Agrupamiento local en grafos dirigidos

VANESA ÁVALOS GAYTÁN*, MARIO RIVERA RAMÍREZ**, ELISA SCHAEFFER*

El agrupamiento de datos es un campo de investigación con numerosas aplicaciones.[1] El objetivo de este trabajo es identificar grupos de elementos estrechamente relacionados en un gran conjunto de datos. Muchos conjuntos de datos tienen una representación natural en forma de *grafos*.

Un grafo $G = (V, E)$ está compuesto por un conjunto de vértices V y un conjunto de aristas E, donde una arista es un *par* de vértices $\{u, v\}$ en V y representa alguna conexión como similitud entre los vértices. El número de vértices es n y el número de aristas es m. Si las aristas son conexiones simétricas, se escribe $(u, v) = (v, u)$. En casos donde la conexión es de un solo sentido, escribimos $[u, v]$ para un vértice de u a v, en tal caso el grafo es dirigido. Se dice que v es *vecino* de u, si están conectados por una arista. El grado de un vértice es el número de aristas que tiene, en grafos dirigidos; nos interesamos por el grado de salida, que es el número de aristas que parten del vértice.

Actualmente existen trabajos acerca de agrupamiento de datos mediante grafos no dirigidos,[2] la mayoría usa técnicas espectrales de grafos.[3,4] En años recientes se han desarrollado métodos de agrupamiento para grafos dirigidos,[5] frecuentemente basados en métodos para grafos no dirigidos. Los métodos mencionados agrupan globalmente los datos de entrada.

En este trabajo se propone un método de agrupamiento local. Con el método propuesto se pueden obtener agrupamientos de alta calidad; no sólo para grafos no dirigidos, sino para grafos dirigidos de gran tamaño.

Agrupamiento global y local

La gran mayoría de los métodos existentes son para agrupamiento global de los datos. Agrupamiento que consiste en identificar, dentro del grafo que representa tales datos, una partición de los datos a subgrupos de tal manera que los datos del mismo grupo estén estrechamente relacionados mediante alguna medida de similitud definida. Este tipo de agrupamiento es muy costoso computacionalmente, cuando se trabaja con grafos de tamaño masivo.

No siempre es necesario conocer el agrupamiento para todos los datos. En algunas aplicaciones y situaciones se tiene interés en conocer sólo el grupo al cual pertenece un vértice de interés (llamado la *semilla).* Por esta razón surge la

* Facultad de Ingeniería Mecánica y Eléctrica, UANL.
** Universidad de Guadalajara.

necesidad de métodos de agrupamiento local: determinar el grupo al cual pertenece el nodo semilla, preferiblemente por computación local, de tal manera que en un caso típico no es necesario procesar el grafo de entrada en su totalidad.

Caminatas aleatorias

Una caminata aleatoria (ciega) es un proceso en el que se realiza un desplazamiento aleatorio de un vértice a su vecino. Dado un vértice semilla s_0, se selecciona un nodo vecino s_1 de s_0, uniformemente al azar y se "mueve" a s_1; luego se mueve uniformemente al azar a un vecino s_2 de s_1, etcétera. Se repiten los movimientos sucesivamente hasta completar un determinado número de pasos. La secuencia de los vértices visitados se llama camino.

El proceso de caminata aleatoria es una cadena de Markov finita. En un grafo no dirigido *conexo,* donde es posible llegar de un vértice semilla a los demás vértices en un tiempo finito, la cadena de Markov correspondiente es ergódica. Esto implica que llegará eventualmente a un estado de equilibrio donde la probabilidad de encontrarse en un cierto vértice ya no cambia. En la caminata ciega, esta probabilidad es directamente proporcional al grado del vértice.

Cada cadena de Markov puede ser vista como una caminata aleatoria, aunque no necesariamente ciega, sobre un grafo dirigido, si admitimos aristas ponderadas: el vértice siguiente se elige con probabilidad que dependen del peso de la arista al vecino. La teoría clásica de caminatas aleatorias se refiere a caminatas aleatorias sobre grafos simples pero infinitos, al igual que en las grandes redes y los estudios cualitativos de su comportamiento.

Recientemente, las caminatas aleatorias sobre grafos finitos han recibido mucha atención en términos cuantitativos: el número de pasos hasta el primer retorno a la semilla, el tiempo de primera llegada a un cierto vértice de interés, el tiempo de visitar a cada vértice por lo menos una vez y el tiempo necesario para llegar al equilibrio (cuando existe). Como resultado, la teoría de caminatas aleatorias está estrechamente relacionada con la teoría de grafos. Este tipo de propiedades básicas de caminatas aleatorias se determina en gran parte por el *espectro* del grafo.[6]

Existe una serie de procesos que pueden modelarse sobre grafos, la mayoría describe algún tipo de difusión (propagación, epidemias, balanceo de carga en redes distribuidas, etcétera), cuyos parámetros básicos están estrechamente relacionados con los parámetros de caminatas aleatorias anteriormente mencionadas. Todas estas conexiones son muy fructíferas y proveen ambas herramientas, para el estudio y oportunidades para aplicaciones de caminatas aleatorias. Las caminatas aleatorias pueden ser usadas para la búsqueda de partes "oscuras" en un gran conjunto de datos, y también para generar elementos aleatorios en grandes y complicados conjuntos.[7]

En la caminata ciega, la probabilidad de transición de vértice v_t a un vértice vecino v_{t+1} es uniforme entre los vecinos, es decir, si denotamos por $d(v_t)$ el grado de v_t, la probabilidad de transición es:

$$1 / d(v_t) \qquad (1)$$

La semilla v_0 de la caminata puede ser fijada o aleatoria: denotamos su distribución por $\eth_0$ y la distribución de v_t por π_t. Denotamos por $P = p_{ij}$ para todo i, j en V la matriz de probabilidades de transición (1). Así:

$$p_{ij} = 1 / d(v_t) \text{ si } [i, j] \hat{\text{I}} E. \qquad (2)$$

y cero en otro caso. Sea A la matriz de adyacencia de G, donde a_{ij} es uno, si $[i, j]$ es una arista en G y cero en otro caso. Sea D una matriz diagonal con elementos diagonales definidos por $d_{ij} = d(v_i)$. Entonces, $P = D^{-1}A$. La regla de la caminata puede expresarse en una simple ecuación: multiplicando P, que es una matriz nxn por la iz-

quierda por el vector de la distribución δ_t que tiene dimensión 1x*n* para obtener la distribución δ_{t+1} del paso siguiente:

$$\pi_{t+1} = \pi_{t+1} P \quad (3)$$

que da por recursión la ecuación

$$\pi_t = \pi_o P^t \quad (4)$$

para la distribución a paso *t*, dada la distribución inicial π_0 y la matriz de transiciones *P*.

Tiempos de absorción

La matriz de transiciones *P* es estocástica: sus elementos son probabilidades y las sumas de sus renglones son todos iguales a uno:

$0 \leq p_{ij} \leq 1$ y $\leq p_{ij} = 1$ donde $1 \leq i, j \leq n$. (5)

Una cadena de Markov es *irreducible*, si es posible llegar desde un vértice *v* a cualquier otro vértice. En un grafo significa que debe existir una sucesión de aristas entre vecinos que comienza en *v* y llega a cada otro vértice *u* en *V*. En un grafo dirigido es importante respetar la dirección de la arista.

En las cadenas de Markov se dice que un vértice *v* es absorbente, si la probabilidad de salir de éste es cero, o sea, $p_{vv} = 1$. En un grafo dirigido sería un vértice sin aristas de salida. Otros vértices son transitorios. Supongamos ahora que *P* está ordenada de tal manera que primero vienen los elementos que corresponden a transiciones entre vértices transitorios, esta submatriz será llamada *Q*, y después vienen los elementos que son transiciones a estados absorbentes:

$$\begin{pmatrix} Q & R \\ I & 0 \end{pmatrix} \quad (6)$$

donde 0 es una submatriz cuyos elementos son todos cero, e *I* es la matriz de identidad donde los elementos diagonales i_{vv} son uno, y todos los otros elementos son cero.

Los valores propios λ_i de una matriz *n*x*n* son los *n* valores para los cuales existe un vector w_i tal que

$$Pw = \lambda_i w_i \quad (7)$$

y para una matriz estocástica el valor propio de mayor valor absoluto es siempre igual a uno. La matriz *Q* es subestocástica, es decir:

$$0 \leq p_{ij} \leq 1 \text{ y} \leq p_{ij} \leq 1 \text{ donde } 1 \leq i, j \leq n. \quad (8)$$

Esto causa que todos los valores propios de *Q* tengan valor absoluto menor o igual que uno. De ahí $I - Q$ es una matriz invertible y se puede definir la matriz:

$$M = (I - Q)^{-1} \quad (9)$$

El tiempo de absorción de un vértice *v* a la semilla *s* es el número esperado de pasos que una caminata iniciada en *v* toma antes de llegar a *s*. Los tiempos de absorción, denotados aquí por m_v, pueden obtenerse de la matriz *M*, como sumas de los renglones.

Por intuición, es posible esperar que los vértices que pertenecen al grupo de *s* tengan valores menores a m_v que los vértices que son estructuralmente lejanos de *s* en el grafo de entrada *G*, lo que se puede comprobar para grafos no dirigidos por la conexión entre los tiempos de absorción y el agrupamiento espectral.[9] Desafortunadamente las matemáticas no se extienden de una manera natural a grafos dirigidos, debido a la asimetría de la matriz de adyacencia.

Solución propuesta

El cálculo de la matriz M es una operación global de complejidad peor que la cuadrática,[10] algo que no es factible para matrices de grandes dimensiones. Para capturar la esencia matemática de la definición de agrupamiento por tiempo de ab-

sorción sin el alto costo computacional, proponemos la aproximación de los tiempos de absorción con base en caminatas aleatorias cortas sobre un grafo *G*.

Es importante que el número de pasos sea pequeño para que la caminata no alcance su equilibrio. En una cadena ergódica, en el equilibrio ya no es estadísticamente observable en cuál vértice inició la caminata. En agrupamiento local, el interés está en los vértices cercanos a un vértice semilla *s*, y no en las propiedades globales, por lo cual la utilidad de las frecuencias de visitas de las caminatas largas se pierde junto con la información del vértice de inicio.

La entrada del algoritmo es un listado de las aristas del grafo, indica en la primera línea el número total de nodos y aristas, junto con el nodo de inicio. Se utilizan listas lineales, simplemente ligadas para almacenar el grafo en forma de listas ordenadas de adyacencia. Los grados de los vértices están preprocesados a un vector *d*.

Asimismo, se utilizan listas lineales simplemente ligadas lineales para almacenar las frecuencias de visitas a los vértices en *V* al repetir *r* caminatas de *k* pasos iniciadas en el vértice *s*. La frecuencia de visitas a un vértice *v* se guarda a partir de la primera visita en ello, por lo cual el tamaño del grafo de entrada no afecta directamente la memoria necesaria para almacenar las frecuencias.

Las caminatas avanzan de la manera siguiente: para cada paso de la caminata se incrementa la frecuencia de visita del vértice actual f_v. Después se obtiene d_v, se genera un número pseudoaleatorio *i* entre cero y d_v y toma el *i*ésimo vecino de *v* como el vértice siguiente, incrementando por uno el número de pasos realizados. Al llegar a *k* pasos, se interrumpe la caminata. Las frecuencias de las *r* repeticiones se acumula en la misma lista. Posteriormente, los vértices con alta frecuencia f_v son interpretados como los vértices cercanos a *s*. Nuestro argumento es que f_v sea inversamente proporcional a m_v: entre más tarda la caminata en llegar a *v*, mayor es su tiempo de absorción m_v y menor es la frecuencia de visitas f_v.

La salida del método es un vector de frecuencias, del cual se detecta el grupo de *s* por seleccionar los vértices con valores altos y eliminar los de valores bajos. Para este propósito se puede utilizar cualquier algoritmo unidimensional de 2-clasificación.[1]

Experimentos

Para realizar la comparación de la relación supuesta entre los tiempos de absorción y las frecuencias de visita, se realizaron experimentos computacionales con un grafo no dirigido y un grafo dirigido. Los grafos son pequeños para poder calcular de forma exacta la matriz *M* para cada vértice semilla. El primer grafo es un caso fácil para el agrupamiento, con una estructura conocida como "hombre de cueva",[8] compuesto por 30 vértices y 120 aristas no dirigidas (ver figura 1).

Para cada vértice calculamos su tiempo de absorción para determinar el grupo al cual pertenece cada uno de los nodos; en este caso; la figura 2 muestra que existen seis grupos, los cuales corresponden a los que forman el grafo de la figura 1.

Para determinar por medio de caminatas aleatorias los grupos, calculamos las frecuencias usando r = 1, 2, ..., 30 empleamos como *s* cada vértice a su turno. Repetimos el proceso 30 veces por cada ,*s* para observar la variabilidad en los resultados. Para determinar el número de pasos *k*, buscamos limitar el procesamiento local del

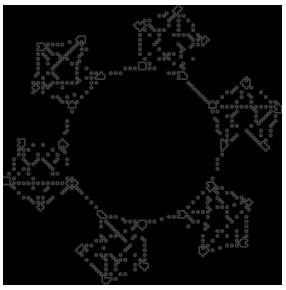

Fig. 1. Grafo no dirigido con estructura "hombre de cueva", con seis cuevas de cinco hombres.

grafo, por lo cual necesitamos que *k* sea menor que el diámetro del grafo que es la distancia máxima en ello: si *k* es igual al diámetro, cada caminata tiene una probabilidad no cero en visitas a cualquier vértice, por lo cual la computación resulta global en un cierto sentido.

En grafos donde no hay grupos naturales presentes, ningún valor de *k* dará una división clara entre el grupo de *s* y el resto del grafo (debido a la ausencia de grupos), pero en grafos que sí cuentan con grupos, el diámetro de un grupo es típicamente menor al diámetro del grafo. Nosotros queremos un valor *k* parecido al diámetro de grupo, lo que convierte a *k* al parámetro más importante del método. Con valores demasiado pequeños, casi todos los vértices visitados tendrán frecuencias altas, como las caminatas raramente salen del grupo, pero es posible que alguna parte del grupo quede sin detectar. Con valores demasiado altos, la caminata empieza a dar preferencia a vértices de alto grado, aunque no pertenezcan el grupo y serán mal clasificados como miembros del grupo.

La figura 2 muestra los tiempos de absorción desde un vértice a todos demás vértices en el grafo de la figura 1, junto con las frecuencias de visita con $k = 2$ y $r = 10$. Los tiempos de absorción revelan la estructura global del grafo con las seis cuevas (un menor valor significa que el vértice es más cercano, con el mínimo en el vértice semilla), mientras las frecuencias separan localmente el grupo del vértice semilla del resto del grafo (un mayor valor significa que el vértice es más cercano, con el máximo en el vértice semilla). Ambos vectores están normalizados al rango [0, 1].

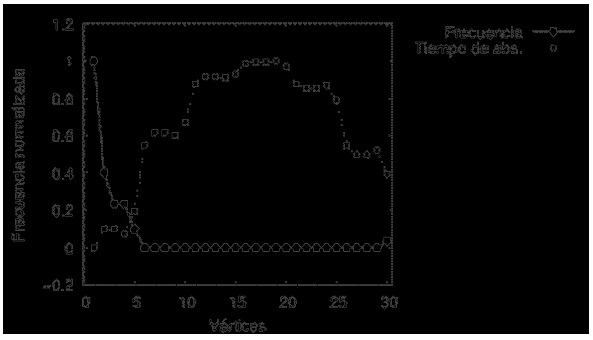

Fig. 2. Las frecuencias de visita f_v (%) en caminatas ciegas con $k = 2$ y $r = 10$ iniciadas en vértice 1 y los tiempos de absorción m_v (%) al vértice 1 en el grafo de la figura 1. Ambos vectores fueron normalizados al rango [0, 1], y los vértices están ordenados por cuevas, de tal manera que los vértice 1-5 forman la primera cueva, en grupo natural del vértice 1.

Como medida de relación entre f_v y m_v, utilizamos la *correlación* entre los tiempos de absorción y las frecuencias de visitas; ambos vectores están sin normalizar con sus valores originales. La correlación es una medida de dependencia lineal que toma valores en [-1, 1], donde -1 significa un valor que crece linealmente cuando otro baja. Para el grafo de la figura 1, mostramos en la figura 3 la correlación para los seis grupos.

El segundo es un grafo generalizado de hombre de cueva[11] de 30 vértices y 248 aristas, donde hay cuatro cuevas con densidad interna 0.95 (es decir, 95% de las aristas posibles entre los vértices de la cueva están presentes) y la densidad entre las cuevas es 0.08. Las aristas fueron orientadas al azar para obtener un grafo dirigido. Realizamos 30 repeticiones del método propuesto con $k = 2$ y $r = 1, 2, \ldots 30$. Los resultados se muestran en la figura 4. El número de repeticiones para obtener una buena correlación es baja (ya con $r = 3$ se obtiene una buena correlación), lo cual indica que el método es eficiente.

Conclusiones y trabajo a futuro

En este trabajo se presenta un método para el agrupamiento local de grafos dirigidos, motivado por agrupamiento espectral y su relación con los tiempos de absorción de caminatas aleatorias. El método es eficiente: basta con una cantidad baja de caminatas muy cortas para determinar el grupo de vértices al cual pertenece un vértice semilla dado.

De interés futuro es el desarrollo de un método que automáticamente ajuste los parámetros *k* y *r*, durante el tiempo de ejecución para eliminar la necesidad de definirlos por parte del usuario.

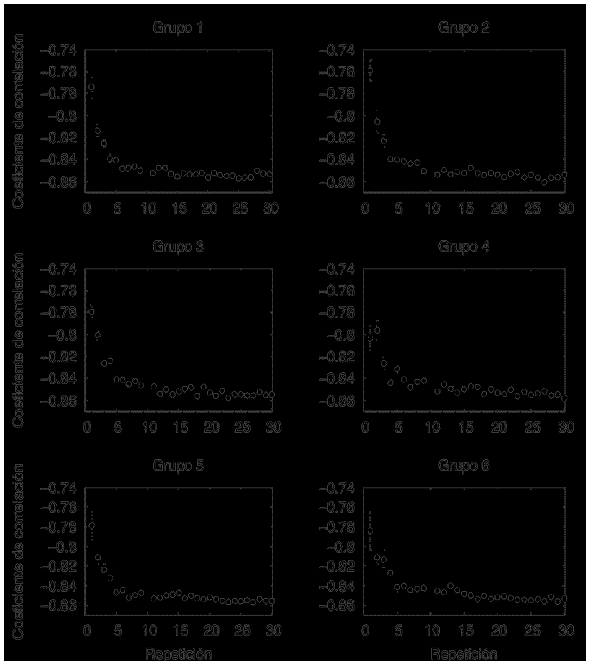

Fig. 3. Correlación entre los tiempos de absorción y las frecuencias de visitas en el grafo de hombre de cueva (figura 1) con parámetros k = 2 y r = 1, 2, ..., 30, promedio y desviación estándar sobre 30 repeticiones del método propuesto de caminatas aleatorias cortas repetidas. Cada gráfica muestra la correlación sobre el conjunto de vértices de una de las seis cuevas, utilizados cada uno como vértice semilla.

También es de interés combinar la salida del método con un algoritmo eficiente de 2-clasificación[1] para la división automática de los vértices visitados, a los que pertenecen al grupo y a los que no pertenecen a él.

Resumen

En este trabajo se presenta un método de agrupamiento local en grafos dirigidos: dado un vértice semilla, se determinó el grupo de vértices al que pertenece de tal forma que los vértices seleccionados sean estructuralmente cercanos de la semilla. A un grafo dirigido se le puede asociar una cadena de Markov que corresponde a una caminata aleatoria ciega en el grafo. Se aprovechó esta conexión para expresar cercanía estructural en términos de los tiempos de absorción para detectar vértices que son "cercanos" al vértice semilla. Se detectó el grupo de un vértice a través de caminatas aleatorias cortas repetidas desde el vértice semilla, analizando la frecuencia de visitas a los otros vértices. Se experimentó con grafos pequeños para comparar el resultado los tiempos exactos de absorción. El agrupamiento local puede ser aplicado a diferentes fenómenos reales, por ejemplo, en propagación de epidemias, balanceo de carga, etc.

Palabras clave: Agrupamiento de grafos, Cadena de Markov, Caminata aleatoria, Computación local, Tiempo de absorción.

Abstract

We propose a method for local clustering in directed graphs: we determine a group of «nearby» vertices to a seed vertex, such that the selected vertices are structurally close to the given vertex. For a directed graph, there is a Markov chain that corresponds to a blind random walk in the graph. We use this connection to express structural closeness in terms of absorption times to detect vertices that are «near» the seed vertex. We detect the cluster of a vertex by repeated short random walks starting at the seed and use the visit frequencies to determine which vertices are nearby. We experiment on small graphs and compare the results with globally computed absorption times. Local clustering can be applied to various real phenomena, such as epidemics propagation, load balancing, etc.

Keywords: Graph clustering, Markov chain, Random walk, Local computation, Absorption time.

Referencias

1. Anil Jain and M. Narasimha Murty y Patrick Flynn, *Data clustering: a review*. ACM Computing Surveys, 31(2): 264-323, 1999.

2. Satu Elisa Schaeffer, *Graph Clustering*, Computer Science Review, 1(1): 27-64, 2007.
3. Charles Fowlkes, Serge Belongie, Fan Chung y Jitendra Malik, Spectral Grouping Using the Nyström *Method*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2): 214-225, 2004.
4. Jinabo Shi y Jitendra Malik, *Normalized Cuts and image segmantation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-901, 2000.
5. Marina Meila y William Pentney, Clustering by weighted cuts in directed graphs, *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007.
6. Peter Doyle y J. Laurie Snell. *Random Walks and Electrical Networks*. Mathematical Association of America, 1984.
7. Lázsló Lovász. Random walks on graphs: A survey. *Bolyai Society Mathematical Studies: Comninatorics -Pál Erdös is Eighty*. Volume 2, pp. 353-397. Bolyai Mathematical Society, 1996.
8. Duncan J. Watts. *Small worlds*, Princeton University Press, 1999.
9. Pekka Orponen, Satu Elisa Schaeffer y Vanesa Ávalos Gaytán. Locally computable approximations for spectral clustering and absorption times of random walks. ILAS 2008, enviado a *Linear Algebra and Applications*.
10. Thomas Cormen, Charles Leiserson, Ronald Rivest y Clifford Stein. *Introduction to Algorithms*, McGraw-Hill, 2001.
11. Satu Elisa Virtanen. *Properties of Nonuniform Random Graph Models*. Informe de investigación HUT-TCS-A77, Helsinki University of Technology, 2003.

Recibido: 2 de abril de 2009
Aceptado: 18 de marzo de 2009