercera y cuarta fila de la tabla 2) es que los algoritmos para la MSFA alcanzan mayores precisiones que los algoritmos exactos y estos últimos no obtienen clasificación alguna en el caso de la colección COIL-100. Respecto a los métodos aproximados, en el caso de la colección COIL-100, se puede observar que VEAM obtiene precisiones mejores que APGM, lo cual indica que el uso de las variaciones en las aristas provee información adicional para la clasificación. En la colección ETH-80, las distorsiones en las aristas no aportan información relevante alguna. Por otro lado, se compara el esquema utilizado con otros métodos de clasificación que no utilizan técnicas de minería. En COIL-100, con el método propuesto por Morales-González y García-Reyes [15] se obtiene una precisión de 91.60% mientras que con el método que usa VEAM se alcanza 92.18%. Para el caso de ETH-80, con VEAM se obtiene 82.03%, resultado comparable con otros métodos del estado del arte reportados en una comparación realizada [15], donde el rango de resultados es desde 79.00% hasta 88.00%.

En resumen, estos resultados muestran que los patrones identificados teniendo en cuenta variaciones semánticas en los vértices y aristas, o sea, utilizando VEAM, son mejores para estas tareas de clasificación que los encontrados por APGM y los algoritmos exactos. Por lo que VEAM obtiene las mejores precisiones en la mayoría de los casos.

3.3 Mejorando en eficiencia

El hecho de encontrar subgrafos frecuentes teniendo en cuenta semejanzas entre las subestructuras tiene la limitante del aumento de la complejidad computa-

cional respecto a los métodos exactos y VEAM no es la excepción. La aparición de candidatos duplicados durante el proceso de minería es uno de los mayores problemas en la mayoría de los enfoques recientes. Un candidato duplicado es un subgrafo que fue considerado en pasos previos, pero aparece nuevamente a partir de varios subgrafos frecuentes durante la búsqueda. El problema de los duplicados se trata representando el subgrafo con un código único conocido como forma canónica (FC) y realizando pruebas de FC, sin embargo, estas pruebas de FC tienen una gran complejidad computacional.

Se han desarrollado varias podas para la MSFA con el objetivo de alcanzar mayor eficiencia en el proceso de minería obteniendo los mismos patrones para mantener la eficacia [1, 4]. Mediante estas podas se logra reducir el espacio de búsqueda de las etiquetas y la cantidad de pruebas de FC en algoritmos que se basan en la propiedad de clausura-descendente, como es el caso de los algoritmos APGM y VEAM. El procesamiento de la información existente en las matrices de sustitución juega un papel importante en la confección de dichas podas.

Table 3. Comparación entre VEAM and VEAMwP en colecciones de imágenes utilizando $\tau = 0.4$.

(a) Cantidad de pruebas de CF realizadas

		Soporte (δ)								
Colección	**Algorithm**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**
Coenen-700	VEAM	350589	114907	33423	11105	5212	2600	2302	2118	1687
	VEAMwP	**259034**	**67405**	**17264**	**5505**	**2522**	**477**	**453**	**428**	**371**
Coenen-400	VEAM	114093	51273	17110	6435	3766	2339	2073	1901	1509
	VEAMwP	**86865**	**32583**	**9542**	**3400**	**1551**	**453**	**432**	**409**	**358**

(b) Tiempo de ejecución (s)

Coenen-700	VEAM	133.78	25.50	6.51	2.43	1.41	0.94	0.88	0.81	0.67
	VEAMwP	**114.93**	**20.34**	**5.53**	**2.15**	**1.24**	**0.42**	**0.40**	**0.39**	**0.34**
Coenen-400	VEAM	20.84	5.37	1.82	0.79	0.52	0.40	0.38	0.35	0.29
	VEAMwP	**13.24**	**4.34**	**1.48**	**0.64**	**0.31**	**0.16**	**0.16**	**0.15**	**0.13**

En la tabla 3 se muestran algunos de los resultados alcanzados al introducir las podas propuestas [1, 4] para el proceso de MSFA de VEAM. Primero se compara el algoritmo original de VEAM con él mismo haciendo uso de dichas podas, denotado por VEAMwP, según la cantidad exhaustiva de pruebas de FC que estos realizan en el proceso de minería (ver la subtabla (a) de la tabla 3). Nótese que en esta comparación la cantidad de estas excesivas pruebas de FC, en la mayoría de los casos, se reduce en un 30% cuando son utilizadas las podas antes mencionadas. Por otro lado, el comportamiento de VEAM con y sin las podas mencionadas anteriormente son comparadas en términos de tiempo de ejecución (ver subtabla (b) de la tabla 3). En esta segunda comparación se muestra que con el uso de las podas se logra reducir en 15% los tiempos de ejecución. Nótese que estas comparaciones se realizan utilizando la colección sintética de imágenes obtenidas mediante el Generador aleatorio de imágenes de Coenen.

El hecho de que estas podas reduzcan la cantidad de candidatos a ser procesados y la reducción del espacio de búsqueda de las etiquetas repercuten posi-

tivamente en el comportamiento de VEAM. Estos resultados permiten afirmar que estas podas son útiles para los procesos de MSFA similares al de VEAM.

4 Aplicaciones de la MSFA

La MSFA puede ser aplicada en cualquier dominio de la ciencia donde se puedan representar los objetos en forma de grafos y exista la suficiente cantidad de datos para realizar la búsqueda de conocimiento. Sin embargo, como el modelo en el que se enfoca este trabajo trata las aproximaciones semánticas sobre el conjunto de etiquetas, solo se hace efectiva la detección de las semejanzas donde las variaciones de las etiquetas juegan un papel importante. Por lo que, la utilización de este tipo de minería es más efectiva en aquellos tipos de datos que pueden ser modelados semánticamente a través de Marcos o Mapas conceptuales [17], Ontologías [7], Redes semánticas [13], entre otros, donde las variaciones en los vértices y aristas contengan un valor semántico interesante para usuarios y sistemas. De esta forma, esos datos se pueden someter a procesos de MSFA para la identificación de nuevos conceptos aproximados o semejantes que tributen a determinadas aplicaciones concretas.

Existen numerosas aplicaciones basadas en los tipos de representaciones antes mencionadas que han utilizado procesos de minería; por ejemplo: clasificación de documentos [10]; análisis de comunidades Web, extracción automática de tópicos desde documentos Web y clasificación de páginas Web mediante su estructura [23]. Sin embargo, no tienen en cuenta las aproximaciones semánticas en la minería. Varias de las aplicaciones basadas en la representación de los datos mediante Marcos conceptuales u Ontologías, en las que se puede utilizar la MSFA, pudieran ser: clasificación de documentos, agrupamiento de documentos por contenido e identificación de tópicos [20], detección de correos sospechosos y clasificación de tópicos [19]. Otras posibles aplicaciones de la MSFA, que más se ajustan a los problemas prácticos y que pudieran utilizar las Redes semánticas como forma de representación, son la detección de tendencias, comportamientos y regularidades sobre Redes Sociales [16]. Estas redes sociales también pueden ser representadas mediante una ontologías de tipo DRO [12], la cual representa la naturaleza semántica que subyace en los datos que se representan en ella ofreciendo un mayor nivel de expresividad respecto a los demás tipos existentes.

Por otro lado, las estructuras o modelos de las bases de datos no son consideradas como una posible aplicación de la MSFA. Estos modelos están compuestos por entidades o tablas (vértices) y las relaciones entre ellas (aristas), lo que se conoce como modelos relacionales. Las relaciones entre las entidades solo son de tres tipos: relación de mucho a mucho, relación de uno a mucho y relación de uno a uno. Como se puede deducir, estas relaciones no poseen información semántica interesante para el proceso de minería aproximado. Por esta razón, hemos llegado a la conclusión de que en esta área no es posible una aplicación satisfactoria de la MSFA tratada en este trabajo. Esto se debe a que las relaciones topológicas no aportan información de valor en el proceso de la minería

y el tipo de minería que se analiza en este trabajo incluye las aproximaciones entre las etiquetas de las aristas.

5 Conclusiones y trabajo futuro

En este trabajo se resumen los resultados alcanzados por la utilización de la MSFA, específicamente los modelos que permiten distorsiones en la semántica de las etiquetas manteniendo la topología de los grafos, en tareas de clasificación de imágenes. Estos resultados muestran la importancia que tiene el tratar las variaciones semánticas en el proceso de minería, las cuales se pueden obtener de forma automática si la ayuda de ningún especialista. También, se puede observar la utilidad del uso de la MSFA en tareas de clasificación de imágenes.

Por otro lado, se muestra también cómo se ha logrado ganar en eficiencia utilizando varias podas en el proceso de minería, las cuales permiten una reducción del espacio de búsqueda de las etiquetas y una disminución de la cantidad de pruebas de FC a realizar. De esta manera se ataca el problema de la eficiencia de la MSFA, el cual es uno de los que más afecta a este tipo de minería. Luego, se presentan algunas de las posibles aplicaciones futuras de este tipo de minería donde se pudieran obtener resultados relevantes.

Como trabajo futuro, se desarrollarán extensiones del algoritmo VEAM para que detecte los subgrafos frecuentes en un gran grafo etiquetado. Esto permitirá la aplicación de la MSFA en dominios donde los datos no sean transicionales como: la mayoría de las aplicaciones sobre redes sociales, análisis de vínculos, detección de tópicos en un solo documento, entre otros.

References

1. N. Acosta-Mendoza, A. Gago-Alonso, J.E. Medina-Pagola. Mejora para la minería de subgrafos frecuentes aproximados mediante la reducción del espacio de búsqueda. in: Memorias del IX Congreso Nacional de Reconocimiento de Patrones, Santa Clara, Villa Clara, Cuba, 2011.
2. N. Acosta-Mendoza, A. Gago-Alonso, J.E. Medina-Pagola. Clasificación de imágenes utilizando minería de subgrafos frecuentes aproximados. Revista Cubana de Ciencias Informáticas, 5, 4(2011), 2012.
3. N. Acosta-Mendoza, A. Gago-Alonso, J.E. Medina-Pagola. Frequent Approximate Subgraphs as Features for Graph-Based Image Classification. Knowledge-Based Systems, 27:381–392, 2012.
4. N. Acosta-Mendoza, A. Gago-Alonso, J.E. Medina-Pagola. On Speeding up Frequent Approximate Subgraph Mining. in: Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12), Springer-Verlag Berlin Heidelberg, Buenos Aires, Argentina, 2012.
5. N. Acosta-Mendoza, A. Morales-González, A. Gago-Alonso, E. B. García-Reyes, J. E. Medina-Pagola. Classification using Frequent Approximate Subgraphs. in: Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12), Springer-Verlag Berlin Heidelberg, Buenos Aires, Argentina, 2012.

6. P.H. Dosch, E. Valveny. Report on the Second Symbol Recognition Contest. Graphics Recognition. Ten years review and future perspectives. Proc. 6th Int. Workshop on Graphics Recognition (GREC'05), 381–397, Springer, 2005.
7. T.R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, Academic Press Ltd., London, UK, UK, 1993.
8. L.B. Holder, D.J. Cook, H. Bunke. Fuzzy substructure discovery. in: Proceedings of the 9th International Workshop on Machine Learning, 218–223, San Francisco, CA, USA, 1992.
9. Y. Jia, J. Zhang, J. Huan. An Efficient Graph-Mining Method for Complicated and Noisy Data with Real-World Applications. Knowledge Information Systems, 28(2):423–447, 2011.
10. C. Jiang, F. Coenen, R. Sanderson, M. Zito. Text Classification using Graph Mining-based Feature Extraction. Knowledge-Based Systems, 23(4):302–308, 2010.
11. C. Jiang, F. Coenen, M. Zito. A Survey of Frequent Subgraph Mining Algorithms. To appear: *Knowledge Engineering Review*, 2012.
12. R. Larin-Fonseca, E. Garea-Llano. Automatic Representation of Geographical Data from a Semantic Point of View through a New Ontology and Classification Techniques. Transaction in GIS, 15(1):61-85, Blackwell Publishing Ltd, 2011.
13. F. Lehmann. Semantic Networks in Artificial Intelligence. Elsevier Science Inc., New York, NY, USA, 1992.
14. B. Leibe, B. Schiele. Analyzing Appearance and Contour Based Methods for Object Categorization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), 409-415, 2003.
15. A. Morales-González, E. B. García-Reyes. Simple object recognition based on spatial relations and visual features represented using irregular pyramids. Multimedia Tools and Applications, 1–23, Springer Netherlands, 1380-7501, 2011.
16. N. Memon, J.J. Xu, D.L. Hicks, H. Chen. Data Mining for Social Network Data. Annals of Information Systems (12), Springer, 2010.
17. M. Minsky. A Framework for Representing Knowledge. Mind Design: Philosophy, Psychology, Artificial Intelligence, 95–128, Cambridge, MA, 1981.
18. S. Nene, S. Nayar, H. Murase. Columbia Object Image Library (COIL-100). Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR & SPR'08, 2008.
19. S. Nizamani, N. Memon, U.K. Wiil, P. Karampelas. CCM: A Text Classification Model by Clustering. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'11), 461-467, Kaohsiung, Taiwan, 2011.
20. A. Pérez-Suárez, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, J.E. Medina-Pagola. A Dynamic Clustering Algorithm for Building Overlapped Clusters. To appear: Journal Intelligent Data Analysis, 16(2), 2012.
21. K. Riesen, H. Bunke. IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning. Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR & SPR'08, 208–297, Orlando, USA, 2008.
22. Y. Song, S. Chen. Item Sets Based Graph Mining Algorithm and Application in Genetic Regulatory Networks. in: Proceedings of the IEEE International Conference on Granular Computing, 337–340, Atlanta, GA, USA, 2006.
23. G. Xu, Y. Zhang, L. Li. Web Mining and Social Networking: Techniques and Applications. Web Information Systems Engineering and Internet Technologies Book Series, Springer, 2010.