DISCOVERY OF FREQUENT CORRELATIONS AMONG DEFORESTATION OBJECTS USING GRAPH MINING

A. M. Maciel, F. S. Franca, M. P. S. Silva

Master's Program in Computer Science (MCC)
Rio Grande do Norte State University (UERN)
Federal University of Semi-Arid Region (UFERSA)
Mossoro, RN, Brazil
{adelsud6,fsoaresmcc}@gmail.com, marcelinopereira@uern.br

KEY WORDS: Remote Sensing, Data Mining, Graph Mining, Frequent Substructures, Deforestation Patterns.

ABSTRACT:

Nowadays, there is a rich diversity and huge volume of data from satellite sensors. Due to their attractiveness and singular characteristics, they are increasingly gaining importance in the research context. Concomitantly, the volume of remote sensing data is quickly growing. Thus, the computational modeling of such data involves many challenges, including a large set of algorithms and techniques to extract strategic information contained in these data. This paper applies a computational mathematical formalism (graphs) to represent deforestation data, in order to perform intelligent search of patterns (graph mining) to discover frequent correlations among deforestation objects.

1 INTRODUCTION

Nowadays, a wide range of satellite data is provided which, periodically, capture images of the Earth's surface. These images are generally stored in organizations and research institutes, eg. the National Institute for Space Research (INPE, in Brazil), which holds more than 1,840,000 distributed images for the more different areas (INPE, 2011a). The Institute uses the information captured through these images to develop researches in different fields, such as the analysis of tropical deforestation.

Due to their attractiveness and exclusive features, their importance is increasing in research context, especially in undergraduate teaching. The amount of data available in repositories is also growing quickly. This scenario demands evolution of the techniques, tools and methodologies that deal with such images.

Combating deforestation in Brazil is a priority for the government, society and environmental organizations. This demands an effective monitoring and supervision, along with actions that can reduce or suppress this problem. Thus, it is necessary the creation of methodologies and techniques that allow monitoring efficiently, and at reasonable costs, areas susceptible to deforestation processes with potential damage to important biomes, especially Amazonia.

In Brazil, one of the great research challenges is the computational modelling of artificial, natural and social complex systems, as well the human-nature interactions (Medeiros, 2008). In this context, the computational modelling of remote sensing images involves important challenges, including the development of algorithms and techniques that aim the extraction of relevant information from these data.

Due to the increasing volume and complexity of databases, the search for new techniques of data mining has been emphasized (Han and Kamber, 2006). Many of these repositories have structural features, i.e., their data are composed of objects and relationships among them, allowing representation through graphs. Therefore, graph mining has provided new resources to the search of strategic information in databases with structural features, allowing pattern discovery from such structures. Given this context, this paper aims to employ graphs to represent relationships among deforestation objects, and hence extract patterns from them applying graph mining, in order to identify frequent substructures among deforestation objects.

2 REMOTE SENSING

Remote sensing is a set of activities to obtain information from objects that constitute the Earth's surface, regardless physical contact with them, using satellites (Schowengerdt, 2007). In these activities the detection, acquisition, storage and analysis (information extraction and interpretation) of the electromagnetic energy (or electromagnetic radiation) that is emitted or reflected by terrestrial objects is carried out by remote sensing systems.

This electromagnetic energy reflected and emitted by objects of the Earth's surface is the basis for the entire image processing and analysis, because though it the spectral energy reflected can be quantified and its relevant features evaluated. This way, the remote sensors are essential tools to perform natural resources mapping and monitoring.

The devices capable of detecting and quantifying the electromagnetic energy are the remote sensors. These devices are able to detect and register the energy, in a specific range of the electromagnetic spectrum, producing information that can be processed and transformed in relevant data to be analyzed and interpreted, in a graphical interface or even like an image or other resource (Moreira, 2007).

2.1 Geographic Data Mining

Similar to the increase in other types of data, for example, biological, astronomical, or commercial area, the geographic data also had a considerable increase. The occurrence of this increase was due to factors, such as the constant capture of images of the terrestrial sphere, which occurs daily and at different periods of time, the increased computational and technological development, growth of researches in the area, among other factors. This all meant that there was a considerable increase in the volume of spatial data and hence the size of the storage repositories to which these data are intended (Miller and Han, 2009). The remote sensing data have characteristics and attributes important,

which when analyzed properly, can reveal implicit information that may be relevant. Given the volume of geographic data, heterogeneity, the difference between data types and complexity is need tools that can find patterns for this type data (Miller and Han, 2009).

2.2 GeoDMA

It was proposed by (Korting et al., 2008), a tool called Geographical Data Mining Analyst (GeoDMA), which is a plug-in for TerraView GIS. This system performs analysis in remote sensing images based on data mining techniques. For this he uses all the processing steps required for handling remote sensing data, including segmentation processes, training, classification, attribute extraction and data exploratory analysis. It emerged from the ideas proposed by (Silva et al., 2005), and that in the elapse of its development became more complete, being used for the more diverse applications.

3 GRAPH MINING

Web data can be represented like graphs, especially hyper documents and other types of links (eg. social networks, senders and receivers of electronic messages, co-authors of academic papers, among others). The representation through graphs demands a specialized data processing, which is an essential step to enable the use of mining algorithms (Santos, 2009).

Due to continuous data growing and their complexity, new techniques of data mining have been researched. Most repositories have structural features, which data are composed of segments and relation among them. In this context, the graph mining area has approached different tasks on structural data.

Several algorithms of search in graph have been developed in computer science, chemistry, computer vision, video indexing and retrieval of text (Cook and Holder, 2007). With the growing demand in the analysis of large amounts of structured data, graph mining has become an important and active topic in data mining. Traditional algorithms of the data mining also are applied in the graphs mining, however due to certain structural features of this representation, techniques have been researched and applied to graph mining, among them are: classification, clustering, frequent patterns mining (frequent substructures) (Aggarwal and Wang, 2010). In this paper we will only cover the mining of frequent patterns, i.e., the search if frequent substructures in the graph.

3.1 Frequent Substructures Mining

Among the various kinds of patterns in graphs, frequent substructures are very basic patterns that can be discovered in a set of graphs. They are useful for the characterization of sets of graphs, differentiating between the different groups, classifying and grouping graphs and in the construction of indices (Cook and Holder, 2007).

Frequent substructures mining searchs similar subgraphs in a set of graphs, where a frequent subgraph may be defined as follows: given a set of vertices of a graph *g* represented by *V(g)* and the set of edges *E(g)*, the graph *g* will be a subgraph of another *g'* if exists an isomorphism of *g* in *g'*. Given a labelled graph $D = G_1, G_2, G_3, ..., G_n$, the support*(g)* (or *frequency(g)*) is defined as the percentage or quantity of graphs in *D* where *g* is the subgraph (Washio and Motoda, 2003). As a consequence, a frequent subgraph can be defined as a graph which support cannot be lower than the established minimum support (Han and Kamber, 2006).

Many algorithms for frequent substructures extraction are referenced, among them we have: *Apriori based Graph Mining* (AGM) (Inokuchi et al., 2000), *Frequent SubGraphs* (FSG) (Kuramochi and Karypis, 2001), *graph-based Substructure pattern* (gSpan) (Yan and Han, 2002), *Fast Frequent Subgraph Mining* (FFSM) (Huan et al., 2003), *GrAph, Sequences and Tree extractiON algorithm* (Gaston) (Nijssen and Kok, 2004) and *SUBstructure Discovery Using Examples* (Subdue) (Cook and Holder, 2000).

The algorithm used in this research is the Frequent SubGraphs (FSG), which uses a candidate generation method based on edge, it increases the size of the substructure on an edge at each iteration of the algorithm (Kuramochi and Karypis, 2001). Two patterns of size k are combined only if they share the same subgraph containing k-1 edges, which is called the core. In the case of FSG, the size of the graph refers to the number of edges it contains. The new candidate the frequent subgraph that was generated includes the core and two edges of the combined patterns. For example, a Figure 1 shows the process of generating candidates for frequent subgraph in the FSG algorithm. After the generation of candidates, their support is given and the process repeats until no more frequent subgraphs found (Han and Kamber, 2006).

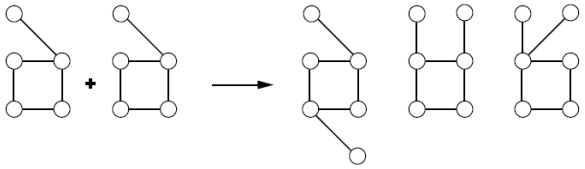

Figure 1: Generation of candidate. Source: (Han and Kamber, 2006).

4 GRAPH MINING IN SPATIAL DATA OF DEFORESTATION

The research developed in data mining conduct an approach in the information analysis, which is obtained by extracting patterns in remote sensing images. For this task are used landscape metrics, which are used as criteria for the classification of some characteristic of the object of the image to a pattern of land use specific. This pattern of land use is derived from the typology of patterns of deforestation for tropical forests, which can be applied to the study region (Saito, 2011, Silva et al., 2005). Other researches cover the crossing of data obtained in field research and diverse databases with spatial data to understand the process of deforestation, thereby elucidating the authors and factors that influence the growth of Amazon deforestation (Lorena and Lambin, 2007).

The work developed by (Tilton et al., 2008) performed an initial approach to representing the RHSEG-produced hierarchical image segmentations in a graphical form understandable by Subdue system. However the results were preliminary and very limited compared to what could potentially be provided. A related approach is shown in (Zamalieva, Aksoy and Tilton, 2009), where they propose a generic unsupervised method for discovering compound objects. The method translates image segmentation into a relational graph, and applies a Subdue graph-based knowledge discovery algorithm to find the interesting and repeating substructures that may correspond to compound objects. In relational graph structure the nodes correspond to the regions and the edges represent the relationships between these regions. The region objects that appear together frequently can be considered as strongly related.

These research initiatives employed to graph modelling, two of them apply graph mining of frequent substructures. However the approaches don't consider the use of a supervised learning method, besides not to perform labeling the nodes of the graph.

In our work, we'll use a supervised learning method for classification of deforestation objects, classified according to a typology of spatial patterns, task performed automatically by the system for spatial data mining, called GeoDMA (Korting et al., 2008). Were performed the creation of homogeneous regions through the creation of cells, allowing, posteriorly, creation of flows among objects. In structure of the graph, both nodes and edges are labeled using a parameter chosen. We also performed the search for frequent substructures using the FSG graph-based knowledge discovery algorithm, which aims at finding all connected subgraphs that appear frequently in a database of graphs (Kuramochi and Karypis, 2001).

5 CASE STUDY: VALE DO ANARI

The chosen study area is the municipality of Vale do Anari, and belongs to Rondonia state, with an area of 3,135 km^2 and, nowadays, has a population of more than 9,000 mil inhabitants (IBGE, 2011). Were used data of the planned rural settlement by the INCRA (National Institute for Colonization and Agrarian Reform) in Vale do Anari municipality in the state of Rondonia. This settlement started in 1982, with land plots sized around 50 ha. Figure 2 represents visualization, through polygons extracted from images using TerraView, of the Rondonia state and Vale do Anari municipality.

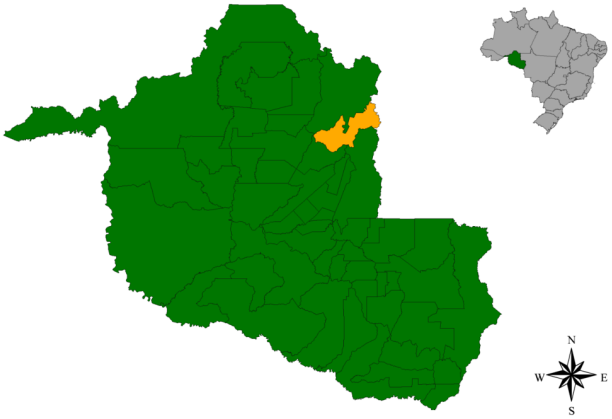

Figure 2: Rondonia State, highlighting Vale do Anari

5.1 Methodology

In this work, we employed deforestation polygons from 1985 to 2000 in the municipality of Vale do Anari, scenes (231/66 and 231/67), available in shape file format, kindly provided by Isabel Escada.

Applications used for the treatment of deforestation data were Geographical Data Mining Analyst (GeoDMA) (Korting et al., 2008), Terraview 4.1.0 (INPE, 2011b), Flow Plugin and PostgreSQL 9.0.3 and FindFRUG, prototype implemented during this work to recover, convert, process and display graphs. Figure 3 schematizes the performed procedures on this work.

Initially, a database was created in TerraView, with data of the municipality (polygons of deforestation from 1985 to 2000). Then, a deforestation pattern classification was performed with the polygons selected using the GeoDMA software (GeoDMA, 2011),

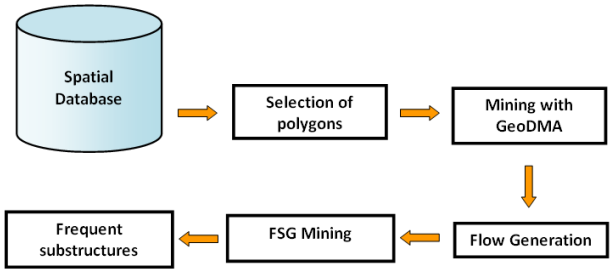

Figure 3: Procedures performed

based on the typology of deforestation patterns from (Silva et al., 2008), which are irregular, linear and geometric. Through of visual models this typology of deforestation patterns can be visualized as shown in Figure 4.

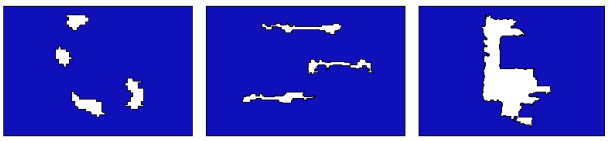

Figure 4: Typology of spatial patterns of tropical deforestation (irregular, linear and geometric). Source: (Silva et. al., 2005)

The result of the classification can be seen in Figure 5. Looking at the graph we can see that there is a large amount of objects classified as irregular from 1994 until 2000. This fact reflects the large amount of deforested areas in the same period for the Legal Amazon, also showing a large amount of objects classified as irregular and geometric patterns for this period in the municipality.

After the classification, the polygon of the municipality was divided into forty-six regions (cells) with a wide quantity of objects, where each cell has an area of 36 km^2. An information plan was created for each region. Figure 6 shows the selected regions and highlights one of them.

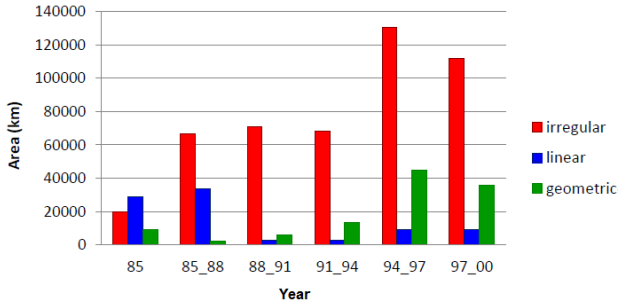

Figure 5: Large farms dynamic in Vale do Anari

A file was generated containing the link among all deforestation objects. It was imported into the Flow Plugin of TerraView, which performed the linkage between the polygons according to the distance between objects. With the results obtained by Flow Plugin a graph file, representing the links among deforestation objects, was generated and mined using FSG (Karypis, 2003), a graph mining software (Kuramochi and Karypis, 2001).

After the extraction of frequent substructures, the output file generated by the miner was mapped into JUNG, a graph viewer developed in Java that allowed the visualization of the discovered substructures (Figure 8 and Figure 9).

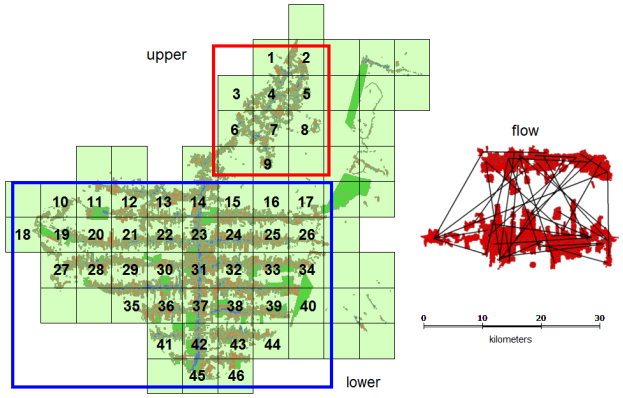

Figure 6: Division of the municipality into cells

6 PRELIMINARY RESULTS

Among the results obtained after analysis of frequent substructures discovered, it was observed that the application of different distances revealed interesting relationships between the patterns of deforestation. With these results we can see relations the neighborhood and aspects related to temporal occupation of some regions (cells), showing the frequent substructures with temporal patterns near and also aspects related to patterns of high frequency in the municipality.

The database has been mined using different values of minimum support, ranging among 10% and 100%, also sought the substructure more frequent in all regions (cells) examined. Analyzing the distances applied to a database, in this case, of 500 meters among each deforestation pattern. And using the lower region of Vale do Anari municipality as the unit of analysis, we verified the existence of large number of frequent substructures to this area.

The lower region had a total of 37 cells. So, we performed the graph mining with FSG algorithm, using for this a minimal support of 90%. As a result, were returned nine frequent substructures, showing only links between the types of irregular patterns. With these results we can see a strong presence of frequent substructures linking the periods of 94_97 and 97_00 the irregular patterns of the same year or belonging to previous years (Figure 7). This result showed that: first, the period 94_97 refers to the years of highest incidence of deforestation in the city of Vale do Anari, reflecting a high number of substructures, and second, that newer patterns, in most cases, are linked to earlier patterns, since the occupation in the region is stabilized, aiding the appearance of other patterns. The Figures 8 and 9 show some of discovered results modeled in graph format.

Number of frequent substructures with distance of 500 meters, minimum support 90% (37 cells)							
Year		Irregular					
		85	85_88	88_91	91_94	94_97	97_00
Irregular	85						
	85_88			33		34	33
	88_91					33	35
	91_94					36	
	94_97					34	37
	97_00						35

Figure 7: Distance from 500m to detect irregular patterns

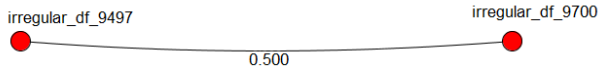

Figure 8: Irregular pattern (94_97 and 97_00) - distance of 500m existing in all 37 cells

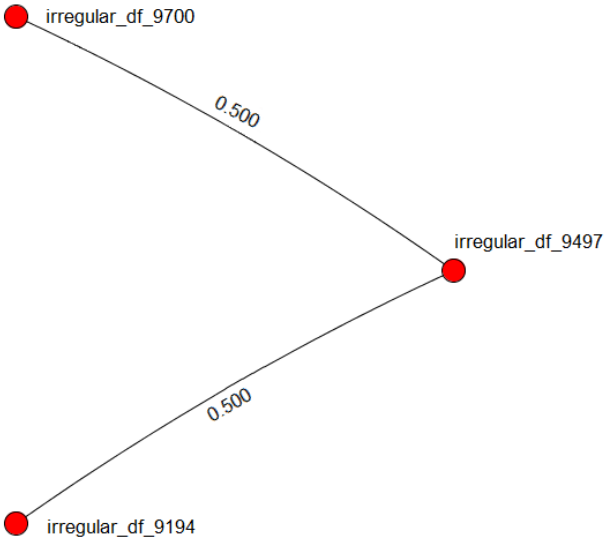

Figure 9: Irregular pattern (91_94, 94_97 and 97_00) - Distance of 500m showing temporal relationship of occupation in 34 of 37 cells.

7 CONCLUSION

This work addresses a new approach to discover relevant knowledge in Amazonia deforestation processes, allowing the search for patterns of correlated areas of deforestation, through the use of graph mining to find frequent substructures in remote sensing data. The technique is relevant to understand, monitor and prevent deforestation processes, tracking events and agents in specific regions.

The application of this proposal on real data was performed using a case study on Vale do Anari, where data, through a prototyped framework, were modeled as graphs and then mined. Analysis through this approach presented relevant relations among patterns of the study area, showing that the developed methodology can be applied to discover and confirm processes in the considered area.

Represent remote sensing data through graphs is an attractive and innovative strategy, once the use of graphs constitutes an well-founded area of research with many resources and applications. Thus, our work brings a new approach for pattern and phenomena discovery, which are implicit and occurring in deforestation areas.

ACKNOWLEDGEMENTS

The authors acknowledge CAPES and CNPq for supporting this research.

REFERENCES

Aggarwal, C. C. and Wang, H., 2010. Managing and Mining Graph Data. Spring, New York.

Cook, D. J. and Holder, L. B., 2000. Graph-based data mining. IEEE Intelligent Systems 15, pp. 32–41.

Cook, D. J. and Holder, L. B., 2007. Mining Graph Data. Wiley-Interscience.

GeoDMA, 2011. *GeoDMA 0.21. Sao Jose dos Campos, SP. National Institute for Space Research - GeoDMA*. Available at http://www.dpi.inpe.br/geodma. Accessed in. 07/10/2011.

Han, J. and Kamber, M., 2006. Data Mining: Concepts and Techniques. 2 edn, Morgan Kaufmann, San Francisco.

Huan, J., Wang, W. and Prins, J., 2003. Efficient mining of frequent subgraphs in the presence of isomorphism. In: Proceedings of the Third IEEE International Conference on Data Mining, ICDM '03, IEEE Computer Society, Washington, DC, USA, pp. 549–.

IBGE, 2011. *The Brazilian Institute of Geography and Statistics -IBGE.* Available at http://www.ibge.gov.br. Accessed in. 07/17/2011.

Inokuchi, A., Washio, T. and Motoda, H., 2000. An apriori-based algorithm for mining frequent substructures from graph data. In: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD '00, Springer-Verlag, London, UK, pp. 13–23.

INPE, 2011a. *National Institute for Space Research: Image Generation Division (DGI)*. Available at http://www.dgi.inpe.br . Accessed in. 06/12/2011.

INPE, 2011b. *National Institute for Space Research: TerraView.* Available at http://www.dpi.inpe.br/terraview/index.php. Accessed in. 09/10/2011.

Karypis, G., 2003. *PAFI Software Package for Finding Frequent Patterns in Diverse Datasets.* Available at http://glaros.dtc.umn.edu/gkhome/pafi/overview. Accessed in. 04/04/2011.

Korting, T., Fonseca, L., Escada, M., Silva, F. and Silva, M. P. S., 2008. Geodma - a novel system for spatial data mining. IEEE International Conference on Data Mining Workshops (ICDMW '08) pp. 975 –978.

Kuramochi, M. and Karypis, G., 2001. Frequent subgraph discovery. IEEE International Conference on Data Mining (ICMD' 01) 0, pp. 313–320.

Lorena, R. B. and Lambin, E. F., 2007. Linking spatial patterns of deforestation to land use using satellite and field data. IEEE International Geoscience and Remote Sensing Symposium (IGARSS '07) pp. 3357–3361.

Medeiros, C. B., 2008. Grand research challenges in computer science in brazil. Computer 41, pp. 59–65.

Miller, H. and Han, J., 2009. Geographic Data Mining and Knowledge Discovery. Chapman & Hall/CRC data mining and knowledge discovery series, CRC Press.

Moreira, M. A., 2007. Fundamentals of Remote Sensing and Application Methods. 3 edn, UFV - Universidade Federal de Vicosa, Sao Jose dos Campos.

Nijssen, S. and Kok, J. N., 2004. A quickstart in frequent structure mining can make a difference. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, ACM, New York, NY, USA, pp. 647–652.

Saito, E. A., 2011. *Patterns trajectories characterization of human occupation in the Legal Amazon through data mining*. Dissertation (Master in Remote Sensing) - National Institute for Space Research - INPE, Sao Jose dos Campos.

Schowengerdt, R. A., 2007. Remote Sensing, Models and Methods for Image Processing. 3 edn, Elsevier.

Silva, M., Camara, G., Souza, R., Valeriano, D. and Escada, M., 2005. Mining patterns of change in remote sensing image databases. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05) 1, pp. 362–369.

Silva, M. P. S., Camara, G., Escada, M. I. S. and de Souza, R. C. M., 2008. Remote-sensing image mining: detecting agents of land-use change in tropical forest areas. International Journal of Remote Sensing 29, pp. 4803–4822.

Tilton, J. C., Cook, D. J. and Ketkar, N. S., 2008. The integraton of graph-based knowledge discovery with image segmentation hierarchies for data analysis, data mining and knowledge discovery. IEEE International Geoscience & Remote Sensing Symposium (IGARSS '08) pp. 491–494.

Washio, T. and Motoda, H., 2003. *State of the art of graph-based data mining. SIGKDD Explor. Newsl. 5(1), pp. 59–68.*

Yan, X. and Han, J., 2002. gSpan: Graph-Based Substructure Pattern Mining. Proceeding of the 2002 international conference on data mining (ICDM' 02) pp. 721–724.