Journal of Informetrics 4 (2010) 291–312

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Graph-based data mining: A new tool for the analysis and comparison of scientific domains represented as scientograms

Arnaud Quirin [a], Oscar Cordón [a,*], Benjamín Vargas-Quesada [b,c], Félix de Moya-Anegón [d]

[a] European Centre for Soft Computing, Edf. Científico Tecnológico, 33600 Mieres, Spain
[b] Communication and Information Science Faculty, University of Granada, 18071 Granada, Spain
[c] CSIC, Unidad Asociada Grupo Scimago, 18071 Granada, Spain
[d] CSIC/CCHS/IPP, 28037 Madrid, Spain

ARTICLE INFO

Article history:
Received 14 October 2009
Received in revised form
25 December 2009
Accepted 20 January 2010

Keywords:
Domain analysis
Social networks
Scientograms
Graph-based data mining
Scientogram mining
Subdue algorithm

ABSTRACT

The creation of some kind of representations depicting the current state of Science (or *scientograms*) is an established and beaten track for many years now. However, if we are concerned with the automatic comparison, analysis and understanding of a set of scientograms, showing for instance the evolution of a scientific domain or a face-to-face comparison of several countries, the task is titanically complex as the amount of data to analyze becomes huge and complex. In this paper, we aim to show that graph-based data mining tools are useful to deal with scientogram analysis. Subdue, the first algorithm proposed in the graph mining area, has been chosen for this purpose. This algorithm has been customized to deal with three different scientogram analysis tasks regarding the evolution of a scientific domain over time, the extraction of the common research categories substructures in the world, and the comparison of scientific domains between different countries. The outcomes obtained in the developed experiments have clearly demonstrated the potential of graph mining tools in scientogram analysis.

1. Introduction

The construction of a great map of sciences[1] has been a persistent idea in the modern ages. This need arises from the general conviction that an image or graphic representation of a domain favors and facilitates its comprehension and analysis. The visualization of scientific information has long been used to uncover and divulge the essence and structure of science (Börner & Scharnhorst, 2009; Chen, 1999a, 2004). Yet despite its ripe age, information display is still in an adolescent stage of evolution in the context of its application to scientific domain analysis. Never before data have been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of scientific data is becoming increasingly difficult. There is a large number of information visualization techniques which have been developed over the last decade within this area (Chen, 1999b; Lucio-Arias & Leydesdorff, 2008; Moya-Anegón et al., 2007, 2005; Small & Garfield, 1985), but none of them has been designed to support the exploration of large datasets. Besides, all the latter approaches require a large amount of expertise from the user, which reduces the chances to automate the analysis procedure. Nevertheless, it is clear that information visualization and visual data mining (Keim, 2002) can provide the theoretical and practical backgrounds to deal with scientific information analysis.

* Corresponding author. Tel.: +34 985 456545; fax: +34 985 456699.
E-mail addresses: arnaud.quirin@softcomputing.es (A. Quirin), oscar.cordon@softcomputing.es (O. Cordón), benjamin@ugr.es (B. Vargas-Quesada), felix_demoya@cchs.csic.es (F. de Moya-Anegón).

[1] In the following we will consider (*visual*) *science maps*, *scientograms*, *graphs*, or simply *maps* as synonyms within our domain.

1751-1577/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2010.01.004

The generation of a *big picture* is something implicit in the process of visualizing scientific information. In an attempt to sum what has taken place to date up, we can say that nowadays there are two proposals for tracking down the *big picture*. On the one hand, one can adopt the traditional units of analysis (authors, documents, and journals) and, through their grouping, identify scientific disciplines following a bottom-up process (Boyack & Klavans, 2008; Klavans & Boyack, 2006; Small & Sweeney, 1985; Small, Sweeney, & Greenlee, 1985). On the other hand, the alternative uses the categories of the documents to the same end, and shows the scientific structure from them in a top-down manner (Moya-Anegón et al., 2004). The former proposal presents all the pros of its fine-grained character, but it runs into difficulties in representing the totality of the panorama on a single plane and in tagging the disciplines. The latter option has its strong points where the former shows weaknesses, and *vice versa*. That is, it is relatively simple to represent the scientific structure of a domain on a single plane by means of a maximum of 300 categories and their interrelation, avoiding tagging problems. However, this implies the acceptance of a classification of science in predefined categories, never transparent and always subjective, as well as the fact that documents are classified by the journals in which they are published and not by their content (coarse-grained character). To our mind, both the former and the latter are valid alternatives for the achievement of the *big picture*. In this contribution, we will consider the use of the latter one to design scientograms in order to delimit and discover scientific disciplines.

Current scientogram analysis techniques (Boyack, Börner, & Klavans, 2009; Chen et al., 2009; Klavans & Boyack, 2006; Leydesdorff & Rafols, 2009; Moya-Anegón et al., 2007) aim to provide a fine, detailed, tight view of a scientogram. To do so, they are based on performing a low-level analysis and comparison of the maps. Statistical techniques, computer algorithms, and macrostructure and microstructure techniques for the identification of thematic areas and scientific disciplines have already been used to analyze and compare scientograms (Boyack, Klavans, & Börner, 2005; Chen, 1999b; Lucio-Arias & Leydesdorff, 2008; Moya-Anegón et al., 2007; Wallace, Gingras, & Duhon, 2009). However, this approach shows a main limitation: only a single or a very reduced set of maps can be analyzed or compared together. In fact, the field lacks an easy-to-use approach allowing the identification and the comparison of scientific structures within scientograms with a higher degree of automation. In our study, graph mining tools are considered to perform a higher level analysis, allowing the joint comparison of a larger number of maps (i.e., performing *scientogram mining*). Thanks to that, the novel high-level analysis methodology introduced in the current contribution and the existing low-level approaches can be used as complementary frameworks for the analysis and comparison of scientograms.

Graph-based data mining (GBDM) (Cook & Holder, 2006; Holder & Cook, 2005; Washio & Motoda, 2003) involves the automatic extraction of novel and useful knowledge from a graph representation of data. By 'novel' we mean that the knowledge retrieved is not directly encoded in the data but deeply masked in it (hence, it requires to be uncovered), and by 'useful' we mean that the discovered patterns have in general an interest for the domain expert: active principles of molecules, common backbones in a communication network, common portions of an electronic circuit, etc. In fact, GBDM techniques have been applied for frequent substructure discovery and graph matching in a large number of domains including chemistry and applied biology (Borgelt & Berthold, 2002; Huan et al., 2004), classification of chemical compounds (Deshpande, Kuramochi, & Karypis, 2002), and unsupervised and supervised pattern learning (Cook & Holder, 2006), among many others. In particular, the first proposal in the topic, Subdue, has proved to be successful in many different real-world applications (Chittimori, Gonzalez, & Holder, 1999; Gonzalez, Holder, & Cook, 2000; Holder, Cook, Coble, & Mukherjee, 2005; Kukluk, You, Holder, & Cook, 2007; Rakhshan, Holder, & Cook, 2004).

Subdue (Cook & Holder, 1994, 2000) is a graph-based knowledge discovery system that finds structural, relational patterns in data representing entities and relationships. It aims to discover interesting and repetitive substructures in a structural database (DB). For this purpose, the minimum description length (MDL) principle (Rissanen, 1989) is used in order to compress the original DB into a hierarchical and shorter version. Since the MDL principle allows the discovery of both large and frequent substructures we think that Subdue, as well as any other GBDM technique based on the same idea (i.e., frequent subgraph mining), is well recommended for scientogram analysis. This paper is actually the first proposal on the use of Subdue in this application domain. In particular, we will describe how this algorithm can be customized to deal with three different scientogram analysis and comparison tasks regarding the evolution of a scientific domain over time, the extraction of the common research categories substructures in the world, and the comparison of scientific domains between different countries.

The structure of the current contribution is as follows. In Section 2, we review the current techniques to design and analyze scientograms as well as the current state of the art of GBDM, detailing the particular case of the Subdue algorithm. In Section 3 we detail the main components of Subdue and we show how several scientogram analysis tasks can be performed by means of this algorithm. The three next sections present experiments related to three different case studies considered. Finally, some concluding remarks are pointed out in the last section.

2. Preliminaries

In this section we will present a state of the art of the current techniques used to design and analyze scientograms. Besides, we will review the GBDM field, describing its scope, the most known techniques (in particular, Subdue), and the application domains.

2.1. Scientogram design

The generation of a scientogram following the top-down approach (Moya-Anegón et al., 2004) requires the sequential application of several techniques. They are reviewed in the next subsections.

2.1.1. Units of analysis

The categories are the units of analysis and representation (Moya-Anegón et al., 2004; Vargas-Quesada & Moya-Anegón, 2007). Each category agglutinates the journals that were categorized under that name, and likewise the documents that were published in those journals. Because we strive to represent and analyze the structure of vast domains, whether they be thematic, geographic or institutional, we fall back on to SCOPUS-SJR co-citation categories as a tool for this purpose.

2.1.2. Unit of measure

Co-citation is a widely used and generally accepted unit of measure for obtaining relational information about documents belonging to a domain. Once the rough information of the SCOPUS-SJR co-citation for the categories present in the domain to be analyzed is obtained, a co-citation measure *CM* is computed for each pair of categories *i* and *j* as follows:

$$CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}} \quad (1)$$

where *Cc* is the co-citation frequency and *c* is the citation frequency.

Notice that, the aim of this scientogram generation method is that the final scientogram obtained is a tree. Hence, in order to avoid the existence of cycles in the pruned network (see the next subsection), the considered measure of association adds the normalized co-citation (divided by the square root of the product of the frequencies of the co-cited documents' citations (Salton & Bergmark, 1979) to the rough category co-citation frequency. In this way, the network weights become real numbers, allowing us to create small differences between similar values for the co-citation frequency. The latter fact allows the pruning algorithm to select between two edges which would have otherwise the same weight (in view of the $Cc(ij)$ value), as in general using the modified formula the weights of the two edges will become different. This allows us to avoid the occurrence of cycles and to achieve the optimal pruning of each link considering the citing conditions of each category. Of course, this technique is not perfect: some edges will keep on having the same weight even if the latter formula is considered. For instance, in the non-pruned version of the European map, on 34 484 weights, around a 50% have the same value. Nevertheless, the new formula actually helps: without the additional factor, less than a 4% of links would be distinct.

2.1.3. Dimensionality reduction

We should take into account the fact that the networks resulting from citation, co-citation, or term co-occurrence analysis are usually dense, when the categories are used as the unit for each node. Hence, the Pathfinder algorithm (Chen, 1998; Dearholt & Schvaneveldt, 1990) is applied to the co-citation matrix to prune the network. Due to the density of the data, and especially in the case of vast scientific domains with a high number of entities (categories in our case) in the network, Pathfinder is usually parameterized to $r = \infty$ and $q = n - 1$. This is done in order to preserve and highlight the salient relationships between categories, and for capturing the essential underlying intellectual structure of a scientific domain. These parameters also allow us to work with quick variants of the original Pathfinder algorithm (Quirin, Cordón, Guerrero-Bote, Vargas-Quesada, & Moya-Anegón, 2008; Quirin, Cordón, Santamaría, Vargas-Quesada, & Moya-Anegón, 2008).

2.1.4. Layout

There are many different methods for the automatic visualization of the Pathfinder networks (PFNETs). The spring embedder family of methods is the most widely used in the area of Information Science. Spring embedders assign coordinates to the nodes in such a way that the final graph will be pleasing to the eye, and that the most important elements are located in the center of the representation (also called its *backbone*). Kamada–Kawai's algorithm (Kamada & Kawai, 1989) is one of the most extended methods to perform this task. Starting from a circular position of the nodes, it generates networks with aesthetic criteria such as the maximum use of available space, the minimum number of crossed links, the forced separation of nodes, the build of balanced maps, etc.

The combination of categories co-citation, PFNETs, and Kamada–Kawai makes the categories that most sources share with the rest, tend to situate themselves toward the center.

2.1.5. Considered scientific data

For strictly research purposes we are using SCOPUS-SJR Data, obtained from the Scimago Journal & Country Rank portal.[2] This gave us a total of 36 millions of documents (comprising articles, biographical items, book reviews, corrections, editorial materials, letters, meeting abstracts, and reviews) since 1996 to 2008 (Vargas-Quesada & Moya-Anegón, 2007). As an

[2] http://www.scimagojr.com/.

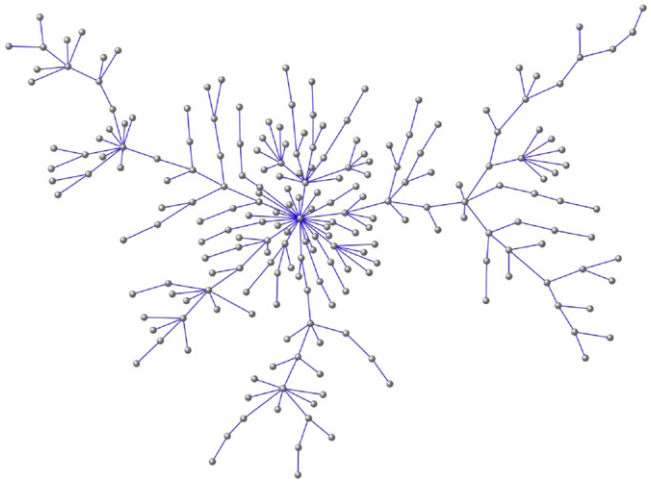


Fig. 1. An example of a scientogram corresponding to the Europe scientific domain in 2002 (category names are not printed to improve the readability).

example, a scientogram resulting from the execution of the different steps mentioned in the previous sections is shown in Fig. 1.

The graph DB considered in the current contribution comprises scientograms of 73 countries (see Table 1). For 60 of them, the data corresponding to their scientific production for each year between 1996 and 2005 is available, thus representing a study period of ten years with one scientogram for each year. Overall, the scientograms in the DB have 159 135 nodes and 172 081 edges.

2.2. State of the art on scientogram analysis

Since their origins (Griffith, Small, Stonehill, & Dey, 1974; Small & Griffith, 1974), scientograms have been only analyzed superficially. This means that from a specific starting point, usually the center, the analysis began taking an excursion (Small and Garfield, 1985) to illustrate the rendered structure of science. This kind of analysis has been performed in this way until today (Boyack, Börner, & Klavans, 2009; Boyack et al., 2005; Moya-Anegón et al., 2005; Samoylenko, Chao, Liu, & Chen, 2006). However, some authors have proposed different methods based on the combination of algorithms, statistics, and/or

Table 1
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	Italy	Slovenia
Brazil	Japan	South Africa
Bulgaria	Jordan (8)	Spain
Canada	Kenya	Sweden
Chile	Kuwait	Switzerland
China	Lebanon (3)	Taiwan
Colombia	Lithuania (8)	Thailand
Croatia	Malaysia	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	

network indicators for the analysis and identification of cognitive structures. For instance, Chen (1999b) and Moya-Anegón et al. (2007) combined PFNETs and factor analysis for the identification and extraction of scientific structures. Leydesdorff et al. (Leydesdorff, 2007; Leydesdorff & Schank, 2008; Lucio-Arias & Leydesdorff, 2008) proposed the use of network measures as structural changes indicators within the scientific disciplines.

More recently, Leydesdorff and Rafols (2009), and Wallace et al. (2009) proposed some statistical and algorithmic techniques for the identification of thematic areas and scientific disciplines in scientograms. Concurrently, Guerrero-Bote, Vargas-Quesada, Espinosa-Calvo, and Moya-Anegón (2009) analyzed and compared, thanks to the information provided by the links, the scientograms of six Spanish scientific domains based on *Materials Science*. Finally, Klavans and Boyack (2009) generated a consensual scientogram from a subjective analysis of twenty existing scientograms.

Hitherto, all the analysis tools put in place are mainly devoted to analyze a single or a couple of scientograms. Hence, they can only analyze a single domain or compare domains with backbones. Therefore, the obtaining of *automatic tools* allowing us to analyze and compare many domains together is strongly encouraged in this field. The main aim of the current contribution is to bridge this gap by means of the use of GBDM techniques.

2.3. Graph-based data mining

In this section, we will review several graph mining techniques with some of their typical applications. We will also detail the particular algorithm used in this study, Subdue.

2.3.1. Introduction

The need of mining structural data to uncover objects or concepts that relates objects (i.e., subgraphs that represent associations of features) has increased in the past ten years, thus creating the area of GBDM (Holder & Cook, 2005; Washio & Motoda, 2003). Nowadays, GBDM has become a very active area and several techniques such as Subdue, the Apriori family of methods (Apriori-based GM (Inokuchi, Washio, & Motoda, 2000), Frequent Subgraph Discovery (Kuramochi & Karypis, 2001), JoinPath (Vanetik, Gudes, & Shimony, 2002), etc.), and the Frequent Pattern-growth family of methods (CloseGraph (Yan & Han, 2003), FFSM (Huan, Wang, & Prins, 2003), Gaston (Nijssen & Kok, 2004), gSpan (Yan & Han, 2002), MoFa/MoSS (Borgelt & Berthold, 2002), Spin (Huan, Wang, Prins, & Yang, 2004), etc.) have been proposed to deal with problems such as graph matching, graph visualization, frequent substructure discovery, conceptual clustering, and unsupervised and supervised pattern learning (Cook & Holder, 2006). Among them, we can highlight Subdue (Cook & Holder, 1994, 2000), a graph-based knowledge discovery system that finds structural, relational patterns in data representing entities and relationships. This algorithm was the first proposal in the topic and has been largely extended through the years. It is able to develop graph shrinking as well as frequent substructure extraction and hierarchical conceptual clustering.

Among the many different GBDM application domains, frequent subgraph discovery, the approach considered in this paper, is in general used in chemistry and applied biology. Huan et al. (2004) applied it to study protein structural families; Borgelt and Berthold (2002) considered it to discover active chemical substructures concerning the human immunodeficiency virus (HIV); Koyutürk, Grama, and Szpankowski (2004) applied it to biological networks in order to find which ones present the larger frequent subpathways; Deshpande et al. (2002) dealt with the classification of chemical compounds using frequent substructures as features; and Yan, Yu, and Han (2004) improved the time of graph searching using frequent patterns as indexing features. Nevertheless, up to our knowledge, the current contribution constitutes the first application of frequent subgraph mining to the analysis of scientific domains.

2.3.2. The Subdue algorithm

Subdue (Cook & Holder, 1994, 2000) is a method for discovering interesting and repetitive substructures in a structural DB. The algorithm uses the MDL principle (Rissanen, 1989) to discover frequent substructures in a DB, extract them and replace them by a single node in order to compress the DB. These extracted substructures represent structural concepts in the data. The Subdue algorithm can be run several times in a sequence in order to extract meta-concepts from the previously simplified DB. After multiple Subdue runs on the DB, we can discover a hierarchical description of the structural regularities in the data (Jonyer, Cook, & Holder, 2001). Subdue can also use background knowledge, such as domain-oriented expert knowledge, to be guided and to discover substructures for a particular domain goal. Through the years, it has been successfully applied to a large range of real-world problems such as aviation (Chittimori et al., 1999), chemistry (Chittimori et al., 1999), geology (Gonzalez et al., 2000), counter-terrorism (Holder et al., 2005), bioinformatics (Kukluk et al., 2007), and web structure mining (Rakhshan et al., 2004).

Fig. 2 shows the outline of the Subdue GBDM algorithm. The algorithm takes as input the original graph from which the substructures (i.e., subgraphs) have to be extracted and four parameters used to limit the search while reducing the runtime. Subdue uses a variant of beam search (Lowerre, 1976) in order to avoid exponential-sized queue: at each step, only *BeamWidth* new children from a given parent are explored (see line 14). Furthermore, only a maximum of *MaxBest* substructures having a maximal size of *MaxSubSize* are returned to the user, and the algorithm does not develop more than *Limit* iterations (see line 6). These parameters ensure that the running time of Subdue is polynomial and is actually constrained by the BeamWidth and the Limit parameters (Jonyer et al., 2001). Extending subgraphs edge by edge is the way Subdue is using to create new subgraphs. This is performed by the function ExtendSubstructure(S) which extends the

```
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
```

Fig. 2. The Subdue GBDM algorithm (reprinted from Cook & Holder, 2000).

substructure *S* in all possible ways, i.e., by adding to *S* a new edge and a vertex from the input graph, or by adding a new edge between two vertices that are already a part of *S*. ChildList and BestList are two ordered lists in which the substructures having the best evaluation values appear first. The evaluation of a substructure (see line 13) can be computed by the MDL-measure (see Section 3.1.1), the Size-measure (see Section 3.1.2), or the Support-measure (see Section 3.1.3). The function *L mod N* (see lines 15 and 18), in which *L* is a list and *N* an integer, returns the list *L* if its size is inferior or equal to *N*, and the first *N* elements of *L* otherwise. The algorithm ends up by returning the best substructures found considering the chosen evaluation measure and the constraint parameters.

Other variants of the original Subdue algorithm, including inexact graph matching, positive and negative graph considerations, and an improved search algorithm were proposed later by the same authors (Jonyer et al., 2001). In our contribution, only the use of positive/negative graphs is an important feature as it is described in the next section.

3. Subdue for scientogram analysis

This section is devoted to describe the use of Subdue as a powerful scientogram analysis tool. As said, the application of Subdue for this domain will rely on its frequent subgraph mining activity (i.e., we will perform scientogram mining). In order to properly understand the customization of this algorithm we needed to develop, some general aspects must be known. In every case, the considered Subdue implementation is that made by the original authors, available at http://ailab.wsu.edu/subdue/. It considers the use of the three existing measures to extract substructures of interest from the graph DB. Besides, it takes the concept of positive/negative graphs into account, thus resulting in two different operation modes, depending on whether there are negative instances in the DB or not. Different combinations of the latter two aspects, substructure evaluation measure and positive/negative substructure operation mode, will be considered to tackle three scientogram analysis tasks. Section 3.1 will introduce the basics of the three Subdue evaluation measures and their interaction with the selected operation mode.

Since the underlying scientogram structure is a social network (i.e., a graph), the uncovering of common subgraphs to different scientograms in an automatic fashion can provide the information analyst with very useful information to explore the characteristics of the scientific domains represented. The latter capability can be applied to many different scientogram analysis and comparison tasks. Notice that, thanks to the structure of the considered scientograms, these common subgraphs will be common research categories substructures (CRCSs). In the current contribution we have considered three possible functions for our novel GBDM-based scientogram analysis tool, although we trust it will give rise to many others in the near future. In particular, we will consider the use of Subdue to: (i) study the evolution of the scientific domain of a single country over time, (ii) extract the world CRCSs in several countries at a given time, and (iii) compare the scientific domains of some countries, in terms of similarities and dissimilarities, at a given time. These three functionalities will be described in Section 3.2. Nevertheless, the diversity of the latter tasks prevent us from defining a common framework for the use of Subdue for scientogram mining. Instead, the particularities of the application of Subdue for each functionality will be discussed in the next section, describing the respective case study.

3.1. Subdue's evaluation measures and operation modes

The three substructure evaluation measures considered by Subdue are described in the following subsections, together with their interaction with the positive/negative substructure operation mode.

3.1.1. Evaluation criterion based on the minimum description length

Rissanen introduced the MDL principle (Rissanen, 1989), which suggests that the best theory to describe a dataset is that which minimizes the description length of the entire dataset. The MDL measure has been used in a rather large number of applications, ranging from decision tree induction (Quinlan & Rivest, 1989) and image processing (Leclerc, 1989; Pednault, 1989; Pentland, 1989), to concept learning from relational data (Derthick, 1991) and learning models of non-homogeneous engineering domains (Rao & Lu, 1992).

In Subdue, when the MDL measure is used, a substructure is evaluated based on how well it can compress the entire dataset. The MDL of a graph is the necessary number of bits for describing completely the graph. This number of bits is usually given by the value $I(S)$, the number of bits required to encode the substructure S. $I(S)$ is computed as the sum of the number of bits to encode the vertices of S, the number of bits to encode the edges of S, and the number of bits to encode the adjacency matrix describing the graph connectivity of S. Subdue looks for the substructure S minimizing $I(S) + I(G|S)$, where G is the input graph, $I(S)$ is the number of bits required to encode the uncovered substructure, and $I(G|S)$ is the number of bits required to encode the graph obtained by compressing G with S, i.e., substituting each occurrence of S in G by a single node (Holder, Cook, & Djoko, 1994). In the following, we renamed the MDLi measure ('i' stands for *inverse*) as we are maximizing its value: Subdue considers a given substructure S is better than another one S' if the MDLi measure $value_{MDLi}(S, G)$ is higher than $value_{MDLi}(S', G)$, where $value_{MDLi}(S, G)$ is computed as follows:

$$value_{MDLi}(S, G) = \frac{I(G)}{I(S) + I(G|S)} \tag{2}$$

Note that by maximizing the MDLi measure, the optimization of two criteria is jointly considered:

- on the one hand, the measure highlights large substructures as a better compression rate (or better MDLi value) is obtained when a bigger substructure can be extracted and replaced (compressed) by a single node;
- on the other hand, the measure highlights substructures having a large support (the support of a substructure is the number of occurrences of this substructure in the DB) as a better compression rate is obtained when many substructures are replaced (compressed) by a single node.

In our case, the graph G on which Subdue is applied is generally a single set of scientograms. However, the alternative operation mode for Subdue considers two distinct sets, a positive set G_p and a negative set G_n, determined by the user. In this operation mode, the goal of Subdue is to find the largest substructures present in the maximum number of graphs in the positive set, which are not included in the negative set. The MDLi measure is thus computed as follows:

$$value_{MDLi}(S, G_p, G_n) = \frac{I(G_p) + I(G_n)}{I(S) + I(G_p|S) + I(G_n) - I(G_n|S)} \tag{3}$$

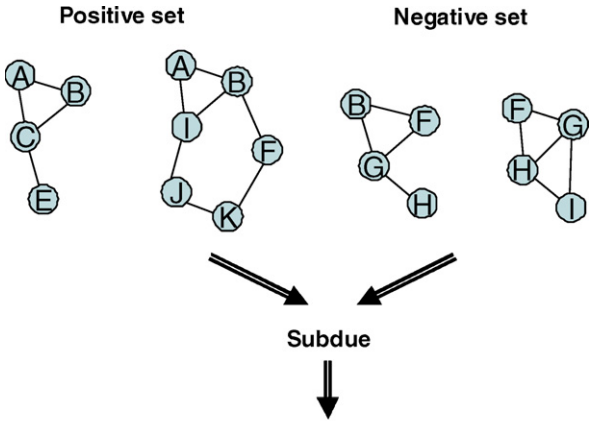
To determine which map should be positive and which map should be negative, the user will have to consider a given discriminative criterion. For instance, if we select the culture as a criterion, the positive set could comprise some graphs corresponding to European country scientograms, and the negative set some scientograms of Asiatic countries. Another example could be to consider the scientograms of a given (historical) time period of a country as a positive set, and the remaining scientograms of the same country as a negative set. Thanks to this new operation mode, much information can be extracted automatically from a scientogram DB: the most characterizing substructures for a given culture, the definition of new specificity and similarity measures between two countries, a set of substructures which illustrate the commonalities or the differences between the scientific CRCSs of two or more countries, the study of the evolution of the scientific production of a country during time, etc. Some of the latter examples will be shown in the next sections.

3.1.2. Evaluation criterion based on the substructure size

The second evaluation measure is based on the size of the substructures, the original graph and the graph compressed with the substructures. The size of an object is not computed from the description length, but from an index based on either the number of nodes, the number of edges or, more usually, the sum of the both values. This measure is faster to compute but less consistent as it does not show the real benefit obtained after the compression of the DB. It is expressed as follows:

$$value_{size}(S, G) = \frac{Size(G)}{Size(S) + Size(G|S)} \tag{4}$$

where, usually, $Size(G) = \#vertices(G) + \#edges(G)$.



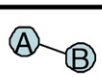
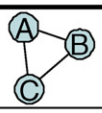
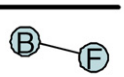
Structure			
Size (nodes/edges)	2 / 1	3 / 3	2 / 1
Support (positive:negative)	2:0	1:0	1:1
$value_{MDL}$	2.111	1.727	1.52
$value_{size}$	2.111	1.727	1.52
$value_{support}$	1	0.75	0.5

Fig. 3. Basic overview of the application of Subdue when considering positive and negative graphs in the DB.

In the case of the second operation mode, in which we have a positive and a negative scientogram set, the Size measure is computed as follows:

$$value_{size}(S, G_p, G_n) = \frac{Size(G_p) + Size(G_n)}{Size(S) + Size(G_p|S) + Size(G_n) - Size(G_n|S)} \tag{5}$$

3.1.3. Evaluation criterion based on the substructure support

The last alternative measure is based on the support of substructure S and it is expressed as follows:

$$value_{support}(S, G) = \frac{\#\text{graphs in } G \text{ including } S}{\text{card}(G)} \tag{6}$$

with card(G) being the cardinal of the set of graphs G composing the DB.

Notice that, this support definition is slightly different from the classical one used by Subdue, which computes the number of occurrences of S in the DB graph. In our case, since the DB is composed or a set of scientograms (i.e., graphs) and the nodes in a scientogram have unique labels (i.e., category names), a substructure can appear at most once in each scientogram. Hence, to compute the support of substructure S it is enough to check if it actually appears or not in each scientogram in the DB, as expressed by the previous equation.

For the second operation mode, this evaluation measure is computed as the sum of the number of positive maps containing S and the number of negative maps not containing S, divided by the total number of maps. Its formulation is as follows:

$$value_{support}(S, G_p, G_n) = \frac{\#G_p\text{graphs including}S + \#G_n\text{graphs not incl.}S}{\text{card}(G_p) + \text{card}(G_n)} \tag{7}$$

The computation of this measure is thus even faster than that of the Size, but it allows only a raw analysis of the DB.

A basic overview of the application of Subdue in its second operation mode is shown in Fig. 3. Notice that, the example substructures extracted from the DB are evaluated using the three existing measures.

3.2. Three specific scientogram analysis and comparison tasks performed through subgraph mining using Subdue

Having in mind the main components of Subdue, a detailed description of the followed aim in each of the proposed scientogram analysis and comparison tasks can now be provided in the next three subsections.

3.2.1. Study of the evolution of the scientific domain of a specific country over time

As a first illustration of the competence of the Subdue algorithm for scientogram analysis, we will depict its use to understand the evolution of a scientific domain through time. An information science expert would be interested in knowing which substructures appear in the analyzed domain, at which time, how big they are, how many they are, where are they located, and so forth. This will allow him to perform at least two kind of studies. On the one hand, an in-deep analysis of the uncovered substructures themselves, which kind of categories are they linking, etc. On the other hand, global statistics about the size and the quantity of these substructures to respectively characterize the importance of the evolution of the domain and its dynamics. This could be very helpful to perform domain comparison or domain evolution analysis (Vargas-Quesada & Moya-Anegón, 2007).

Thus, the goal of the first analysis task is to present a framework for the study of the evolution of a scientific domain over time using Subdue. First, a scientific domain is chosen: in our study, the scientific production of a whole country is considered. As we want to look for CRCSs which were appearing at a given time, we also need to pick two ranges of years, the negative range and the positive range (see a detailed explanation of Subdue's positive/negative examples operation mode in Section 3.1.1). The negative range is usually a set of years from the past, in which these substructures (i.e., CRCSs) are not meant to exist. The positive range is usually a set of years dated after the negative range, in which the substructures are meant to be present. Subdue's MDLi evaluation criterion will be considered for this aim. As Subdue will be run to extract the substructures present in the maps of the positive years but not in the maps of the negative years, it will effectively uncover the CRCSs that appeared at least once during the positive years. We should notice that if the positive range is too large, the CRCSs which appeared and disappeared several times would be also shown. This is why we will always use small positive ranges.

3.2.2. Identification of the common research categories substructures in the world

The aim of the second scientogram analysis task is to uncover the CRCSs in the world by analyzing the scientograms of a large number of different countries. To detect CRCSs, which is a localized artifact within time, there is also a need to pick a single year. All the selected maps representing the scientific production of those countries for that given year will be viewed as positive examples, so the goal of Subdue will be to extract the substructures with the best support among all of them. Notice that, no negative examples are considered in this case. As the user will be specially interested on the extracted CRCSs to be as specific as possible, the MDLi measure will be again considered to extract both frequent and large substructures.

3.2.3. Comparison of the scientific domains of different countries

The goal of the last scientogram analysis functionality is to estimate the specificity of the research developed in a given country with respect to a set of previously selected countries (i.e., to perform scientogram comparison). To do so, the scientogram of each country in a given set is compared against the remaining ones in that set, the current country viewed as a positive map and the others as negative maps. Apart from the countries list, we also need to select a specific year. The goal of Subdue is thus to extract the substructures contained in the single positive map, having the best (smaller) negative support among all the remainder for the given year in order to highlight the specific research connections defining that country. This experiment is similar to the leave-one-out cross-validation technique, well known in statistical analysis (Fukunaga, 1990). Actually, in our case, it should be called the leave-one-*in* technique. Note that this experiment could also be done using time periods larger than a single year, or more than one country in the positive set each time, thus allowing an expert to extract the substructures highlighting the possible similarities between these countries (similar to the *k*-fold cross-validation in statistical analysis). As the run time of the comparison is a key issue, the Size measure will be considered to extract both frequent and large substructures (see Section 3.1.2).

The previously described analysis tasks are now detailed in the next section as three different case studies.

4. Case study 1: evolution of a scientific domain over time

Two of the countries listed in Table 1 have been selected for this case study, Ukraine and United States. The ten scientograms corresponding to the 1996–2005 period are considered for each country. We have set up the parameters of Subdue so that it finds the best 300 substructures regarding their MDLi-based evaluation (see Section 3.1.1), and a BeamWidth of 4 to allow small response times. We performed our tests on an Intel Quad-Core 2.40 GHz CPU with 2 GB of memory, obtaining a computation time inferior to 3 seconds. In all the following discussions the substructure support is reported using two numbers (such as 3:4, for instance), with the first number being the support in the positive set (corresponding to the scientograms in the positive years), and the second number being the support in the negative set (corresponding to the scientograms in the negative years). We consider a substructure having a larger positive support and a smaller negative support as having a better quality. In the same way, substructures having a larger size are preferred over smaller ones as they are more specific.

First of all, we will look the Ukrainian scientograms domain with 7 negative years (between 1996 and 2002) and 3 positive years (between 2003 and 2005). Using Subdue, we have uncovered 300 substructures sizing from 1 to 23 nodes, having a maximum support of 3 in the positive set and a minimum support of 0 in the negative set (i.e., the best possible values for both sets).

Table 2
Support and size of the substructures extracted from the Ukrainian dataset.

Support (pos:neg)	#Subs.	Size (nodes)			Size (edges)		
		min	max	avg	min	max	avg
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
Total	**300**			**4.45**			**3.45**

Table 2 shows the global statistics of the substructures found for this experiment. The substructures have very diverse size, ranging from 1 to 23 nodes and from 0 to 22 edges. Substructures having only one node are the most common (a 70% of the total). Among them, 3 substructures have a support of 3:0. These nodes are respectively *Leadership and Management*, *Philosophy*, and *Media Technology*, indicating the Ukrainian researchers developed exclusively research in these categories after 2003. On the other hand, 71 substructures were found with a support of 3:1, among them 5 have the maximal size of 23 nodes. Overall, the most interesting substructures, those having a null negative support as well as the largest ones, are not numerous, thus allowing an expert to quickly browse and analyze all of them.

As an example, Fig. 4 shows one of these substructures comprised by 23 nodes and 22 edges, and its location within the full scientogram of the Ukrainian scientific production in 2005. As can be seen, this substructure is quite large and appears only during the last three years (actually the negative support of 1 comes from the fact that it also appears in the scientogram of 1998). This large substructure has in fact two main clusters, *Biochemistry* and *Physics and Astronomy*, suggesting the research focuses on these topics during the three last years. It occupies the center of the map, where the backbone of the Ukrainian research is concentrated. Note also that, even if *Biochemistry* occupies in general the central part of the scientograms (Vargas-Quesada & Moya-Anegón, 2007), the fact that it lies in the central part of an extracted common substructure is irrelevant.

We can make a comparison with a totally different country to see what kind of differences can be observed. Exploring what happens in the United States for the same period shows us that significantly more smallest substructures are highlighted.

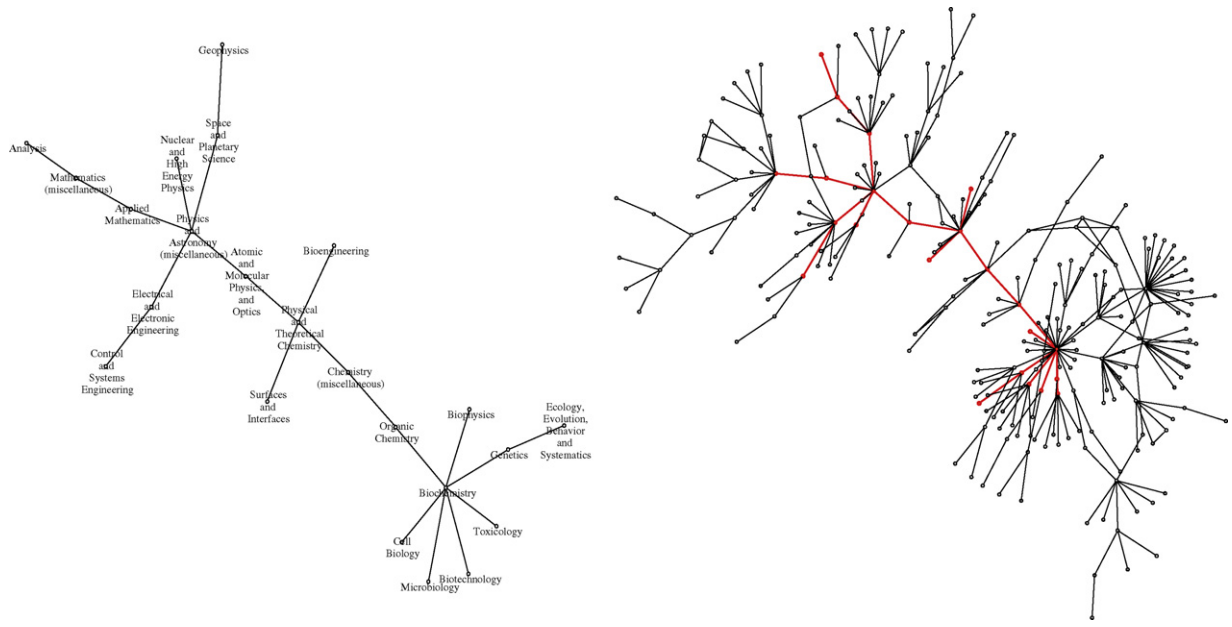


Fig. 4. One of the substructures uncovered in Ukrainian scientograms during period 2003–2005 (on the left), and its location within the 2005 scientogram (on the right).

Table 3
Support and size of the substructures extracted from the United States dataset.

Support (pos:neg)	#Subs.	Size (nodes)			Size (edges)		
		min	max	avg	min	max	avg
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
Total	**300**			**2.04**			**1.04**

Fig. 5. Some substructures uncovered in the United States scientograms during years 2003–2005.

300 substructures have been extracted, ranging from 1 to 13 nodes and from 0 to 12 edges, having an average size of 2 nodes instead of 4.5 nodes as in the Ukrainian case (see Table 3). Three substructures were obtained with the best maximum support (that is, 3:0), but they are similar to those observed in the Ukrainian domain, as they have only one node. Fig. 5 shows more interesting substructures which appear during period 2003–2005 in the United States, all of them having a support of 2:1 and a size of only 13 nodes. We could presumably assume that differences in the form of smaller substructures is an evidence of countries having a more established research track.

The study we conducted did not exactly show the real evolution over the years as we were limited by the choice of the year ranges, which is made by the information science expert. In order to have a deeper insight of the data, we have conducted another study in which this range is not fixed by the user, but it is defined by moving windows. We start with five negative years and two positive years, and we add a new positive year and remove the oldest negative year at each step. Note that the use of a moving window of only one year size does not generate any substructure due to the fact that there is not enough data to process.

As a matter of comparison with the previous study, we will use the United States dataset to detect smaller changes within the years, using a moving window of two positive years. Many substructures are extracted following this approach, but we kept only those corresponding to a support of 2:1 or 2:0, i.e., the maximal possible support for this experiment. Table 4 presents some statistics for this experiment. In general, all the uncovered substructures present a small size, ranging from

Table 4
Support and size for some substructures extracted from the United States dataset using a moving window of two positive years.

Year ranges		Support (pos:neg)	#Inst.	Size (nodes)		
(negative)	(positive)			min	max	avg
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

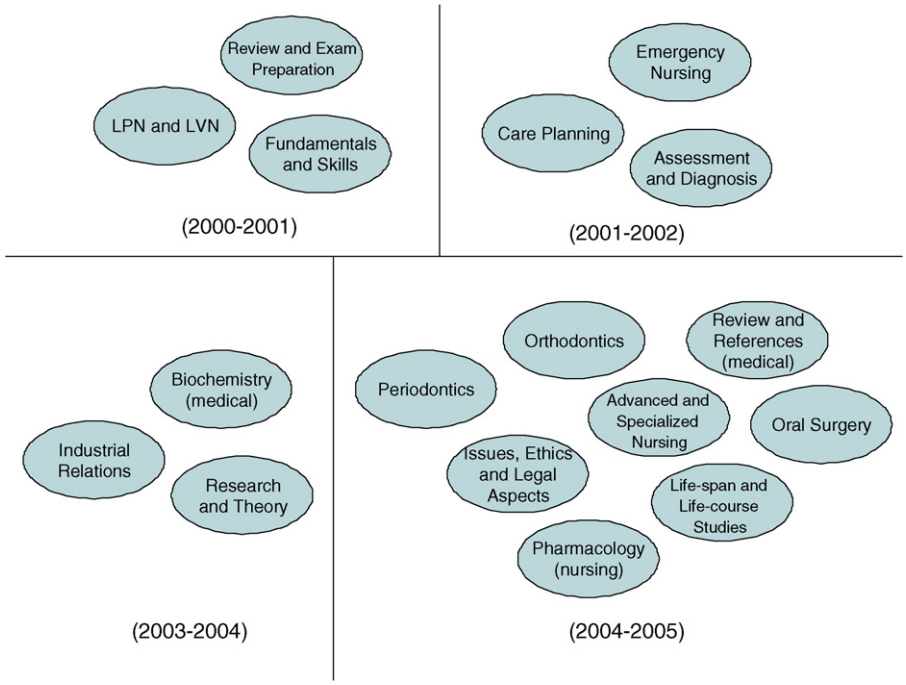


Fig. 6. Some substructures which appear repeatedly between 2000 and 2005 in the United States scientograms.

1 to 15 nodes but being equal to 1 in a 79% of the cases. All the substructures having a support of 2:0 are presented in Fig. 6. These substructures are small as they are composed of only one node. However, even if they are independent, some relationships could be found between them. For instance, during period 2001–2002 research areas focused on care, diagnosis, and emergency appeared. During period 2004–2005, more research areas focused on medical specialities (orthodontics, periodontics, oral surgery, pharmacology, etc.) made their apparition.

We should also remark an unusual fact, the high number of instances obtained considering periods 2001–2002 and 2004–2005 with a support of 2:1. We respectively obtained 55 and 32 substructures for those periods, two quite large numbers when compared with the remaining statistics. During these periods, the research in the United States evolved enough to produce a lot of changes in the corresponding maps. These concerned categories mainly belong to the medical domain, such as *Emergency Nursing*, *Care Planning*, *Oral Surgery*, *Orthodontics*, etc. Note also that only an automatic approach can quickly find and highlight those periods with larger changes.

To conclude this first case study, we can say that Subdue is a useful tool to identify the new CRCSs in a given country and during a given set of years. By looking into the specific research topics developed from one year to another one, or even looking at the global statistics, one can figure out some relevant information about the evolution of research in that country. Notice how the extracted substructures are not always located in the scientogram backbone but in other different parts of the map, thus making the use of Subdue become a complementary analysis tool to the existing low-level approaches (see Section 2.2).

5. Case study 2: common research categories substructures in the world

The full list of 73 considered countries (see Table 1) was selected for the current case study. Since data of their scientific production for year 2005 is available for all of them, this was the period chosen. 2005 corresponds to the last year we have in our DB. The parameters we used for Subdue remain the same than in the previous experiment. On our computer, we obtained a computation time of 59.39 s for this experiment.

Again, 300 substructures have been found, ranging from 1 to 12 nodes and from 0 to 11 edges (having an average size smaller than 2). For the sake of clarity, we removed all the substructures having a single node, only reporting the remaining ones in Table 5. 44 substructures have thus been kept, with a size ranging from 2 to 12 nodes (showing an average of 7) and with a support ranging from 10 to 73. As expected, the higher the support, the smaller the substructure, and *vice versa*. The largest substructures (let say, at least 10 nodes) have a small support, between 10 and 15, which comprises a perfectly expected behavior as they are more complex. On the opposite, only one CRCS is found with the best maximum support (73) and it is composed of two nodes and a single link (*Physics and Astronomy (miscellaneous) —Condensed Matter Physics*). In such way, we can conclude the latter is the most representative CRCS of the research developed in the 73 considered countries in 2005 as it is the only one existing in every scientogram.

Table 5
Support and size of the substructures extracted for the 2005 world CRCSs case study.

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**

Table 6 collects detailed statistics of the twelve largest CRCSs in the world scientific research in 2005. As can be seen, all these substructures concern a relatively small set of countries, only 20 of the 73 countries share all of them. Three main clusters of countries can also be easily extracted. Belgium, Jordan, India and Morocco scientograms share substructures #19, #20, #25, #31, and #35. Mexico, Austria and Hungary scientograms include substructures #18, #21, #30, #37, #38, and

Table 6
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

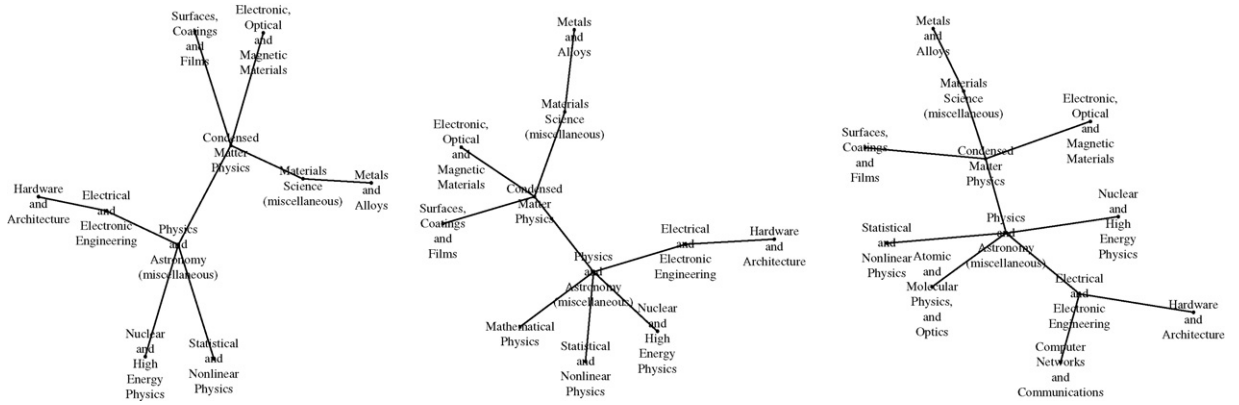


Fig. 7. Some of the largest CRCSs extracted from the world scientific research in 2005: left: index: #20, support: 15, size: 10 nodes, 9 edges; center: index: #25, support: 13, size: 11 nodes, 10 edges; right: index: #40, support: 10, size: 12 nodes, 11 edges.

#39. Finally, Czech Republic, Algeria, Turkey, Poland and Pakistan scientograms contain all the twelve largest uncovered substructures, while Bulgaria is only missing one CRCS (#38) and Ukraine is missing two of them (#35 and #39). Some geographic proximity can be identified within the countries in this list: Austria and Hungary; Czech Republic, Bulgaria, Poland, and Ukraine; Morocco and Algeria; India and Pakistan; and Jordan and Turkey.

We should also remark the specific case of Niger, which only appears in one of the twelve extracted CRCS (#31). This is an interesting case that deserves a specific analysis. Niger's scientific production in 2005 only comprises 1785 documents distributed in only 27 categories. In this way, it is unlikely that a common CRCS could be extracted relating the research developed in this country and in the remainder of the world, although it is not impossible, as it is the case in our experiment. In view of the available scientific production data for the country, there is some chance that, regardless the fact that it is a low developed country, the government could be applying any kind of scientific policy trying to mimic the research patterns of the more developed countries. Nevertheless, the very low number of scientific documents and categories are not enough to corroborate that supposition. In view of that, the fact that the scientific structure of a country like Niger shares a CRCS with other countries seems to be a coincidence, especially having in mind the fact that the handled structures are a consequence of the authors' citations. The composition of the twelve CRCSs is actually similar as they are mainly formed by a common group of linked nodes plus some additional specific ones. The common nodes are *Materials Science (miscellaneous)*; *Electronic, Optical and Magnetic Materials*; *Condensed Matter Physics*; *Physics and Astronomy (miscellaneous)*; *Electrical and Electronic Engineering*; *Nuclear and High Energy Physics*; *Hardware and Architecture*; *Metals and Alloys*; and *Statistical and Nonlinear Physics*. These nodes are all connected in the same way in every substructure and they form what we can call the *main backbone* of the research scientogram of the 73 considered countries in 2005 (i.e., the most complex CRCS in the world research in that year). As an example, Fig. 7 shows three representative substructures from this group.

Some medium-sized substructures are also worth to be analyzed and they are thus presented in Fig. 8. They only have 6 nodes but a high support of 31, 32, and 38 (more than the half of the considered countries), respectively. Of course, all of them concern Physics research and include a significant part of the said "main backbone": CRCS #9 shares four nodes with that backbone (*Electronic*, *Optical and Magnetic Materials*; *Condensed Matter Physics*; *Physics and Astronomy* (*miscellaneous*);

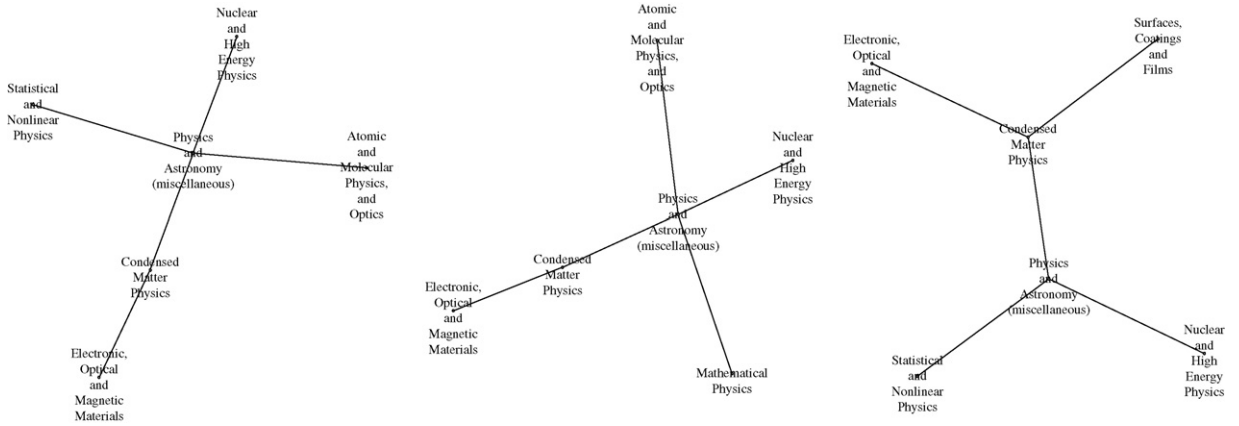


Fig. 8. Some medium-sized CRCSs extracted from the world scientific research in 2005: left: index: #5, support: 38, size: 6 nodes, 5 edges; center: index: #9, support: 32, size: 6 nodes, 5 edges; right: index: #11, support: 31, size: 6 nodes, 5 edges.

and *Nuclear and High Energy Physics*), while CRCSs #5 and #11 share five nodes (the latter four plus *Statistical and Nonlinear Physics*). Besides, they also show their particularities representing some scientific specializations of the countries containing the respective CRCS in their scientograms: *Atomic and Molecular Physics, and Optics* in CRCS #5; *Atomic and Molecular Physics, and Optics*, and *Mathematical Physics* in CRCS #9; and *Surfaces, Coatings and Films* in CRCS #11.

To conclude this second case study, we can say that Subdue allows us to detect the CRCSs in the entire world in an automatic fashion. The consideration of a larger number of maps combined with the size of these maps did not slow down the algorithm too much when following an appropriate parameterization for BeamWidth (it took around one minute of computation in the same computer used for the previous experiment). The algorithm allows us to obtain various statistics, ranging from individually extracted substructures and averaged node size or support, to global information such as countries or substructure groups. Such data can help an expert to discover relationships between world research and information extracted from current geopolitical sources.

6. Case study 3: comparison of scientific domains between different countries

For this third case study, ten maps of 2005 were selected (see Table 7) from different parts of the world (4 European countries, 3 Asiatic countries, 1 African country, 1 American country, and 1 Eurasia country). We are expecting to extract substructures highlighting specificities between these countries, but also common points depending on the part of the world in which they are located. We have designed the experimental setup in order to test this hypothesis. To determine the specificity of each country, we will perform a leave-one-in experiment (see Section 3.2.3). To evaluate the similarities between several countries, we will perform a pair-comparison experiment. They both are presented in the next two sections.

6.1. Detecting specificities: the leave-one-in experiment

For this experiment, we run Subdue ten times, each time taking one of the ten selected countries as the single positive map and all the others as negative maps. The total amount of runtime was 29.3 s. Fig. 9 represents the support and size histograms of the 3000 obtained substructures (300 for each run). In view of the former, the most interesting substructures uncovered (having the ideal support of 1:0) are not so frequent (only 55, less than a 2% of the total) although there is a significant number of interesting substructures with an 1:1 support (630, around a 21% of the total). The size histogram (which directly collects the 1138 uncovered substructures which were globally unique) shows us how every possible single node substructure, corresponding to the 288 unique SCOPUS-SJR categories, was extracted. Another important conclusion drawn from this histogram is that most of the country-specific substructures identified have a significant complexity (6.12 nodes in average), thus being quite informative and justifying the current experiment.

Table 7
List of the ten countries selected for the scientific domain comparison experiment.

Countries		
European	Asiatic	Others
France	China	Algeria
Italy	India	Cuba
Poland	Japan	Russia
Spain		

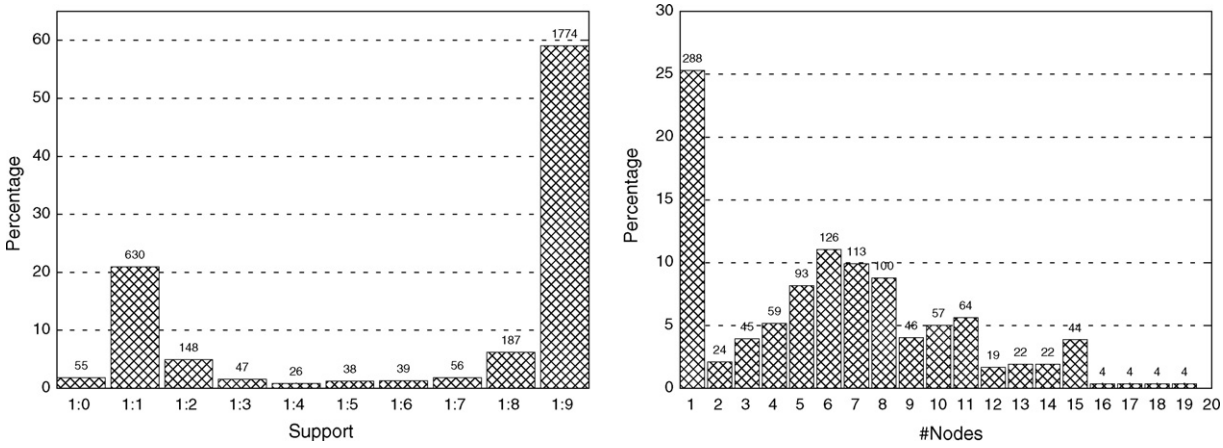


Fig. 9. Support and size histograms of the extracted substructures in the leave-one-in experiment.

Table 8
Detailed statistics of the best substructures extracted from the leave-one-in experiment with ten countries in 2005. Those graphically represented in Fig. 11 are marked with an asterisk.

Country	Support (pos:neg)	#Subs.	Size (nodes)			Size (edges)		
			min	max	avg	min	max	avg
Algeria	1:0	28	7	9 (*)	8.32	6	8	7.32
	1:1	32	3	8	5.5	2	7	4.5
	(Others)	240	1	3	1.86	0	2	0.86
China	1:1	115	1	8 (*)	6.91	0	7	5.91
	1:2	8	1	4	3.38	0	3	2.38
	(Others)	177	1	3	1.27	0	2	0.27
Cuba	1:0	16	8	11 (*)	10.31	8	10	9.56
	1:1	42	2	10	5.95	1	9	4.95
	(Others)	242	1	3	1.75	0	2	0.75
France	1:0	1	1	1	1	0	0	0
	1:1	48	1	14 (*)	8.54	0	13	7.54
	(Others)	251	1	7	1.63	0	6	0.63
India	1:1	5	1	3 (*)	2.2	0	2	1.2
	1:2	101	1	15	11.65	0	14	10.65
	(Others)	194	1	2	1.17	0	1	0.17
Italy	1:0	6	7	7 (*)	7	6	6	6
	1:1	104	1	7	5.45	0	6	4.45
	(Others)	190	1	3	1.34	0	2	0.34
Japan	1:0	1	3	3 (*)	3	2	2	2
	1:1	49	2	8	6.27	1	7	5.27
	(Others)	250	1	4	1.33	0	3	0.33
Poland	1:1	134	2	11 (*)	8.81	1	10	7.81
	1:3	1	2	2	2	1	1	1
	(Others)	165	1	2	1.08	0	1	0.08
Russian Federation	1:0	2	1	1	1	0	0	0
	1:1	70	2	19 (*)	10.89	1	18	9.89
	(Others)	228	1	2	1.11	0	1	0.11
Spain	1:0	1	1	1	1	0	0	0
	1:1	31	1	14 (*)	10.39	0	13	9.39
	(Others)	268	1	14	2.78	0	13	1.78

Table 8 reports specific statistics about the group of substructures having the two most interesting support values for each selected country. In general, every country has at least one substructure with the ideal 1:0 support. The only exceptions are China, India, and Poland, which share all their substructures with at least another (not common) country. All these countries are around Russia, which could probably explain a local sharing and a common scientific specificity. In the case of France, Spain, and Russia, the ideal support substructures (which are unique in the former two countries) correspond to a single node, i.e., a single specific category for these countries. For all the remainder, there are more than one substructure with the ideal support and these substructures are of a larger size in average (seven or more nodes in five of the seven cases), thus becoming much more informative.

We have focused our study on the most interesting country-specific substructures uncovered, i.e., those having a support of 1:0 or 1:1, and a size larger than 2 nodes. Fig. 10 shows a graphical representation of this subset, in which the color of each circle represents the number of extracted substructures (that we could assimilate to a *specificity* measure) ranging from 4 to 134, and the size of each circle represents the average size of these substructures, ranging from 2.5 to 10.9 nodes. The map also shows clearly that European countries have the biggest substructures (especially Spain, France, and Russia), which are mainly located in the center (backbone) of the scientograms, while the smallest ones are associated to India. Besides, Poland is the country with the largest number of specific substructures while India has also the smallest number in this indicator.

Fig. 11 shows the largest country-specific substructure for each country. Substructures of various sizes, with the number of nodes ranging from 3 to 19, involving diverse thematics are observed. For example, the two largest substructures, from Poland and Russia, deal with *Immunology* and *Microbiology*, two categories usually located in the center of the scientograms. However, the differences in both scientific domains can be clearly observed. While in Poland the latter two categories are related to "general Medicine" (*Medicine (miscellaneous)*), in Russia they build a bridge with *Genetics*, *Biochemistry*, and *Chemistry*, showing a completely different line of research. Besides, France and Spain share more categories in central parts of their country-specific substructures, i.e., *Physics and Astronomy (miscellaneous)*, *Electrical and Electronic Engineering*, *Control and Systems Engineering*, *Applied Mathematics*, and four other categories related to *Chemistry*, and thus the proximity of their scientific domains shows to be larger. Nevertheless, the two extracted substructures also allow us to identify the different

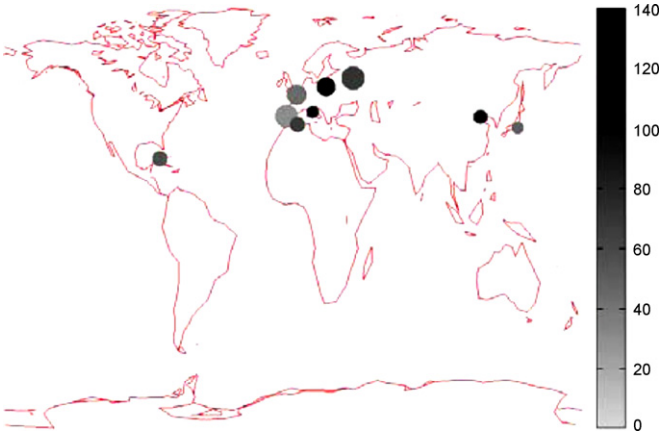


Fig. 10. Substructure specificity (circle gray scale) and average size (circle size) for each country in the leave-one-in experiment.

specializations: while in France it tends towards *Biochemistry* and *Genetics*, in Spain the latter main track of research is related to *Computer Science* categories. Notice also the huge difference of the latter two country-specific substructures with that of Italy, which is focused on *Psychology* and *Psychiatry*, regardless the geographical proximity between the three countries.

Another interesting pair-wise comparison to show the particularities of the country-specific substructures uncovered is that between Algeria and China. Notice that, both substructures show *Electrical and Electronic Engineering* in their center and focus on *Computer Science* topics. However, the categories inter-relations are completely different. The only common link is that between the former category and *Computer Networks and Communications*, while the other two shared categories, *Software* and *Computer Graphics and Computer-aided Design*, are related in a different way: direct link to the central category in the case of Algeria and indirect link between them in China. Besides, Japan's substructure is also related to *Computer Science* topics, although it is so small (showing the lack of specificity of this country's scientific domain). Its main novelty with respect to the latter two countries is the direct link between *Software* and *Control and Systems Engineering*, which can be justified by the large amount of technological research developed in Japan.

Finally, the last two countries, India and Cuba, show very specific substructures in comparison with the remaining countries studied. On the one hand, the former is related to "Public Health" and *Education*, although it shows a very small size, thus showing a possible lack of fully specific research branches in the country. On the other hand, the most representative Cuba-specific substructure is quite large and focuses on the domain of "pure and applied Mathematics", a very distinguishing research field.

6.2. Detecting similarities: the pair-comparison experiment

The goal of this experiment is to propose and study a measure of similarity between a couple of countries. We aim that this measure could characterize the proximity between the scientific domains of two countries by giving a real-valued score between 0 and 1. That score will depend on the amount of common substructures uncovered by Subdue for the scientograms of these countries. We are aware that many measures exist to compute the distance between two graphs (Cook & Holder, 2006), but they are really tricky to introduce. We propose a new domain-specific measure, directly based on the results obtained by Subdue in scientogram mining and quicker to compute.

For this experiment, we run Subdue 90 times (less than two seconds were required for each run), each time taking a different pair of the selected countries as positive maps. So each run considers only two maps. As the nodes have unique labels within the same map, the support of substructure S in a DB composed of a single scientogram N, $value_{support}(S, N)$, can only be equal to 0 or 1 (see Section 3.1.3). After each run, the following statistic was computed[3]:

$$\begin{aligned} &distance(X, Y) = 1 - proximity(X, Y) \\ &= 1 - \frac{\text{card}(\{S | value_{support}(S, X) = 1 \wedge value_{support}(S, Y) = 1\})}{\text{card}(\{S | value_{support}(S, X) = 1 \vee value_{support}(S, Y) = 1\})} \end{aligned} \tag{8}$$

where X and Y are the graphs corresponding to the scientograms of the first and the second country.

In short, we state that the proximity between two countries X and Y is given by the ratio of the number of substructures present in the intersection of X and Y (i.e., the number of CRCSs identified), and the number of substructures present in the union of X and Y (i.e., the total number of CRCSs identified). Thus, when two countries do not share any CRCSs, the numerator

[3] Note that this statistic was obtained using the limited set of substructures found by Subdue with a given set of parameters, and it is not related to a theoretical measure concerning an unlimited set of substructures. We consider it as a *quick-to-compute* approximation of what could be a real distance measure. For instance, the term card $(\{S | value_{support}(S, X) = 1\})$ is difficult to compute directly on a large graph.

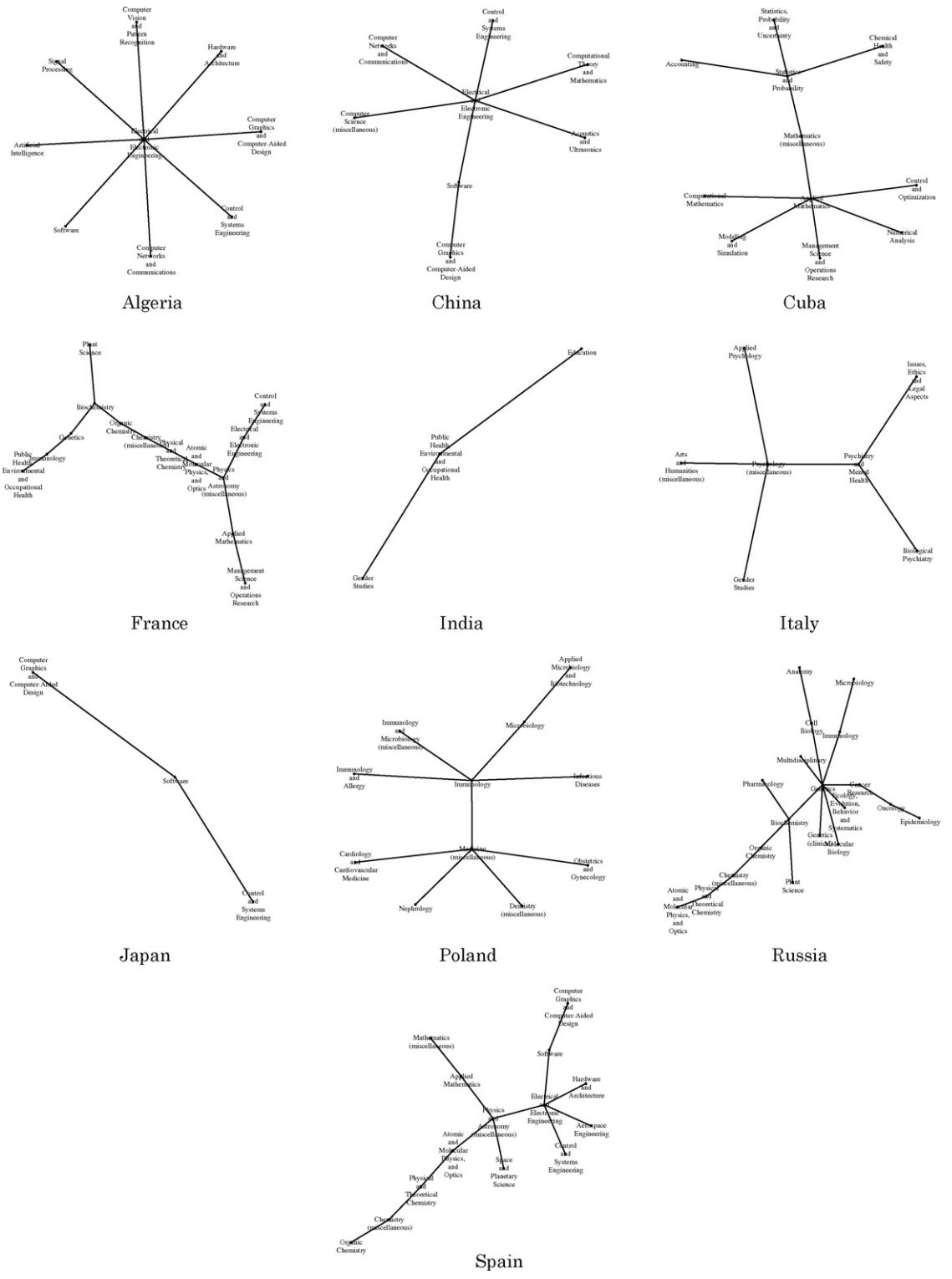


Fig. 11. The best country-specific substructures of the ten countries in 2005.

Table 9
Distance matrix (symmetric) obtained from the pair-comparison experiment.

	France	Spain	Italy	Poland	China	Japan	India	Russia	Algeria	Cuba
France	–	0.631	0.543	0.683	0.671	0.630	0.703	0.722	0.818	0.790
Spain	0.631	–	0.527	0.672	0.661	0.673	0.682	0.742	0.828	0.801
Italy	0.543	0.527	–	0.616	0.639	0.597	0.648	0.729	0.804	0.788
Poland	0.683	0.672	0.616	–	0.662	0.643	0.676	0.732	0.801	0.791
China	0.671	0.661	0.639	0.662	–	0.550	0.611	0.704	0.794	0.817
Japan	0.630	0.673	0.597	0.643	0.550	–	0.624	0.683	0.810	0.824
India	0.703	0.682	0.648	0.676	0.611	0.624	–	0.737	0.805	0.786
Russia	0.722	0.742	0.729	0.732	0.704	0.683	0.737	–	0.854	0.829
Algeria	0.818	0.828	0.804	0.801	0.794	0.810	0.805	0.854	–	0.868
Cuba	0.790	0.801	0.788	0.791	0.817	0.824	0.786	0.829	0.868	–

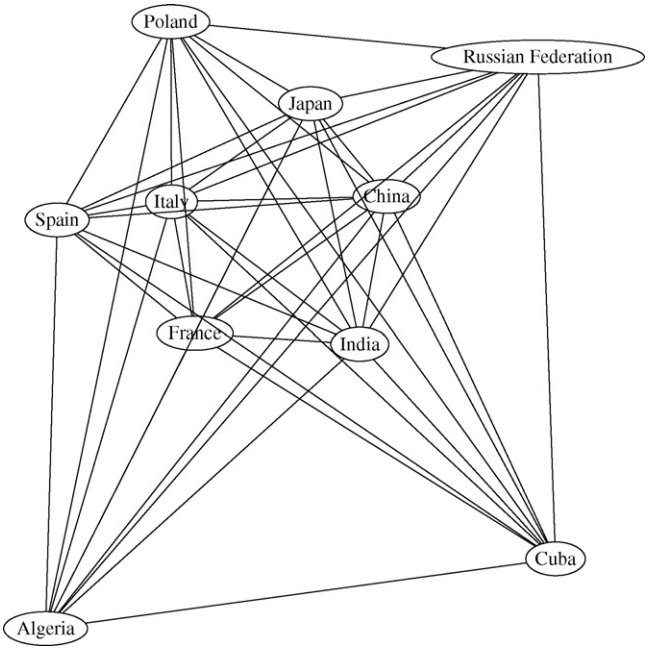


Fig. 12. A *proximity map*: the representation of the distances between the ten countries used in the pair-comparison experiment.

is equal to 0, as the proximity value, and the distance is maximal (having value 1). On the opposite, if they share many CRCSs, these two numbers become equal, the ratio tends to 1, and we can say that these two countries have a high proximity and a distance close to 0. Notice that, the *proximity* metric[4] is equivalent to the Jaccard index (Jaccard, 1912), frequently used in information retrieval (Grossman & Frieder, 1998; Rousseau, 1998).

The count of every substructure S having a support of 1 in a network $N(\{S|value_{support}(S, N) = 1\})$ could require a high computation time for large networks. So, we limited the search to substructures having between 2 and 3 nodes. This provided us with an approximation of the distance between two countries that is far enough for a graphical exploitation of the results. With these parameters, the total runtime was about 170.83 s. Between 214 and 1965 substructures were obtained for each run. The support for each of them could obviously be either 2:0 or 1:0.

The final result of this experiment is a square matrix of dimension $N \times N$ ($10\times$ 10 in our case), in which each cell corresponds to *distance*(X, Y), and where the diagonal is equal to 0 (see Table 9). A graphical representation can be extracted from this distance matrix in which countries having a larger proximity value appear close together. This representation can be done using a two dimensional spring-layout visualization method, for instance the Kamada–Kawai algorithm (see Section 2.1.4). Fig. 12 shows this representation by setting, for the weight of each link, the cube of the real distance measure value in order to emphasize the distance and the proximity between the countries, which would not be the case using the distance value directly. We call this representation a *proximity map*.

Even if obtained using only the co-citation matrices, this *proximity map* demonstrates very surprising geographical coincidences. Japan and China appear close together, as well as Poland and Russia, and Spain, France, and Italy; while Russia and

[4] This measure verifies the triangular inequality, the non-negativity, the identity of indiscernible, and the symmetry, and thus it can be considered as a metric.

Cuba, or Russia and Algeria are far away from each other. Apart from the trivial confirmation of the geographical proximity due to the evident collaborations between countries which share or at least have any historical, socio-economical or ideological likeness, other relationships appear. For instance, Italy appears in the center of the map, close to non-related countries such as Japan, China, or India. As suggested and demanded by this map, further analysis should be performed to understand better this behavior, employing other kind of tools (Moya-Anegón et al., 2007; Vargas-Quesada & Moya-Anegón, 2007).

We should also point out some cons of this representation. Basically, a larger matrix (e.g., one with a size of 73×73 to extend this study to our complete database) would make the representation, as well as the drawn conclusions, more realistic. However, a so large map would be probably complex to be drawn in 2D and to be understood, even if no edges are drawn. Interactive applets or further simplifications such as the use of an additional pruning algorithm would be needed for that purpose. Another issue with this representation is the notion of *projected view*. Just by considering the distances in a subgraph composed of three or more nodes, the visualization would already be wrong since the represented distances would never perfectly match the actual ones, as expected when a perfect replication of the distance matrix is wanted. This is due to the fact that it is impossible in an n-dimensional space to solve all the geometric constraints to make the Euclidean distance between two points equal to the distance value given in the matrix. In fact, the Kamada–Kawai algorithm is used to provide a projection as close as possible to the distance matrix for some countries, but this carries some side effects for others. In general, only the global idea of the distance matrix (the shortest and the largest distances) is reflected in the graphical map.

In a nutshell, and to conclude this case study, we can see that different protocols can be used to explore different aspects of the same dataset. The leave-one-in experiment is useful to detect specificities within a set of countries, while the pair-comparison experiment is useful to detect similarities between them. For the latter experiment, note that the defined distance is very simple, easy and fast to compute and gives coherent results with respect to the geographical proximity between the selected countries. Of course, other metrics and a deeper analysis of these statistics should be performed, but this gives a first insight of the possibilities provided by the use of GBDM approaches (in particular, the Subdue algorithm) to analyze and compare scientific domains.

7. Conclusion

In this paper, we showed how a GBDM technique can be successfully applied to the complex task of scientogram analysis and comparison. A large amount of data (73 countries over 10 years, covering scientograms with 159 135 nodes and 172 081 edges in total) have been processed to extract a great number of different units of analysis (global statistics, local substructures, histograms, proximity map, etc.). From the diversity of the results obtained, as well as from the different ways of applying the methodology in order to address diverse requests (evolution of a research domain, identification of world CRCSs, and uncovering of similarities and specificities in the scientific production of different countries), we can see that the proposed approach is an efficient and powerful tool which is able to filter, reduce, and provide a help to analyze such data.

The methodology is scalable and will not suffer if applied to an increased volume of data. It has been shown that the generation of the graph visualizations, graph highlights (see Fig. 4), tables and histograms is fully automatic. Even if only the Subdue algorithm was used in this proposal, the measures (the MDLi principle, the size, the support, and the distance metric) are generic and thus other GBDM algorithms can be considered. For these reasons, GBDM can be viewed as a novel scientogram analysis tool developed in complement to the current state-of-the-art techniques. Its ability to detect and identify micro-substructures (at the disciplines level) from as many scientograms as wanted, makes it become an essential tool for the comparison and the study of the evolution of scientific domains.

Probably the most interesting challenge would be the extraction of non intuitive substructures that will bring a real added value to the experts. Anyway, the possibilities of GBDM are not only limited to the applications considered in this paper. It can also be used for the detection of institutions with similar interests and goals in a given scientific domain or discipline, for the identification of potential collaborators, either at a personal or an institutional level, for the study of the scientific collaboration at a institutional or national level, etc. Our next works will be devoted to the latter issues.

Acknowledgments

This work has been supported by the Spanish Ministerio de Ciencia e Innovación under project TIN2009-07727, including EDRF fundings. We would like to thank Elsevier for its permission to use the SCOPUS-SJR data in order to build and compare the scientograms.

References

Borgelt, C., & Berthold, M. R. (2002). Mining molecular fragments: Finding relevant substructures of molecules. In *IEEE international conference on data mining (ICDM'02)* IEEE Computer Society, Washington, (pp. 51–58).
Börner, K., & Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics*, *3*(3), 161–172.
Boyack, K. W., Börner, K., & Klavans, R. (2009). Mapping the structure and evolution of chemistry research. *Scientometrics*, *79*(1), 45–60.
Boyack, K. W., & Klavans, R. (2008). Measuring science–technology interaction using rare inventor-author names. *Journal of Informetrics*, *2*(3), 173–182.
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.
Chen, C. (1998). Bridging the gap: the use of Pathfinder networks in visual navigation. *Journal of Visual Languages and Computing*, *9*, 267–286.
Chen, C. (1999a). *Information visualization and virtual environments*. Berlin: Springer.
Chen, C. (1999b). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, *35*, 401–420.

Chen, C. (2004). *Information visualization: Beyond the horizon*. Berlin: Springer.
Chen, C., Chen, Y., Horowitz, H., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191–209.
Chittimori, R., Gonzalez, J., & Holder, L. B. (1999). Structural knowledge discovery in chemical and spatio-temporal databases. In *Sixteenth national conference on artificial intelligence and eleventh conference on innovative applications of artificial intelligence (AAAI/IAAI)* Orlando, (p. 959).
Cook, D. J., & Holder, L. B. (1994). Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, *1*, 231–255.
Cook, D. J., & Holder, L. B. (2000). Graph-based data mining. *IEEE Intelligent Systems*, *15*(2), 32–41.
Cook, D. J., & Holder, L. B. (Eds.). (2006). *Mining graph data*. New Jersey: Wiley.
Dearholt, D., & Schvaneveldt, R. (1990). Properties of Pathfinder networks. In R. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 1–30). Ablex Publishing Corporation.
Derthick, M. (1991). A minimal encoding approach to feature discovery. In *Ninth national conference on artificial intelligence* Anaheim, (pp. 565–571).
Deshpande, M., Kuramochi, M., & Karypis, G. (2002). Automated approaches for classifying structures. In *Second workshop on data mining in bioinformatics (BIOKDD)* Alberta, (pp. 11–18).
Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). New York: Academic Press.
Gonzalez, J. A., Holder, L. B., & Cook, D. J. (2000). Structural knowledge discovery used to analyze earthquake activity. In *Thirteenth annual Florida artificial intelligence research symposium (FLAIRS)* Florida, (pp. 86–90).
Griffith, B. C., Small, H., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literature, part II: Toward a macro and microstructure for science. *Science Studies*, *4*(4), 339–365.
Grossman, D. A., & Frieder, O. (1998). *Information retrieval: algorithms and heuristics*. New York: Springer.
Guerrero-Bote, V. P., Vargas-Quesada, B., Espinosa-Calvo, M. E., & Moya-Anegón, F. (2009). Estudio comparativo de seis dominios cientificos nacionales (in Spanish). *Revista Española de Documentación Científica*, *32*(3), 9–28.
Holder, L. B., Cook, D. J. (2005). Graph-based data mining. In: J. Wang (Ed.), *Encyclopedia of data warehousing and mining: Vol. II. Information Science Reference, Hershey* (pp. 943–949).
Holder, L. B., Cook, D. J., Coble, J., & Mukherjee, M. (2005). Graph-based relational learning with application to security. *Fundamenta Informaticae, Special Issue on Mining Graphs, Trees and Sequences*, *6*(1–2), 83–101.
Holder, L. B., Cook, D. J., & Djoko, S. (1994). Substructure discovery in the SUBDUE system. In *AAAI workshop on knowledge discovery in databases* Seattle, (pp. 169–180).
Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., & Tropsha, A. (2004). Mining protein family specific residue packing patterns from protein structure graphs. In *Eighth annual international conference on research in computational molecular biology (RECOMB'04)* ACM, New York, (pp. 308–315).
Huan, J., Wang, W., & Prins, J. (2003). Efficient mining of frequent subgraphs in the presence of isomorphism. In *Third IEEE international conference on data mining (ICDM'03)* IEEE Computer Society, Washington, (pp. 549–552).
Huan, J., Wang, W., Prins, J., & Yang, J. (2004). Spin: mining maximal frequent subgraphs from graph databases. In *Tenth ACM SIGKDD international conference on knowledge discovery and data mining (KDD'04)* ACM, New York, (pp. 581–586).
Inokuchi, A., Washio, T., & Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *Fourth European conference on principles of data mining and knowledge discovery (PKDD'00)* Springer-Verlag, London, (pp. 13–23).
Jaccard, P. (1912). The distribution of flora in the alpine zone. *The New Phytologist*, *11*(2), 37–50.
Jonyer, I., Cook, D. J., & Holder, L. B. (2001). Graph-based hierarchical conceptual clustering. *Journal of Machine Learning Research*, *2*, 19–43.
Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15.
Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1–8.
Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics*, *68*(3), 475–499.
Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.
Koyutürk, M., Grama, A., & Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, *20*(1), 200–207.
Kukluk, J., You, C., Holder, L. B., & Cook, D. J. (2007). Learning node replacement graph grammars in metabolic pathways. In *International conference on bioinformatics and computational biology (BIOCOMP-07)* Las Vegas, (pp. 44–50).
Kuramochi, M., & Karypis, G. (2001). Frequent subgraph discovery. In *IEEE international conference on data mining (ICDM'01)* IEEE Computer Society, Washington, (pp. 313–320).
Leclerc, Y. G. (1989). Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, *3*(1), 73–102.
Leydesdorff, L. (2007). Mapping interdisciplinarity at the interfaces between the science citation index and the social science citation index. *Scientometrics*, *71*(3), 391–405.
Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.
Leydesdorff, L., & Schank, T. (2008). Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *Journal of the American Society for Information Science and Technology*, *59*(11), 1810–1818.
Lowerre, B. T. (1976). *The HARPY speech recognition system*. Ph.D. thesis. Pittsburgh: Carnegie Mellon University.
Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite(TM)-based historiograms. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962.
Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology*, *58*(14), 2167–2179.
Moya-Anegón, F., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Herrero-Solana, V., Corera-Álvarez, E., & Munoz-Fernández, F. J. (2005). Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. *Information Processing & Management*, *41*(6), 1520–1533.
Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, *61*(1), 129–145.
Nijssen, S., & Kok, J. N. (2004). A quickstart in frequent structure mining can make a difference. In *Tenth ACM SIGKDD international conference on knowledge discovery and data mining (KDD'04)* ACM, New York, (pp. 647–652).
Pednault, E. P. D. (1989). Some experiments in applying inductive inference principles to surface reconstruction. In *Eleventh international joint conference on artificial intelligence* Detroit, (pp. 1603–1609).
Pentland, A. (1989). Part segmentation for object recognition. *Neural Computation*, *1*(1), 82–91.
Quinlan, J. R., & Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, *80*(3), 227–248.
Quirin, A., Cordón, O., Guerrero-Bote, V. P., Vargas-Quesada, B., & Moya-Anegón, F. (2008). A quick MST-based algorithm to obtain Pathfinder networks. *Journal of the American Society for Information Science and Technology*, *59*(12), 1912–1924.
Quirin, A., Cordón, O., Santamaría, J., Vargas-Quesada, B., & Moya-Anegón, F. (2008). A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. *Information Processing and Management*, *44*(4), 1611–1623.
Rakhshan, A., Holder, L. B., & Cook, D. J. (2004). Structural web search engine. *International Journal of Artificial Intelligence Tools*, *13*(1), 27–44.
Rao, R. B., & Lu, S. C. (1992). Learning engineering models with the minimum description length principle. In *Tenth national conference on artificial intelligence* Menlo Park, (pp. 717–722).
Rissanen, J. (1989). *Stochastic complexity in statistical inquiry theory*. River Edge: World Scientific Publishing Co., Inc.

Rousseau, R. (1998). Jaccard similarity leads to the Marczewski-Steinhaus topology for information retrieval. *Information Processing & Management*, *34*(1), 87–94.
Salton, C., & Bergmark, D. (1979). A citation study of computer science literature. *IEEE Transactions on Professional Communication*, *22*, 146–158.
Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*, *57*(11), 1461–1469.
Small, H., & Garfield, E. (1985). The geography of science: disciplinary and national mappings. *Information Science*, *11*(4), 147–159.
Small, H., & Griffith, B. C. (1974). The structure of scientific literature, part I: Identifying and graphing specialties. *Science Studies*, *4*(1), 17–40.
Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations, part I: A comparison of methods. *Scientometrics*, *7*(3–6), 391–409.
Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations, part II: Mapping science. *Scientometrics*, *8*(5–6), 321–340.
Vanetik, N., Gudes, E., & Shimony, S. E. (2002). Computing frequent graph patterns from semistructured data. In *IEEE international conference on data mining (ICDM'02)* IEEE Computer Society, Washington, (pp. 458–465).
Vargas-Quesada, B., & Moya-Anegón, F. (2007). *Visualizing the structure of science*. New York, Secaucus: Springer-Verlag.
Wallace, M. L., Gingras, Y., & Duhon, R. (2009). A new approach for detecting scientific specialties from raw cocitation networks. *Journal of the American Society for Information Science and Technology*, *60*(2), 240–246.
Washio, T., & Motoda, H. (2003). State of the art of graph-based data mining. *SIGKDD Explorations*, *5*(1), 59–68.
Yan, X., & Han, J. (2002). gSpan: Graph-based substructure pattern mining. In *IEEE international conference on data mining (ICDM'02)* IEEE Computer Society, Washington, (pp. 721–724).
Yan, X., & Han, J. (2003). CloseGraph: mining closed frequent graph patterns. In *Ninth ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03)* ACM, New York, (pp. 286–295).
Yan, X., Yu, P. S., & Han, J. (2004). Graph indexing: a frequent structure-based approach. In *ACM SIGMOD international conference on management of data (SIGMOD'04)* ACM, New York, (pp. 335–346).