Structural Knowledge Discovery Used to Analyze Earthquake Activity

Jesus A. Gonzalez, Lawrence B. Holder and Diane J. Cook
Department of Computer Science and Engineering
University of Texas at Arlington
Box 19015, Arlington, TX 76019-0015
{gonzalez,holder,cook@cse.uta.edu}

Abstract

The Subdue structural discovery system is being used as the Data Mining tool to study the "Orizaba Fault" located in Mexico, as part of a research project of the geologist Dr. Burke Burkart. We analyze the information of the Earthquake Database to discover if the earthquake activity in the area is related to the fault. We experimented with different samples of data mainly using two heuristics to guide Subdue through the substructure discovery process. We also added some spatio-temporal information as background knowledge. The results show how Subdue can successfully be used as a Data Mining tool in Real World Domains.

Introduction

The advancement of technology has allowed not only the automation of complex processes but also the accumulation of large amounts of process information in databases. But having the information is useless if we do not take advantage and learn from it by extracting knowledge that helps to improve a process or identifying a possible failure manifested in the stored information. However this is a difficult task to achieve using standard tools due to the large amount and complexity of data.

That is the reason why different approaches in the field of Knowledge Discovery (Fayyad et al. 1996) have been developed to extract hidden information from those databases. In this project we use the Knowledge Discovery process (Fayyad et al. 1996) and a specific Data Mining tool applied to a real-world domain problem. The Data Mining tool is the Subdue program, and the domain is the Earthquake database that consists of reports of earthquakes. In the case of this domain we worked with a geology expert, Dr. Burke Burkart who helped us to analyze the results and to guide the geology-related research.

We experimented with different samples of data mainly using two heuristics to guide Subdue through the substructure discovery process. We also added some geographical and time knowledge to connect earthquakes that occurred close to each other in time and distance. The results show how Subdue was able to effectively find patterns with a logical interpretation, and how it can be used as a research tool in the geological domain.

Substructure Discovery Using Subdue

Subdue (Cook, Holder and Djoko 1995) is a Data Mining tool that achieves the task of clustering using an algorithm categorized as an example based and relational learning method. This tool was first developed in 1990 and has been expanded and optimized to generate better results. It is a general tool that can be applied to any domain that can be represented as a graph. Subdue has been successfully used on several domains like CAD circuit analysis, chemical compound analysis, and scene analysis (Cook, Holder and Djoko 1996, Cook, Holder and Djoko 1995, Cook and Holder 1994, Chittimoori, Gonzalez and Holder 1999, and Djoko, Cook and Holder 1995).

Subdue implements two model evaluation criteria as a means to decide which patterns are going to be chosen as important knowledge or structures. The first model evaluation method is called “Minimum Encoding” that is a technique derived from the minimum description length principle (Cook and Holder 1994) and chooses as best substructures those that minimize the description length metric that is the length in number of bits of the graph representation. The number of bits is calculated based on the size of the adjacency matrix representation of the graph. According to this, the best substructure is the one that minimizes $I(S) + I(G|S)$, where $I(S)$ is the number of bits required to describe substructure S, and $I(G|S)$ is the number of bits required to describe graph G after being compressed by substructure S. The second method chooses the substructures according to how well they compress the graph in terms of its size in number of vertices and edges. Another method used consists of finding large substructures in spite of their low number of instances.

The main discovery algorithm is a computationally constrained beam search. The algorithm begins with the substructure matching a single vertex in the graph. Each iteration the algorithm selects the best substructure and incrementally expands the instances of the substructure. The algorithm searches for the best substructure until all possible substructures have been considered or the total amount of computation exceeds a given limit. Evaluation of each substructure is determined by how well the substructure compresses the input graph according to the heuristic being used (MDL or Graph size). The best substructure found by Subdue can be used to compress the input graph, which can then be input to another iteration of Subdue. After several iterations, Subdue builds a hierarchical description of the input data where later substructures are defined in terms of substructures discovered on previous iterations.

There are other components that make Subdue more powerful. We can specify predefined substructures that Subdue looks for in the data. This allows Subdue to use previous knowledge as a starting point and guide the discovery process. Subdue uses an inexact graph match technique so that instances of substructures that are slightly different can be matched. We can also iterate Subdue's discovery process in order to find more substructures in new iterations that might contain substructures found in previous iterations. Figure 1 shows a simple example of Subdue's operation. Subdue finds four instances of the triangle-on-square substructure in the geometric figure. The graph representation used to describe the substructure, as well as the input graph, is shown in the middle.

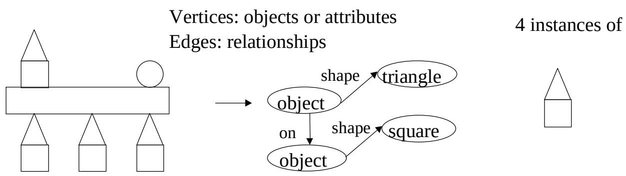

Figure 1: Subdue's Example

The Earthquake Database

The earthquake database contains information collected from several catalogs (http://wwwneic.cr.usgs.gov). These catalogs were provided by sources like the National Geophysical Data Center of the National Oceanic and Atmospheric Administration (NOAA). The database has records of earthquakes from 2000 B. C. through the current week. An earthquake record consists of 35 fields: source catalog, date, time, latitude, longitude, magnitude, intensity and seismic related information such as cultural effects, isoseismal map, geographic region and stations used for the computations. Earthquakes of magnitude below 1.0 are not stored in the database; most of the magnitudes of earthquakes range from 2.5 to 9.5.

There are some differences between catalogs, e.g. it is possible to find the same earthquake with a slightly different epicenter or magnitude in two catalogs. This is due to the methods and instruments used to compute the data. As an example we mention that currently epicenters and magnitudes are calculated with computer programs using seismographic data. The problem is that the computer programs contain assumptions about the earth in the formulae they use. If those assumptions are violated then the results can be different.

The size of the Earthquake database is extremely large (e.g. 2.2 MB only for 1995 data), so we could not use all the information in our experiments; we just used subsets of information corresponding to periods of time between 6 months and 1 year. We created a relational database containing the earthquake information (the 35 fields). This eased the extraction of information for the experiments, because we can use SQL queries to extract the desired subset of the database. We use the Data Mining approach instead of queries because we do not pre-set the information to be included in the result. This means that we prepare a query that can uncover novel structural patterns in the same way as the Subdue system.

Earthquake Database Knowledge Representation

Every record in the database represents an earthquake event. In this domain we used two kinds of edges to connect the events (earthquakes). The first type of edge is the "near_in_distance" edge, which is set between two events if the distance between them is equal or less than 75 kilometers. The second type of edge is the "near_in_time" edge that is set between two events if they happened with a difference of time equal or less than 36 hours. We chose those parameters because of two reasons. First, they were a good combination that generates enough edges so that the system may find them, and not too many to overload the graph so that those were the only substructures found. Second, our geology specialist told us that 75 kilometers was reasonable for the size of the area of study and that the effects between one earthquake and another are usually shown within 36 hours. An earthquake event in graph form is shown in figure 2. All the fields of the Earthquake database are included except for the empty fields, which would bias the system because of the large amount of them.

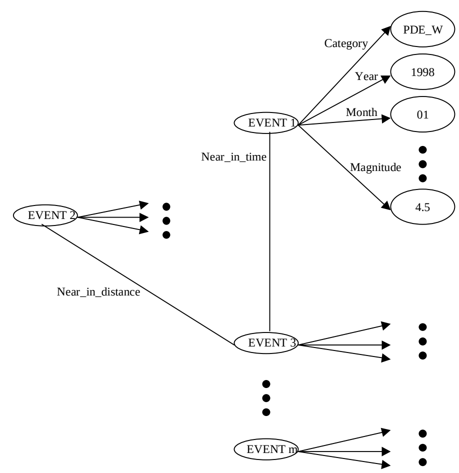

Figure 2: Earthquake Knowledge Representation

Earthquake Database Experimental Results

We chose only a subset of the database to run the experiments. For example, we took 6 months of information and ran Subdue on it, so the query to extract the information from the database included the year and month of the earthquakes that we wanted. We started using all the fields of the database, but the year field affected our results because the values were all the same, so we decided to exclude that field.

We wanted to take a random sample from the database (from the 5 years of information and keeping the same graph size) but that would affect the "near_in_time" edges, because the sampled earthquakes would have a larger range over time and cause a loss of important information (there would be less near_in_time related records). So we just randomly sampled from the information collected in one year creating a graph with 10135 events, 136,077 vertices, 125,941 attribute edges and 757,417 undirected "near_in_distance" and "near_in_time" connections and a size of 26,963,605 bytes.

Minimum Encoding Heuristic Results

With the minimum encoding heuristic Subdue was able to find structures that linked events with the "near_in_time" and "near_in_distance" edges. The first substructure (substructure 1, not shown) linked one event to four others with near_in_time edges and to a fifth event with a "near_in_distance" edge. The second substructure (substructure 2, not shown) linked one event to three other events with "near_in_distance" edges and to the category field "PDE-W" that corresponds to the source of an earthquake's catalog entry. The third substructure (substructure 3, not shown) linked one event to another event and to one substructure_2 with "near_in_distance" edges. The fourth substructure is more complex and is shown in figure 3. This substructure is interesting because several earthquakes happened in a short period of time and could be related to a fault placement.

The interesting issue here is the potential to find important relations between earthquakes that happened in a localized region within a short period of time.

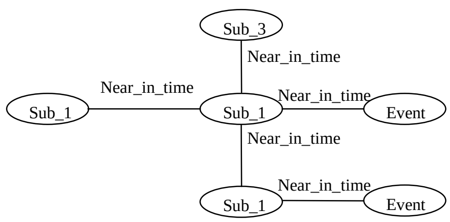

Figure 3: Substructure 4, 90 Instances

Graph Size Heuristic Results

With the graph size heuristic we found more substructures in the Earthquake database. The reason is because it works faster and we could go deeper in the number of iterations. Subdue found relations between events and substructures with the "near_in_time" and "near_in_distance" edges, but it also found relations that included some other fields like "Catalog", "Month", "Mag1 Scale", and "Depth". Here, it was possible to conclude that the earthquakes related by the substructure were provided by the "PDE-W" catalog which lists the most recent weeks in events and the "PDE-Q" catalog that lists the most recent events that are still subject to change. It was also possible to conclude from the data that more earthquakes occurred in the months of "June" and "May" and that a frequent depth for the related earthquakes was "33.0000" and "10.0000" kilometers. The fact that Subdue found the depth characteristic of "33.0000" kilometers is validated in the Earthquakes database description where it is mentioned that this is the most common depth for an earthquake. As an example, figure 4 shows how in the eighth iteration Subdue found that 140 of the instances of substructure 1 happened in a depth of 33 kilometers. Substructure 1 in figure 4 has 9465 instances and connects an earthquake event to the category value "PDE_W". Substructure 7 with 141 instances, connects an event to substructures found in previous iterations with "near-in-distance" and "near-in-time" edges and also contains the "PDE_Q" attribute.

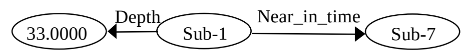

Figure 4: Substructure 8, 140 Instances

Determining Earthquake Activity

We already mentioned how we used Subdue to find patterns in the earthquakes database. Now we are going to describe a project in which we used Subdue to determine the earthquake activity of a specific area of Mexico. Dr. Burke Burkart, a Geologist at the University of Texas at Arlington, who has studied Mexican geology and seismology for years, is interested in the study of the seismology caused by the Orizaba Fault (Burkart 1994, Burkart and Self 1985). This fault runs from the Vulcan "Pico de Orizaba" located in the state of Veracruz through the "Itsmo de Tehuantepec" in the state of Oaxaca.

A fault is defined as a fracture in a surface where a displacement of rocks also happened. Faults are caused by forces acting over the rock bodies. When a rupture occurs, there is going to be two walls forming the fault. Faults receive a different name according to the rocks' movement (Hamblin and Christianses 1998).

When the movement among the rocks happens in the vertical plane, the fault is called a Dip-Slip Fault, where the Hanging-wall is the one above the fault and the Foot-wall is the one below the fault. Dip-Slip Faults are classified according to the direction of the rocks' movement. A Normal Dip-Slip Fault is created when a pulling force generates the fracture, then by the gravity force, the hanging wall is displaced downwards. Reverse Dip-Slip Faults are created when a compression force forms the fracture. In this case the hanging wall moves upward due to the compression force. A Thrust Fault is a reverse fault with an inclination of less than 45^o.

If the movement among the rocks happens in the horizontal plane, the fault is called a Strike-Slip Fault. This type of fault is described as Left-Lateral Fault or Right-Lateral Fault.

Oblique Faults are those with the characteristics of both, Dip-Slip and Strike-Slip Faults, that is, the rocks move in both planes, the vertical and horizontal. The "Orizaba Fault" is a Strike-Slip Fault. We want to know the location of the active zone of earthquakes, which will be located at the weakest point of the fault. This is more complex than it appears, because the fault is not continuous. It is interrupted in some locations, changes direction and is probably connected to other faults. This means that the earthquakes might take place in a location out of the fault, but still as a consequence of this fault.

Figure 5: Area of Study of the Orizaba Fault.

This study started with the identification of the area with more possibilities of being affected by the fault. We started by selecting two rectangles. The first has coordinates 94.5W Longitude through 101.0W Longitude and 17.0N Latitude through 18.0N Latitude. The second has coordinates 94.0W Longitude through 98.0W Longitude and 18.0N Latitude through 19.0N Latitude. The area includes parts of the states of Guerrero, Oaxaca, Puebla and Veracruz. We can see this area in figure 5.

We ran Subdue over the graph representation of the earthquakes in these two rectangles. Subdue helped us to find not only a subarea with a high concentration of earthquakes, but also some of the area's characteristics. The most representative substructures found with Subdue are shown in figure 6. In Substructure 1 we can see that an earthquake Event is related to another earthquake Event with a Near_in_distance edge. We also see that one of the earthquake Events is linked to a node representing the region number "59" and to another node representing the Catalog "PDE." What this substructure is telling us is that region number 59, which is located in the state of Guerrero, is the one with more earthquake activity (it has more occurrences than other regions), in this case with 556 earthquakes. Finally the substructure tells us that there is a distance relation between some of the earthquakes, identified by the Near_in_distance edges, that means that there is a distance of less than 75 km. between the events. Dr. Burkart identified this area as very active. However, the cause of these earthquakes is not related to the fault in study, at least this is not yet clear. Substructure 2 links two substructure 1s with a "Near_in_distance" edge. It also links one of the Substructure 1s to a vertex describing the depth of one of its events (Substructure 1 contains two events as can be seen in figure 6) at 33 km. This substructure tells us about a common depth among some of the earthquakes in the area of study.

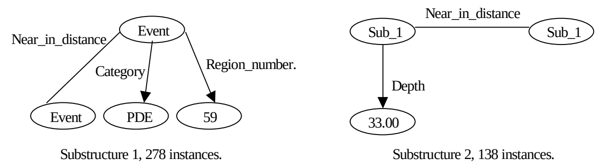

Figure 6: Substructures Found in the Whole Area of Sudy.

Next, we decided to divide the area and study the sub-areas. We divided the rectangles in small pieces of one half of a degree in both longitude and latitude. For example, one of those rectangles has coordinates 101.0W to 100.5W of Longitude and 17.0N to 17.5N of Latitude. We divided the total area into 44 sub-areas. After we divided the area of study, we got all the available information about earthquakes in each sub-area from the Earthquake Database described before (the database contains earthquake information from 1973 up to the present date). Table 1 shows how many earthquakes we found in each of these sub-areas.

Area Number	Area Coordinates				Area Name	Number of Events
	Latitude		Longitude			
1	101.0W	100.5W	17.0N	17.5N	Gue1	62
2	101.0W	100.5W	17.5N	18.0N	Gue2	40
3	100.5W	100.0W	17.0N	17.5N	Gue3	57
4	100.5W	100.0W	17.5N	18.0N	Gue4	13
5	100.0W	99.5W	17.0N	17.5N	Gue5	71
6	100.0W	99.5W	17.5N	18.0N	Gue6	15
7	99.5W	99.0W	17.0N	17.5N	Gue7	35
8	99.5W	99.0W	17.5N	18.0N	Gue8	16
9	99.0W	98.5W	17.0N	17.5N	Gue9	13
10	99.0W	98.5W	17.5N	18.0N	Gue10	14
⋮						
26	95.0W	94.5W	17.5N	18.0N	Ver1	43
27	94.5W	94.0W	17.0N	17.5N	Oaxver4	35
28	94.5W	94.0W	17.5N	18.0N	Ver2	23
29	98.0W	97.5W	18.0N	18.5N	Pue1	6
30	98.0W	97.5W	18.5N	19.0N	Pue2	0
⋮						
42	95.0W	94.5W	18.5N	19.0N	Vergolf5	1
43	94.5W	94.0W	18.0N	18.5N	Vergolf4	3
44	94.5W	94.0W	18.5N	19.0N	Vergolf6	1

Table 1: Sub-Areas of Study for the Orizaba Fault.

Once we collected the information about earthquakes registered in each sub-area, we were ready to study its characteristics (e.g. common depth and intensity per region). Here Subdue takes part in this research again. We took the information of each sub-area where more than ten earthquakes were registered and converted it into the graph representation used by Subdue. Then we ran Subdue to find out the characteristics of the earthquakes in that sub-area. As an example lets take sub-area 26 of table 1 labeled with the name of "Ver1." Figure 7 shows the first two substructures found by Subdue in that sub-area. Substructure 1 in the figure shows that the events happened in the region number "61," which corresponds to the selected sub-area. We can also see that the events' information was taken from the "PDE" catalog. In substructure 2 we find a pattern of some of the events at a depth of "33 Km." This is a very interesting pattern, because it might give us information about the cause of those earthquakes. If the earthquake is not caused by subduction (a force caused by the Pacific plate, which effects depth based on the closeness to the Pacific Ocean), then there is more possibility that it is related to the fault. However, we first have to evaluate and determine the depth of earthquakes caused by subduction in that zone.

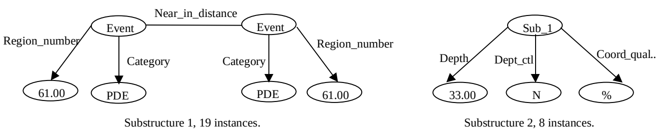

Figure 7: Substructures Found in Sub-Area 26 from Table 1.

As we see in this study, Subdue is capable of finding not only the shared characteristics of the events, but also space relations between them. In the case of the identification of shared characteristics, we used the pattern containing the region number specification to recognize the area being studied. The pattern containing the depth node at 33 km. gave us information that the Geology specialist Dr. Burkart is studying so that he can use it to give direction to this research. In the case of the space relations, we expect to find patterns that represent parts of the paths of the involved fault. The time relations ("near_in_time" edges) were not considered by Subdue, because the earthquakes in the area are not close in time. However, there are other areas with different characteristics where "near_in_time" connections provide important information, and we hope to uncover these relations in future studies.

Conclusions

In this research, we showed that Subdue was able to successfully analyze the real-world earthquake database when applied as the Data Mining tool of the Knowledge Discovery process. It was found that Subdue can be used to find interesting patterns that might represent new knowledge or that might lead to new knowledge.

It was also shown how Subdue used prior knowledge to guide the search with temporal and spatial relations provided by the "near_in_time" and " near_in_distance" edges. Subdue was able to find substructures that included those edges. Using this knowledge representation, the system not only found repetitive patterns in the data, but also provided temporal and distance relations that made possible the discovery of more interesting patterns. As an example in the Earthquake database, spatial relations were incorporated through the "near_in_distance" edges. Subdue was able to find substructures containing these edges, and these substructures are being used to help study the "Orizaba Fault" in Mexico.

Something very important about the temporal and spatial relations is the definition of the "near_in_time" and "near_in_distance" edges. We need to establish the meaning of "near" in both cases. This is not a simple task, because it depends directly on the domain and the semantics of the relation to be represented.

In our future work we will be working on a concept learning approach that will learn substructures distinguishing two sets of sub-areas so that we can study their geological behavior based on earthquake activity. We will continue the analysis of earthquake activity in collaboration with Dr. Burkart. We have also used the spatio-temporal relation annotations to study the Aviation Safety Reporting System Database (Chittimoori, Gonzalez and Holder 1999), and we plan to work with other domains including a graph representation of program source code. We are also working on a theoretical analysis of Subdue based on the PAC learning theory (Jappy and Nock 1998) and conceptual graphs (Sowa 1984).

References

Burkart, Burke 1994. Geology of northern Central America, Book chapter for Geology of the Caribean,, Jamaican Geological Society, Kingston, S.Donovan Ed. p. 265-284.

Burkart, Burke and Self, S. 1985. Extension and rotation of crustal blocks in northern Central America and its effect upon the volcanic arc, Geology, v 13, p 2226.

Cook, Diane J. and Holder, Lawrence B. 1994. Substructure Discovery Using Minimum Description Length and Background Knowledge, Journal of Artificial Intelligence Research, Vol. 1, pp. 231-255.

Cook, Diane J.; Holder, Lawrence B.; and Djoko, Surnjani 1994. Knowledge Discovery from Structural Data, Journal of Intelligence and Information Sciences, Vol. 5, Number 3, pp. 229-245.

Cook, Diane J.; Holder, Lawrence B.; and Djoko, Surnjani 1996. Scalable Discovery of Informative Structural Concepts Using Domain Knowledge, IEEE Expert vol. 11 number 5, pp. 59-68, October.

Sowa, J. F. 1984. Conceptual Structures – Information Processing in Mind and Machine, Addison-Wesley.

Jappy, Pascal and Nock, Richard 1998, PAC Learning Conceptual Graphs, Proceedings of the 6^{th} International Conference on Conceptual Structures, pp. 303-315.

Chittimoori, Ravindra N.; Gonzalez, Jesus A.; and Holder, Lawrence B. 1999. Structural Knowledge Discovery in Chemical and Spatio-Temporal Databases, Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp. 959.

Djoko, Surnjani; Cook, Diane J.; and Holder, Lawrence B. 1995. Analyzing the Benefits of Domain Knowledge in Substructure Discovery, Proceedings of the first Int. Conf. on Knowledge Discovery and Data Mining, pp. 75-80.

Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; and Uthurusamy, Ramasamy 1996. Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, Menlo Park, California.

Hamblin, W. Kenneth and Christianses, Eric H. 1998. Earth's Dynamic Systems, Prentice Hall.