VI SIMPOSIO DE TEORÍA Y APLICACIONES DE MINERÍA DE DATOS (TAMIDA 2013)

Minería de Grafos Multiobjetivo usando Algoritmos Evolutivos Aplicación al Análisis de Mapas Visuales de la Ciencia

Oscar Cordón

ocordon@decsai.ugr.es, oscar.cordon@softcomputing.es

Contenidos

1. Introducción
2. Diseño y Análisis de Cienciogramas
3. Minería de Grafos y Algoritmos Evolutivos Multiobjetivo
4. Aplicación de la Minería de Grafos al Análisis de Cienciogramas
5. Minería de Grafos Multiobjetivo
6. Conclusiones y Agradecimientos

1. Introducción

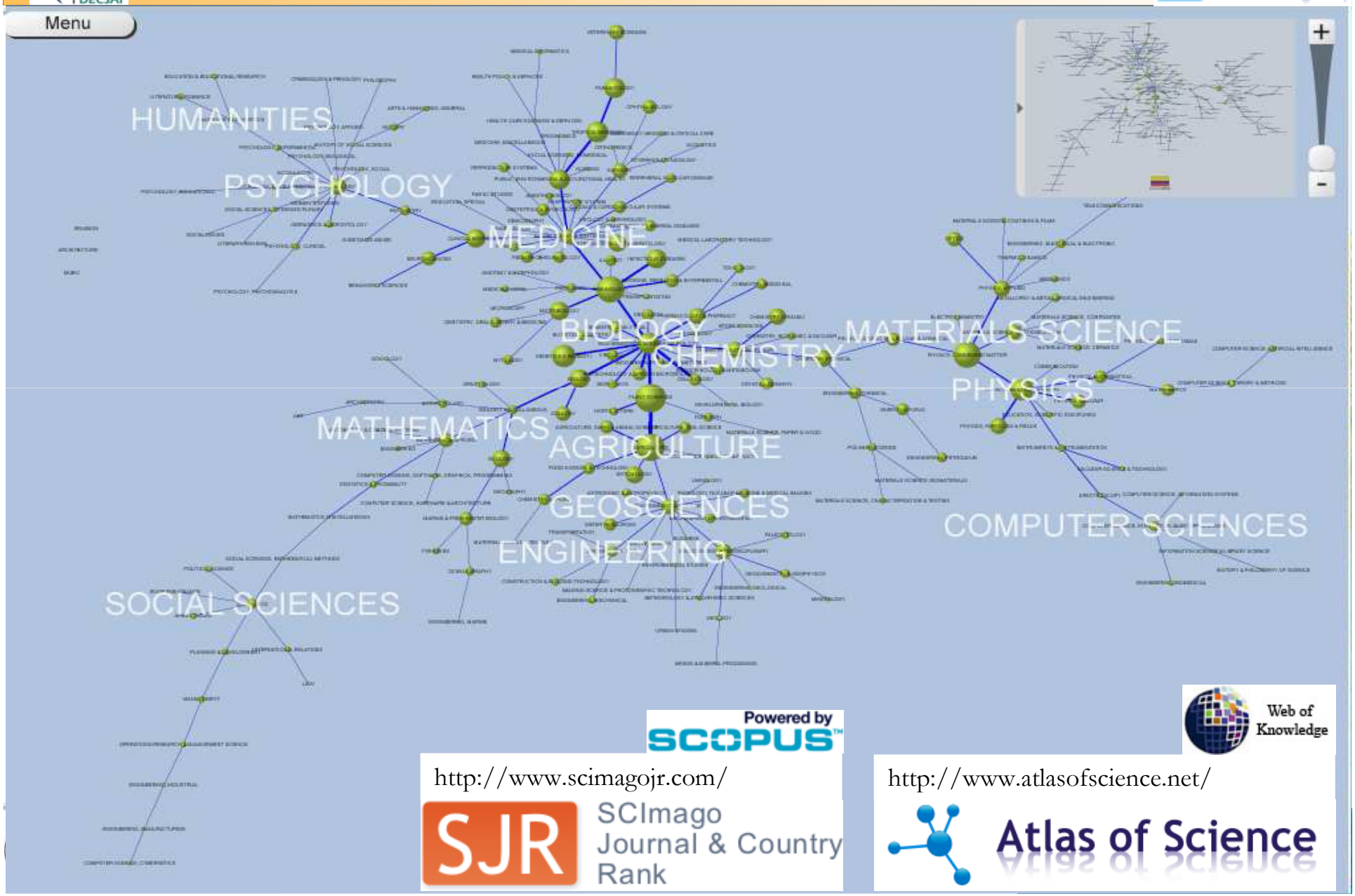
- Los **cienciogramas** son modelos de representación visual del estado de la Ciencia en un determinado dominio
- El diseño de cienciogramas es un procedimiento fuertemente interdisciplinar: *Data Mining/Knowledge Discovery/Big Data*, *Information Visualization*, *Social Networks*, etc.
- La **comparación y análisis automático de conjuntos de cienciogramas** es una tarea de alta dificultad
- Nuestra propuesta se basa en emplear herramientas de Minería de Grafos (GBDM), mono- y multiobjetivo, para ayudar al profesional de la información automatizando esta tarea

2. Diseño y Análisis de Cienciogramas

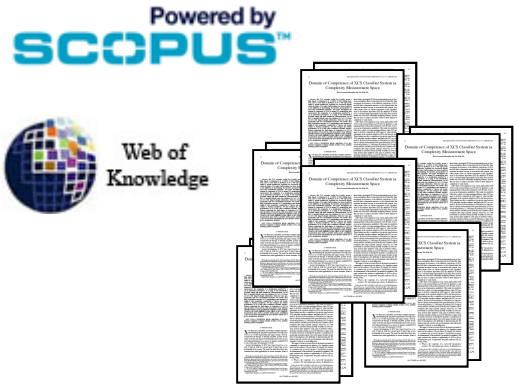
- Un cienciograma es una representación visual (mapa) del estado de la investigación científica de un dominio concreto (institución, región, país, continente o nivel mundial) en un instante de tiempo concreto

- Aplicaciones:
  - Análisis de la estructura de campos científicos y frentes de investigación
  - Desarrollo de interfaces visuales de recuperación de información
  - Representación de la evolución de la producción científica en dominios institucionales/de conocimiento
  - Evaluación cualitativa de la producción científica de instituciones/ regiones/países y del impacto de políticas científicas

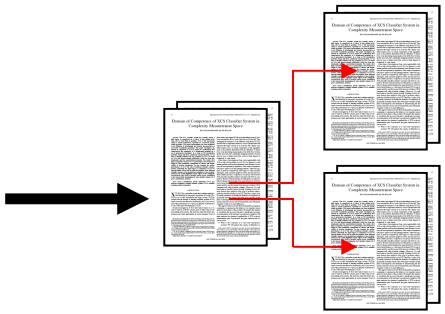
Ejemplo de Cienciogramas (SRI visual)



Diseño de Cienciogramas



Base de Datos de producción científica



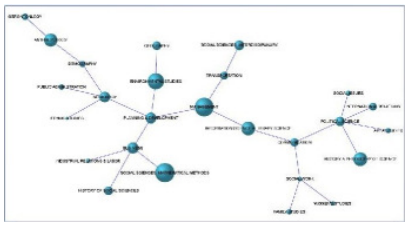
- Definición de las unidades de análisis
- Extracción de las relaciones de co-cit
- Filtrado

$$CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}}$$

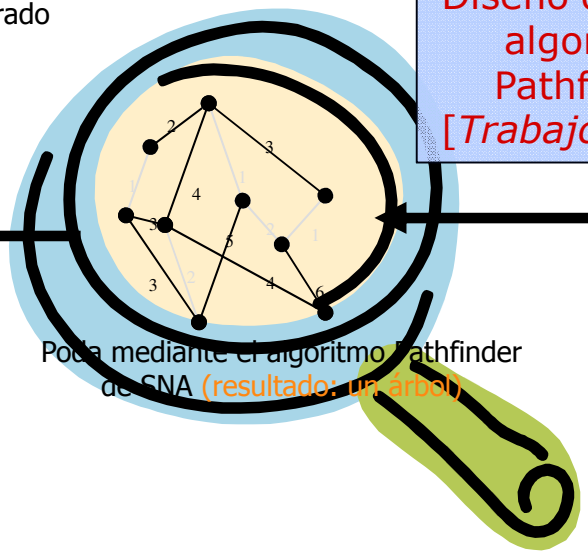
	A	B	C	D	E
A	8.2	2.0	9.7	1.4	8.7
B	1.1	5.9	8.5	9.4	6.2
C	9.8	3.3	0.9	2.2	7.3
D	9.3	4.4	8.0	6.8	6.4
E	2.7	7.5	4.0	3.5	1.0

Creación de una matriz de co-citación

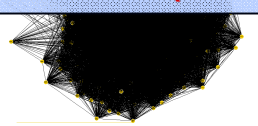
Diseño de variantes eficientes del algoritmo Pathfinder: Fast-Pathfinder y MST-Pathfinder [*Trabajos de Quirin et al. (2009)*]



Visualización con algoritmo de dibujo de grafos (Kamada-Kawai)
(Interfaz gráfico en SVG)



Poda mediante el algoritmo Pathfinder de SNA (resultado: un árbol)



Representación como grafo ponderado
(**red social**)

Análisis de Cienciogramas

- El análisis de cienciogramas permite realizar tareas como:
  - Explorar automáticamente las características de un dominio
  - Comparar dominios (pej. diferencias entre países o en el tiempo)

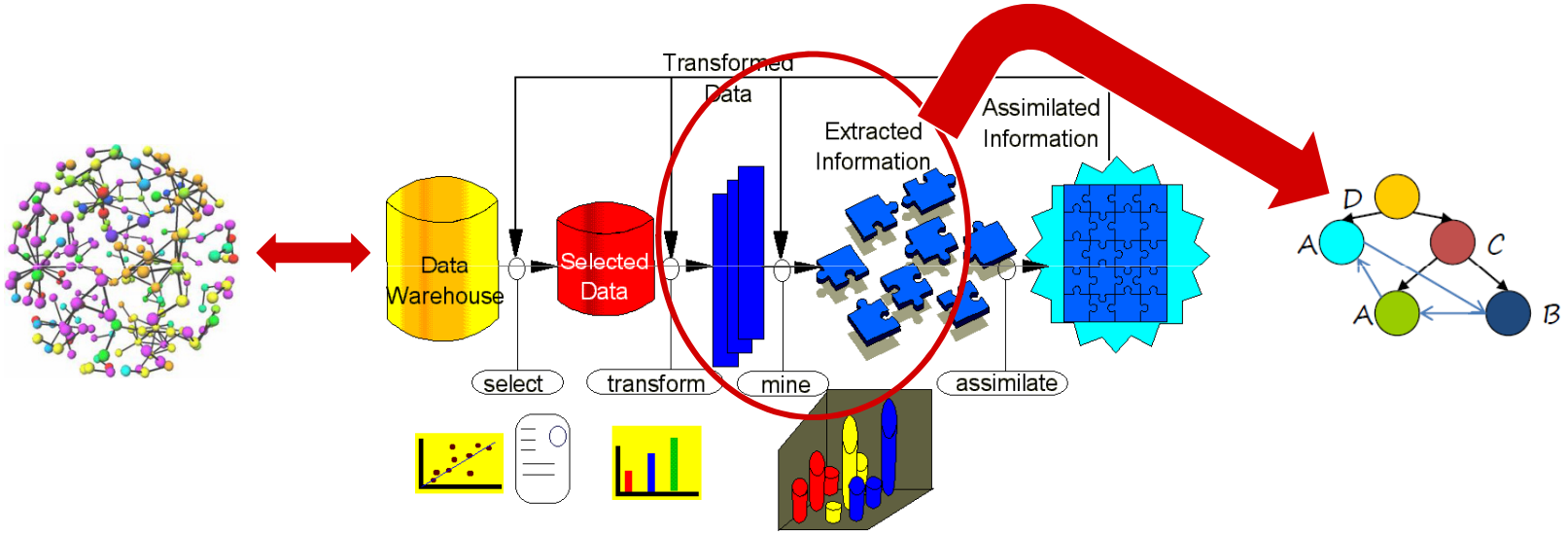
- Posibles aplicaciones:
  - Evolución temporal del dominio científico de un país concreto
  - Comparación de los dominios científicos de distintos países,
  - etc.

- **Las técnicas existentes sólo pueden manejar un único o muy pocos cienciogramas a la vez**

- Hemos diseñado una metodología para analizar automáticamente conjuntos grandes de cienciogramas usando métodos de GBDM

3. Minería de Grafos y Algoritmos Evolutivos

- La Minería de Grafos (GBDM) es un área de la Minería de Datos basada en la extracción de conocimiento (DMKD) de bases de datos estructuradas en relaciones (i.e. datos no planos)



- La GBDM y la Minería de Redes Sociales son 2 de las 10 retos actuales de la DMKD [*Yang et al., 10 Challenging Problems in DM Research, IJITDM 5:4 (2006) 597-604*]

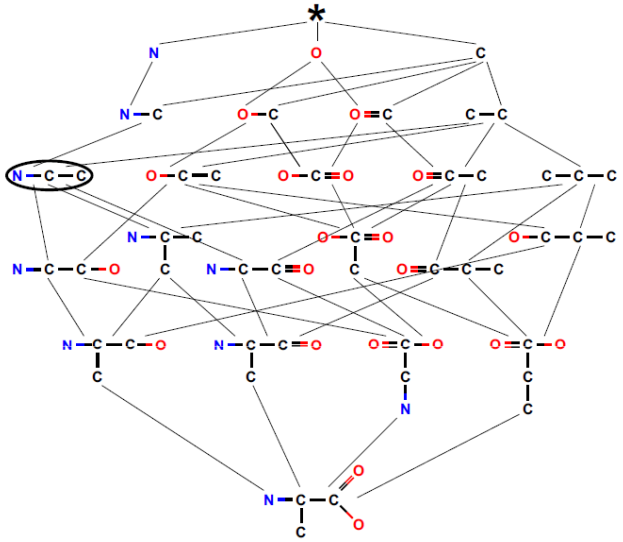
Tareas y Aplicaciones de la Minería de Grafos

- Posibles tareas de DMKD en Bases de Datos de grafos:
  - Identificar patrones repetitivos en una BD estructurada en forma de grafo (minería de subgrafos frecuentes)
  - Encontrar grupos de grafos similares (clustering)
  - Construir modelos predictivos para los grafos (clasificación) …

- Campos de aplicación:
  - -ÓMICA (Proteómica, Genómica, etc.): Descubrimiento de motivos en genes y proteínas, Redes biológicas, Compuestos químicos, …
  - Dinámica de fluidos
  - Análisis de redes sociales
  - Redes de telecomunicaciones
  - Web semántica, …

El Algoritmo Subdue (1)

- Los algoritmos de GBDM se basan en efectuar una búsqueda (exacta o aproximada) en el espacio (**retículo**) de todos los subgrafos posibles

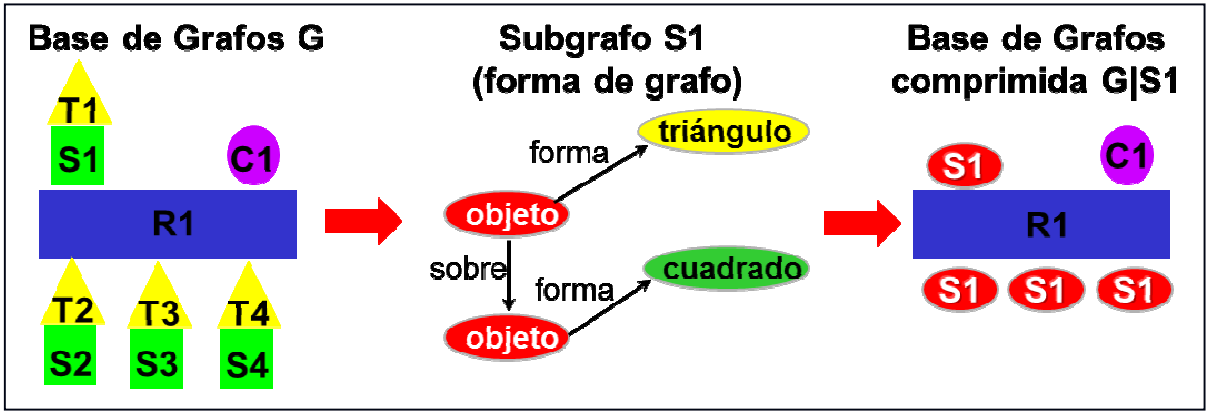


- Subdue fue la primera propuesta de algoritmo GBDM y es una de las más extendidas [*Cook and Holder, IEEE Intelligent Systems 15 (2000) 32–41*]

- Explora el espacio con una técnica de búsqueda de Inteligencia Artificial (Beam search) guiada por un criterio de Minimum Description Length (MDL)

El Algoritmo Subdue (2)

- Se busca extraer el subgrafo S que mejor comprime un grafo de entrada G, sustituyendo todas las ocurrencias de S en G por un solo nodo (G|S)



$$\text{value}_{\text{MDLi}}(S,G) = \frac{I(G)}{I(S)+I(G\mid S)}$$

- MDL es una combinación de la frecuencia y el tamaño de G
- Con esta formulación, es capaz de extraer subgrafos frecuentes en una Base de Datos de grafos (entre otras muchas tareas de DMKD)
- Aplicaciones: circuitos CAD, compuestos químicos, pathways metabólicos, análisis de imágenes, detección de ciber-crímenes, …

El Algoritmo Subdue (3)

Evaluación de un subgrafo:

- Se persigue minimizar la "longitud de descripción" (*description length*) de los datos comprimidos (i.e. subgrafo + BD comprimida)
- Para ello, se considera la maximización de la MDL:

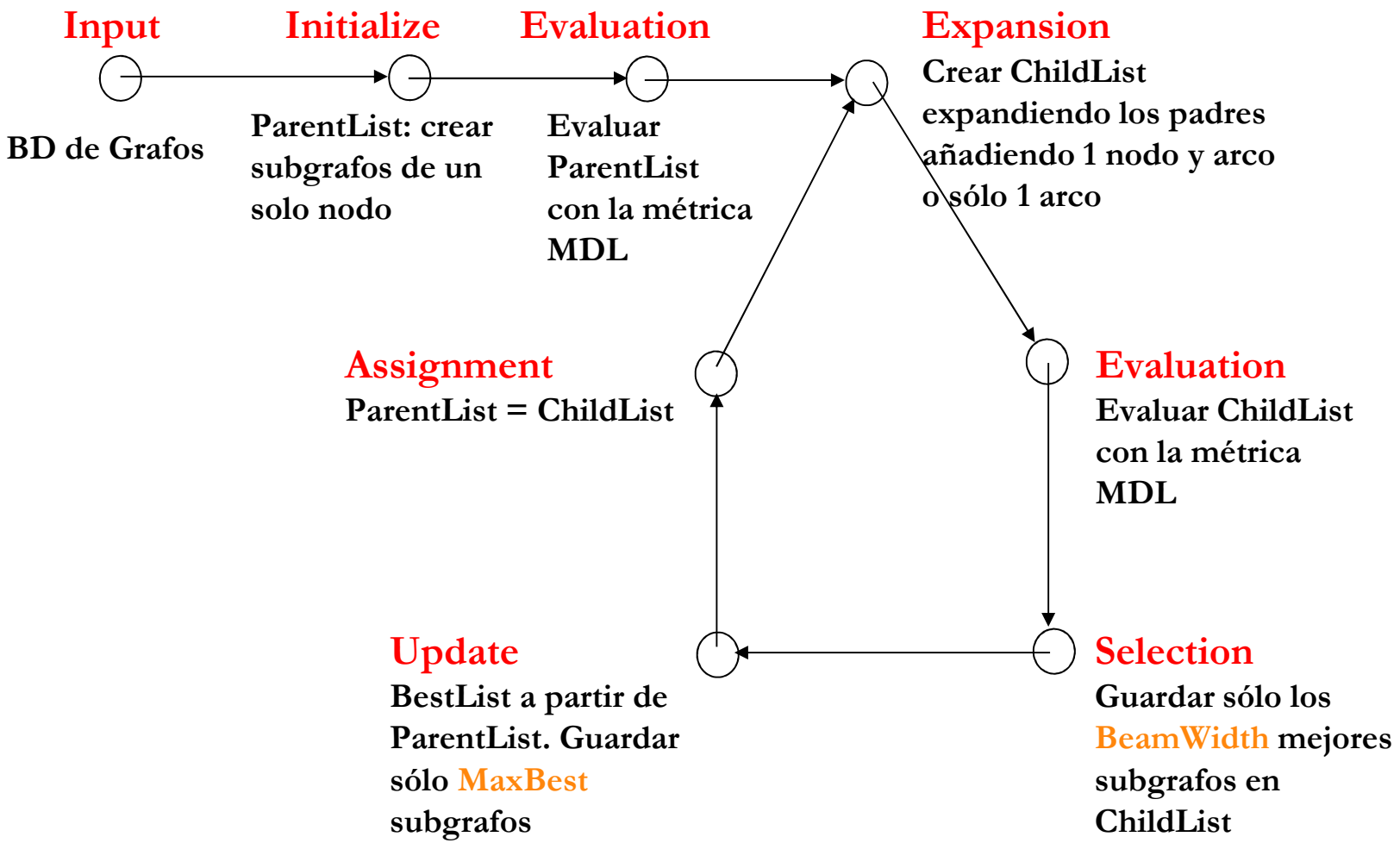
$$\text{value}_{\text{MDLi}}(S,G) = \frac{I(G)}{I(S) + I(G \mid S)}$$

  - *G*: grafo de entrada
  - *S*: subgrafo candidato
  - *I(G)*: número de bits necesarios para codificar G
  - *I(G|S)*: número de bits necesarios para codificar la compresión de *G* usando *S*

El Algoritmo Subdue (4)

```
1.  Subdue(Graph, BeamWidth, MaxBest, MaxSubSize, Limit)
2.      ParentList = {Vertex v | v has a unique label in Graph}
3.      Evaluate each vertex in ParentList
4.      ChildList = {}
5.      BestList = {}
6.      ProcessedSubs = 0
7.      WHILE ProcessedSubs ≤ Limit and ParentList ≠ ∅ DO
8.          WHILE ParentList ≠ ∅ DO
9.              Parent = RemoveHead(ParentList)
10.             CandidateList = ExtendSubstructure(Parent)
11.             FOR EACH Child ∈ CandidateList DO
12.                 IF SizeOf(Child) ≤ MaxSubSize THEN
13.                     Evaluate the Child
14.                     Insert Child in ChildList in order by value
15.                     ChildList = ChildList mod BeamWidth
16.             ProcessedSubs = ProcessedSubs+1
17.             Insert Parent in BestList in order by value
18.             BestList = BestList mod MaxBest
19.         Switch ParentList and ChildList
20.     Return BestList
```

El Algoritmo Subdue (5)



Algoritmos Evolutivos Multiobjetivo

- Muchos problemas reales se caracterizan por la existencia de múltiples medidas de actuación que deben ser optimizadas simultáneamente

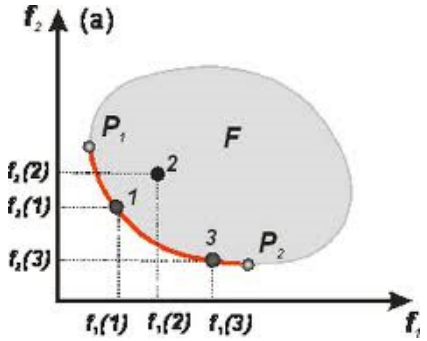
¡EL MUNDO REAL ES MULTIOBJETIVO!

- Ejemplo: Diseño de un sistema de control de aire acondicionado, minimizando consumo y maximizando estabilidad y confort del usuario

- En estos casos, no existe una única solución óptima. Hay un conjunto de soluciones con distintos equilibrios en la satisfacción de los objetivos y que son igualmente preferibles entre sí (Conjunto de Pareto)

- Los métodos de resolución son mucho más complejos. Se persigue proporcionar diversas opciones optimales al experto

- Los algoritmos evolutivos (AEs) son una de las mejores técnicas existentes para problemas multiobjetivo



4. Aplicación de la Minería de Grafos al Análisis de Cienciogramas

- Propuesta de una Metodología Automática de Análisis y Comparación de Cienciogramas basada en técnicas de GBDM

- Basada en formular tareas de análisis y comparación de cienciogramas como problemas de GBDM aprovechando que este tipo de mapas visuales de la ciencia son redes sociales (i.e. grafos)

- La extracción automática de subgrafos comunes (subestructuras de categorías científicas comunes (CRCSs)) a distintos cienciogramas puede proporcionar información muy útil al experto para explorar las características de los dominios científicos representados

A. Quirin, O. Cordón, B. Vargas-Quesada, F. de Moya-Anegón, Graph-based Data Mining: A New Tool for the Analysis and Comparison of Scientific Domains Represented as Scientograms, Journal of Informetrics 4:3 (2010) 291-312. FI : 3.119. Cat: INFORMATION SCIENCE & LIBRARY SCIENCE. Order: 3/76. Q1

4. Aplicación de la Minería de Grafos al Análisis de Cienciogramas

- La metodología permite extraer (de forma automática)…
  - ¿Qué subestructuras aparecen en el dominio analizado?
  - ¿Cuándo (en qué año)?
  - ¿Cuántas son/ Cuán grandes son?
  - ¿Dónde están localizadas en el mapa?
- Con ello, posibilita al experto el…
  - Analizar la composición de las propias estructuras (categorías, …)
  - Obtener estadísticas globales
  - Realizar comparaciones de los dominios y la dinámica de los mismos
  - Analizar la importancia del dominio objetivo en el tiempo
- Tres tareas concretas de análisis de cienciogramas:
  - Evolución temporal del dominio científico de un país concreto
  - Identificación de CRCSs en un conjunto de países o a nivel mundial
  - Comparación de los dominios científicos de distintos países

Evolución temporal del dominio científico de un país concreto

Objetivo:

- Determinar qué subestructuras aparecen en un instante de tiempo concreto en el dominio analizado
- ¿Cómo son de grandes?, ¿cuántas son?, ¿dónde están localizadas?, etc.
- Análisis en profundidad de las CRCSs descubiertas, tipos de categorías que relacionan, etc.
- Estadísticas globales sobre el tamaño y la cantidad de estas subestructuras para caracterizar la importancia de la evolución del dominio y su dinámica, respectivamente

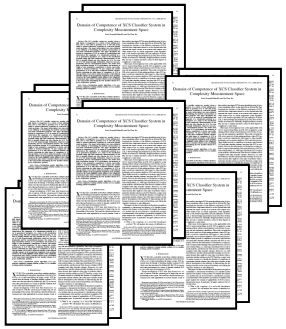
Configuración:

- Se escoge un dominio científico: por ejemplo, un país concreto
- Se determinan dos rangos de años, el rango negativo y el rango positivo
- La tarea de GBDM implicar extraer subgrafos presentes en los mapas del rango positivo y que no existan en los mapas del rango negativo

Evolución temporal del dominio científico de un país concreto

1. Construcción de los cienciogramas

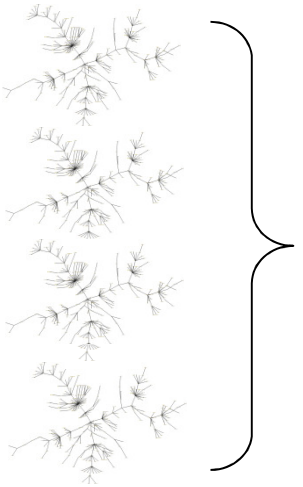
Dado un país y un año concreto…

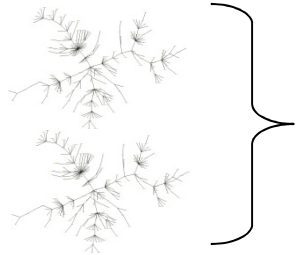


A. Extraer la información de co-citaciones de los documentos científicos

B. Representarla como red social de categorías y podar el grafo

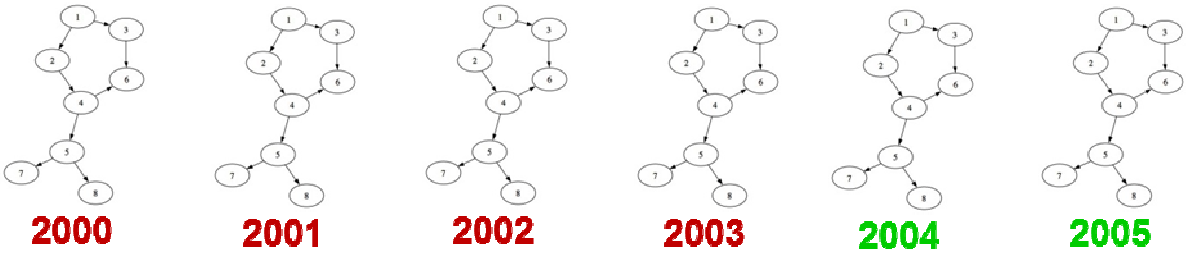
2. Selección del conjunto de mapas y de los periodos de análisis





3. Uso de un método de GBDM (Subdue) para identificar las subestructuras comunes (CRCSs) que aparecen en los años positivos y no en los negativos

Modo de trabajo alternativo de Subdue:
uso de grafos positivos/negativos



- Puede verse como el ratio de compresión obtenido si se comprime la BD positiva y no se comprime la BD negativa
- Criterio a maximizar:

$$\text{value}_{\text{MDLi}}(S, Gp, Gn) = \frac{I(Gp) + I(Gn)}{I(S) + I(Gp \mid S) + I(Gn) - I(Gn \mid S)}$$

Identificación de subestructuras en un conjunto de países

Objetivo:

- Determinar qué subestructuras de investigación (CRCSs) se repiten en la producción científica de un número significativo de países en un año concreto
- ¿Cómo son de grandes?, ¿cuántas son?, ¿dónde están localizadas?, etc.
- ¿Influyen las características de los países (nivel de desarrollo, situación geográfica, políticas científicas, etc.)?

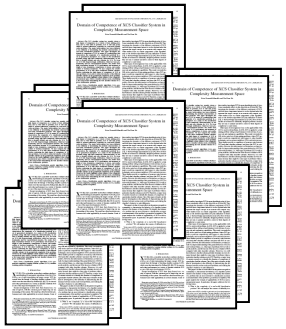
Configuración:

- Se seleccionan un número grande de países
- Se determina el año o rango de años concretos de producción científica concretos para ser representada en los cienciogramas
- La tarea de GBDM implica la extracción de subgrafos específicos, frecuentes y grandes

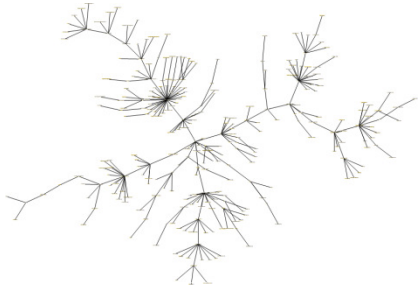
Identificación de subestructuras en un conjunto de países

1. Construcción de los cienciogramas

Dados un(os) año(s) concreto y varios países…

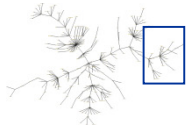


A. Extraer la información de co-citaciones de los documentos científicos



B. Representarla como red social de categorías y podar el grafo

2. Selección del conjunto de mapas



España

Francia

Reino Unido

Alemania

Italia

Portugal

3. Uso de un método de GBDM (Subdue) para identificar las subestructuras comunes (CRCSs) que aparecen más frecuentemente en los cienciogramas de los distintos países

Experimentos de identificación de subestructuras en un conjunto de 73 países

List of the 73 countries contained in the DB. For 60 of them, the data for the ten years period between 1996 and 2005 is available. If that is not the case, the number of available years are indicated in parenthesis.

Countries		
Algeria (4)	Hungary	Puerto Rico (2)
Argentina	Iceland (1)	Republic of Korea
Armenia (1)	India	Romania
Australia	Indonesia (6)	Russian Federation
Austria	Ireland	Saudi Arabia
Bangladesh (7)	Islamic Republic of Iran	Singapore
Belarus	Israel	Slovakia
Belgium	[illegible]	Slovenia
Brazil	[illegible]	South Africa
Bulgaria	Jordan (8)	Spain
Canada	[illegible]	Sweden
Chile	[illegible]	Switzerland
China	Lebanon (3)	Taiwan
Colombia	[illegible]	Thailand
Croatia	[illegible]	Tunisia
Cuba	Mexico	Turkey
Czech Republic	Morocco	Ukraine
Denmark	Netherlands	United Arab Emirates (4)
Egypt	New Zealand	United Kingdom
Estonia	Nigeria	United States
Finland	Norway	Uruguay (1)
France	Pakistan	Venezuela
Germany	Philippines (4)	Viet Nam (3)
Greece	Poland	
Hong Kong	Portugal	

36M documentos
159135 vértices
172081 aristas

Experimentos de identificación de subestructuras en un conjunto de 73 países

Support and size of the substructures extracted for the 2005 world CRCSs case study.

19253 vértices; 19709 aristas

Support (positive)	#Subs.	Size (nodes)			Size (edges)		
		min	max	avg	min	max	avg
10	3	12	12	12	11	11	11
11	1	12	12	12	11	11	11
12	1	11	11	11	10	10	10
13	1	11	11	11	10	10	10
14	4	10	11	10.5	9	10	9.5
15	2	10	10	10	9	9	9
16	4	9	9	9	8	8	8
17	1	8	8	8	7	7	7
18	2	8	8	8	7	7	7
19	1	7	7	7	6	6	6
20	2	7	8	7.5	6	7	6.5
23	1	7	7	7	6	6	6
24	1	7	7	7	6	6	6
28	1	6	6	6	5	5	5
31	1	6	6	6	5	5	5
32	1	6	6	6	5	5	5
38	1	6	6	6	5	5	5
39	1	5	5	5	4	4	4
48	1	5	5	5	4	4	4
50	1	5	5	5	4	4	4
55	2	2	3	2.5	1	2	1.5
58	4	3	5	4	2	4	3
59	1	4	4	4	3	3	3
62	1	2	2	2	1	1	1
68	1	2	2	2	1	1	1
70	1	4	4	4	3	3	3
71	2	3	3	3	2	2	2
73	1	2	2	2	1	1	1
Total	**44**			**7.0**			**6.0**

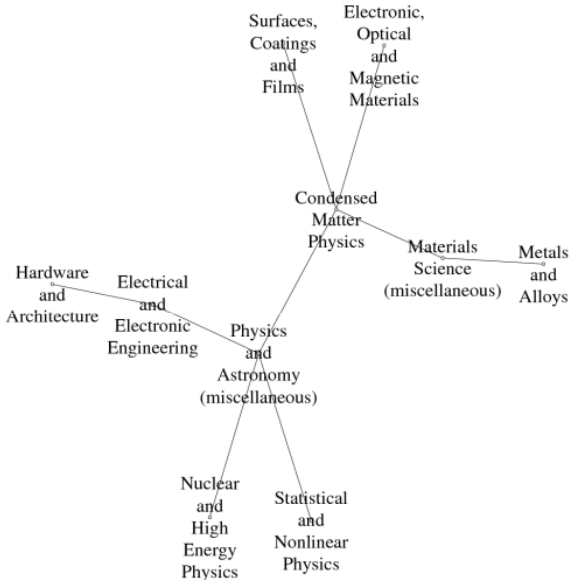
Experimentos de identificación de subestructuras en un conjunto de 73 países

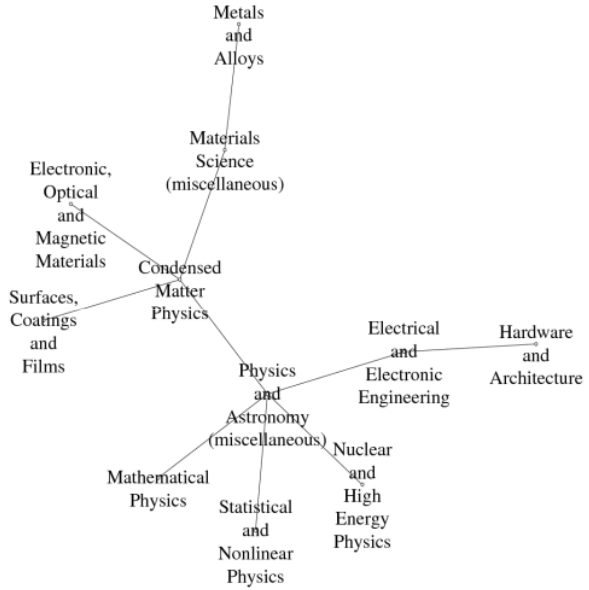
Detailed statistics about the CRCSs uncovered in the world scientific domain case study in 2005.

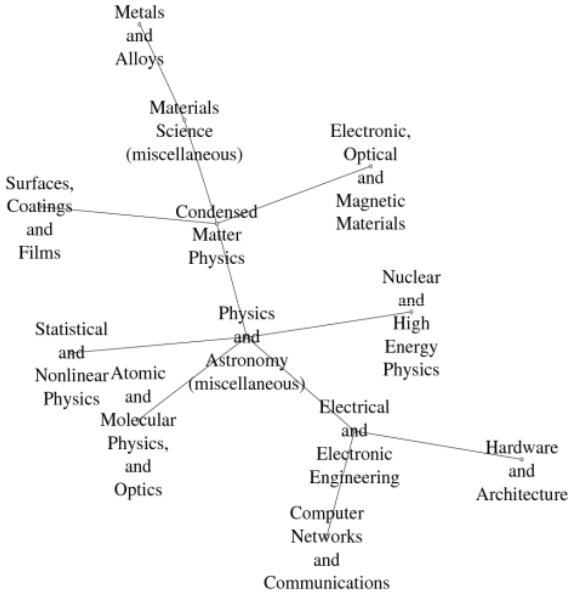
Index	20	21	30	31	18	19	25	35	37	38	39	40
Substructure statistics												
Support	15	15	14	14	14	14	13	12	11	10	10	10
Size (nodes)	10	10	10	10	11	11	11	11	12	12	12	12
Size (edges)	9	9	9	9	10	10	10	10	11	11	11	11
Substructure repartition within the countries												
Niger				YES								
Lithuania	YES	YES		YES			YES	YES				
Belgium	YES			YES		YES	YES	YES				
Jordan	YES			YES		YES	YES	YES				
India	YES			YES		YES	YES	YES				
Morocco	YES			YES		YES	YES	YES				
Croatia		YES	YES		YES							
Mexico		YES	YES		YES				YES	YES	YES	
Austria		YES	YES		YES				YES	YES	YES	
Hungary		YES	YES		YES				YES	YES	YES	
Slovakia	YES	YES	YES		YES	YES						YES
Viet Nam	YES	YES	YES	YES	YES	YES	YES		YES			YES
Finland	YES	YES	YES		YES	YES		YES		YES	YES	YES
Ukraine	YES	YES	YES	YES	YES	YES	YES		YES	YES		YES
Bulgaria	YES	YES	YES	YES	YES	YES	YES	YES	YES		YES	YES
Czech Republic	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Algeria	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Turkey	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Poland	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Pakistan	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Experimentos de identificación de subestructuras en un conjunto de 73 países

CRCSs de mayor tamaño extraídas de la BD de la producción científica de los 73 países en 2005







Support	15	13	10
Size (nodes)	10	11	12

Experimentos de identificación de subestructuras en un conjunto de 73 países

Algunas CRCSs de tamaño medio extraídas de la BD de la producción científica de los 73 países en 2005

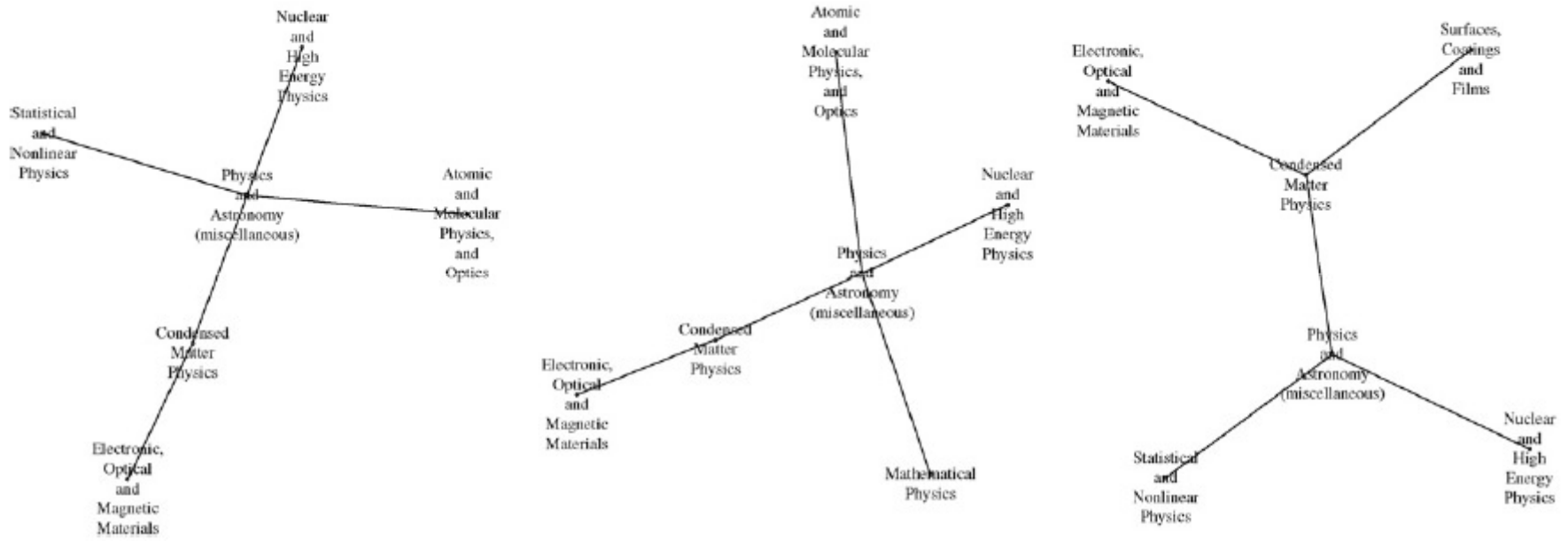


Fig. 8. Some medium-sized CRCSs extracted from the world scientific research in 2005: left: index: #5, support: 38, size: 6 nodes, 5 edges; center: index: #9, support: 32, size: 6 nodes, 5 edges; right: index: #11, support: 31, size: 6 nodes, 5 edges.

5. Minería de Grafos Multiobjetivo

- En una tarea habitual de GBDM se devuelven subgrafos con:
  - un umbral de frecuencia (soporte) mínimo (al menos *s* grafos), o
  - un umbral de tamaño (complejidad) mínimo (al menos *n* vértices)

- Estos objetivos están en conflicto porque:
  - Los subgrafos más frecuentes suelen estar asociados a descubrimientos más obvios, que proporcionan poca información al usuario
  - Los subgrafos más complejos suelen representar descubrimientos más significativos pero tienen poco soporte, con lo que no son demasiado útiles

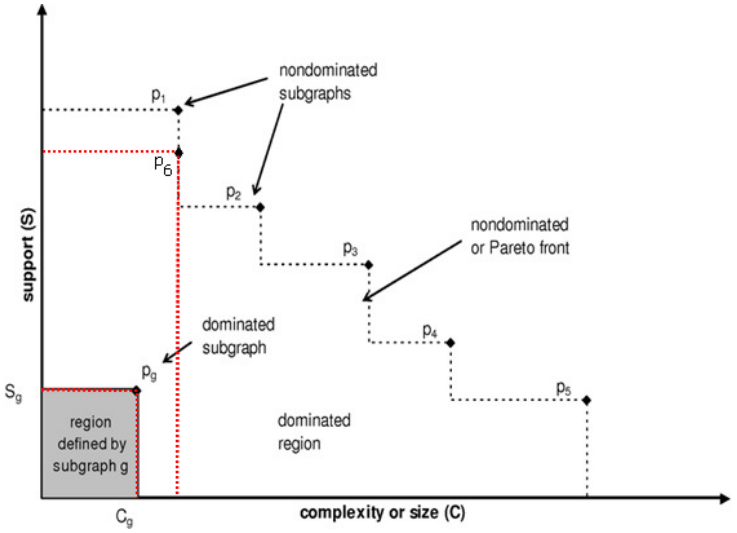
- Éste es un concepto habitual en DMKD (reglas de asociación, identificación de patrones frecuentes, etc.): a mayor complejidad de la unidad de información extraída, menor soporte y viceversa

5. Minería de Grafos Multiobjetivo

- Usando AEs multiobjetivo es posible optimizar simultáneamente los dos criterios para obtener subgrafos con distintos equilibrios entre ambos (conjunto de soluciones Pareto-optimales):

- Se usan los dos objetivos para evaluar cada subgrafo g:
  - Max Soporte (G,g) = #subgrafos de G que se emparejan con g
  - Max Complejidad (G,g) = #vértices(g) + #aristas(g)

- Devuelve un conjunto completo de soluciones, dando información de mejor calidad (un número mayor de subgrafos de mayor interés) al experto



- p_1 domina a p_g (mejor en S y C)
- p_1 domina a p_6 (igual en C pero mejor en S)
- p_1 no domina a p_2
- p_1 y p_2 están en la frontera Pareto-optimal
- p_3, p_4, p_5 también son no-dominadas

5. Minería de Grafos Multiobjetivo

- Hemos diseñado tanto una variante multiobjetivo de Subdue basada en componentes de AEs multiobjetivo:

  [*Shelokar et al., MOSubdue: A Pareto Dominance-based Multiobjective Subdue Algorithm for Frequent Subgraph Mining. Knowledge and Information Systems 34:1 (2013) 75-108*]

  como AEs multiobjetivo puros para GBDM:

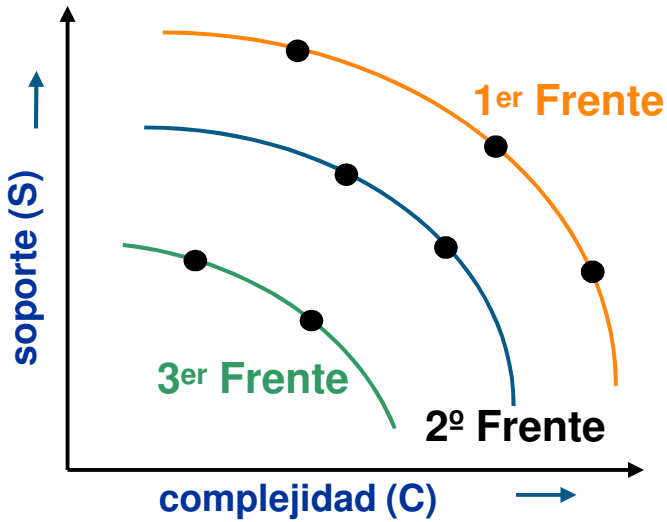
  [*Shelokar et al., A Multiobjective Evolutionary Programming Framework for Graph-based Data Mining. Information Sciences 273:1 (2013) 118–136*]

  [*Shelokar et al., Three-Objective Subgraph Mining using Multiobjective Evolutionary Programming. Journal of Computer and System Sciences (2013), en prensa*]

- Los métodos son flexibles. Se pueden aplicar a distintas tareas de GBDM MO personalizando las funciones objetivo y pueden considerar más de dos objetivos

Subdue Multiobjetivo (MOSubdue) (1)

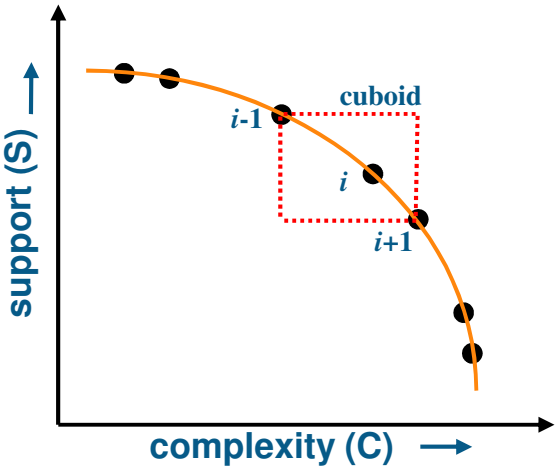
- Algoritmo GBDM multiobjetivo basado en incluir componentes de Optimización Evolutiva Multiobjetivo (EMO), NSGA-II, en Subdue
- Se escogen los objetivos para evaluar cada subgrafo g: Soporte, orden, tamaño, densidad, etc.
- Se asigna fitness a cada subgrafo candidato usando el *non-dominated sorting* de NSGA-II*:
  - Las soluciones del primer frente no-dominado tienen el fitness más alto (rank 1)
  - Todas las soluciones del mismo frente tienen el mismo fitness (rank)
  - Las mejores soluciones tienen rank 1, las de rank 2 son las segundas mejores y así⁰



* **K. Deb et al. *IEEE TEC*, vol. 6, pp. 182–197, 2002**

MOSubdue (2)

- Se seleccionan los *BeamWidth* mejores subgrafos candidatos considerando el *crowding distance procedure* de NSGA-II*
- Los subgrafos de mayor diversidad son los de mayores valores de *crowding distance*
- Uso de un archivo externo de Pareto, actualizado en cada iteración



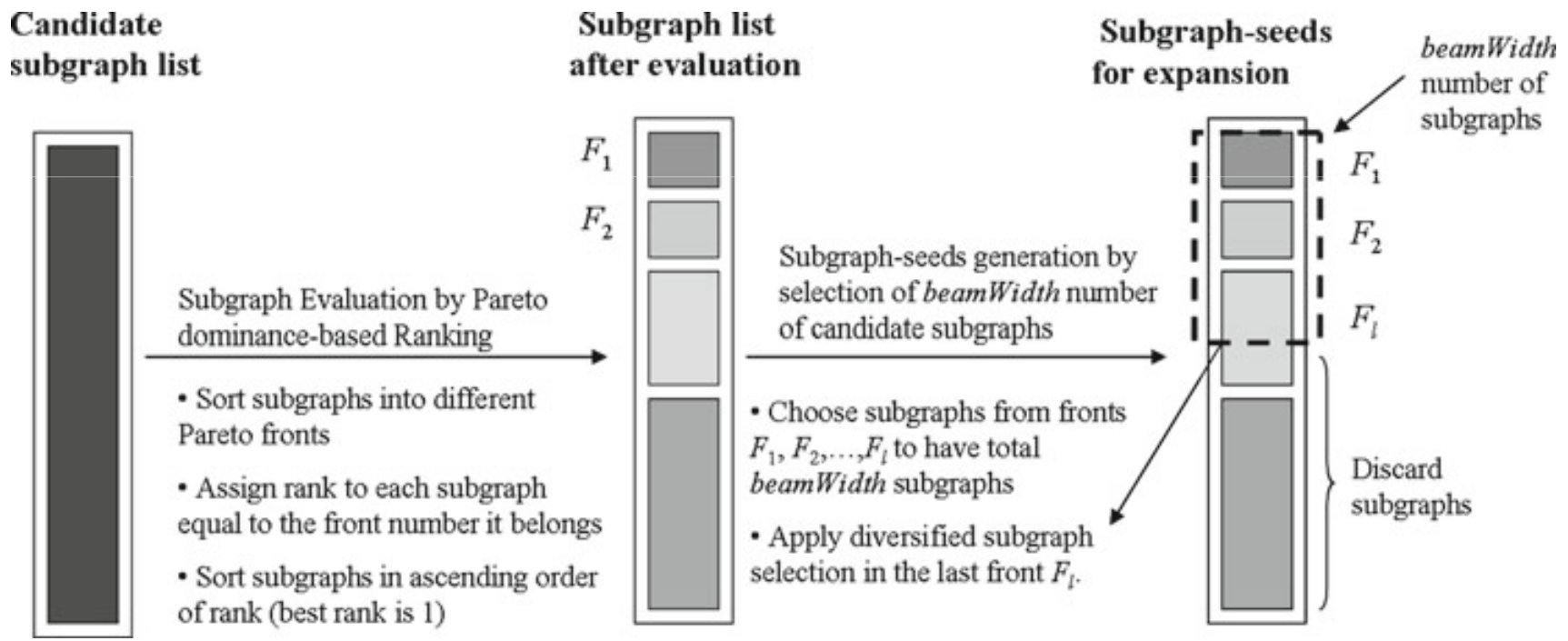
Búsqueda *Beam Search* Multiobjetivo basada en componentes EMO (NSGA-II)

* **K. Deb et al. *IEEE TEC*, vol. 6, pp. 182–197, 2002**

MOSubdue (3)

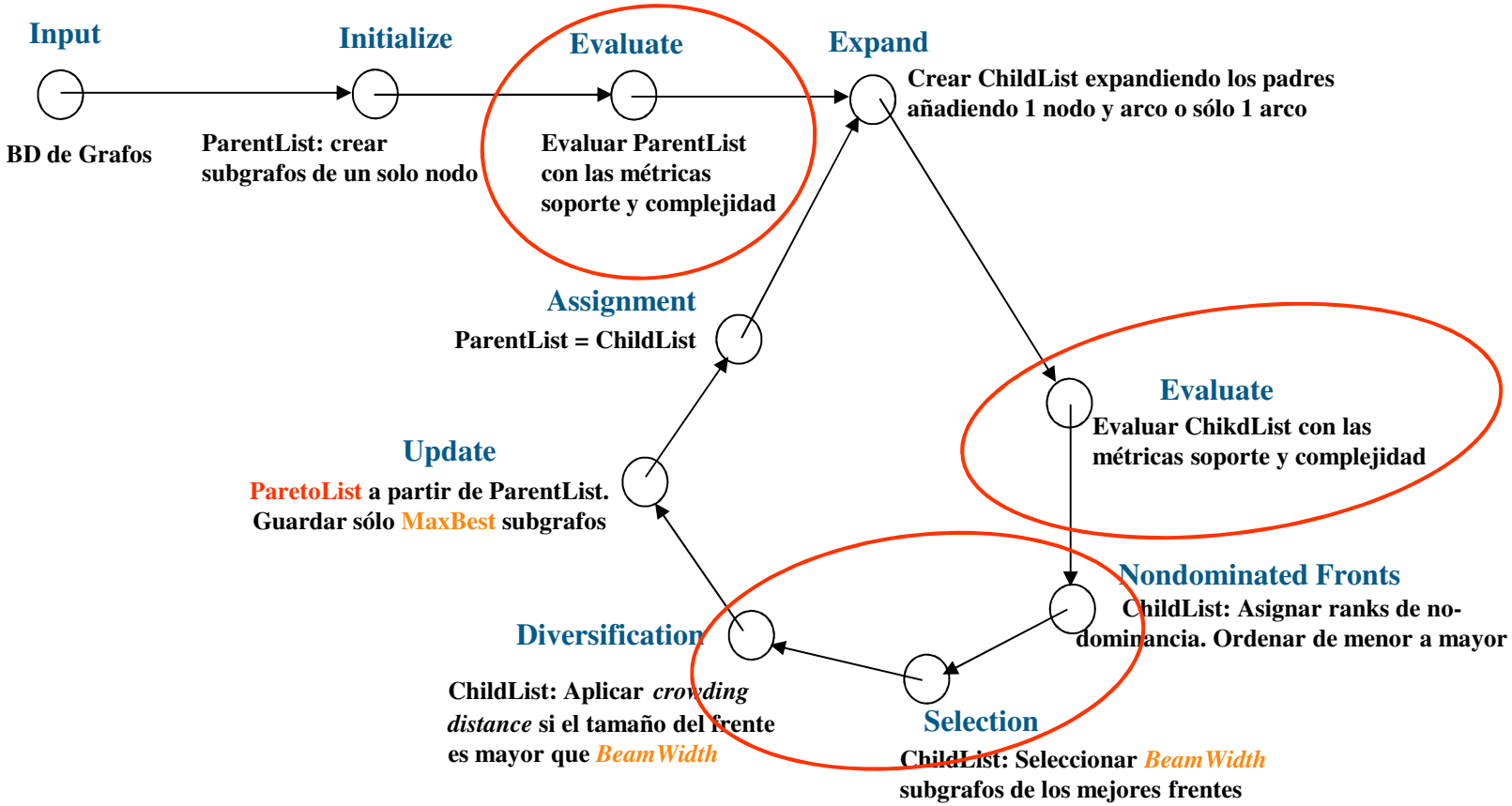
MOSubdue:

Búsqueda *Beam Search* Multiobjetivo basada en componentes EMO (NSGA-II)



MOSubdue (4)

P. Shelokar, A. Quirin, O. Cordón, MOSubdue: A Pareto Dominance-based Multiobjective Subdue Algorithm for Frequent Subgraph Mining. Knowledge and Information Systems 34:1 (2013) 75-108



Experimentos (1)

Descripción de las BDs y los objetivos empleados:

Datasets	#Graphs	#Nodes	#Edges	#Unique labels
Chemical	340	9,189	9,317	66
Scientograms	73	19,253	19,709	296

- *Chemical Compound Data*: base de datos con 340 compuestos químicos de la *Predictive Toxicology Evaluation* (PTE) *challenge*
- *Scientograms*: Cienciogramas de la producción mundial (73 países) en 2005. Identificación de CRCSs
- Dos objetivos: Max Soporte y Max Orden (número de nodos)

Experimentos (2)

Algoritmos de comparación:

- Subdue-I: Subdue clásico mono-objetivo para generar un Pareto de subgrafos. Determinístico. 3 ejecuciones con 3 objetivos distintos (MDL, soporte y complejidad) combinadas.
- Subdue-II: Subdue mono-objetivo con una función de evaluación multicriterio basada en una combinación con pesos. Determinístico. 11 ejecuciones con diferentes vectores de pesos combinadas
- MOSubdue-I: MOSubdue con *non-dominated sorting*. Det. 1 ej.
- MOSubdue-II: MOSubdue con *nondominated sorting* y *crowding-distance selection*. Probabilístico. 10 ejecuciones aleatorias
- MOGaston: Variante MO del algoritmo clásico de GBDM Gaston con un archivo de Pareto externo. Búsqueda exhaustiva basada en umbrales, **inabordable** en BDs grandes. Determinístico. 1 ej.

Experimentos (3)

Parámetros:

- BeamWidth = 5, 10 y 20
- MaxBest = MaxParetoSubs = 100 (tamaño del Archivo de Pareto)
- Criterio de parada = hasta que se vacíe la ParentList
- Subdue-II: 11 pesos distintos (0,1) a (1,0) en pasos de 0.1
- MOGaston: Tres tiempos de ejecución distintos:
  - El tiempo empleado por la mejor variante de Subdue
  - Dos veces el tiempo anterior
  - Cinco veces el tiempo anterior

Indicadores de Rendimiento (Métricas) Multiobjetivo:

- Hipervolumen (HVR) (unaria) y Cobertura (C) (binaria)

Experimentos (4)

Comparación métrica HVR y tiempos (Chemical Compound):
(a mayor valor, mejor rendimiento)

Methods	*beamWidth*
	10
Subdue-I	0.9392 (–)
Subdue-II	0.3009 (0.3129)
MOSubdue-I	**0.9898** (–)
MOSubdue-II	0.9662 (0.0012)
Subdue-I	79.2 (–)
Subdue-II	15.69 (4.87)
MOSubdue-I	**49.47** (–)
MOSubdue-II	40.64 (0.49)

Notación: valor (desviación típica). Tiempos de ejecución en segundos

Experimentos (5)

Comparación métrica HVR y tiempos (Scientograms):
(a mayor valor, mejor rendimiento)

Methods	*beamWidth*
	10
Subdue-I	0.7990 (–)
Subdue-II	0.1052 (0.0242)
MOSubdue-I	0.8491 (–)
MOSubdue-II	**0.8968 (0.0000)**
Subdue-I	1,289.62 (–)
Subdue-II	99.38 (32.27)
MOSubdue-I	684.77 (–)
MOSubdue-II	**484.65 (4.14)**

Notación: valor (desviación típica). Tiempos de ejecución en segundos

Experimentos (6)

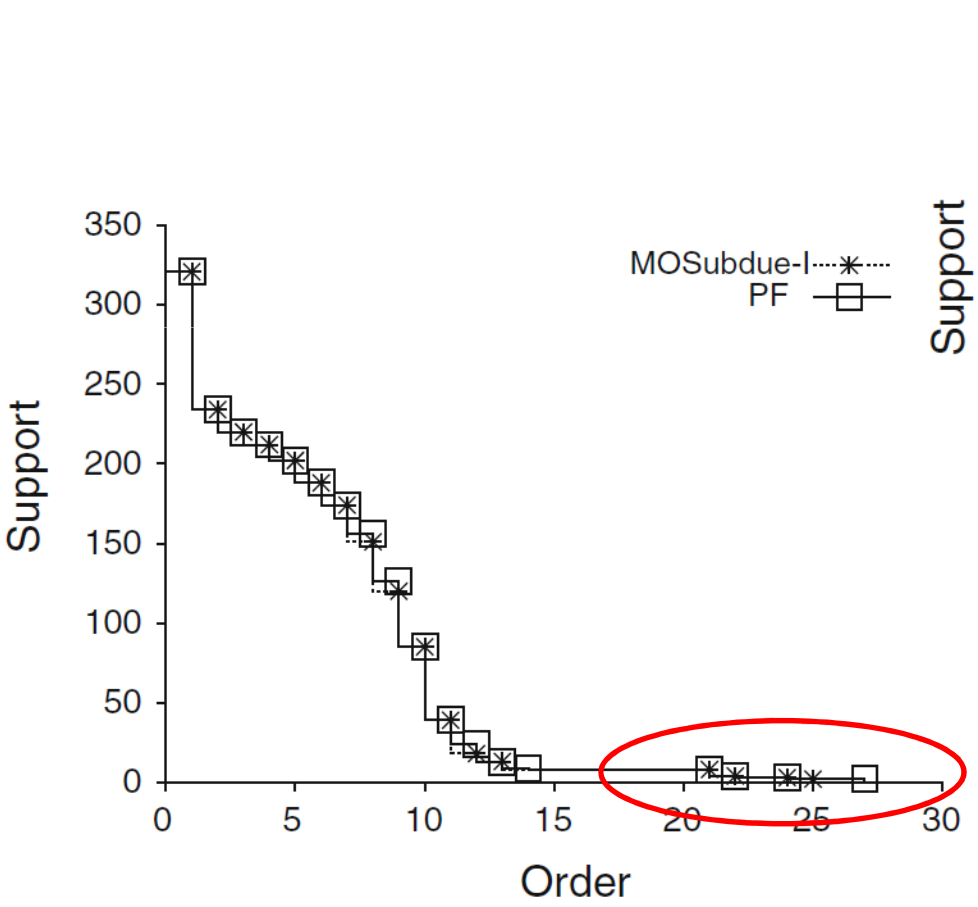
Resultados de MOGaston en HVR en ambos conjuntos:
(a mayor valor, mejor rendimiento)

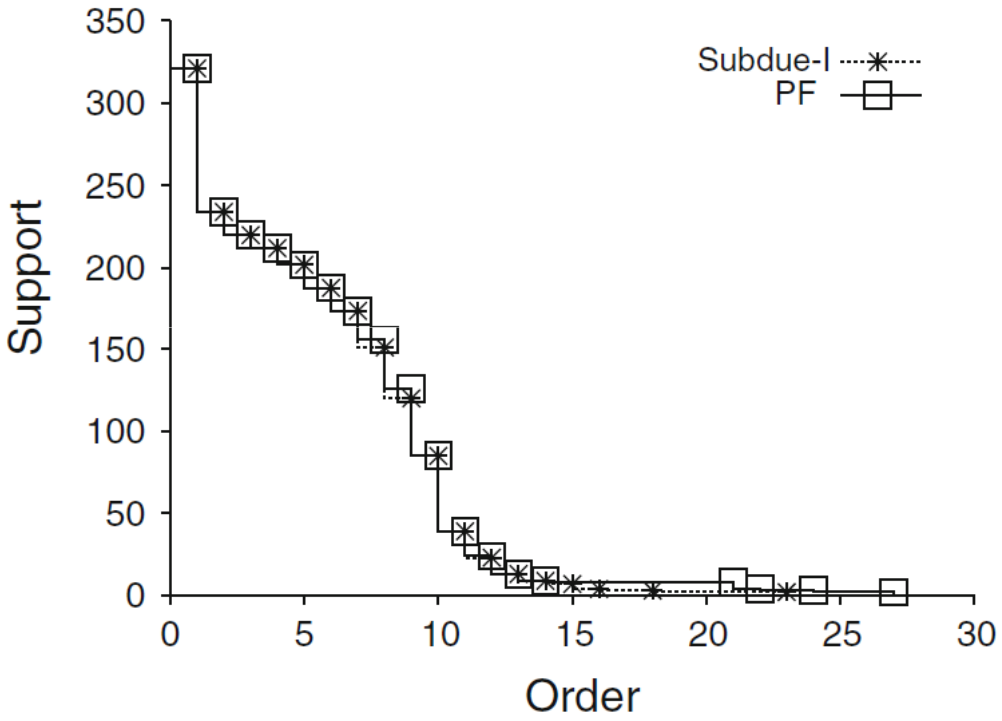
Dataset	Run 1	Run 2	Run 3
Chemical	0.0583 [50]	0.0583 [100]	0.0612 [250]
Scientogram	0.0746 [485]	0.0762 [970]	0.0762 [2,425]

Notación: entre corchetes, el tiempo de ejecución (segundos)

Experimentos (7)

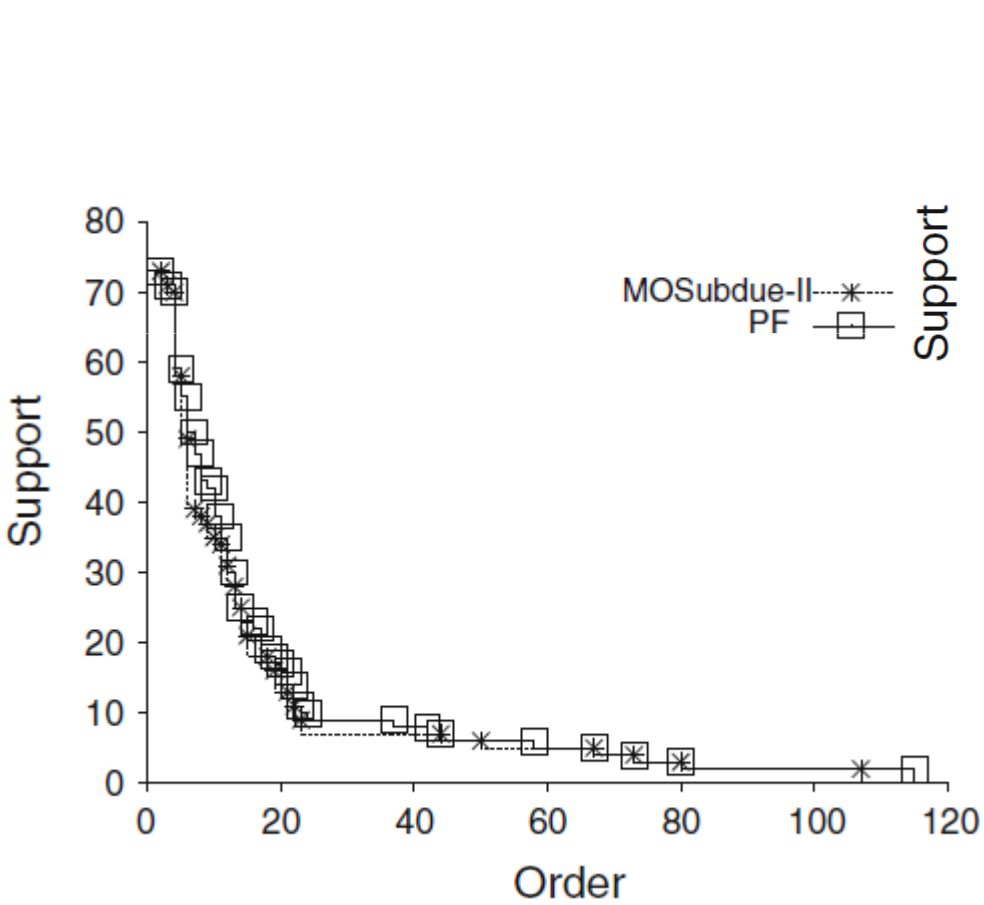
Comparativa visual de los conjuntos de Pareto generados Subdue-I vs. MOSubdue-I (Chemical Compound):

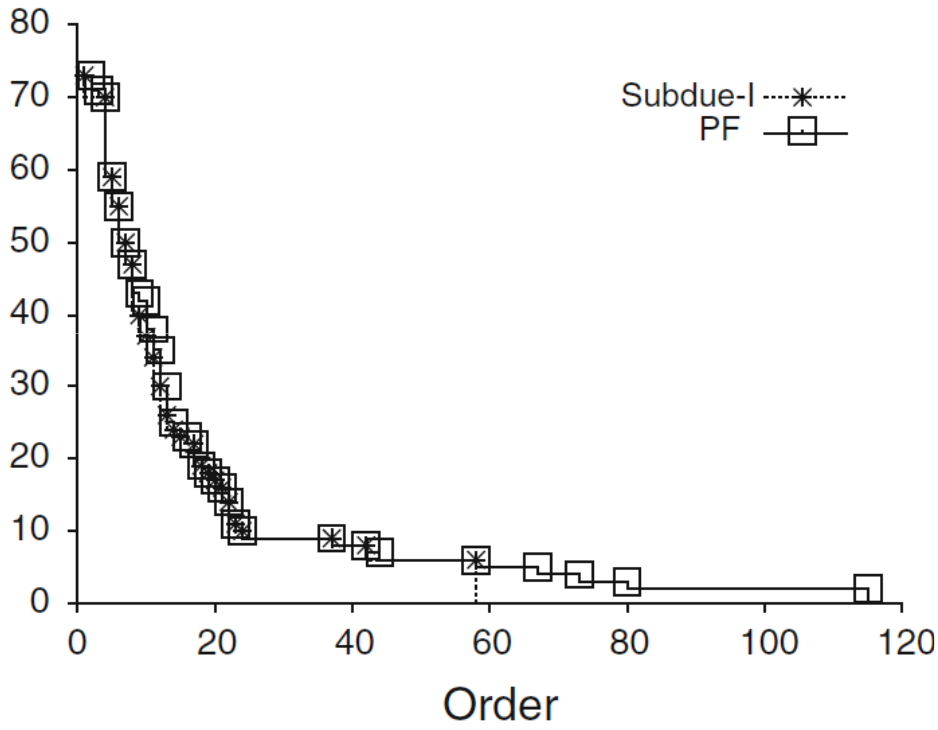




Experimentos (8)

Comparativa visual de los conjuntos de Pareto generados Subdue-I vs. MOSubdue-II (Scientograms):





Experimentos (8)

Segunda experimentación: BDs y objetivos empleados:

Dataset	#Graphs	#Nodes	#Edges	#Unique Labels
*gd*01	62	15539	16354	276
*gd*02	67	16921	17531	276
*gd*03	69	17633	18144	279
*gd*04	70	18184	18597	291
*gd*05	73	19253	19709	295

- Cada conjunto de datos contiene los cienciogramas de los países disponibles para un año concreto entre 2001 y 2005
- P.e. la base *gd01* corresponde al año 2001 y comprende 62 cienciogramas de otros tantos países con un número total de 15539 nodos, 16354 aristas y 276 etiquetas únicas (*gd05*=Scientograms)
- Dos objetivos: Max Soporte y Max Complejidad (#nodos+#aristas)

Experimentos (9)

Comparación métrica HVR:
(a mayor valor, mejor rendimiento)

Dataset	Subdue-I	Subdue-II	MOSubdue-I	MOSubdue-II
*gd*01	0.6978(-)	0.0273(0.00)	0.7482(-)	**0.8445(0.06)**
*gd*02	0.6286(-)	0.3589(0.11)	0.8028(-)	**0.8185(0.03)**
*gd*03	0.6347(-)	0.0228(0.00)	0.8641(-)	**0.8717(0.02)**
*gd*04	0.6902(-)	0.0257(0.00)	0.8616(-)	**0.9083(0.01)**
*gd*05	0.6449(-)	0.0295(0.00)	0.7339(-)	**0.7763(0.02)**
average	0.6592(0.59)	0.0928(0.09)	0.8021(0.46)	**0.8438(0.59)**

Notación: valor (desviación típica)

Experimentos (10)

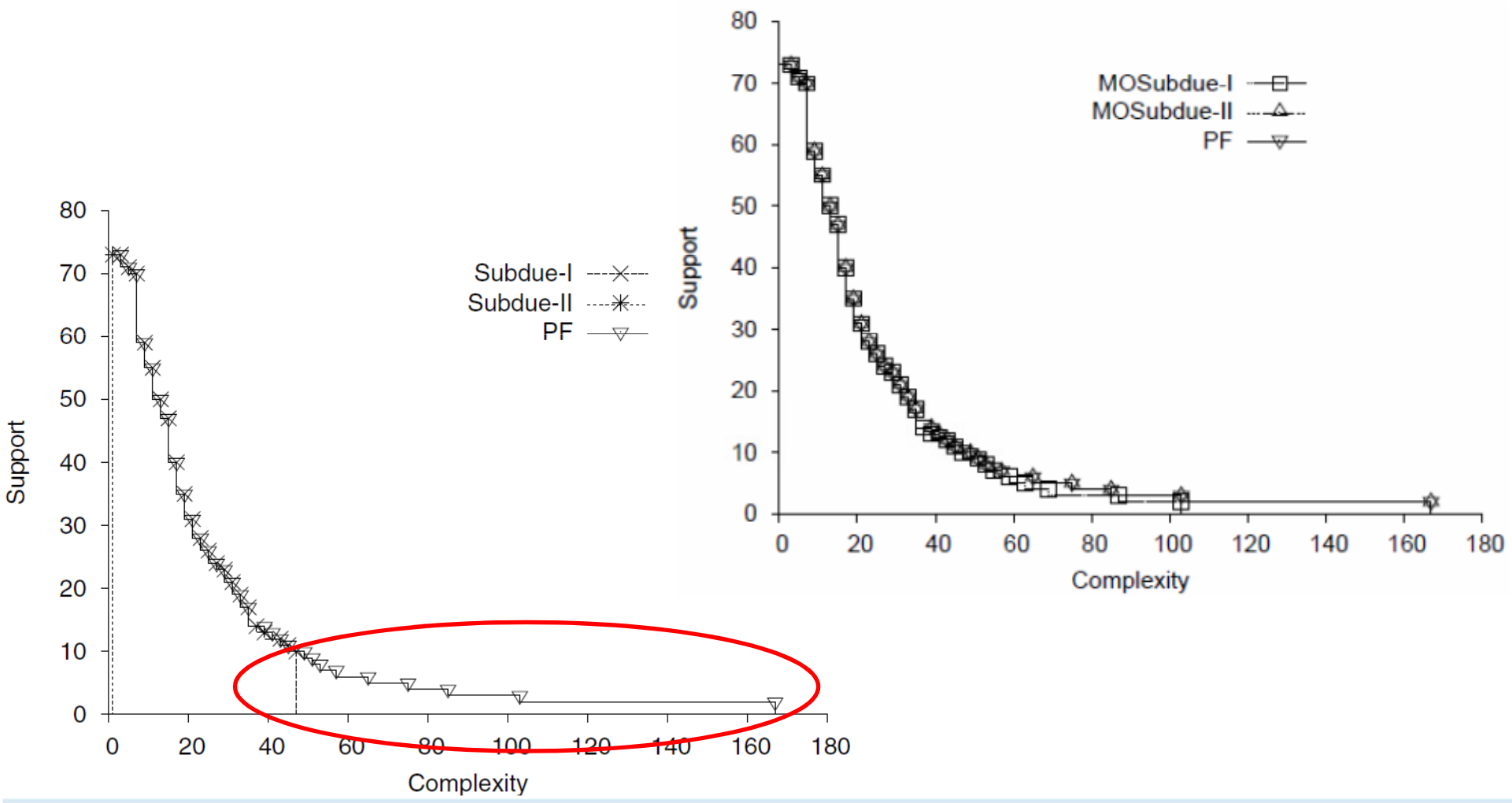
Comparación métrica C: (a mayor valor, mejor rendimiento)

- Box-plots con los valores de C
- La métrica C es tanto mejor cuanto más tienda a 1
- C(A,B)=1 indica que la aproximación de Pareto del algoritmo A cubre todas las soluciones de la de B

Cada rectángulo contiene 5 box-plots que representan la distribución de los valores de C para una pareja de algoritmos; el de más a la izquierda corresponde al conjunto *gd*01 y el de más a la derecha al *gd*05

Experimentos (11)

Comparativa visual de los conjuntos de Pareto generados (*gd05*):



Experimentos (12)

Comparación de las CRCSs extraídas en *gd05*-Scientograms

#support	#Complexity (#nodes+#edges)	MOSubdue-I #subs	Subdue-I #subs
2	103	1	-
3	87	1	-
4	69	1	-
5	63	1	-
6	59	1	-
7	55	2	-
8	53	2	-
9	51	1	-
10	47	9	5
11	45	4	4
12	43	1	1
13	39	2	2
14	37	15	5
17	35	1	1
19	33	2	2
21	31	1	1
23	29	1	1
24	27	3	3
26	25	1	1
28	23	2	2
31	21	1	1
35	19	1	1
40	17	2	2
47	15	1	1
50	13	1	1
55	11	1	1
59	9	1	1
70	7	1	1
71	5	2	2
73	3	2	1
total subgraphs		64	40

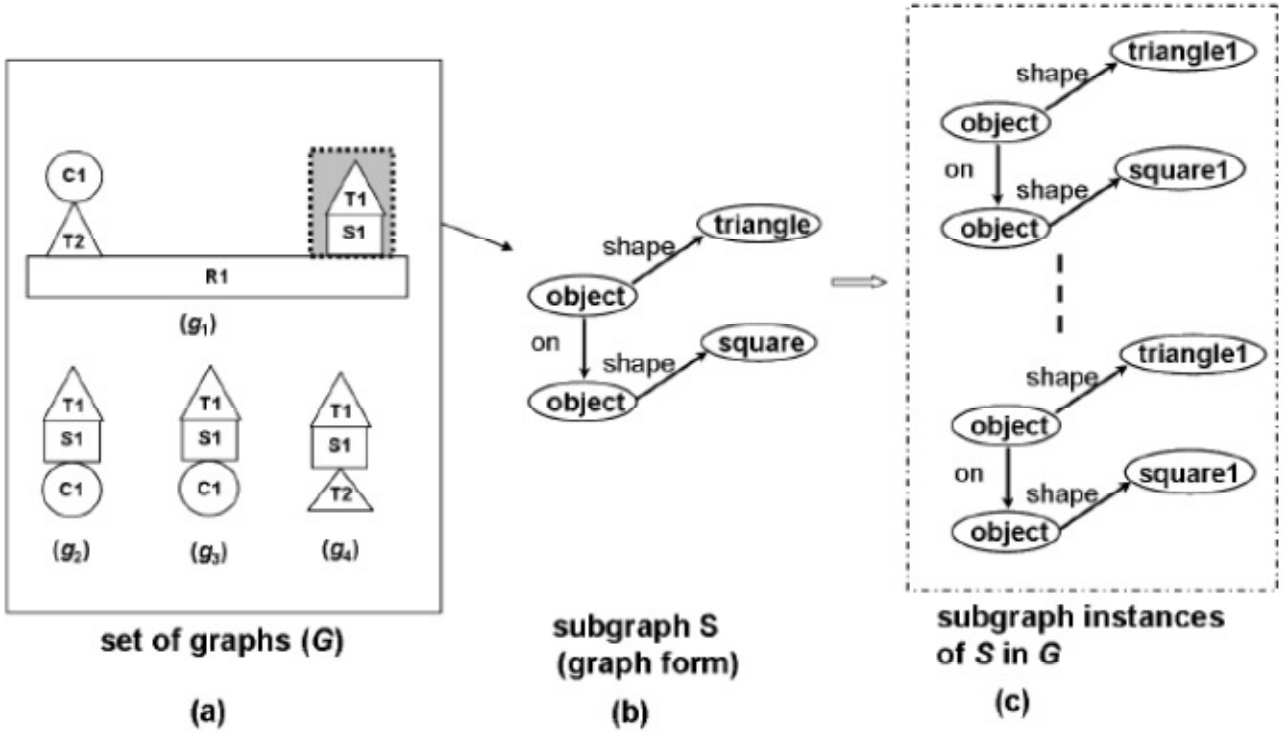
MOEP-GBDM (1)

- MOSubdue tiene un buen rendimiento pero presenta el problema de que la búsqueda Beam Search multiobjetivo es una búsqueda en profundidad en grafos que no permite backtracking
- Por ello, pierde buenas soluciones que al inicio de la exploración no son lo suficientemente prometedoras para entrar en la *ChildList*
- Para solucionar este problema, hemos diseñado métodos de GDBM multiobjetivo basados en algoritmos EMO puros, en concreto algoritmos de programación evolutiva multiobjetivo (MOEP)
- No se aplica operador de cruce. Los subgrafos se generan sólo por mutación en distintas zonas del espacio de búsqueda en paralelo

P. Shelokar, A. Quirin, O. Cordón, A Multiobjective Evolutionary Programming Framework for Graph-based Data Mining, Information Sciences 273:1 (2013) 118–136. FI 2012: 3.643. Cat: COMPUTER SCIENCE, INFORMATION SYSTEMS. Orden: 6/132. Q1

Componentes de MOEP-GDBM:

1. **Codificación**: Representación basada en grafos. El individuo incluye el subgrafo candidato (prototipo genérico) y todas las instancias que cubre en G:



Componentes de MOEP-GDBM:

2. Inicialización:

- La población debe contener subgrafos de distintos niveles del retículo de subgrafos (espacio de búsqueda)
- Primero se crean subgrafos de un solo nodo a partir de las etiquetas únicas de G
- Luego se explora el siguiente nivel del retículo expandiendo todas las instancias de esos subgrafos en G con un nodo y un arco de todas las formas válidas posibles
- Se genera la población inicial escogiendo aleatoriamente entre los subgrafos candidatos anteriores

Componentes de MOEP-GDBM:

3. Generación de subgrafos candidatos (operador de mutación):

- Se expanden todas las instancias del subgrafo padre en G añadiendo un arco (y un nodo si no se crea un ciclo) para crear los subgrafos hijo candidatos

- Se escoge aleatoriamente un subgrafo hijo **con al menos dos instancias cubiertas** en G como individuo resultante de la mutación

Componentes de MOEP-GDBM:

4. Evaluación:

- Se evalúa los subgrafos de la población mediante las funciones objetivo consideradas

5. Mecanismo de Selección Multiobjetivo:

- Se emplea un mecanismo MOEP para seleccionar la nueva población a partir de la población unión de la actual y la de descendientes
- Hemos considerado tres variantes:
  - MOEP-Nondominated Sorting (MOEP-NS), MOEP-Summation of Objectives (MOEP-SO) y MOEP-Nondominance (MOEP-ND)

  aunque podría usarse cualquier otra

Experimentos (1)

BDs y objetivos:

Dataset	#Graphs	#Nodes	#Edges	#Unique Labels	MOEP Run Time (secs)
random1	100	2954	3009	7	100
random2	200	5876	6015	7	145
shapes	100	500	400	8	1
www1	5	832	885	511	450
www2	4	2178	2539	1156	625
US	10	2762	2769	294	2000
UK	10	2732	2748	292	1300
Japan	10	2635	2680	278	3100
Germany	10	2676	2702	284	900
scientograms73	73	19253	19709	296	265

- Dos objetivos: Max Soporte y Max Orden (número de nodos)

Experimentos (2)

Métodos de Comparación:

- Los anteriores más EP-Subdue [Bandyopadhyay et al. Proc. Florida Artif Intell Res Symp, pp. 232-236, 2002]

Parámetros:

- BeamWidth = 5
- MaxBest = PopulationSize = MaxParetoSubs = 100 (tamaño de la Población MOEP y del Archivo de Pareto MOSubdue, MOEP-SO y ND)
- Criterio de parada = Tiempo fijo. Se ejecuta 10 veces MOSubdue-II hasta que se vacíe la ParentList y se toma la media del tiempo

Indicadores de Rendimiento (Métricas) Multiobjetivo:

- Hipervolumen (HVR) (unaria) y Cobertura (C) e I_ε (binarias)

Experimentos (3)

Comparación métrica HVR:
(a mayor valor, mejor rendimiento)

Dataset	Subdue	EP-Subdue	MOSubdue-I	MOSubdue-II	MOEP-NS	MOEP-SO	MOEP-ND
random1	0.9552(-)	0.8724(0.02)	0.9536(-)	0.9623(0.00)	**0.9721(0.01)**	0.9708(0.01)	0.9614(0.01)
random2	0.9663(-)	0.8626(0.02)	0.9747(-)	**0.9795(0.01)**	0.9163(0.02)	0.9723(0.01)	0.8675(0.03)
shapes	**1.0000(-)**	0.9920(0.00)	**1.0000(-)**	0.9954(0.01)	**1.0000(0.00)**	**1.0000(0.00)**	**1.0000(0.00)**
www1	0.7788(-)	0.8467(0.06)	0.8391(-)	0.9899(0.03)	0.9663(0.01)	0.9969(0.01)	**1.0000(0.00)**
www2	0.7567(-)	0.5939(0.07)	0.7432(-)	**0.9162(0.03)**	0.5824(0.03)	0.8790(0.06)	0.8054(0.10)
US	0.6114(-)	0.5796(0.15)	0.9673(-)	0.9013(0.11)	0.9740(0.03)	**0.9791(0.03)**	0.9741(0.03)
UK	0.6318(-)	0.4123(0.10)	0.7302(-)	0.7635(0.18)	**0.9326(0.04)**	0.9316(0.05)	0.9217(0.03)
Japan	0.6429(-)	0.6030(0.09)	0.8021(-)	0.9850(0.03)	0.9857(0.01)	**0.9878(0.01)**	0.9857(0.01)
Germany	0.7406(-)	0.5204(0.11)	0.8083(-)	0.8920(0.06)	**0.9501(0.04)**	0.9452(0.03)	0.9329(0.03)
scientograms73	0.8096(-)	0.5067(0.06)	0.7775(-)	**0.8336(0.01)**	0.8072(0.05)	0.8026(0.07)	0.7940(0.04)
Average	0.7893(0.14)	0.6770(0.20)	0.8596(0.10)	0.9219(0.08)	0.9087(0.13)	**0.9465(0.06)**	0.9336(0.08)

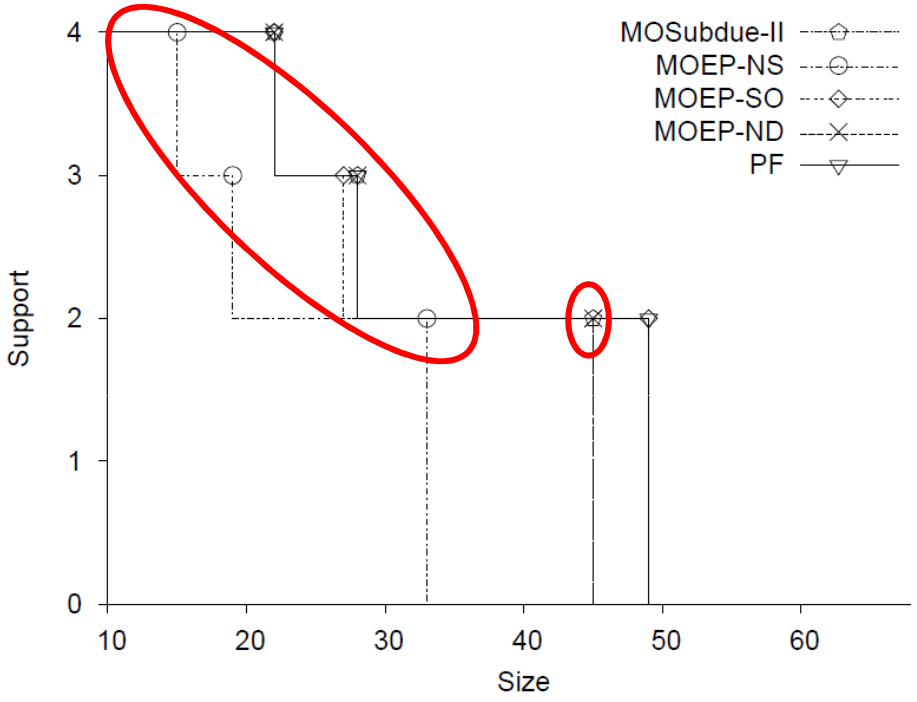
Notación: valor (desviación típica)

Experimentos (4)

Comparación métrica C:

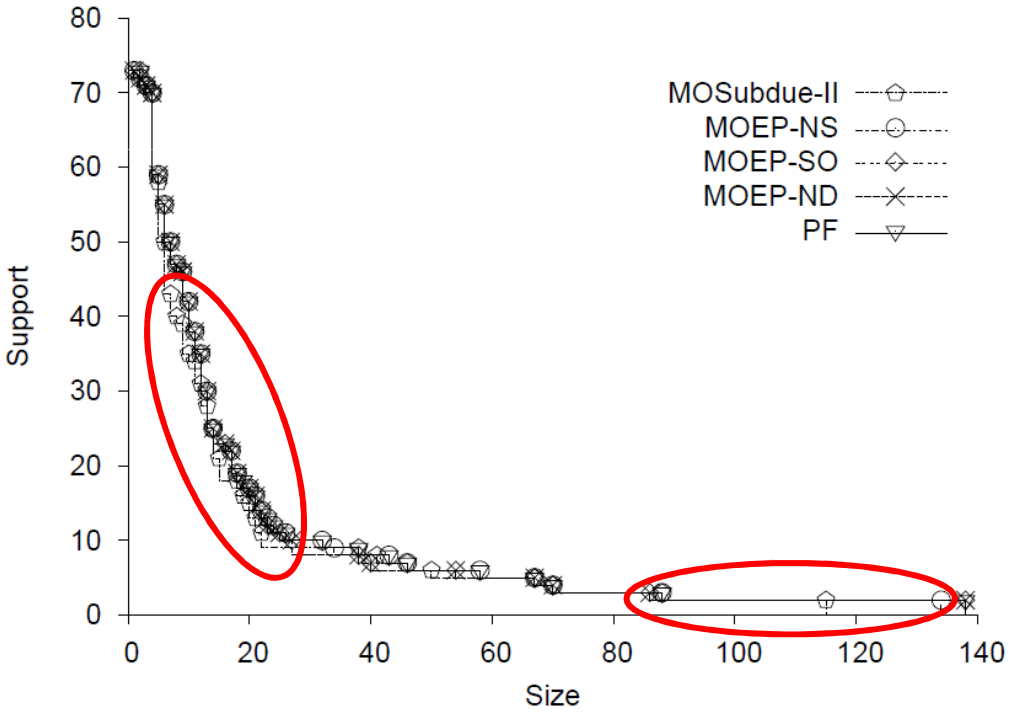
Experimentos (5)

Comparativa visual de los conjuntos de Pareto generados:



www2

Scientograms73 (gd05)



Experimentos (6)

Análisis de Resultados:

- Los resultados de las métricas MO indican que **MOEP-SO** muestra el mejor rendimiento, seguido por MOEP-ND, MOSubdue-II y MOEP-NS
- Aún así, algunas de las aproximaciones de Pareto generadas por MOEP-SO están dominadas por MOSubdue-II y viceversa
- Por ello, hemos realizado un test estadístico por pares sobre los resultados de la métrica I_ε para los tres mejores algoritmos:
  - MOEP-SO domina significativamente a MOSubdue-II y MOEP-ND en 4 BDs
  - Dominado significativamente por MOSubdue-II en *www2* y por MOEP-ND en *www1*
  - La comparación entre MOSubdue-II y MOEP-ND muestra que dichos algoritmos se dominan en 3 y 2 BDs, respectivamente
- MOEP-GBDM tiene problemas en ejecuciones concretas de la BD *www2*: el objetivo soporte no procesa correctamente la redundancia de los subgrafos en este conjunto durante la selección

Experimentos (7)

Test Estadístico:

	(A,B)	(B,A)	(A,C)	(C,A)	(B,C)	(C,B)
random1	1.0000	1.0000	0.0840	0.9362	0.0008	0.9994
random2	1.0000	1.0000	0.0840	0.9362	0.0168	0.9873
shapes	7.969e-06	1.0000	1.0000	1.0000	1.0000	7.969e-06
www1	0.5290	0.5290	1.0000	7.969e-06	1.0000	7.969e-06
www2	0.9998	0.0002	0.0002	0.9999	5.1704e-05	1.0000
US	0.0071	0.9943	0.1290	0.8867	0.6764	0.3541
UK	0.0032	0.9975	0.0029	0.9977	0.8985	0.1164
Japan	0.0558	0.9525	0.0015	0.9988	0.0725	0.9383
Germany	0.0029	0.9978	0.0074	0.9944	0.8643	0.1841
scientograms73	1.0000	1.0000	0.0840	0.9362	1.0000	1.0000

A = MOEP-SO, B = MOSubdue-II, and C = MOEP-NS algorithms

Uso de más de dos objetivos

- Los métodos son flexibles y se pueden aplicar a problemas de GBDM con más de dos objetivos:
  - Max Soporte (G,g) = #subgrafos de G que se emparejan con g
  - Max Orden (G,g) = #vértices(g)
  - Max Densidad (G,g) = 2 · #arcos(g) / (#vértices(g) · (#vértices(g)-1))

- No requieren cambios significativos más alla de la definición de las funciones objetivo

P. Shelokar, A. Quirin, O. Cordón, MOSubdue: A Pareto Dominance-based Multiobjective Subdue Algorithm for Frequent Subgraph Mining. Knowledge and Information Systems 34:1 (2013) 75-108

P. Shelokar, A. Quirin, O. Cordón, Three-Objective Subgraph Mining using Multiobjective Evolutionary Programming, Journal of Computer and System Sciences (2013), en prensa. FI 2012: 1.000. Cat: COMPUTER SCIENCE, THEORY & METHODS. Orden: 34/100. Q2

Uso de más de dos objetivos (2)

Comparación métrica HVR y tiempos (Scientograms):

Methods	*beamWidth*		
	5	10	20
Subdue-I	0.7334 (–)	0.5315 (–)	0.4815 (–)
Subdue-II	0.0368 (0.0010)	0.0620 (0.1002)	0.0341 (0.0756)
MOSubdue-I	0.9482 (–)	0.9486 (–)	0.9209 (–)
MOSubdue-II	0.9508 (0.0036)	**0.9540 (0.0028)**	0.9356 (0.0171)
MOGaston	0.0615 [587]	0.0617 [1,174]	0.0617 [2,935]

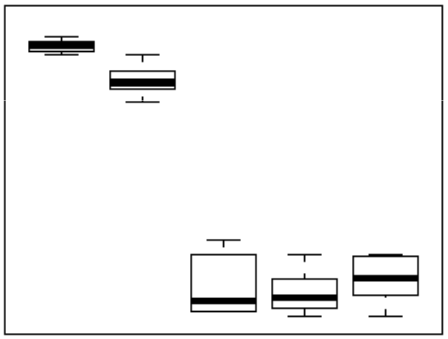
The numbers in the brackets represent the run times in seconds

	5	10	20
Subdue-I	661.72 (–)	1,289.62 (–)	5,674.45 (–)
Subdue-II	40.60 (22.05)	85.10 (47.27)	2,438.18 (1,388.57)
MOSubdue-I	132.55 (–)	681.10 (–)	197.50 (–)
MOSubdue-II	262.29 (48.17)	**587.28 (88.42)**	587.98 (603.86)

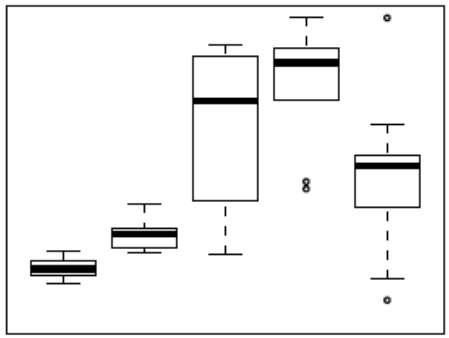
Uso de más de dos objetivos (3)

Dataset	Subdue	EP-Subdue	MOSubdue	MOEP-SO
*random*1	0.7421	0.6955(0.01)	**0.9933(0.0)**	0.9456(0.0)
*random*2	0.7446	0.6904(0.01)	**0.9902(0.0)**	0.9522(0.0)
US	0.3166	0.2291(0.02)	0.4219(0.10)	**0.8446(0.04)**
UK	0.3636	0.2580(0.03)	0.4785(0.11)	**0.8616(0.10)**
Germany	0.4190	0.2678(0.02)	0.4641(0.07)	**0.8291(0.05)**
Average	0.5171(0.47)	0.4281(0.24)	0.6696(0.55)	**0.8866(0.56)**

MOSubdue



Comparación métrica HVR y C (varios):

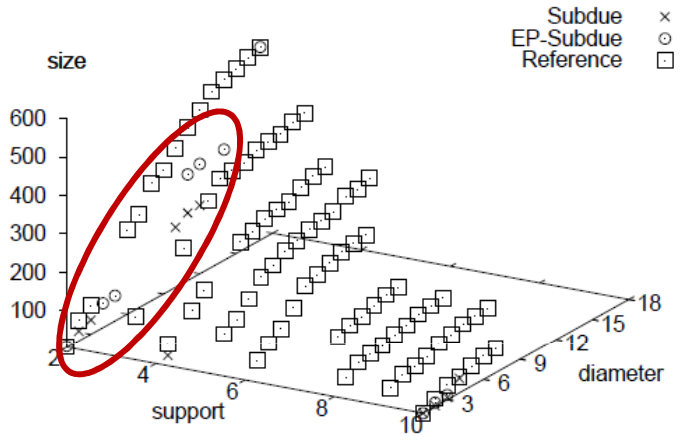


MOEP-SO

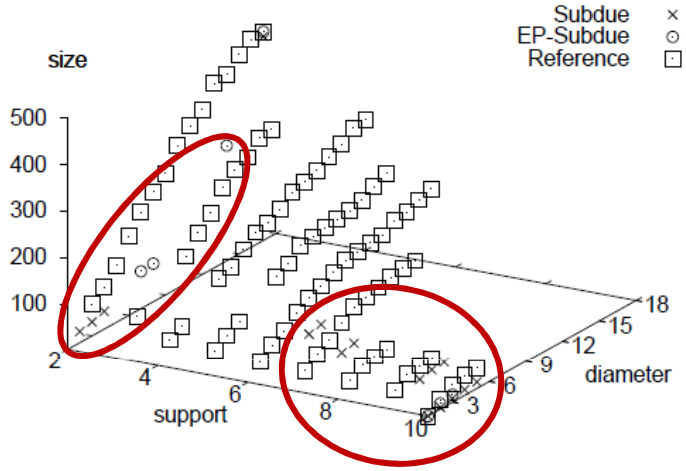
Uso de más de dos objetivos (4)

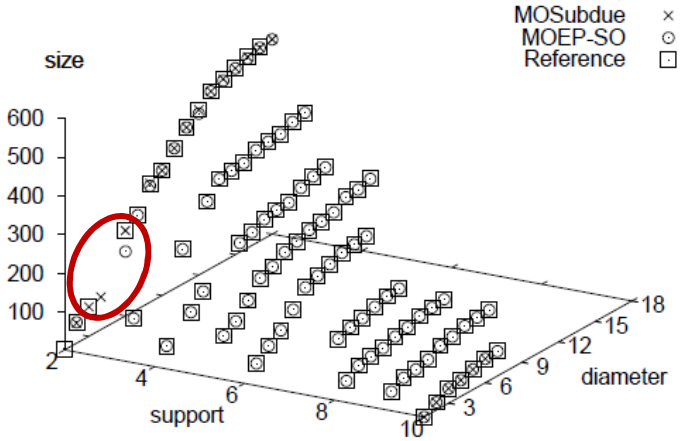
Comparativa visual de los conjuntos de Pareto (*US* y *gd05*):

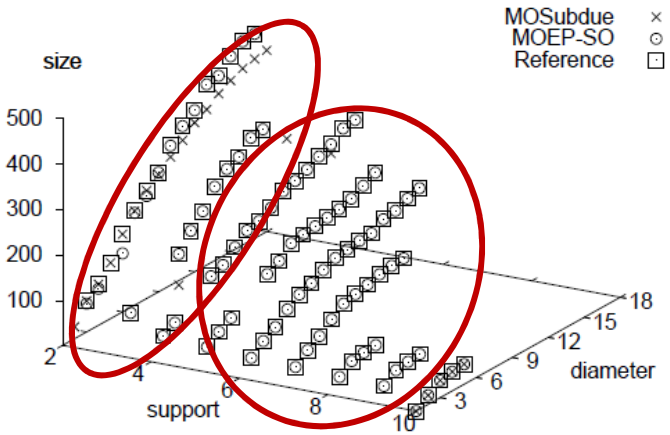
US



gd05







6. Conclusiones

- Hemos propuesto una **metodología automática para el análisis de cienciogramas** basada en GBDM
- Hemos desarrollado **métodos de GBDM MO** basados en EMO, ya sea incorporando componentes de AEs MO a un algoritmo clásico, MOSubdue, como considerando algoritmos EMO puros, MOEP-GBDM
- Hemos aplicado Subdue, MOSubdue y las distintas variantes de MOEP-GBDM a la tarea compleja de extraer CRCSs de cienciogramas a nivel mundial, así como a otros problemas
- La comparación entre los métodos mono-objetivos clásicos y las nuevas propuestas MO demuestra claramente el mejor rendimiento de estas últimas
- **Los trabajos futuros implican un análisis experto para verificar la calidad y la utilidad de los subgrafos extraídos por el enfoque MO**
- También aplicar la GBDM MO para otras tareas de análisis de cienciogramas como la evolución temporal o la comparación entre dos o más países

6. Agradecimientos

Dr. Oscar Cordón
Profesor UGR
Asesor Científico ECSC

Dr. Arnaud Quirin
Antiguo Investigador
Postdoctoral ECSC
Investigador Postdoctoral Gradiant

Dr. Prakash Shelokar
Investigador
Postdoctoral ECSC

Dr. Félix de Moya
Profesor Investigación CSIC

Dr. Benjamín Vargas
Profesor UGR