Analysis of the Time Evolution of Scientograms Using the Subdue Graph Mining Algorithm

Arnaud Quirin, Oscar Cordón, Prakash Shelokar, and Carmen Zarco

European Centre for Soft Computing,
Edf. Científico Tecnológico, 33600 Mieres, Spain
{arnaud.quirin,oscar.cordon,prakash.shelokar,carmen.zarco}
@softcomputing.es

Abstract. Scientograms are a kind of graph representations depicting the state of Science in a specific domain. The automatic comparison and analysis of a set of scientograms, to show for instance the evolution of a scientific domain of a given country, is an interesting but challenging task as the handled data is huge and complex. In this paper, we aim to show that graph mining tools are useful to deal with scientogram analysis. We have chosen Subdue, a well-known graph mining algorithm, as a first approach for this purpose. Its operation mode has been customized for the study of the evolution of a scientific domain over time. Our case study clearly shows the potential of graph mining tools in scientogram analysis and it opens the door for a large number of future developments.

1 Introduction

The generation of a map of sciences or *scientogram* has been a persistent idea in the modern ages. For instance, this could be achieved by the drawing of a graph linking together different scientific research fields, topics or categories, using the co-citation rate between the papers of these categories to denote the strength of the links. It has been a persistent idea as the visualization of such information graph has long been used to uncover and divulge the structure of Science [1,2]. However, analyzing scientific data is becoming increasingly difficult due to the vast volumes of data generated nowadays. Up to our knowledge, no previous fully automatic approaches have been designed to support the exploration of large datasets in scientogram mining.

In general, the current scientogram analysis techniques perform a low-level, non-automatic analysis and comparison of the maps [3,4,5]. To do so, they are based on statistical techniques and macro- and micro-structure analysis for the identification of thematic areas and scientific disciplines [6]. 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

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, LNAI 6178, pp. 310–319, 2010.
© Springer-Verlag Berlin Heidelberg 2010

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) [7] involves the automatic extraction of novel and useful knowledge from a graph representation of data. It has been applied for frequent substructure discovery and graph matching in a large number of domains including chemistry and applied biology, classification of chemical compounds, and unsupervised and supervised pattern learning, among many others. In particular, the first proposal in the topic, Subdue [8], based on the use of the minimum description length (MDL) principle [9], has proved to be successful in many different real-world applications. 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.

The structure of the current contribution is as follows. In the second section, we review the current techniques to design and analyze scientograms as well as the current state of the art of GBDM and the particular case of the Subdue algorithm. In the third section we show how a scientogram analysis task, the study of the evolution of a scientific domain over time, can be performed by means of this algorithm. The fourth section presents the obtained results. Finally, some concluding remarks and future works 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, as well as a review of the GBDM field, describing its scope and the most known techniques.

2.1 Scientogram Design

The generation of a scientogram following a top-down approach based on the existence of a previous document category structure requires the sequential application of several techniques. The scientograms considered in this contribution are built following De Moya-Anegón et al.'s methodology [10,5]. The SCOPUS-SJR co-citation categories are used as units of analysis. Each category agglutinates the journals that were categorized under that name, and likewise the documents that were published in those journals. A co-citation measure is used to compute the relational similarity between two categories i and j. It is defined as $CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}}$, where Cc is the co-citation frequency and c is the citation frequency. The Pathfinder algorithm [11,12] is considered to prune the co-citation matrix. As a result, only the salient relationships between categories are kept, capturing the essential underlying intellectual structure of the studied scientific domain. The pruned network is then graphically represented using the Kamada-Kawai's graph drawing algorithm [13], chosen for its ability to represent naturally the most important elements in the center of the representation (called the scientogram *backbone*).

The rough considered data have been extracted from the Scimago Journal & Country Rank portal[1] (SCOPUS-SJR data) [5]. In this contribution, we will deal with the United States and the Ukrainian maps from 1996 to 2005, based on respectively 4 307 536 and 74 248 documents. Overall, the 20 scientograms used have 4991 nodes and 5304 edges.

2.2 Graph-Based Data Mining and the Subdue Algorithm

The need of mining structural data to uncover objects or concepts that relate objects (i.e., subgraphs that represent associations of features) has increased in the past decade, thus creating the area of GBDM [7]. Nowadays, many GBDM algorithms (Apriori-based GM, Frequent Subgraph Discovery, MoFa/MoSS, 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 [14].

Among them, we can highlight Subdue [8], 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 uses the MDL principle [9] to discover interesting and repetitive (frequent) substructures in a structural database (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. Through the years, it has been successfully applied to a large range of real-world problems such as aviation, chemistry, geology, counter-terrorism, bioinformatics, and web structure mining.

Fig. 1 shows the outline of the Subdue GBDM algorithm. It takes as input the original graph DB (comprised by a single graph or a set of graphs) from which the substructures (i.e. subgraphs) have to be extracted and four parameters used to limit the search while reducing the runtime. These parameters (*BeamWidth*, *MaxBest*, *MaxSubSize*, and *Limit*) constrain the number of considered substructures and the total number of iterations of the algorithm. ChildList and BestList are two ordered lists in which the substructures having the best evaluation values appear first which guide the beam search process applied. The algorithm ends up by returning the best substructures found considering the chosen evaluation measure and the constraint parameters.

The evaluation of a substructure (see line 13) can be computed by different measures, but the MDL-measure is the most popular. It measures how well a substructure can compress the entire dataset. Hence, the algorithm aims to maximize the following measure: $value_{MDLi}(S,G) = \frac{I(G)}{I(S)+I(G|S)}$ where G is the input graph, S is the candidate substructure, $I(G)$ and $I(S)$ are the number of bits required to encode G and S, 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.

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

```
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. 1. The Subdue GBDM algorithm (reprinted from [8])

3 Subdue for Scientogram Analysis. Case Study: Evolution of a Scientific Domain over Time

The application of Subdue as a powerful scientogram analysis tool will rely on its frequent subgraph mining activity (i.e., we will perform scientogram mining). Since the underlying scientogram structure is a social network (i.e., a graph), the uncovering of common subgraphs (named *Common Research Categories Substructures* or CRCSs in the following) 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. In the current contribution we have considered the use of Subdue to study the evolution of the scientific domain of a single country over time. The considered Subdue implementation is that made by the original authors, available at `http://ailab.wsu.edu/subdue/`.

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

- 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 DB 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:

$$value_{MDLi}(S, G_p, G_n) = \frac{I(G_p) + I(G_n)}{I(S) + I(G_p|S) + I(G_n) - I(G_n|S)} \quad (1)$$

The use of negative maps allows the user to consider a given discriminative criterion. For instance, for a given country, we can consider the scientograms of a given (historical) time period as a positive set, and the remaining scientograms as a negative set, to extract relevant information about the substructures appearing or disappearing during this historical transition.

When considering the latter analysis of 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, the study of some 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 [5].

To do so, a scientific domain is first 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 positive range and the negative range. 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 in equation (1) 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.

4 Experiments and Analysis of Results

Two countries 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, considering 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 values (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.

Table 1. Support and size of the substructures extracted from the Ukrainian dataset

Support		Size (nodes)			Size (edges)		
(pos:neg)	#subs.	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
TOT.	**300**			**4.45**			**3.45**

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). Table 1 shows the global statistics of the 300 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 the optimal support of 3:0. These nodes are respectively *Leadership and Management*, *Philosophy*, and *Media Technology*, indicating the Ukraine-based researchers developed research in these categories exclusively 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. 2 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

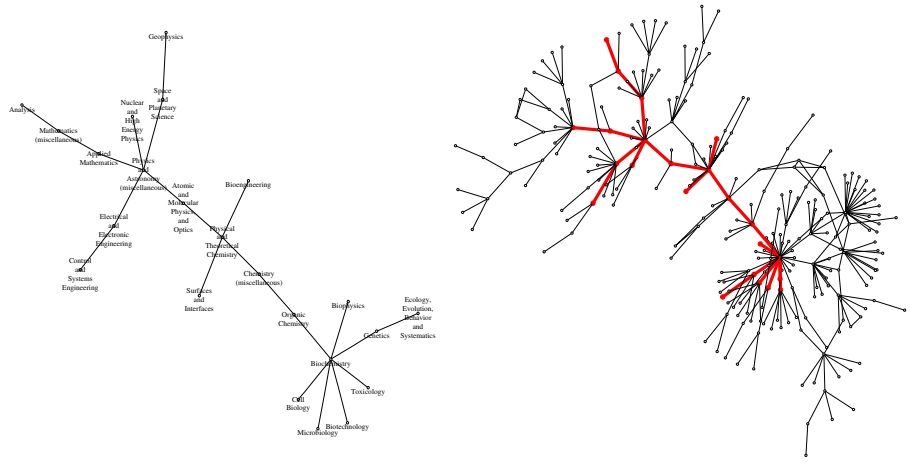

Fig. 2. One of the substructures uncovered in Ukrainian scientograms during period 2003-2005 (on the left), and its location within the 2005 scientogram (on the right)

the central part of the scientograms [5], it could perfectly appear at any location in an extracted substructure, or even not appear at all, depending on how the maps are compared together and if this node is relevant given the chosen criterion.

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

Support		Size (nodes)			Size (edges)		
(pos:neg)	#subs.	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
TOT.	**300**			**2.04**			**1.04**

On the other hand, exploring what happens in the United States for the same period shows us that significantly more smallest substructures are highlighted (see Table 2). 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. 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 only have one node. Fig. 3 shows three more interesting substructures, all of them having a support of 2:1 and a size of only 13 nodes. We interpret that differences in the form of smaller substructures is an evidence of countries having a more stable research infrastructure.

In order to have a deeper insight of the data, we have conducted another study in which the time range is not fixed by the user, but it is defined by moving

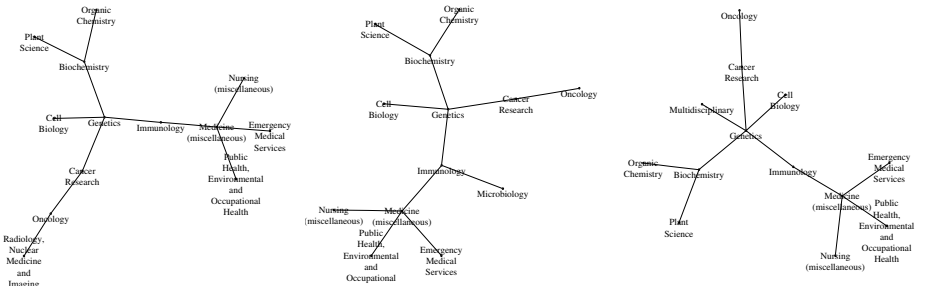

Fig. 3. Some substructures uncovered in the USA scientograms during years 2003-2005

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.

Table 3. Support and size for some substructures extracted from the United States dataset using a moving window of two positive years

Year ranges		Support		Size (nodes)		
(negative)	(positive)	(pos:neg)	#inst.	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

As a matter of comparison with the previous study, we will use the United States dataset to detect smaller changes within the 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 3 presents some statistics for this experiment. In general, all the uncovered substructures present a small size, ranging from 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. 4. 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 are found. 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*,

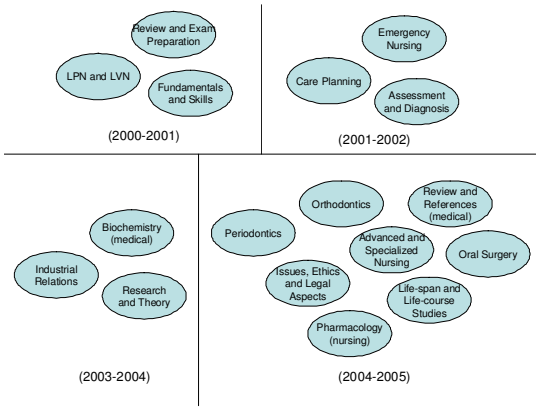

Fig. 4. Some substructures which appear repeatedly between 2000 and 2005 in the USA scientograms

etc. Note also that only an automatic approach can quickly find and highlight those periods with larger changes.

In view of the developed experiments, 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, 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.

5 Conclusions

In this paper, we showed how a GBDM technique, namely Subdue, can be successfully applied to the complex task of scientogram analysis and comparison. The scientific domains of two countries have been processed to study the evolution of research during time by extracting some interesting substructures as well as some statistical parameters.

This methodology is scalable and it 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. 2), tables, and histograms is fully automatic. Even if only the Subdue algorithm was used in this proposal, 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. In the future, we plan to use other GBDM techniques (notably multiobjective-optimization-based ones) and discover other uses of Subdue for the analysis and comparison of scientograms.

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 and Drs. Félix De Moya-Anegón and Benjamín Vargas-Quesada for their permission to use the SCOPUS-SJR data to build the scientograms.

References

1. Börner, K., Scharnhorst, A.: Visual conceptualizations and models of science. Journal of Informetrics 3(3), 161–172 (2009)
2. Chen, C.: Information Visualization: Beyond the Horizon. Springer, Berlin (2004)
3. Chen, C., Chen, Y., Horowitz, H., Hou, H., Liu, Z., Pellegrino, D.: Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics 3(3), 191–209 (2009)
4. Leydesdorff, L., Rafols, I.: A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology 60(2), 348–362 (2009)
5. Vargas-Quesada, B., Moya-Anegón, F.D.: Visualizing the Structure of Science. Springer, New York (2007)
6. Wallace, M.L., Gingras, Y., Duhon, R.: 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 (2009)
7. Washio, T., Motoda, H.: State of the art of graph-based data mining. SIGKDD Explorations 5(1), 59–68 (2003)
8. Cook, D.J., Holder, L.B.: Graph-based data mining. IEEE Intelligent Systems 15(2), 32–41 (2000)
9. Rissanen, J.: Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Inc., River Edge (1989)
10. Moya-Anegón, F.D., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F.J.: A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics 61(1), 129–145 (2004)
11. Dearholt, D., Schvaneveldt, R.: Properties of Pathfinder networks. In: Schvaneveldt, R. (ed.) Pathfinder Associative Networks: Studies in Knowledge Organization, pp. 1–30. Ablex Publishing Corporation, Greenwich (1990)
12. Quirin, A., Cordón, O., Guerrero-Bote, V.P., Vargas-Quesada, B., Moya-Anegón, F.D.: A quick MST-based algorithm to obtain Pathfinder networks. Journal of the American Society for Information Science and Technology 59(12), 1912–1924 (2008)
13. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Information Processing Letters 31(1), 7–15 (1989)
14. Cook, D.J., Holder, L.B. (eds.): Mining Graph Data. Wiley, New Jersey (2006)