Agrupamiento mediante Q-análisis

Q-analysis clustering

Luis Eduardo Múnera Salazar, Ph.D.
Grupo i2T, Universidad Icesi, Colombia
lemunera@icesi.edu.co

Fecha de recepción: 15-02-2010 Fecha de selección: 30-04-2010 Fecha de aceptación: 12-04-2010

ABSTRACT

This article presents a novel clustering algorithm, based upon topology concepts and Q-analysis. The algorithm allows to create a structural clustering, which does not depend on the knowledge of similarity measures, distances, etc. The clustering is directed in a natural fashion by the internal topology of the network/graph. Such topology becomes evident by applying Q-analysis, by means of the structure vector, which shows the network's internal Q-connectivity.

KEYWORDS

Clustering, Q-analysis, Topology, Networks, Graphs

RESUMEN

En este trabajo se presenta un algoritmo de agrupamiento basado en conceptos topológicos y fundamentado en el q-análisis, el cual permite desarrollar un agrupamiento estructural independiente del conocimiento o no de medidas de similitud, distancias, etc. El agrupamiento está dirigido de manera natural por la topología interna de la red o grafo, la cual se hace explícita por medio del q-análisis mediante el vector de estructura, que muestra la q-conectividad interna de la red.

PALABRAS CLAVE

Agrupamiento, Q-análisis, Topología, Redes, Grafos

Clasificación Colciencias: Tipo 1

1. INTRODUCCIÓN

Existen muchas aproximaciones a las técnicas de agrupamiento (en inglés, clustering), en gran medida debido también a la diversidad de aplicaciones, que han incentivado la investigación y el desarrollo de propuestas. Por ejemplo, la detección de comunidades en redes sociales y biológicas, la investigación de la interacción de usuarios de internet, la obtención de módulos de software con alta cohesión y bajo acoplamiento, el agrupamiento jerárquico en sociología y otras disciplinas, indicadores de citación y colaboración en índices bibliométricos, etc.

Las diversas técnicas de agrupamiento las podemos clasificar en dos grandes categorías. La primera consiste en realizar los agrupamientos basándose en el conocimiento de unas valoraciones numéricas que miden el grado de afinidad entre los elementos a agrupar, y que pueden ser medidas de similitud, pesos, distancias euclidianas, distancias ultramétricas, etc.[1] Las técnicas de esta categoría pueden ser en general acumulativas o divisivas, y suelen tener aplicaciones en el análisis de redes sociales,[2] en el diseño y fragmentación de bases de datos distribuidas,[3] etc.

La segunda categoría consiste en obtener los agrupamientos inducidos por la partición de una estructura, usualmente una red o grafo que puede estar o no dirigido y sin necesidad de contar con información adicional como el peso de los enlaces entre los nodos de la red. En esta categoría cabe destacar las técnicas de partición espectral de redes o grafos,[4,5] las cuales hacen uso de la matriz de adyacencia o de la matriz laplaciana que representa el grafo o red y utilizan los vectores y valores propios de esas matrices para realizar una partición de la estructura que indirectamente induce un agrupamiento.

En particular, se suele utilizar el vector propio asociado al segundo valor propio más pequeño de la matriz de adyacencia o de la matriz laplaciana y que en la literatura se conocen con los nombres de vector y valor propio de Fiedler, respectivamente, en honor del pionero de estos trabajos a principio de los años setenta.[6,7] Las técnicas de esta categoría se suelen utilizar en diversas aplicaciones en ciencias físicas e ingeniería, para permitir el análisis de redes eléctricas, redes de comunicaciones, redes de osciladores, etc.[8,9,10,11]

Este trabajo propone una aproximación al tema del agrupamiento (algoritmo de agrupamiento), que en cierto sentido está emparentado con las dos categorías anteriores, pues por una parte se puede usar para realizar agrupamientos a partir del conocimiento de una matriz de valoraciones numéricas tipo similitudes, afinidades, pesos, distancias, etc. Y por otra parte también se puede usar para generar la partición de un grafo o red e inducir con ello un agrupamiento.

La propuesta se basa en el q-análisis (también conocido como análisis poliédrico) desarrollado por el matemático británico Ronald Atkin,[12,13] y que grosso modo consiste en calcular un invariante topológico de un complejo simplicial. Dicho invariante del complejo es un vector de números enteros, llamado vector de estructura, el cual da cuenta de la conectividad interna de los símplices del complejo, lo que

puede ser visto como un caso particular de análisis de agrupamiento del tipo vecino más cercano (en inglés, nearest-neighbor).

El artículo está organizado de la siguiente manera: La sección 2 está dedicada a los elementos matemáticos, y comprende básicamente dos cosas. Por una parte, las nociones fundamentales de topología algebraica que son necesarias, y por otra parte los conceptos básicos de q-análisis. La sección 3 se ocupa de presentar el algoritmo de agrupamiento y su aplicación tanto para casos donde se cuente con una matriz de valoraciones numéricas, como para casos donde se trate de una red o grafo. La sección 4 compara el algoritmo propuesto contra otros algoritmos de ambas categorías.

2. ELEMENTOS MATEMÁTICOS

Un complejo simplicial abstracto sobre un conjunto finito cuyos elementos se llaman vértices $V = \{ a_0,...,a_n \}$ es un subconjunto no vacío de partes de V (excluyendo el vacío) cuyos elementos son llamados símplices con las siguientes propiedades:

(P1) Si $\sigma \in K$ y $\tau < \sigma$ entonces $\tau \in K$. Decimos que σ y τ son símplices y que τ es una cara de σ.

(P2) Si σ y τ pertenecen a K, entonces $\sigma \cap \tau$ o bien es vacía, o bien es una cara común de σ y τ.

(P3) Si a_i pertenece a V entonces $\{ a_i \}$ pertenece a K.

La dimensión de un símplice es el número de sus vértices menos uno. La dimensión de K es el máximo de las dimensiones de todos sus símplices.

A cada complejo simplicial abstracto le podemos asociar un complejo simplicial geométrico y viceversa. El procedimiento es sencillamente asociar a cada símplice de dimensión n(n-símplice) del complejo abstracto un símplice geométrico que es una generalización a "n" dimensiones de espacios geométricos muy conocidos como segmentos, triángulos, tetraedros, etc.

De esta manera un 0-símplice es un punto, un 1-símplice es una línea de segmento, un 2-símplice es una región triangular, un 3-símplice es un tetraedro sólido, etc. En general podemos decir que si $\{a_0,....., a_n\}$ es un conjunto de puntos independientes en R^m, entonces el n- símplice geométrico σ_n generado por $\{a_0,.....a_n\}$ es el conjunto de todos los puntos de R^m, x, tales que,

$$x = \sum_{i=0} t_i a_i, \text{ donde } \sum_{i=0} t_i = 1 \text{ y } t_i > 0 , \forall_i$$

Los puntos $a_0, a_1,.....a_n$ son llamados los vértices de σ_n.

Los números t_i, son llamados las coordenadas baricéntricas del punto x de σ_n con respecto a $\{a_0,....., a_n\}$. El número n de σ_n es llamado la dimensión de σ_n. El subespacio de σ_n generado por un subconjunto de los vértices $\{a_0,....., a_n\}$ de σ_n, se denomina una cara de σ_n.

La realización geométrica de un complejo simplicial abstracto K, es un poliedro que denotaremos por $< K >$.

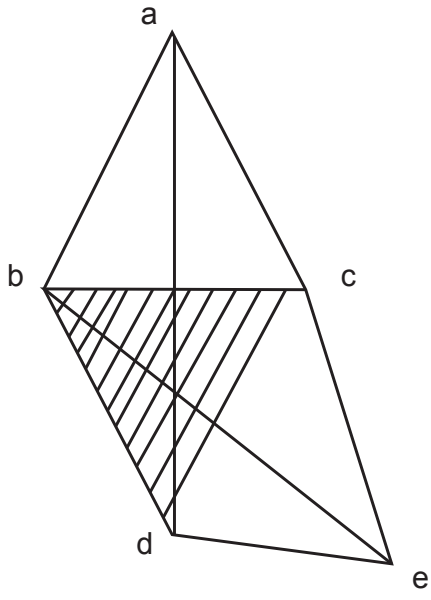
Un teorema conocido en topología algebraica,[14] prueba que es seguro que un complejo simplicial abstracto de dimensión n tiene una realización geométrica en R^{2n+1}. Aunque muchas veces la realización puede ser en

espacios euclidianos de dimensión menor a (2n+1).

La dualidad entre complejos abstractos y geométricos nos da la flexibilidad suficiente para representar inicialmente un complejo abstracto que surge de relaciones abstractas mediante una matriz de incidencia y poderlo interpretar geométricamente y representarlo gráficamente después. Sin perder la posibilidad de trabajar en espacios multidimensionales donde no es posible una representación gráfica.

Ejemplo 1:

Consideremos un complejo geométrico en R^3, cuya representación está dada por la figura:



Asociado a este complejo geométrico tenemos un complejo simplicial abstracto, definido como:

$K = \{ \sigma_0^i \} \cup \{ \sigma_1^j \} \cup \{ \sigma_2^k \}$ con i = 1,....,5, j = 1,.....,9 , k = 1, en donde:

$\sigma_2^1 = \{b,c,d\}$, $\sigma_1^1 = \{a,b\}$, $\sigma_1^2 = \{a,c\}$, $\sigma_1^3 = \{b,c\}$, $\sigma_1^4 = \{b,d\}$, $\sigma_1^5 = \{b,e\}$, $\sigma_1^6 = \{a,d\}$, $\sigma_1^7 = \{c,d\}$, $\sigma_1^8 = \{c,e\}$, $\sigma_1^9 = \{d,e\}$, $\sigma_0^1 = \{a\}$, $\sigma_0^2 = \{b\}$, $\sigma_0^3 = \{c\}$, $\sigma_0^4 = \{d\}$, $\sigma_0^5 = \{e\}$.

Una representación mediante una matriz de incidencia puede ser:

MI	a	b	c	d	e
σ_2^1	0	1	1	1	0
σ_1^1	1	1	0	0	0
σ_1^2	1	0	1	0	0
σ_1^5	0	1	0	0	1
σ_1^6	1	0	0	1	0
σ_1^8	0	0	1	0	1
σ_1^9	0	0	0	1	1

Las columnas son etiquetadas por los vértices, y las filas por los símplices. No hay necesidad de incluir filas correspondientes a los símplices que son caras.

A un complejo simplicial K le podemos asociar arreglos numéricos que son invariantes topológicos (todos los poliedros equivalentes topológicamente poseen los mismos arreglos).

El primero de ellos se conoce con el nombre de primer vector de estructura del complejo y permite ver la conectividad interna del complejo (visión local) recurriendo a la noción de q-conectividad. Esta noción y sus aplicaciones en ciencias sociales fueron desarrolladas por el matemático Ronald Atkin.[12,13]

Dados dos símplices de un complejo K, σ_p y σ_r. Decimos que σ_p y σ_r son "q-adyacentes" si existe al menos una cara común entre ellos que es un q- símplice. Obviamente si σ_p y

σ_r son q-adyacentes entonces son q-1, q-2,...,1,0 adyacentes.

Sea δ_q la relación definida como "es q-adyacente con ". Dicha relación es reflexiva y simétrica pero no transitiva. La representaremos por una matriz cuadrada de NxN, siendo N el número de símplices de K de dimensión ≥ q y que no son caras de otros símplices en K.

Asociado a un complejo K tendremos m+1 matrices de q-adyacencia, siendo m la dimensión de K.

Dados dos símplices de un complejo K, σ_p y σ_r. Decimos que σ_p y σ_r están "q- conectados" si σ_p y σ_r son q-adyacentes o si existe una secuencia de símplices en K, σ_1,....., σ_n tal que σ_p es q-adyacente a σ_1, σ_1 es q-adyacente a σ_2 y así sucesivamente hasta llegar a que σ_n es q-adyacente a σ_r . Si σ_p y σ_r son q-conectados entonces ellos también son (q-1),...1,0-conectados en K.

Si definimos la relación γ_q como significando "es q-conectado con" entonces $_q$ es una relación de equivalencia sobre los símplices de K. Las clases de K/ γ_q son ahora las piezas de K las cuales son separadamente q-conectadas.

La relación γ_q puede ser representada por una matriz cuadrada de NxN que denominamos matriz de "q-conexión". La matriz de q-conexión se obtiene a partir de la matriz de q-adyacencia, mediante un cierre transitivo.

Los "unos" de la matriz de q-conexión, determinan las clases de equivalencia de K/γ_q que son ahora las piezas de K que son separadamente q-conectadas. Q_q es la cardinalidad de K/γ_q.

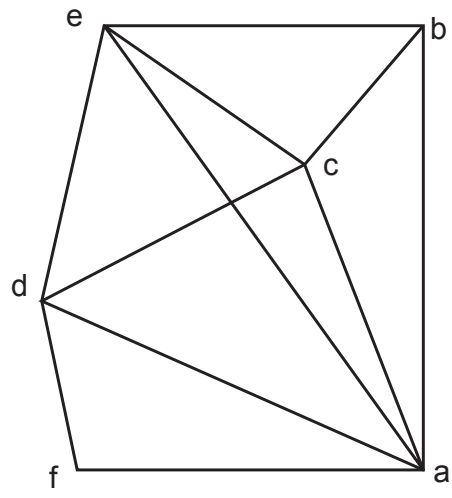
Si K es un complejo finito no vacío de dimensión n, le podemos asociar el arreglo $\mathbf{Q} = <Q_0, Q_1,..., Q_n>$ en donde Q_i es la cardinalidad de K/γ, , siendo $Q_i \geq 1, \forall_i = 0,1,.., n$.

Las clases de Q_0 son las componentes arco-conexas del complejo.

Ejemplo 2: Consideremos un complejo finito K cuyos vértices son a, b, c, d, e, f y cuyos símplices vienen dados por la matriz de incidencia:

MI	a	b	c	d	e	f
σ_1^1	1	1	0	0	0	0
σ_2^1	0	0	1	1	1	0
σ_2^2	0	1	1	0	1	0
σ_2^3	1	0	1	0	1	0
σ_2^4	1	0	0	1	0	1

La representación geométrica del complejo es:



Las Matrices de Conexión asociadas son:

0	σ_1^1	σ_2^1	σ_2^2	σ_2^3	σ_2^4
σ_1^1	1	1	1	1	1
σ_2^1	1	1	1	1	1
σ_2^2	1	1	1	1	1
σ_2^3	1	1	1	1	1
σ_2^4	1	1	1	1	1

1	σ_1^1	σ_2^1	σ_2^2	σ_2^3	σ_2^4
σ_1^1	1	0	0	0	0
σ_2^1	0	1	1	1	0
σ_2^2	0	1	1	1	0
σ_2^3	0	1	1	1	0
σ_2^4	0	0	0	0	1

2	σ_2^1	σ_2^2	σ_2^3	σ_2^4
σ_2^1	1	0	0	0
σ_2^2	0	1	0	0
σ_2^3	0	0	1	0
σ_2^4	0	0	0	1

Por lo tanto **Q** = < 1, 3, 4 >.

3. ALGORITMO DE OBTENCIÓN DE LOS CLÚSTER

Presentamos un algoritmo de obtención de los clúster que puede ser aplicado tanto para el caso en que se cuente con una matriz de ponderaciones numéricas como similitudes, distancias, etc. **;** como para el caso de que se trate de una red o grafo.

Entrada: Una matriz cuadrada simétrica S que representa las similitudes. Un parámetro MIN que representa el tamaño mínimo de los clúster.

Salida: K = El conjunto de los clúster.

Método:

Paso 1. Sea $\{ V_i \}$ con i desde 1 hasta n, la sucesión de valores de la matriz de similitudes S, ordenados de mayor a menor.

Paso 2. Hacer i = 1

Paso 3. Generamos la matriz de incidencia MI a partir de la matriz de similitudes S, haciendo:

IF $s(m,n) \in S$ y $s(m,n) \geq V_i$ THEN MI(m,n) = 1

ELSE MI(m,n) = 0

Paso 4. For q = 0, … , DIM (donde DIM es la dimensión del complejo de MI) hacer:

i) Calcular las q-clases conectadas

ii) IF para q = DIM, hay clases de equivalencia q-conectadas de tamaño ≥ MIN

THEN 1) Incluir esas clases en K

2) Eliminar de S las filas y columnas que contengan elementos de las clases de K incluidas en 1).

3) IF la dimensión de S = 0 THEN fin del algoritmo

ELSE

IF la dimensión de S < MIN

THEN las filas de S son un residuo que incluimos en la última clase introducida en K (en caso de que sea más de una, lo agregamos a la clase más afín) y fin del algoritmo.

ELSE regresar al paso 1.

ELSE hacer i = i+1 y regresar al paso 3.

Ejemplo 3: Consideremos la siguiente matriz de afinidades tomada de [3],

	1	2	3	4	5	6	7	8	9	10
1	75	25	25	0	75	0	50	25	25	0
2	25	110	75	0	25	0	60	110	75	0
3	25	75	115	15	25	15	25	75	115	15
4	0	0	15	40	0	40	0	0	15	40
5	75	25	25	0	75	0	50	25	25	0
6	0	0	15	40	0	40	0	0	15	40
7	50	60	25	0	50	0	85	60	25	0
8	25	110	75	0	25	0	60	110	75	0
9	25	75	115	15	25	15	25	75	115	15
10	0	0	15	40	0	40	0	0	15	40

La secuencia de valores ordenada de mayor a menor es { V_1=115, V_2=110, V_3=85, V_4=75, V_5=60, V_6=50, V_7=40, V_8=25, V_9=15, V_{10}=0 }.

Consideremos el parámetro MIN = 3.

Para i = 1, V_1 = 115, así que la matriz MI_0 viene dada por,

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0

Esta matriz de incidencia representa un complejo simplicial de dimensión 1, cuyo vector de estructura es **Q** = < 1, 1 > , que corresponde a las clases de equivalencia, 0-clases conectadas:{{3,9}} y 1-clases conectadas: {{3,9}}.

Dado que el tamaño (cardinalidad) de {3,9} es 2 < MIN=3, es necesario incrementar el valor de i y de j, hasta obtener una matriz de incidencia MI_j que corresponda a un complejo simplicial con D-clases conectadas de tamaño ≥ 3, siendo D la dimensión del complejo.

Para i = 4, con V_4 = 75 (tomamos los cuatro valores más grandes de la matriz de afinidad, 115,110, 85 y 75) con lo cual tenemos la siguiente matriz de incidencia MI_3 :

	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	1	0	0	0	0	0
2	0	1	1	0	0	0	0	1	1	0
3	0	1	1	0	0	0	0	1	1	0
4	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0
8	0	1	1	0	0	0	0	1	1	0
9	0	1	1	0	0	0	0	1	1	0
10	0	0	0	0	0	0	0	0	0	0

Esta matriz de incidencia representa un complejo simplicial de dimensión 3, cuyo vector de estructura es **Q** = < 3, 2, 1, 1> , que corresponde a las clases de equivalencia, 0-clases conectadas: {{7}, {1,5}, {2,3,8,9}},1-clases conectadas: {{1,5} , {2,3,8,9}}, 2-clases conectadas: {{2,3,8,9}}, 3-clases conectadas: {{2,3,8,9}}.

Dado que el tamaño (cardinalidad) de {2,3,8,9} es 4 >MIN=3, el primer clúster es {2,3,8,9}.

Eliminando las filas y columnas denotadas por 2,3,8,9; la matriz de afinidad S queda reducida a:

	1	4	5	6	7	10
1	75	0	75	0	50	0
4	0	40	0	40	0	40
5	75	0	75	0	50	0
6	0	40	0	40	0	40
7	50	0	50	0	85	0
10	0	40	0	40	0	40

Para i = 3, con V_3 = 50 (tomamos los tres valores más grandes de la matriz de afinidad, 85 , 75 y 50) con lo

cual tenemos la siguiente matriz de incidencia MI_2:

	1	4	5	6	7	10
1	1	0	1	0	1	0
4	0	0	0	0	0	0
5	1	0	1	0	1	0
6	0	0	0	0	0	0
7	1	0	1	0	1	0
10	0	0	0	0	0	0

Esta matriz de incidencia representa un complejo simplicial de dimensión 2, cuyo vector de estructura es **Q** = < 1, 1, 1> , que corresponde a las clases de equivalencia, 0-clases conectadas: {{1,5,7}} ,1-clases conectadas: {{1,5,7}} , 2-clases conectadas: {{1,5,7}}.

Dado que el tamaño (cardinalidad) de {1,5,7} es 3 = MIN, el segundo clúster es {1,5,7}.

Dado que en S quedarían 4,6,10; el tercer clúster es {4,6,10}.

Finalmente K = { {2,3,8,9}, {1,5,7}, {4,6,10} }.

Ejemplo 4: Consideremos un grafo cuya matriz de adyacencia aumentada con unos en la diagonal es la siguiente,

	MI	A	B	C	D	E	F	G	H	I
A	1	1	1	1	0	0	0	0	0	0
B	1	1	1	0	1	0	0	0	0	0
C	1	1	1	1	1	1	0	0	0	0
D	1	0	1	1	1	0	0	0	1	1
E	0	1	1	1	1	0	0	0	0	0
F	0	0	1	0	0	1	1	1	1	1
G	0	0	0	0	0	1	1	1	1	1
H	0	0	0	0	0	1	1	1	1	1
I	0	0	0	1	0	1	1	1	1	1

Esta matriz se corresponde con la matriz de incidencia de un complejo Simplicial asociado al grafo, llamado complejo de vecindad del grafo.

El producto de MI por su traspuesta MI', es la matriz de caras compartidas(S),

S	A	B	C	D	E	F	G	H	I
A	4	3	4	3	3	1	0	0	1
B	3	4	4	3	3	1	0	0	0
C	4	4	6	4	4	2	1	1	2
D	3	3	4	5	3	2	1	1	2
E	3	3	4	3	4	1	0	0	1
F	1	1	2	2	1	5	4	4	4
G	0	0	1	1	0	4	4	4	4
H	0	0	1	1	0	4	4	4	5

A partir de esta matriz podemos aplicar el algoritmo. El conjunto de valores ordenado de mayor a menor es { V_1=6, V_2=5, V_3=4, V_4=3, V_5=2, V_6=1, V_7=0 }.

Consideremos un valor de MIN=3. Para i=4, con V_4=3, la matriz de incidencia que le corresponde es,

MI	A	B	C	D	E	F	G	H	I
A	1	1	1	1	1	0	0	0	0
B	1	1	1	1	1	0	0	0	0
C	1	1	1	1	1	0	0	0	0
D	1	1	1	1	1	0	0	0	0
E	1	1	1	1	1	0	0	0	0
F	0	0	0	0	0	1	1	1	1
G	0	0	0	0	0	1	1	1	1
H	0	0	0	0	0	1	1	1	1
I	0	0	0	0	0	1	1	1	1

Esta matriz de incidencia representa un complejo simplicial de dimensión 4, cuyo vector de estructura es **Q** = < 2, 2, 2, 2, 1 > .

Las correspondientes clases de equivalencia son, 0-clases conectadas: {{A,B,C,D,E}, {F,G,H,I}} ,1-clases conectadas: {{A,B,C,D,E}, {F,G,H,I}}, 2-clases conectadas: {{A,B,C,D,E}, {F,G,H,I}}, 3-clases conectadas: {{A,B,C,D,E}, {F,G,H,I}}, 4-clases conectadas: {{A,B,C,D,E}}.

Por lo tanto el primer clúster es {A,B,C,D,E}.

Eliminando en S las filas y columnas correspondientes al primer clúster, obtenemos la siguiente matriz reducida,

S	F	G	H	I
F	5	4	4	4
G	4	4	4	4
H	4	4	4	4
I	4	4	4	5

El conjunto de valores es { V_1=5, V_2=4 }. Para i = 2, con V_2=4, la matriz de incidencia que le corresponde es,

MI	F	G	H	I
F	1	1	1	1
G	1	1	1	1
H	1	1	1	1
I	1	1	1	1

Esta matriz de incidencia representa un complejo simplicial de dimensión 3, cuyo vector de estructura es **Q** = < 1, 1, 1, 1 > .

Las correspondientes clases de equivalencia son, 0-clases conectadas: {{F,G,H,I}} ,1-clases conectadas: {{F,G,H,I}}, 2-clases conectadas: {{F,G,H,I}}, 3-clases conectadas:{{F,G,H,I}}.

Por lo tanto el segundo clúster es {F,G,H,I}.

Finalmente K = { {A,B,C,D,E}, {F,G,H,I} }.

4. COMPARACIÓN CON OTROS ALGORITMOS

Cuando comparamos este algoritmo con el algoritmo propuesto en [3], encontramos que en general, ambos producen agrupamientos muy similares.

Sin embargo hay que tener en cuenta que el algoritmo gráfico de particionamiento [3], produce agrupamientos que son ciclos de longitud mínima de 3 y permite la extensión de los ciclos cuando hay al menos un enlace con un valor numérico mayor o igual al mínimo de los valores numéricos en el ciclo.

Esto tiene el inconveniente que puede disminuir notablemente la cohesión del clúster (ciclo). Por ejemplo, consideremos la siguiente matriz de afinidad,

	A	B	C	D	E	F
A	3	3	2	2	1	1
B	3	3	1	1	1	1
C	2	1	3	1	1	1
D	2	1	1	3	0	0
E	1	1	1	0	3	0
F	1	1	1	0	0	3

El algoritmo en [3] produce un solo clúster, { A, B, C, D, E, F }.

Si aplicamos nuestro algoritmo con conjunto de valores { V_1=3, V_2=2, V_3=1 }, obtenemos la siguiente matriz de incidencia,

	A	B	C	D	E	F
A	1	1	1	1	1	1
B	1	1	1	1	1	1
C	1	1	1	1	1	1
D	1	1	1	1	0	0
E	1	1	1	0	1	0
F	1	1	1	0	0	1

Esta matriz de incidencia representa un complejo simplicial de dimensión 5, cuyo vector de estructura es **Q** = < 1, 1, 1, 1, 1,1 > .

Las correspondientes clases de equivalencia son, 0-clases conectadas: {{A,B,C,D,E,F}} ,1-clases conectadas: {{A,B,C,D,E,F}}, 2-clases conectadas: {{A,B,C,D,E,F}}, 3-clases conectadas:{{A,B,C,D,E,F}}, 4-clases conectadas: {{A,B,C}}, 5-clases conectadas: {{A,B,C}}.

Por lo tanto el primer clúster es { A,B,C }, un segundo clúster sería { D,E,F }, lo cual es más razonable.

Cuando aplicamos las técnicas de partición espectral [4,5] al grafo del ejemplo 4, obtenemos que el traspuesto del vector de Fiedler es,

< -0.29,-0.32,-0.28,-0.18,-0.29,0.34,0.41,0.41,0.36 > con valor propio de 13,6472.

Los valores negativos corresponden al clúster { A,B,C,D,E } y los valores positivos corresponden al clúster { F,G,H,I }, lo que se corresponde con el resultado de nuestro algoritmo.

5. CONCLUSIONES

Existen varias propuestas para trabajar el tema de clustering, el cual tiene importancia en muchas áreas por sus aplicaciones.

En este artículo se presenta una nueva propuesta que consiste en utilizar la topología a través del q-análisis de Atkin, en lo que podríamos denominar un clustering estructural.

El algoritmo presentado produce resultados semejantes a otros algoritmos conocidos, los cuales tienen aproximaciones diferentes como son, la descomposición espectral de grafos, agrupamientos basados en similitudes, distancias, ultramétricas, etc.

Una ventaja del algoritmo propuesto es la posibilidad de utilizarse tanto si se tiene una matriz de ponderaciones numéricas, como si se trata de un grafo, el cual puede o no estar dirigido.

Quedan pendientes algunos estudios adicionales sobre el algoritmo. El primero de ellos, es determinar formalmente la eficiencia computacional del mismo y el segundo es mejorar el manejo de los residuos (grupos de objetos de tamaño menor que el parámetro MIN).

6. BIBLIOGRAFÍA

1 J. Barthélemy, A. Guenoche. Les arbres et les représentations des proximites. Editorial Masson, Paris, 1988.

2 J. Scott. Social Network Analysis: a Handbook. Sage publications, London, 2000.

3 S. Navathe, M. Ra. Vertical Partitioning for Database Design: a Graphical Algorithm. Proceedings of the 1989 ACM SIGMOD International Conference of the Management of the Data, Portland, Oregon, Vol. 18, No.2.

4 B. Mohar. Some Applications of Laplace Eigenvalues Graphs. University of Ljubljana Slovenia, preprint series, Vol. 35, 1997.

5 A. Seary, W. Richards. Partitioning Networks by Eigenvectors. INSNA Sunbelt XVI, London, July, 1995.

6 M. Fiedler. Algebraic Connectivity of Graphs. Czechoslovak Mathematical Journal, 23 (98): 298-305, 1973.

7 M. Fiedler. A property of Eigenvectors of nonnenegative Symmetric matrices and its applications to Graph Theory. Czechoslovak Mathematical Journal, 25 (100): 619-633, 1975.

8 F. Chung. Spectral Graph Theory. Number 92 in CBMS regional conference series in Mathematics. American Mathematical Society, Providence, R.I., 1997.

9 B. Bollobás. Modern Graph Theory. Springer, New York, 1988.

10 M. Barahona, L. Pecora. Synchronization in Small-World Systems. Phys. Rev. Lett. 89, 2002.

11 T. Nishikawa, A. Motter, Y. Lai, F. Hoppensteadt. Heterogeneity in Oscillator Networks: Are Smaller Worlds easier to Synchronize ? Phys. Rev. Lett. 91, 2003.

12 R. Atkin. From Cohomology in Physics to q-connectivity in Social Science. Int. J. Man-Machine Studies 4, 139-167, 1972.

13 P. Gould. Q-analysis or a Language of Structure: an Introduction for Social scientists, Geographers and Planners. Int. J. Man-Machine Studies 13, 169-199, 1980.

14 J. Munkres. Elements of Algebraic Topology. Addison-Wesley, 1984.

CURRÍCULO

Luis Eduardo Múnera. Matemático de la Universidad del Valle, Cali-Colombia. Máster y Doctor en Informática de la Universidad Politécnica de Madrid-España. Profesor Titular de la Universidad Icesi de Cali-Colombia, miembro del grupo de investigación I2T.