Un algoritmo incremental para la obtención de cubrimientos con datos mezclados

Aurora Pons-Porrata[1], José Ruiz-Shulcloper[2], Rafael Berlanga-Llavori[3], Yovanis Santiesteban Alganza[1]
[1] Universidad de Oriente, Santiago de Cuba (Cuba)
aurora@app.uo.edu.cu, yosva@csd.uo.edu.cu
[2] Instituto de Cibernética, Matemática y Física (Cuba)
recpat@cidet.icmf.inf.cu
[3]Universitat Jaume I, Castellón (España)
berlanga@lsi.uji.es

Resumen

En este trabajo se introduce un algoritmo incremental eficiente para estructurar un conjunto de datos en conjuntos fuertemente compactos. Este algoritmo se apoya en algunas propiedades de los conjuntos fuertemente compactos que son demostradas en el trabajo. El algoritmo propuesto crea un cubrimiento único del conjunto de datos, por lo que no depende del orden de presentación de los objetos. La complejidad computacional del algoritmo incremental presentado sigue siendo la misma que la de su variante no incremental y, por tanto, mucho más eficiente que la aplicación reiterada de la variante no incremental. El algoritmo propuesto puede ser utilizado en todas las tareas que requieran del agrupamiento de objetos en fuertemente compactos y del procesamiento dinámico de la información, tales como, por ejemplo, la identificación y seguimiento de noticias en un flujo de documentos; el estudio de la morbilidad de una población; la estructuración de la información para estudios socio-económicos; entre otras posibles aplicaciones. Es bueno resaltar que este algoritmo puede utilizarse, además, en problemas donde se deseen agrupar objetos de cualquier naturaleza, descritos por rasgos cuantitativos y cualitativos mezclados, incluso con ausencia de información.

Palabras clave: algoritmos incrementales, algoritmos de agrupamiento, conjuntos fuertemente compactos.

1. Introducción.

El agrupamiento de datos es uno de los problemas centrales en la Minería de Datos. Entre las técnicas de agrupamiento existentes, el criterio de agrupamiento en conjuntos fuertemente compactos tiene la propiedad de que la semejanza entre un par de objetos de un mismo grupo es máxima. Los grupos obtenidos por este criterio son solapados y relativamente pequeños y densos. Existen muchas aplicaciones donde se necesita formar grupos de objetos de los que sabemos que por su naturaleza son solapados, por lo que este criterio resulta de utilidad. Existen,

además, otros problemas en los que el conjunto de datos se incrementa en el tiempo. Por ejemplo, la identificación y seguimiento de noticias en un flujo de documentos. Un documento que trate acerca de la mortalidad por el virus VIH en un país dado, aparecerá junto a otros que aborden el tema de la medicina, pero también formará parte de otros conjuntos de documentos, digamos por caso, que hablen de problemas sociales tales como la esperanza de vida al nacer o los que traten sobre los derechos humanos, entre otros. Estas estructuraciones pueden ser de utilidad para analistas información de diferentes perfiles. Para este tipo de aplicaciones se requiere de algoritmo incremental que obtenga los conjuntos fuertemente compactos.

En este trabajo se introduce un algoritmo incremental eficiente para estructurar un conjunto de datos en conjuntos fuertemente compactos. Este algoritmo se apoya en algunas propiedades de los conjuntos fuertemente compactos que demostradas en el trabajo.

2. Conceptos y resultados necesarios.

Sea ζ una colección de objetos y S una función de semejanza entre objetos simétrica. Aquí sólo consideraremos este tipo de función de semejanza. además, β_0 un umbral de semejanza definido por el usuario.

Definición 1. *Se dice que dos objetos O_i y O_j son β_0-semejantes si $S(O_i,O_j)\geq\beta_0$. para todo O_j de la colección de objetos ζ se cumple que $S(O_i,O_j)<\beta_0$ entonces O_i denomina β_0-aislado.*

Definición 2. *Diremos que $NU\subseteq\zeta$, $NU\neq\emptyset$ es una componente conexa si se cumple que* [1]:

1. $\forall O_i,O_j\in NU,\ O_i\neq O_j,\ \exists O_{i_1},\ldots,O_{i_q}\in NU[O_i=O_{i_1}\wedge O_j=O_{i_q}\wedge\forall p=1,\ldots,q\text{-}1\ S(O_p,O_{p+1})\geq\beta_0]$.
2. $\forall O_i\in\zeta[O_j\in NU\wedge S(O_i,O_j)\geq\beta_0]\Rightarrow O_i\in NU$.
3. *Todo elemento β_0-aislado es una componente conexa (degenerada).*

Definición 3. *Diremos que $NU\subseteq\zeta$, $NU\neq\emptyset$, es un conjunto compacto si* [1]:

1. $\forall O_j\in\zeta[O_i\in NU\wedge \max\limits_{\substack{O_t\in\zeta\\ O_t\neq O_i}}\{S(O_i,O_t)\}=S(O_i,O_j)\geq\beta_0]\Rightarrow O_j\in NU$.
2. $[\max\limits_{\substack{O_i\in\zeta\\ O_i\neq O_p}}\{S(O_p,O_i)\}=S(O_p,O_t)\geq\beta_0\wedge O_t\in NU]\Rightarrow O_p\in NU$.
3. $|NU|$ *es mínimo.*

Todo elemento β_0-aislado constituye un conjunto compacto (degenerado).

El criterio de agrupamiento basado en los conjuntos compactos forma, al que el de las componentes conexas, una partición.

Definición 4. Diremos que $NU\subseteq\zeta$, $NU\neq\emptyset$, es un *conjunto fuertemente compacto* si:

1. $\forall O_j \in \zeta\ [O_i \in NU \wedge \max_{\substack{O_t \in \zeta \\ O_t \neq O_i}} \{S(O_i, O_t)\} = S(O_i, O_j) \geq \beta_0] \Rightarrow O_j \in NU.$

2. $\exists O_i \in NU\ \forall O_j \in NU\ \exists O_{i_1}, \ldots, O_{i_q} \in NU\ [O_i = O_{i_1} \wedge O_j = O_{i_q} \wedge \forall p<q$
$[\max_{\substack{O_t \in \zeta \\ O_t \neq O_{i_p}}} \{S(O_{i_p}, O_t)\} = S(O_{i_p}, O_{i_{p+1}}) \geq \beta_0]].$

3. No existe NU' que cumpla 1. y 2. y $NU \subset NU'$.

Todo elemento β_0-aislado constituye un conjunto fuertemente compacto (degenerado) [1].

La condición 1 dice que todo objeto de *NU* tiene en *NU* al objeto màs β_0-semejante. La condición 2 significa que en *NU* existe un objeto tal que para cualquier otro que pertenezca a *NU* existe una sucesión de objetos de *NU* tales que uno es el más β_0-semejante al siguiente. La condición 3 significa que *NU* es el conjunto más grande que cumple las condiciones 1 y 2. Este criterio forma grupos no disjuntos.

Ejemplo 1. Sean $\zeta=\{O_1,O_2,O_3,O_4,O_5,O_6,O_7,O_8,O_9,O_{10},O_{11},O_{12}\}$ y la matriz de semejanza siguiente, obtenida a partir de una función de semejanza *S*:

$$MS = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \\ O_9 \\ O_{10} \\ O_{11} \\ O_{12} \end{array} \begin{pmatrix} 1.0 & 0.85 & 0.75 & 0.78 & 0.63 & 0.70 & 0.51 & 0.38 & 0.43 & 0.27 & 0.22 & 0.13 \\ & 1.0 & 0.90 & 0.90 & 0.76 & 0.62 & 0.58 & 0.45 & 0.56 & 0.38 & 0.36 & 0.27 \\ & & 1.0 & 0.88 & 0.83 & 0.50 & 0.60 & 0.47 & 0.65 & 0.47 & 0.47 & 0.33 \\ & & & 1.0 & 0.85 & 0.58 & 0.68 & 0.55 & 0.63 & 0.45 & 0.42 & 0.30 \\ & & & & 1.0 & 0.48 & 0.73 & 0.65 & 0.80 & 0.65 & 0.58 & 0.50 \\ & & & & & 1.0 & 0.51 & 0.36 & 0.27 & 0.11 & 0.05 & 0.00 \\ & & & & & & 1.0 & 0.83 & 0.66 & 0.58 & 0.48 & 0.51 \\ & & & & & & & 1.0 & 0.66 & 0.63 & 0.53 & 0.61 \\ & & & & & & & & 1.0 & 0.83 & 0.80 & 0.70 \\ & & & & & & & & & 1.0 & 0.88 & 0.85 \\ & & & & & & & & & & 1.0 & 0.80 \\ & & & & & & & & & & & 1.0 \end{pmatrix}$$

Si consideramos β_0=0.8, entonces los conjuntos compactos son $NU_1=\{O_1,O_2,O_3,O_4,O_5\}$, $NU_2=\{O_9,O_{10},O_{11},O_{12}\}$, $NU_3=\{O_7,O_8\}$ y $NU_4=\{O_6\}$ y los fuertemente compactos son: $NU_1=\{O_1,O_2,O_3,O_4\}$, $NU_2=\{O_2,O_3,O_4,O_5\}$, $NU_3=\{O_9,O_{10}, O_{11}\}$, $NU_4=\{O_{10},O_{11}, O_{12}\}$, $NU_5=\{O_7,O_8\}$ y $NU_6=\{O_6\}$

Definición 5. *Llamaremos grafo basado en la máxima β_0-semejanza según S, y lo denotaremos máx-S, al grafo orientado $G=(\zeta,E)$ cuyos vértices son los objetos de la colección ζ y existe un arco del vértice O_i al O_j si se cumple que O_j es el objeto más β_0-semejante a O_i.*

Proposición 1. *El conjunto de todos los conjuntos compactos de ζ coincide con el conjunto de todas las componentes conexas del grafo máx-S asociado a ζ sin tener en cuenta la orientación.*

Entre las componentes conexas, los conjuntos compactos y fuertemente compactos definidos anteriormente se satisfacen las siguientes relaciones [1]:

1. Toda componente conexa es la unión finita de conjuntos compactos.
2. Todo conjunto compacto es la unión finita de conjuntos fuertemente compactos.
3. Toda componente conexa es la unión finita de conjuntos fuertemente compactos.

Definición 6. *Una componente fuertemente conexa de un grafo orientado es un conjunto maximal de vértices en el cual para todo vértice en el conjunto existe un camino a cualquier otro vértice del conjunto* [2]. *Todo objeto β_0-aislado es una componente fuertemente conexa (degenerada).*

3. Nuevas propiedades de los conjuntos fuertemente compactos.

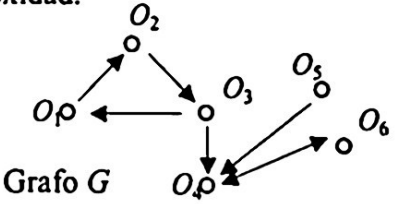
Definición 7. *Sea $G=(X,E)$ un grafo orientado y $B \subset X$. B es una base del grafo G si se cumplen las dos condiciones siguientes* [3]:

1. *Todo vértice del conjunto X es descendiente de al menos un vértice de B.*
2. *No existe en el conjunto X un subconjunto de menor número de elementos con dicha propiedad.*

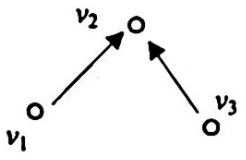
Definición 8: *Sean $G=(X,E)$ un grafo orientado y P una partición del conjunto de vértices X, donde V_i denota a cada elemento de P. Se llama grafo reducido de G, y lo denotaremos como $G_r=(V,E_r)$, al grafo cuyos vértices son los subgrafos generados por los elementos V_i de P y existe un arco del vértice v_i al v_j si existe un vértice x en V_i y un vértice y en V_j tal que existe el arco de x a y en G* [3].

Definición 9. *Sea G un grafo orientado. Llamaremos grafo reducido según la fuerte-conexidad de G, y lo denotaremos como G/f-C, al grafo reducido de G asociado a la partición inducida por las componentes fuertemente conexas de G. En el grafo G/f-C los vértices son las componentes fuertemente conexas de G y existe un arco de un vértice v a otro v' en el grafo reducido, $v \neq v'$, si existe al menos un arco en G de un vértice en la componente fuertemente conexa que representa v a algún vértice en la componente fuertemente conexa que representa v'.*

Ejemplo 2. En la figura siguiente se muestra un grafo *G* y su grafo reducido según la fuerte-conexidad.



Grafo *G*



Grafo reducido según la fuerte-conexidad de *G*

Las componentes fuertemente conexas del grafo G son $v_1=\{O_1,O_2,O_3\}$, $v_2=\{O_4,O_6\}$ y $v_3=\{O_5\}$. Las bases del grafo G son $\{O_1,O_5\}$, $\{O_2,O_5\}$ y $\{O_3,O_5\}$. La base del grafo reducido según la fuerte-conexidad de G es $\{v_1,v_3\}$.

Proposición 2. *Sea $G=(\zeta,E)$ el grafo máx-S. El grafo $G_r/f\text{-}C=(V,U)$ tiene una única base formada sólo por los vértices de grado interior nulo.*

Sean $G=(\zeta, E)$ el grafo *máx-S*, $G_r/f\text{-}C=(V,U)$ su grafo reducido, $B=\{b_1,...,b_r\}$ la base de $G_r/f\text{-}C$ y $D(b_i)$ el conjunto de vértices descendientes de b_i en $G_r/f\text{-}C$. Sea, además, F: $V\rightarrow\zeta$ tal que $F(v_i)$ es una componente fuertemente conexa en G. F es sobreyectiva porque $G_r/f\text{-}C$ es el grafo reducido de G.

Lema 1. $K=\bigcup_{v_j\in D(b_i)} F(v_j)$ *es un conjunto fuertemente compacto en* ζ.

Demostración. $K\neq\phi$, ya que $D(b_i)\neq\phi$. Probemos que K cumple las siguientes condiciones:

1. $\forall x\in\zeta\,[x'\in K\wedge \max_{\substack{x_j\in\zeta\\ x_j\neq x}}\{S(x',x_j)\}=S(x',x)\geq\beta_0]\Rightarrow x\in K.$

Como $x'\in K$ existe al menos un j tal que $x'\in F(v_j)$, $v_j\in D(b_i)$. Luego, un elemento cualquiera x, más β_0-semejante a x', pertenece a $F(v_j)$ o pertenece a $F(v_t)$, $t\neq j$, $v_t\in V$.

Si $x\in F(v_j)$, entonces $x\in K$. Por el contrario, si $x\in F(v_t)$, implica que en el grafo reducido $G_r/f\text{-}C$ tiene que existir un arco de v_j a v_t, por lo que v_t sería descendiente de v_j y, por tanto, de b_i. Luego, $v_t\in D(b_i)$ y $x\in K$.

2. $\exists x'\in K\ \forall x\in K\ \exists x_{i_1},...,x_{i_q}\in K\,[x'=x_{i_1}\wedge x=x_{i_q}\wedge\forall p<q$

$$[\max_{\substack{x_t\in\zeta\\ x_t\neq x_{i_p}}}\{S(x_{i_p},x_t)\}=S(x_{i_p},x_{i_{p+1}})\geq\beta_0]].$$

Sea $x'\in K$ tal que $x'\in F(b_i)$. Entonces, para todo $x\in K$ se cumple que $x\in F(b_i)$ ó $x\in F(v_j)$, $v_j\neq b_i$, $v_j\in D(b_i)$. En el primer caso, existen $x_{i_1},...,x_{i_q}\in F(b_i)$ tal que $x'=x_{i_1}\wedge x=x_{i_q}\wedge\forall p<q\,[\max_{\substack{x_t\in\zeta\\ x_t\neq x_{i_p}}}\{S(x_{i_p},x_t)\}=S(x_{i_p},x_{i_{p+1}})\geq\beta_0]]$ porque $F(b_i)$ es una componente fuertemente conexa en el grafo G y, por tanto, $x_{i_1},...,x_{i_q}\in K$. En el segundo caso, como $v_j\in D(b_i)$, existen $v_{i_1},...,v_{i_r}\in D(b_i)$ tal que $b_i=v_{i_1}\wedge v_j=v_{i_r}\wedge\forall p<r\ \exists<v_{i_p},v_{i_{p+1}}>\in U$. Basta con escoger los elementos x_{i_t} $t=1,...,q$ de la manera siguiente: $x_{i_1}=x'$, $x_{i_q}=x$. Como $<v_{i_p},v_{i_{p+1}}>$ y $<v_{i_{p+1}},v_{i_{p+2}}>\in U$ $\exists w,y,z,t$ $w\in F(v_{i_p})$, $y,z\in F(v_{i_{p+1}})$ y $t\in F(v_{i_{p+2}})$ tal que y es el más β_0-semejante a w, t es el más β_0-semejante a z y como $F(v_{i_{p+1}})$ es

una componente fuertemente conexa en G existe un camino en G entre y y z. Luego, existe un camino entre w y t en G. Por tanto, existe un camino entre x' y x en G.

3. No existe $K' \supset K$ que cumple las condiciones 1 y 2.

Supongamos que $\exists K' \supset K$ que cumple las condiciones 1 y 2. En $K' \setminus K$ no puede existir x que sea el más β_0-semejante a un $x' \in K$, pues tendría que estar en K. Luego, cualquier elemento de $K' \setminus K$ es el más β_0-semejante de uno de $K' \setminus K$ y/o tiene su más β_0-semejante en K. No puede ocurrir que todos los elementos de $K' \setminus K$ tengan su más β_0-semejante en $K' \setminus K$, ya que no se cumpliría la condición 2 (no habría manera de llegar de un elemento de $K' \setminus K$ a K ni tampoco habría manera de llegar de un elemento de K a uno de $K' \setminus K$). Luego, tiene que existir $x \in K' \setminus K$ que tenga su más β_0-semejante en K y no sea el más β_0-semejante de un $y \in K$. En virtud de lo anterior y de la condición 2 que también cumple K', tiene que existir un $x \in K' \setminus K$ tal que se pueda llegar a cualquier elemento de K. Luego, existe v_j tal que $x \in F(v_j)$ y $b_i \in D(v_j)$ y b_i no sería un elemento de la base del grafo reducido $G_r/f\text{-}C$. Por tanto, no puede existir K'. ∎

Teorema. *El conjunto de todos los conjuntos fuertemente compactos de* ζ *es:*

$$\bigcup_{i=1}^{r} \left\{ \bigcup_{v_j \in D(b_i)} F(v_j) \right\}$$

Demostración. Supongamos que existe K' conjunto fuertemente compacto en ζ tal que no existe $b_j \in B$ que cumpla que $K' = \bigcup_{v_t \in D(b_j)} F(v_t)$. Sean $\{P_1,...,P_n\}$ la partición en componentes fuertemente conexas de G. Consideremos los $P_{i_1},...,P_{i_m}$ tal que $P_{i_j} \cap K' \neq \phi \ \forall j = 1,...,m$. Demostremos que $P_{i_j} \subset K' \ \forall j = 1,...,m$. Sea P_{i_j} una componente fuertemente conexa de G con $j \in \{1,...,m\}$. Supongamos que $\exists x \in P_{i_j}$ tal que $x \notin K'$. Como $x \in P_{i_j}$, el grado interior de x debe ser diferente de cero. Pero si esto se cumple significa que $\exists x' \in K'$ tal que x es el más β_0-semejante a x y, en ese caso, K' no sería un conjunto fuertemente compacto. Por lo tanto, $\bigcup_{j=1}^{m} P_{i_j} = K'$. Sean $v_{i_j}, j = 1,...,m$ los vértices del grafo $G_r/f\text{-}C$ correspondientes a $P_{i_j}, j = 1,...,m$. Por condición 2 de la definición de conjunto fuertemente compacto $\exists x' \in K'$ tal que existe un camino entre x' y cualquier elemento de K'. Por tanto, entre el v_{i_t} tal que $x' \in F(v_{i_t})$ y cualquier otro v_{i_t}, $j \neq t$ existe un camino en $G_r/f\text{-}C$. Luego, $v_{i_j}, j = 1,...,m$, $j \neq t$ son descendientes de v_{i_t} en $G_r/f\text{-}C$. Pero, por la suposición, $v_{i_t} \notin B$, lo que significa que v_{i_t} tiene que ser descendiente de un $b_i \in B$ por

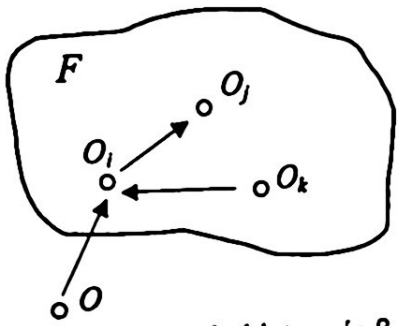
definición de base. Luego, $K' \subset K = \bigcup_{v_j \in D(b_i)} F(v_j)$ y K' no sería fuertemente compacto, pues incumple la tercera condición de la definición. Luego, $v_{i_t} \in B$. ∎

Corolario. *El número de conjuntos fuertemente compactos de ζ es igual al cardinal de la base de G_F/f-C.*

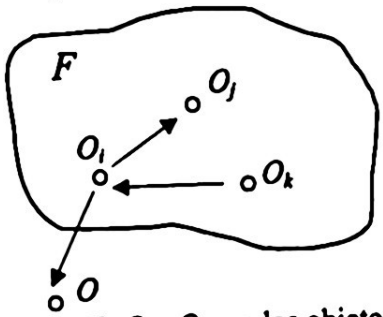
Definición 10. *Sea F un conjunto fuertemente compacto (compacto) de una colección de objetos ζ. Decimos que un objeto O está conectado con F si existe al menos un O'∈ F que cumpla una de las dos condiciones siguientes:*

1. *O' es el objeto más β_0-semejante a O según S.*
2. *O es el objeto más β_0-semejante a O' según S.*

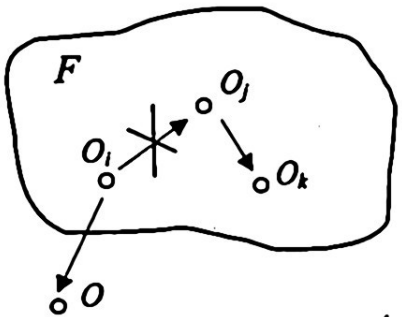
Un objeto O puede conectarse con un conjunto fuertemente compacto (compacto) F de varias formas. La siguiente figura muestra todos los casos que pueden presentarse. En los casos I y II el objeto O no rompe ningún arco del grafo *máx-S* asociado a F. En los casos III y IV sí se rompen arcos.



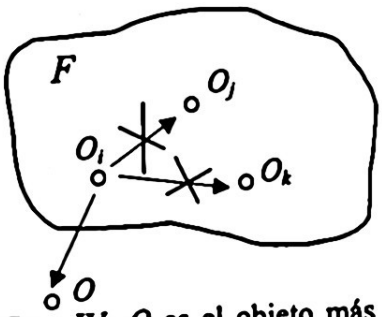
Caso I: O_i es el objeto más β_0-semejante a O.



Caso II: O y O_j son los objetos más β_0-semejantes a O_i.



Caso III: O es el objeto más β_0-semejante a O_i y O_j deja de serlo.



Caso IV: O es el objeto más β_0-semejante a O_i y O_j y O_k dejan de serlo.

Sean $\wp$ el conjunto de todos los conjuntos fuertemente compactos de ζ, $F \in \wp$ un conjunto fuertemente compacto y $G_F=(F,U)$ el grafo *máx-S* asociado a F. Sea, además, O un objeto tal que $O \notin \zeta$.

Proposición 3. *Si O no está conectado con F, entonces F será un conjunto fuertemente compacto de $\zeta \cup \{O\}$.*

Corolario. *Si O es β_0-aislado, entonces el conjunto de todos los conjuntos fuertemente compactos de $\zeta\cup\{O\}$ es $\wp\cup\{\{O\}\}$.*

Proposición 4. *Si O es el objeto más β_0-semejante a un objeto de F y O no rompió ningún arco de G_F, entonces O y todos los objetos de F pertenecerán a un mismo conjunto fuertemente compacto de $\zeta\cup\{O\}$.*

Demostración. Como F es un conjunto fuertemente compacto $\exists x\in F\ \forall y\in F$ $\exists x_{i_1},\ldots,x_{i_q}\in F$ [$x=x_{i1}\wedge\ y=x_{i_q}\wedge\ \forall p<q$ [$\max\limits_{\substack{x_l\in\zeta\\ x_l\neq x_{i_p}}}\{S(x_{i_p},x_l)\}=S(x_{i_p},x_{i_{p+1}})$ $\geq\beta_0$]] . Luego, en particular, existe un camino en G_F de x a O' y, por tanto, de x a O. Por tanto, $F\cup\{O\}$ satisface la condición 2 de la definición de conjunto fuertemente compacto.

Como O es el más β_0-semejante a un objeto $O'\in F$ pueden ocurrir dos cosas:

a) Existe al menos $O''\in F$ tal que O'' es el más β_0-semejante a O y, en ese caso, $F\cup\{O\}$ cumple la condición 1 y también la condición 3 de la definición de conjunto fuertemente compacto. Por tanto, $F\cup\{O\}$ sería un conjunto fuertemente compacto y O y todos los objetos de F estarían en el mismo conjunto fuertemente compacto.

b) Existe al menos $O'''\in\zeta\backslash F$ tal que O''' es el más β_0-semejante a O y, en ese caso, existe F' conjunto fuertemente compacto tal que $F'\supset F\cup\{O\}$. Luego, O y todos los objetos de F estarían en el mismo conjunto fuertemente compacto. ∎

Definición 11. *Un punto de articulación de un grafo no orientado es un vértice v tal que cuando removemos v y todos los arcos incidentes en él, la componente conexa de v se divide en dos o más partes* [2].

Sean G' el grafo *máx-S* asociado a $\zeta\cup\{O\}$, C un conjunto compacto, tal que $C=\bigcup\limits_{F_i\in\wp}F_i$, $G_c=(C,U)$ el grafo *máx-S* asociado a C, NG_c el grafo G_c sin tener en cuenta la orientación y H, el subconjunto de objetos de C que pertenecen a la componente conexa de O en G' sin tener en cuenta la orientación.

Proposición 5. *Si O está conectado con C y O rompe uno y sólo un arco de G_c (caso III), entonces si el conjunto $C\backslash H\neq\emptyset$, $C\backslash H$ es un conjunto compacto de $\zeta\cup\{O\}$.*

Demostración. Sea $O_i\in C$ el objeto al cual O es el más β_0-semejante y tal que se rompe su único arco a que incide al exterior. Si O_i no es un punto de articulación en NG_c, entonces al eliminar a el grafo sigue siendo conexo y, por tanto, $H=C$. Si O_i es un punto de articulación en NG_c, entonces el grafo $G''=<C,U\backslash\{a\}>$ sin tener en cuenta la orientación no es conexo. Si $C\backslash H\neq\emptyset$, entre cualesquiera dos objetos de $C\backslash H$ existe un camino en NG_c porque, de lo contrario, C no sería un conjunto compacto. Como O_i es un punto de articulación en NG_c, no existe en H ningún otro objeto conectado con $C\backslash H$. Por tanto, $C\backslash H$ es una componente conexa en G' sin tener en cuenta la orientación, es decir, $C\backslash H$ es un conjunto compacto en $\zeta\cup\{O\}$. ∎

Note que el único arco que se rompe en el conjunto compacto C se romperá también en los conjuntos fuertemente compactos F_i donde pertenece el objeto O_l.

4. Algoritmo fuertemente compacto incremental

Paso 1. Cada vez que se presenta un nuevo objeto O se calcula la similaridad con todos los objetos de los conjuntos fuertemente compactos existentes.

Paso 2. Se seleccionan los objetos que son más β_0-semejantes a O y aquellos a los que O es el más β_0-semejante. Si no existen tales objetos, en virtud del corolario de la proposición 3, el conjunto $\{O\}$ es un nuevo conjunto fuertemente compacto y terminar.

Paso 3. a) Se construye el conjunto compacto $C(O)$ al que pertenece O, con todos los objetos que pertenecen a la componente conexa de O en el grafo *máx-S* de $\zeta \cup \{O\}$ sin tener en cuenta la orientación (proposición 1). Se coloca el conjunto compacto construido en la lista de compactos a procesar.

b) Estos objetos son eliminados de todos los conjuntos fuertemente compactos a los que pertenecían. Sean estos conjuntos $FC_1,...,FC_m$ los cuales son eliminados de $\wp$.

Paso 4. Para cada objeto del conjunto $K = \left(\bigcup_{i=1}^{n} FC_i \right) \setminus C(O)$:

a) Se construye el conjunto compacto al que él pertenece y se coloca en la lista de compactos a procesar.

b) Se eliminan esos objetos del conjunto K.

Paso 5. Para cada compacto C en la lista de compactos a procesar se construye su cubrimiento en conjuntos fuertemente compactos:

a) Hallar las componentes fuertemente conexas del grafo *máx-S* asociado a C, G_c.

b) Construir el grafo reducido de G_c.

c) Determinar la base B del grafo reducido de G_c, que estará formada por todos los vértices del grafo reducido cuyo grado interior sea nulo.

d) Para cada elemento b_l de B:

i. Construir el conjunto fuertemente compacto F a partir de b_l, esto es, a F pertenecerán todos los vértices de G_c cuyos vértices correspondientes en el grafo reducido sean descendientes de b_l.

ii. Agregar F al conjunto de conjuntos fuertemente compactos $\wp$.

Observación: Conservar la información de cuáles son los conjuntos fuertemente compactos que conforman un compacto resulta de mucha utilidad en el paso 4 dado que en virtud de la proposición 5 no sería necesario procesar a los objetos que pertenecen a conjuntos fuertemente compactos que conforman un compacto donde sólo se perdió un arco. Bastaría con colocar directamente a dicho compacto en la lista de compactos a procesar.

Note que en este algoritmo se aprovecha la propiedad que cumplen los conjuntos fuertemente compactos de ser un cubrimiento de conjuntos compactos.

Por eso, en lugar de hallar los conjuntos fuertemente compactos en todo el conjunto de objetos que modificaron sus arcos producto de la llegada del nuevo objeto se construye el grafo reducido asociado a cada conjunto compacto modificado por separado y se aplica el teorema demostrado anteriormente.

5. Análisis de la complejidad computacional

La complejidad computacional de este algoritmo es $O(n^2)$, pues en el paso 1 cada objeto se compara con todos los existentes. El paso 3 conforma el conjunto compacto (componente conexa) al que el nuevo objeto pertenece. Este paso tiene complejidad $O(e)$, pues se trabaja con listas de adyacencia [4]. Aquí e es la cantidad de aristas del grafo. Aunque desde el punto de vista teórico e es del orden n^2, en la práctica los grafos de máxima semejanza no son grafos completos y lo que ocurre con mucha frecuencia es que la cantidad de objetos más semejantes a uno dado es 1. Por eso, en nuestro caso, hemos estimado experimentalmente que $e=cn$, donde c es la cantidad máxima de objetos más semejantes a uno dado. En el paso 4 se reconstruyen las componentes conexas de los grupos que perdieron arcos, lo cual, es $O(n+e)$. El paso 5 se encarga de construir las componentes fuertemente conexas para, a partir de ellas, conformar los conjuntos fuertemente compactos. Este paso es $O(n+e)$, por lo que no sube la complejidad total del algoritmo. La complejidad espacial (teniendo en cuenta lo anterior) es $O(n)$, pues para cada objeto sólo tendríamos que almacenar el valor de su máxima semejanza y la lista de los objetos más β_0-semejantes, cuyo cardinal nuevamente podría aproximarse por una constante.

6. Experimentación

La efectividad del algoritmo incremental propuesto ha sido evaluada usando una colección de 526 artículos de periódicos publicados en la sección de noticias internacionales del periódico español "El País" durante el mes de junio de 1999. En este tipo de problema, es claro el carácter dinámico de la colección de datos (en este caso, de documentos) y, por tanto, imprescindible la utilización de un algoritmo incremental que identifique los tópicos que son abordados en las noticias. Los documentos, denotados aquí por su código y el título del artículo, fueron representados en función de la frecuencia relativa de los términos que ocurren en dichos documentos y como medida de semejanza se utilizó la medida del coseno, ampliamente usada en este tipo de aplicaciones. Al aplicar el algoritmo se obtuvieron 297 grupos. A continuación mostramos algunos de los grupos formados utilizando como umbral de semejanza 0.25:

Grupo#1: {i101-Un observador ruso se incorpora a la reunión, i100-Los generales yugoslavos se resisten a firmar el plan para la retirada de sus tropas, i50-Calendario del conflicto, i181-Los serbios tienen 11 días para dejar Kosovo}.

Grupo#2: {i103-EE UU teme que se produzca un éxodo de población serbia, i24-Clinton dice que Europa asumirá el peso de la intervención y la reconstrucción de Kosovo, i50-Calendario del conflicto, i46-Clinton deja abiertas todas las opciones,

i181-Los serbios tienen 11 días para dejar Kosovo, i290-Francia se opone a que Europa pague la reconstrucción de los Balcanes}
Grupo#3: {i318-El anuncio oficial de la candidatura de Gore abre la campaña presidencial en Estados Unidos, i103-EE UU teme que se produzca un éxodo de población serbia, i50-Calendario del conflicto, i181-Los serbios tienen 11 días para dejar Kosovo}
Grupo#4: {i50-Calendario del conflicto, i258-El mando de la OTAN deja a los soldados rusos el control del aeropuerto de Pristina, i239-La OTAN confirma la retirada de 11.000 soldados serbios, i305-Los últimos soldados serbios dejan Pristina bajo control de la Kfor, i181-Los serbios tienen 11 días para dejar Kosovo}
Grupo#5: {i181-Los serbios tienen 11 días para dejar Kosovo, i184-Clinton se alegra del avance hacia los 'objetivos' aliados en Kosovo, i50-Calendario del conflicto}
Grupo#6: {i181-Los serbios tienen 11 días para dejar Kosovo, i4-Calendario del conflicto, i50-Calendario del conflicto}
Grupo#7: {i181-Los serbios tienen 11 días para dejar Kosovo, i141-Calendario del conflicto, i50-Calendario del conflicto}
Grupo#8: {i181-Los serbios tienen 11 días para dejar Kosovo, i214-Las tropas aliadas prevén entrar hoy o mañana en Kosovo, i239-La OTAN confirma la retirada de 11.000 soldados serbios, i50-Calendario del conflicto}
Grupo#9: {i181-Los serbios tienen 11 días para dejar Kosovo, i135-Los puntos de la discordia, i50-Calendario del conflicto}
Grupo#10: {i134-Silencio en Kumanovo tras el parón en las conversaciones, i50-Calendario del conflicto, i181-Los serbios tienen 11 días para dejar Kosovo}
Grupo#11: {i181-Los serbios tienen 11 días para dejar Kosovo, i500-España se integrará en el Grupo de países Amigos de Kosovo, i225-Calendarido del conflicto, i50-Calendario del conflicto}

Como puede notarse a estos 11 conjuntos fuertemente compactos pertenecen las noticias i181 e i50. Precisamente estas noticias constituyen holotipos de este conjunto, es decir, son los elementos a los cuales más se parecen el resto de los objetos en el conjunto dado. La base de este grafo es: {i500, i101, i290, i318, i305, i184, i141, i4, i214, i135, i134}. Note cómo cada elemento de la base genera un conjunto fuertemente compacto. Cada elemento de la base pudiera interpretarse como la noticia "primaria" más particular y cada conjunto fuertemente compacto generado por éste pudiera verse como el "desarrollo" de dicha noticia. Hemos observado también que el grado de similaridad aumenta monótonamente de la noticia "primaria" al holotipo.

7. Conclusiones

En este trabajo se presenta un algoritmo incremental de agrupamiento que crea un cubrimiento en conjuntos fuertemente compactos de la colección de objetos. La variante no incremental de este criterio de agrupamiento necesitaba almacenar la matriz de similaridad (de orden n^2) de los objetos de la colección lo que, sin dudas, constituye una gran limitación cuando se desea trabajar con colecciones dinámicas o

grandes de objetos. La variante incremental desarrollada no necesita almacenar esta matriz.

Otra ventaja del algoritmo propuesto es que permite encontrar grupos con formas arbitrarias. Además el agrupamiento obtenido por el algoritmo propuesto es único, es decir, no depende del orden de presentación de los objetos, a diferencia del *algoritmo de las estrellas* [5] que pueden producir diferentes agrupamientos al cambiar el orden de entrada de los objetos. Esta unicidad hay que entenderla de la siguiente manera: dados m objetos la estructuración en conjuntos fuertemente compactos es única, independientemente del orden en que los m objetos sean considerados. Es claro que en una población dinámica, la llegada de un nuevo objeto puede producir nuevas estructuraciones, pero de poblaciones diferentes. Sin embargo, en el caso de la estructuración en fuertemente compactos, una vez que ha sido considerada una población determinada, los objetos que la conforman pudieron haber arribado en cualquier orden produciendo la misma estructuración. El algoritmo propuesto no necesita fijar a priori el número de grupos a obtener, es decir, es aplicable a problemas de clasificación no supervisada libre. La complejidad computacional del algoritmo incremental presentado sigue siendo la misma que la de su variante no incremental, lo que es muy superior a la aplicación reiterada del algoritmo no incremental.

El algoritmo propuesto puede ser utilizado en todas las tareas que requieran del agrupamiento de objetos en fuertemente compactos y del procesamiento dinámico de la información, tales como, por ejemplo, navegación, filtrado, ruteo, detección y seguimiento de noticias en un flujo de documentos; el estudio de la morbilidad de una población; la estructuración de la información para estudios socio-económicos; entre otras aplicaciones. Elementos como la base y los conjuntos fuertemente compactos pueden resultar de mucha utilidad para el analista de información. Es bueno resaltar que este algoritmo puede utilizarse, además, en problemas donde se deseen agrupar objetos de cualquier naturaleza, descritos por rasgos cuantitativos y cualitativos mezclados, incluso con ausencia de información.

Referencias

[1] Martínez-Trinidad, J. F., Ruiz-Shulcloper J., Lazo-Cortés, M. (2000). Structuralization of Universes, *Fuzzy Sets and Systems*, Vol. 112 (3), 485-500.

[2] Aho, A. V. , Hopcroft, J. E., Ullman, J. D. (1983). *Data Structures and Algorithms*, Addison-Wesley Publishing Company.

[3] Kakes, A., Casas, O. (1987). *Teoría de grafos*, Editorial Pueblo y Educación.

[4] Horowitz, E., Sahni, S. (1975). *Fundamentals of Data Structures*, Computer Science Press, Woodland Hills, California.

[5] Aslam, J., Pelekhov, K., Rus, D. (1998). Static and Dynamic Information Organization with Star Clusters, *in Proceedings of the 1998 Conference on Information Knowledge Management*, CIKM'98, Baltimore, MD.