Análisis de la evolución temporal de Cienciogramas mediante el algoritmo Subdue de minería de grafos

Arnaud Quirin, Oscar Cordón, Prakash Shelokar, Carmen Zarco

European Centre for Soft Computing
Edificio Científico Tecnológico, 33600 Mieres, Asturias
[arnaud.quirin, oscar.cordon, prakash.shelokar, carmen.zarco]@softcomputing.es

Resumen

Los Cienciogramas son una clase de representaciones gráficas que representan el estado de la Ciencia en un determinado dominio. La comparación y análisis automático de un conjunto de cienciogramas, que muestren por ejemplo la evolución de un dominio científico de un país determinado, es una tarea interesante pero de alta dificultad debido a la gran dimensionalidad y complejidad de los datos manejados. En este trabajo, pretendemos demostrar que las herramientas de minería de grafos son muy útiles para llevar a cabo análisis de cienciogramas. Hemos elegido Subdue, un conocido algoritmo de minería de grafos, como primera opción para este propósito. Su modo de operación ha sido adaptado para el estudio de la evolución de un dominio científico en el tiempo. Nuestro caso de estudio muestra claramente el potencial de las herramientas de minería de grafos en el análisis de cienciogramas y además abre la puerta a un gran número de trabajos futuros.

1. Introducción

La generación de un mapa de la ciencia o cienciograma ha sido una idea persistente en los últimos tiempos. Este concepto viene del hecho de que la visualización de la información científica ha sido utilizada durante mucho tiempo para descubrir y divulgar la esencia y estructura de la Ciencia [1,2]. Sin embargo, analizar toda la producción científica se vuelve cada vez más difícil debido a los grandes volúmenes de datos generados hoy en día. Hasta donde nosotros conocemos, no se ha desarrollado previamente ningún enfoque completamente automático que ayude en la exploración de un gran conjunto de datos en la minería de cienciogramas.

En general, las técnicas actuales de análisis de cienciogramas realizan un análisis y comparación a bajo nivel y no automático de los mapas [3, 12]. Para ello, se basan en técnicas estadísticas y análisis de macro y micro estructuras para la identificación de áreas temáticas y disciplinas científicas [13]. Sin embargo, esta filosofía muestra una importante limitación: sólo se puede analizar o comparar un mapa o un conjunto muy reducido de mapas. De hecho, este campo carece de un enfoque de fácil manejo que permita la identificación y comparación de estructuras científicas dentro de cienciogramas con un mayor grado de automatización. En nuestro estudio, hemos utilizado herramientas de minería de grafos para realizar un análisis a mayor nivel, permitiendo la comparación conjunta de un importante número de mapas (i.e. realizando minería de cienciogramas). Gracias a ello, la nueva metodología de análisis de alto nivel presentada en esta propuesta y las aportaciones ya existentes de bajo nivel, pueden utilizarse como marcos de trabajo complementarios para el análisis y comparación de cienciogramas.

La minería de grafos (Graph-based data mining – GBDM) [14] implica la extracción automática de conocimiento novedoso y útil de una base de datos estructurada en forma de grafos. Esta técnica se ha aplicado al descubrimiento de subestructuras frecuentes y correspondencia de grafos en un gran número de disciplinas incluidas la química o la biología aplicada, la clasificación de compuestos químicos, y el desarrollo de modelos de aprendizaje supervisados y no supervisados, entre otros. Concretamente, la primera propuesta en el tema, Subdue [4], basada en el uso del principio de longitud de descripción mínima (Minimum Description Length – MDL) [11], se ha aplicado con éxito a distintos problemas del mundo real. El hecho de que el

principio MDL permita el descubrimiento de subestructuras grandes y frecuentes nos lleva a pensar que Subdue, al igual que otras técnicas de GBDM basadas en la misma idea (i.e. minería de subgrafos frecuentes), es una técnica recomendable para el análisis de cienciogramas.

La estructura del presente trabajo es la siguiente. En la segunda sección hacemos una revisión de las actuales técnicas para diseñar y analizar cienciogramas así como el actual estado del arte de GBDM y, en particular, del algoritmo Subdue. En la tercera sección mostramos como desarrollar una tarea concreta de análisis del cienciograma, el estudio de la evolución a través del tiempo de una determinada materia científica, por medio de este algoritmo. La cuarta sección presenta los experimentos desarrollados y resultados obtenidos. Finalmente, se indicarán algunas conclusiones finales y trabajos futuros en la última sección.

2. Preliminares

En esta sección presentaremos el estado del arte de las actuales técnicas para diseñar y analizar cienciogramas, así como una revisión del campo de la minería de grafos, describiendo su alcance y las técnicas más conocidas.

2.1. Diseño de un cienciograma

La generación de un cienciograma siguiendo una estructura top-down basada en la existencia de una estructura de categorías documentales previa requiere la aplicación secuencial de distintas técnicas. Los cienciogramas considerados en este trabajo se construyen siguiendo la metodología de De Moya-Anegón et al. [9,12]. Se emplean las categorías de co-citación SCOPUS-SJR como unidades de análisis. Cada categoría aglutina las revistas que son categorizadas bajo ese nombre, así mismo los documentos que fueron publicados en esas revistas. Se utiliza una medida de co-citación para computar la similitud relacional entre dos categorías i y j, definida como: $CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i)\cdot c(j)}}$, donde Cc es la frecuencia de co-citación y c es la frecuencia de citación. Se considera el algoritmo Pathfinder [6,10] para podar la matriz de co-citación. Como resultado, sólo se guardan las relaciones más destacadas entre categorías, capturando la estructura intelectual esencial subyacente del dominio científico estudiado. La red podada se representa gráficamente utilizando el algoritmo de dibujo de grafos de Kamada-Kawai [7], elegido por su capacidad para representar naturalmente los elementos más importantes en el centro del mapa (llamado "columna vertebral" del cienciograma).

Los datos considerados han sido extraídos del portal Scimago Journal & Country Rank (SCOPUS-SJR) y comprenden un conjunto de 36 millones de documentos desde 1996 a 2008 sobre 73 países [12]. En este trabajo, sólo trataremos con los mapas de Estados Unidos y Ucrania desde 1996 a 2005. En conjunto, los 20 cienciogramas utilizados tienen 4991 nodos y 5304 aristas.

2.2. Minería de datos basada en grafos y el algoritmo Subdue

La necesidad de minería de datos estructurales para descubrir objetos o conceptos que se relacionen entre sí (i.e., subgrafos que representen asociaciones de características) se ha incrementado en la pasada década, creando así el área de la GBDM [14]. Actualmente, se han propuesto muchos algoritmos de minería de grafos (A priori-based GM. Frequent Subgraph Discovery, MoFa/MoSS, etc.) para tratar con problemas tales como correspondencia de grafos, visualización de grafos, descubrimiento de subestructuras frecuentes, clasificación conceptual, y aprendizaje de patrones supervisados y no supervisados [5].

Entre ellos, podemos destacar Subdue [8], un sistema de descubrimiento de conocimiento basado en grafos que encuentra estructuras y patrones relacionales en datos que representan entidades y relaciones. Este algoritmo fue la primera propuesta en este área y ha sido desarrollado significativamente a lo largo de los años. La técnica usa el principio MDL [11] para descubrir subestructuras interesantes y repetitivas (frecuentes) en una base de datos estructural (DB), extrayéndolos y reemplazándolos por un nodo simple para comprimir la base de datos. Estas

subestructuras extraídas representan conceptos estructurales en los datos. A lo largo de los años, se ha aplicado con éxito a un gran número de problemas del mundo real como aviación, química, geología, antiterrorismo, bioinformática y minería en estructuras web.

La Fig. 1 muestra el perfil del algoritmo Subdue GBDM. Toma como entrada el grafo original DB (compuesto por un único grafo o un conjunto de grafos) desde el cual tienen que extraerse las subestructuras (i.e. subgrafos) y cuatro parámetros empleados para limitar la búsqueda reduciendo el tiempo de ejecución. Estos parámetros (BeamWidth, MaxBest, MaxSubSize, y Limit) restringen al número de subestructuras consideradas y al número total de iteraciones del algoritmo. ChidList and BestList son dos listas ordenadas en las que las subestructuras que tienen los mejores valores de evaluación aparecen primero para guiar en el proceso de búsqueda. El algoritmo termina devolviendo las mejores subestructuras encontradas considerando la medida de evaluación y los parámetros seleccionados.

```
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.         SwitchParentList and ChildList
20.   Return BestList
```

Figura 1. El algoritmo Subdue [4]

La evaluación de la subestructura (ver línea 13) puede computarse con distintas medidas, siendo la medida MDL la más popular. La MDL mide el grado en una subestructura puede comprimir la base de datos completa. Por lo tanto, el algoritmo trata de maximizar la siguiente medida:

$$valor_{MDLi}(S,G) = \frac{I(G)}{I(S)+I(G \mid S)} \qquad (1)$$

donde G es el grafo de entrada, S es la subestructura candidata, I(G) y I(S) son el número de bits requeridos para codificar G y S, I(G|S) es el número de bits requeridos para codificar el grafo obtenido comprimiendo G con S, i.e. sustituyendo cada ocurrencia de S en G por un nodo simple.

3. Subdue para el análisis de cienciogramas. Caso de estudio: evolución de un dominio científico a lo largo del tiempo

La aplicación de Subdue como una herramienta potente de análisis de cienciogramas se basa en su capacidad de minería de subgrafos frecuentes (i.e. realizaremos minería de cienciogramas). Puesto que la estructura subyacente de los cienciogramas es una red social (i.e. un grafo), el descubrimiento automático de subgrafos comunes (denominados Subestructuras de Categorías de Investigación Comunes o CRCSs a partir de ahora) en distintos cienciogramas puede proporcionar al analista una información muy útil para explorar las características de los dominios científicos representados. Esta habilidad puede aplicarse al análisis y comparación de diferentes cienciogramas. En el presente trabajo hemos considerado el uso de Subdue para estudiar la evolución del dominio científico de un país a lo largo del tiempo. La implementación de Subdue considerada es la original de los autores, disponible en http://ailabl.wsu.edu/subdue/.

Obsérvese que, maximizando la medida MDLi, se considera conjuntamente la optimización de dos criterios distintos dentro de Subdue:

- por un lado, la medida potencia las subestructuras más grandes puesto que se obtienen mejores índices de compresión (o mejor valor MDLi) cuando una subestructura mayor puede extraerse y reemplazarse (comprimida) por un nodo simple;
- por otro lado, la medida beneficia a las subestructuras que tienen un mayor soporte (el soporte de una subestructura es el número de ocurrencias de la misma en DB) al obtenerse un mejor índice de compresión cuando muchas

subestructuras son remplazadas (comprimidas) por un nodo simple.

En nuestro caso, la DB de grafos G sobre la cual se ha aplicado Subdue es generalmente un conjunto único de cienciogramas. Sin embargo, existe un modo de operación alternativo para Subdue que considera dos conjuntos distintos, un conjunto positivo y un conjunto negativo , determinados por el usuario. En este modo de operación, el objetivo de Subdue es encontrar las subestructuras más grandes presentes en el máximo número de grafos en el conjunto positivo, las cuales no estén incluidas en el conjunto negativo. La medida MDLi asociada se computa de la siguiente forma:

$$value_{MDLi}(S, G_p, G_n) = \frac{I(G_p) + I(G_n)}{I(S) + I(G_p \mid S) + I(G_n) - I(G_n \mid S)} \quad (2)$$

El uso de mapas negativos permite al usuario considerar un criterio discriminativo determinado. Por ejemplo, para un país en concreto, consideramos los cienciogramas de un determinado periodo (histórico) como conjunto positivo, y el resto de cienciogramas como conjunto negativo, para extraer información relevante sobre subestructuras que aparecen y desaparecen durante una transición histórica.

Cuando se considera este análisis de la evolución de un dominio científico a lo largo del tiempo, un experto en información científica podría estar interesado en conocer qué subestructuras aparecen en el dominio analizado, en qué periodo, cuales son sus tamaños, cuántas son, dónde están localizadas, etc. Esto le permitiría realizar al menos dos tipos de estudios. Por un lado, un análisis en profundidad de las subestructuras descubiertas en sí mismas, qué tipos de categorías están uniendo, etc. Por otro lado, un estudio de algunas estadísticas globales sobre el tamaño y cantidad de esas subestructuras para caracterizar respectivamente la importancia de la evolución del dominio y su dinámica. Esto puede resultar muy útil para realizar una comparación del dominio o un análisis de la evolución del dominio [12].

Para ello, primero se elige un dominio científico. En nuestro estudio, se ha considerado la producción científica de un país entero. Como lo que queremos es buscar los CRCSs que hayan aparecido en un determinado periodo, necesitamos dos periodos anuales, el periodo positivo y el periodo negativo. El periodo negativo normalmente es un conjunto de años del pasado, en los cuales esas subestructuras (i.e. CRCSs) no deberían aparecer. El periodo positivo es normalmente un conjunto de años datados después del periodo negativo, en los cuales las subestructuras sí deben estar presentes. Para tal fin, se considerará el criterio de evaluación MDLi de Subdue en la ecuación (2). Como Subdue se ejecutará para extraer las subestructuras presentes en los mapas de los años positivos pero no en los años negativos, descubrirá de forma eficaz aquellos CRCs que aparezcan al menos una vez a lo largo de los años positivos.

4. Experimentos y análisis de resultados

Se seleccionaron dos países para este estudio, Ucrania y Estados Unidos. Los diez cienciogrmas considerados para cada país corresponden al periodo 1996-2005. Hemos establecido los parámetros de Subdue para que encuentre las 300 mejores subestructuras en cuanto a su evaluación basada en MDLi, considerando un BeamWidth de 4 que permita tiempos cortos de respuesta. Realizamos nuestras pruebas en un Intel Quad-Core 2.40 GHz CPU con 2GB de memoria, obteniendo tiempos de cálculo inferiores a 3 segundos. En todas las siguientes propuestas las subestructuras se presentan usando dos valores (como 3:4, por ejemplo), refiriéndose el primer número al conjunto positivo (correspondiente a los cienciogramas de los años positivos) y el segundo número al conjunto negativo (correspondiente a los cienciogramas de los años negativos). Consideramos que las subestructuras que tengan una mayor representación positiva y una menor representación negativa son las de mejor calidad. Igualmente, se prefieren las subestructuras que tengan un mayor tamaño a las que sean más pequeñas porque las primeras son más específicas.

Soporte (pos:neg)	#subs.	Tamaño (nodos)			Tamaño (aristas)		
		min	max	media	min	max	media
1:1	10	3	8	5.6	2	7	4,6
2:0	6	1	1	1	0	0	0
2:1	2	1	2	1.5	0	1	0,5
2:2	3	1	1	1	0	0	0
2:4	1	1	1	1	0	0	0
3:0	3	1	1	1	0	0	0
3:1	71	1	23	14.63	0	22	13,63
3:2	7	1	5	2.57	0	4	1,57
3:3	11	1	4	1.55	0	3	0,55
3:4	13	1	1	1	0	0	0
3:5	23	1	2	1.04	0	1	0,04
3:6	32	1	2	1.03	0	1	0,03
3:7	118	1	1	1	0	0	0
TOT.	**300**			**4,45**			**3,45**

Tabla 1. Representación y tamaño de las subestructuras extraídas del conjunto de datos de Ucrania

En primer lugar analizaremos el dominio de los cienciogramas ucranianos con 7 años negativos (entre 1996 y 2002) y 3 años positivos (entre 2003 y 2005). La Tabla 1 muestra las estadísticas globales de las 300 subestructuras encontradas para este experimento. Las subestructuras tienen tamaños muy distintos, presentando entre 1 y 23 nodos y de 0 a 22 aristas. Las subestructuras que sólo tienen un nodo son las más comunes (un 70% del total). De éstas, tres subestructuras tienen la proporción óptima de 3:0. Estos nodos corresponden a las disciplinas Leadership and Management, Philosophy, y Media Technology, lo que indica que los investigadores ucranianos realizaron exclusivamente investigaciones en estas categorías después del 2003. Por otro lado, se encontraron 71 subestructuras con una proporción de 3:1, entre las cuales 5 tenían el tamaño máximo encontrado de 23 nodos. En general, las estructuras más interesantes, aquellas que tienen una proporción negativa o nula así como las que son más grandes, no son muy numerosas, lo que permite al experto identificarlas rápidamente y analizarlas todas.

Como ejemplo, la Figura 2 muestra una de esas subestructuras compuestas por 23 nodos y 22 aristas, así como su localización dentro del cienciograma completo de la producción científica ucraniana en 2005. Como puede observarse, esta subestructura es bastante grande y aparece sólo durante los últimos tres años (en realidad la proporción negativa de 1 viene del hecho de que está también presente en el cienciograma de 1998). Esta gran subestructura tiene de hecho dos clasificaciones principales, Biochesmistry y Physics and Astronomy, sugiriendo que la investigación se centró en esos campos durante los últimos tres años. La subestructura ocupa el centro del mapa, donde se encuentra la columna vertebral de la investigación ucraniana. Destacaremos que, incluso si Biochesmitry ocupa en general la parte central de los cienciogramas [12], el hecho de que esté situada en la parte central de una subestructura común extraída es irrelevante.

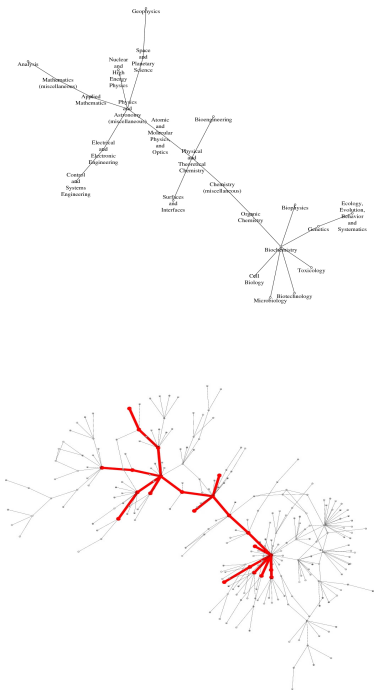
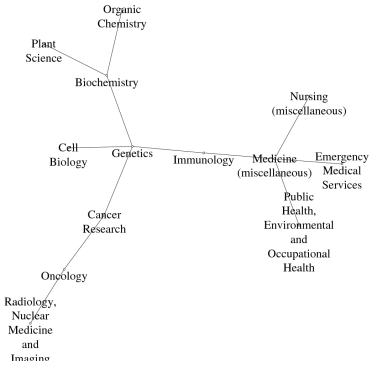


Figura 2. Una de las subestructuras descubiertas en los cienciogramas Ucranianos durante el periodo 2003-2005 (arriba) y su localización en el cienciograma del 2005 (abajo)

Soporte		Tamaño (nodos)			Tamaño (aristas)		
(pos:neg)	#subs.	min	max	media	min	max	media
1:2	2	2	3	2.5	1	2	1,5
2:0	8	1	1	1	0	0	0
2:1	32	4	13	9.41	3	12	8,41
3:0	3	1	1	1	0	0	0
3:2	3	1	1	1	0	0	0
3:3	7	4	6	5	3	5	4
3:4	1	1	1	1	0	0	0
3:7	244	1	4	1.05	0	3	0,05
TOT.	**300**			**2,04**			**1,04**

Tabla 2. Proporción y tamaño de las subestructuras extraídas del conjunto de datos de EEUU.

Por otro lado, el estudio de lo que sucede en los EEUU para el mismo período de tiempo nos muestra de forma significativa que se obtienen las subestructuras más pequeñas (véase Tabla 2). Se han extraído 300 subestructuras, presentando de 1 a 13 nodos y de 0 a 12 aristas, con un tamaño medio de 2 nodos en lugar de los 4,5 nodos del caso ucraniano. Se han obtenido tres subestructuras con el soporte máximo (esto es, 3:0), pero son similares a aquellas obtenidas en el dominio ucraniano, al tener un solo nodo. La Fig. 3 muestra tres subestructuras más interesantes, teniendo todas ellas un soporte de 2:1 y un tamaño de sólo 13 nodos. Presumiblemente podríamos suponer que esas diferencias en forma de subestructuras más pequeñas es una evidencia de los países que tienen una trayectoria investigadora más sólida.



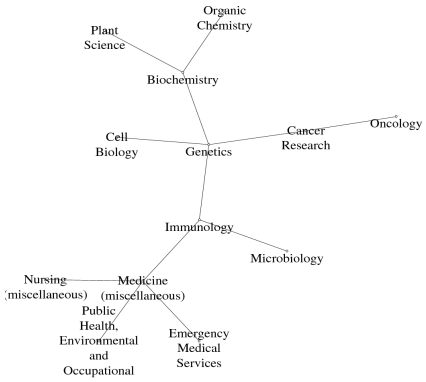


Figura 3. Algunas subestructuras descubiertas en los cienciogramas de EEUU durante los años 2003-2005

Con objeto de tener un mejor conocimiento de los datos, hemos realizado otro estudio en el que el periodo de tiempo no viene fijado por el usuario, sino que se define por medio de ventanas deslizantes. Comenzamos con cinco años negativos y dos años positivos, y añadimos un año positivo y eliminamos el año negativo más antiguo en cada paso.

Periodos años		Soporte		Tamaño (nodos)		
(negativos)	(positivos)	(pos:neg)	#inst.	min	max	media
1996-1999	2000-2001	2:0	3	1	1	1
1996-1999	2000-2001	2:1	1	1	1	1
1996-2000	2001-2002	2:0	3	1	1	1
1996-2000	2001-2002	2:1	55	3	15	8,82
1996-2001	2002-2003	2:1	3	1	1	1
1996-2002	2003-2004	2:0	3	1	1	1
1996-2003	2004-2005	2:0	8	1	1	1
1996-2003	2004-2005	2:1	32	1	11	8,69

Tabla 3. Soporte y tamaño de algunas subestructuras extraídas del conjunto de datos de EEUU utilizando ventanas deslizantes para dos años positivos

Como punto de comparación con el estudio previo, utilizaremos el conjunto de datos de

EEUU para detectar pequeños cambios dentro de los años. Se han extraído muchas subestructuras siguiendo este procedimiento, pero sólo guardamos aquellas que correspondan a las proporciones de 2:1 o 2:0, es decir, el soporte máximo posible para este experimento. La Tabla 3 presenta algunas estadísticas del mismo. En general, todas las subestructuras descubiertas presentan un tamaño pequeño, en un rango de 1 a 15 nodos pero siendo iguales a 1 en el 79% de los casos. Todas las subestructuras encontradas con un soporte 2:0 se muestran en la Figura 4. Dichas subestructuras son pequeñas ya que están compuestas por un solo nodo. Sin embargo, incluso si son independientes, se pueden encontrar algunas relaciones entre ellas. Por ejemplo, durante el periodo 2001-2002 se han encontrado las áreas de investigación centradas en cuidados médicos, diagnósticos y emergencias. Durante el periodo 2004-2005, aparecieron más áreas de investigación centradas en especialidades médicas (orthodontics, periodontics, oral surgery, pharmacology, etc.).

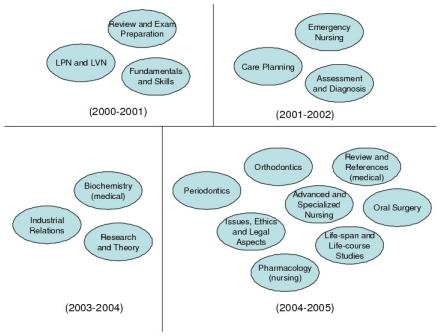


Figura 4. Algunas subestructuras que aparecen repetidamente entre el año 2000 y 2005 en los cienciogramas de EEUU

Debemos destacar un hecho inusual, el alto número de instancias obtenidas con un soporte de 2:1 para los periodos 2001-2002 y 2004-2005. Obtuvimos 55 y 32 subestructuras respectivamente para esos periodos, dos datos bastante altos comparados con las estadísticas restantes. Durante esos periodos, la investigación en los EEUU se desarrolló lo suficiente como para producir suficientes cambios en sus correspondientes mapas. Dichos cambios concernían a categorías pertenecientes principalmente al campo médico, tales como Emergency Nursing, Care Planning, Oral Surgery, Orthodontics, etc. Obsérvese además que sólo un enfoque automático puede encontrar y destacar rápidamente aquellos periodos con cambios significativos.

A la vista de los experimentos realizados, podemos decir que Subdue es una herramienta muy útil para identificar los nuevos CRCSs en un país determinado y durante un periodo temporal concreto. Fijándonos en las investigaciones específicas desarrolladas de un año a otro, o incluso fijándonos en las estadísticas globales, podemos obtener información relevante sobre la evolución de la investigación en ese país. Obsérvese cómo las subestructuras extraídas no están siempre localizadas en la columna vertebral del cienciograma sino en otras partes diferentes del mapa, por tanto Subdue se convierte en una herramienta de análisis complementaria a los enfoques de bajo nivel existentes.

5. Conclusiones

En este trabajo, hemos mostrado cómo una técnica de GBDM, concretamente Subdue, puede aplicarse con éxito al complejo campo del análisis y comparación de cienciogramas. Se han procesado los dominios científicos de dos países para estudiar la evolución de la investigación a lo largo del tiempo extrayendo algunas subestructuras muy interesantes así como algunos parámetros estadísticos.

Esta metodología es escalable y no presentará problemas significativos al ser aplicada a un gran volumen de datos. Se ha visto que la generación de representaciones gráficas de subestructuras (CRCSs) (véase Fig. 2), tablas e histogramas es completamente automática. Aunque solo se ha utilizado el algoritmo Subdue para este propósito, se podrían haber tenido en cuenta otros algoritmos de GBDM. Por estas razones, la minería de grafos puede considerarse como una nueva herramienta de análisis de cienciogramas desarrollada como complemento a las otras técnicas ya existentes. En el futuro, tenemos pensado utilizar otras técnicas GBDM (específicamente, técnicas basadas en optimización multiobjetivo) y descubrir otras utilidades de Subdue para el análisis y comparación de cienciogramas.

Agradecimientos

Este trabajo está soportado por el Ministerio de Ciencia e Innovación dentro del proyecto TIN2009-07727, incluyendo fondos FEDER. Nos gustaría agradecer a Elsevier y a los Dres. Félix de Moya-Anegón y Benjamín Vargas-Quesada por autorizar el uso de los datos SCOPUS-SJR para construir los cienciogramas.

Referencias

[1] Börner, K., Scharnhorst, A.: Visual conceptualizations and models of science. Journal of Informetrics. 3(3) (2009) 161-172

[2] Chen, C.: Information Visualization: Beyond the Horizon. Springer, Berlin (2004)

[3] Chen, C., Chen, Y., Horowitz, H., Hou, H., Liu, Z., Pellegrino, D.: Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics 3 (3)(2009) 191-209

[4] Cook, D.J., Holder, L.B.: Graph-Based data mining. IEEE Intelligent Systems 15(2) (2000) 32-41

[5] Cook, D.J., Holder, L.B., eds.: Mining Graph Data. Wiley, New Jersey (2006)

[6] Dearholt, D., Schvaneveldt, R.: Properties of Pathfinder networks. In Schvaneveldt, R., ed.: Pathfinder Associative Networks: Studies in Knowledge Organization. Ablex Publishing Corporation (1990) 1-30

[7] Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Inormation Processing Letters 31(1) (1989) 7-15

[8] Leydesdorff, L., Rafols, I.: A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology 60(2) (2009) 348-362

[9] Moya-Anegón, F.D., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, A., Corera-Álvarez, E., Muñoz-Fernández, F.J.: A new technique for building maps of large scientific domains base don the cocitation of classes and categories. Scientometrics 61(1) (2004) 139-145

[10] Quirin, A., Cordón, O., Guerrero-Bote, V.P., Vargas-Quesada, B., Moya-Anegón, F.D.: A quick MST-based algorithm to obtain Pathfinder Networks. Journal of the American Society for Information Science and Technology 59(12) (2008) 1912-1924

[11] Rissanen, J.: Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Ind., River Edge (1989)

[12] Vargas-Quesada, B., Moya-Anegón, F.D.: Visualizating the Structure of Science. Springer-Verlag: New Yotk, Secaucus, (2007)

[13] Wallace, M.L., Gingras, Y., Duhon, R.: A new approach for detecting scientific specialties from raw cocitation networks. Journal of American Society for Information Science and Technology, 60(2) (2009) 240-246

[14] Washio, T., Motoda, H.: State of the art of graph-based data mining. SIGKDD Explorations 5(1) (2003) 59-68