Proceedings of the International Conference on
Electrical Engineering and Informatics
Institut Teknologi Bandung, Indonesia June 17-19, 2007

Spatial Data Preprocessing for Mining Spatial Association Rule with Conventional Association Mining Algorithms

Imam Mukhlash[1], Benhard Sitohang[2]

[1]*Departement of Mathematics, ITS Surabaya Jl. Arif Rahman Hakim 100 Sukolilo Surabaya, 60111*
[2] *School of Electrical Engineering and Informatics, ITB Bandung Jl. Ganeca 10 Bandung 40132*

Abstract

The increasing usage of Geographical Information Systems (GIS) for various problems makes the volume of spatial data is growing fast. Spatial data mining is one of the several ways to find the new knowledge from data collection. One of spatial data mining tasks is spatial association rule. There are numerous association rule algorithms have been developed for mining association. Unfortunately, the most algorithms can only used for mining non-spatial and specific formatted data. Therefore, spatial data preprocessing is needed in order conventional association algorithms can be used for spatial data.

In this paper, we will propose methodology and implementation of spatial data preprocessing and implementation of conventional association rule algorithms to discover spatial association rules. Spatial data preprocessing is performed to spatial data with their non-spatial attributes. This process delivers specific formatted data based on spatial relation that specified by user, e.g. topological relation and distance relation. The other task is performed data categorizing for non-spatial data according to parameters specified by user. All results are saved in a table and used as data source for mining spatial association. Two conventional association rule mining algorithms are implemented for mining spatial association. Those are Apriori and FP- Growth (FP-Tree). From software testing, it is indicated that preprocessing time is time consuming. In addition, for the small data volume, Apriori algorithm process is faster than FP-Growth algorithm.

Keywords: spatial data preprocessing, spatial data mining, association rules

1. Introduction

Information technology, remote sensing and GIS (Geographic Information System) grow considerably fast and have been applied in various areas. As a result, data related to geographic also growing fast. These things inspired how to explore these complex geographical data. One of research study related to exploring of data with large volume is spatial data mining. Following definition from data mining, spatial data mining is invention of knowledge from a large amount of spatial data (7)(11). Spatial database consists of geographic database, CAD database, multimedia database, and image database. Thereby, one branch of spatial data mining is geographic data mining (GDM). Geographic data mining is the invention of new knowledge from a large amount of geo-spatial data (geo-reference) (11).

A major difference between data mining in ordinary relational databases and in spatial databases is that attributes of the neighbors of some object of interests may have an influence on the object and therefore have to be considered as well. The explisit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relation) which are used by spatial data mining algorithms (8).

There are various data mining algorithms developed to discover the new and interesting pattern from data sources. These algorithms can be classified into several task, that is, generalization, classification, association, clustering, and trend analysis (4). Mostly, these algorithms working in non-spatial data and usually have a special formatted input, a single table or a single file. For other data-type, adjustment to special data format with data preprocessing is required. Data preprocessing consist of some sub-processes: aggregation, sampling, dimensional reduction, feature selection, discretiza-tion and transformation (12). This task is hardly required in order to conventional data mining algorithms can be applied to spatial data. Thereby, data preprocessing is a very importance task in knowledge discovery.

2. Spatial Association Rule

Spatially, association is a relationship between spatial objects. Association analysis is one of the most widely research topics in data mining. The main focus of association rule mining is to generate hypothesis rather than to test them as is commonly achieved using statistical techniques (15). The concept of association rule, introduced by Agrawal (1), was used for analyzing market basket data to mine customer shopping patterns. This algorithm has been extended by Koperski and Han (7) to spatial data. A spatial association rule is a rule of the form

$$P_1 \wedge P_2 \wedge P_3 \wedge \ldots \wedge P_n \rightarrow Q_1 \wedge Q_2 \wedge Q_3 \wedge \ldots \wedge Q_n \quad (s\%, c\%)$$

where at least one of the predicates P_1, P_2, P_3, ... , P_n , Q_1, Q_2, Q_3, ... , Q_n is a spatial predicate, s% and c% is support and confidence of the rule. For example, a rule

$$\forall X \in DB\ \exists\ Y \in DB: \text{is-a}(X,\text{town}) \rightarrow \text{close-to}(X,Y) \wedge \text{is-a}(Y, \text{Water})\ (c=80\%)$$

express that 80% of the town is close-to water (sea, river, etc).

They have proposed an algorithm to discover multi-level (hierarchical) spatial association rule. The spatial and non-spatial data are organized into hierarchies. Furthermore, the rules are searched initially from the most general level to most specific level (top-down). Rules in the most general level indicated pattern on the widely scale. Based on the strong implication/rule, rule searching is continued to more specific level. Spatial predicates that are used in this algorithm is spatial locality, by means, pattern discovery considered objects contiguity in the space.

Salleb and Vrain (14) extended Koperski algorithm and applied it to find spatial association rule from many layers of geografic data from mineral exploration problems. Their extension is adding non-spatial predicates and searching inter-level association rule. Clementini, *et al.*(3) modified Koperski algorithm and applied to spatial object with broad boundary (spatial object included inaccurately informations). Srinivas and Lam (15) applied spatial association rule mining to find demographics (socioeconomic) and healthy (cancer mortality) data association.

Malerba and Lissi (9) developed Inductive Logic Programming (ILP) to find spatial association rule on demographic database. Basic idea of the algorithm is that spatial database can be reduced by transforming it into deductive database (DDB). Besides, they added background knowledge, for example spatial hierarchie and spatial constraint, and qualitative reasoning. Furthermore, frequent pattern generation and rule generation are performed.

Ickjai Lee proposed other approach to find multivariate association rule in the GIS environment. In this approach, the main added process is preprocessing data included conversion and categorization. Conversion process performed transformation of all layers to numeric value hold areal aggregate and then categorization process grouped into several categories. After that, mining association rule is performed to the preprocessed results.

Lizhen Wang, *et al.*(16) proposed multilevel association rule mining with partition algorithm approach. Basic idea of this approach is storing separately spatial predicates that are obtained from spatial query and then partition-based algorithm is used to find multilevel association rule.

3. Data Preprocessing

Spatial data preprocessing is one of main process in spatial data mining. This process is the most expensive and effort consuming step in the knowledge discovery process because it entangled many operations to get spatial relation (spatial predicates). Therefore, it is required comprehensive understanding to determine the type of preprocessing result.

Spatial relations (topology, direction and distance) along with operations and functions that support geographical data processing generally have been provided by GIS. Thereby, operations that are needed to be added are operations to handle addition constraints. Based on this thing and spatial data preprocessing methodology proposed in (2) and (7), we proposed methodology for spatial data preprocessing (Fig. 1).

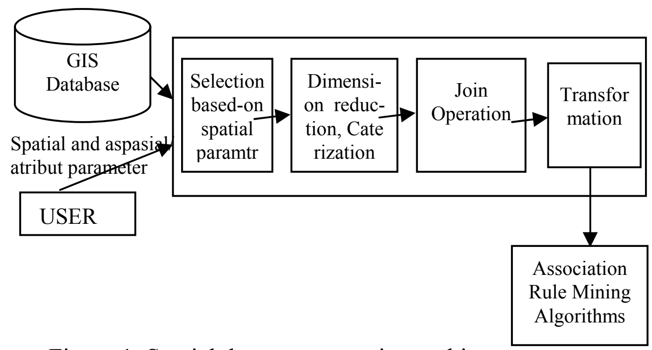


Figure 1. Spatial data preprocessing architecture

In general, spatial data preprocessing can be formulated as follows:

Input:

1. Spatial data(base)
2. Conventional association rule mining algorithms
3. Target feature
4. Spatial relation (parameters)
5. Non-spatial parameters

Process:

Find spatial relation based on based on parameters determined of input.

This process consisted of:

1. Feature (spatial and non-spatial) selection based on spatial parameters.
2. Perform dimension reduction and selection of non-spatial attributes
3. Perform data categorization based on non-spatial data parameters
4. Perform join operations for spatial objects based on spatial parameters.
5. Transform into output form.

Output:

Spatial and non-spatial relations that ready to mine with conventional association rule algorithms.

Spatial feature selection is performed to find spatial layers and non-spatial attributes which will be mined. After that, dimensional reduction and attributes selection by the way of choosing required fields by mining process.

ISBN 978-979-16338-0-2

Data categorization for non-spatial data is performed by dividing attribute values into three categories, those are low, medium and high. For example, these attributes are density population, level of prosperity and number of DBD patients in a certain regional, etc.

Join operations is performed to spatial objects. To apply conventional data mining algorithms, spatial data have to be defined in terms of spatial predicates rather than item (2). Spatial predicates can be in the form of topology, distance and direction. In this paper, spatial relations used are distance (close-to) relation specified by user and topology relation. The result of all processes above is a table that contains spatial and non-spatial relations. Fields in this table contains data categorization of non-spatial and spatial relationship (close-to or far) for spatial objects, shown in Table 1. SR_1, ..., SR_M express spatial relationship, while AttCat_1, ..., AttCat_n express categorization of non-spatial attribute.

Table 1. Table design for preprocessing result

No.	SR_1	SR_m	AttCat_1	AttCat_n
1.	Closed	Far	Low	Mid
2.	Closed	Closed	High	High
...	...	...	...	...

4. Result and Analysis

Based on software architecture above, a software prototype has been developed to handle spatial data preprocessing. Case study that used in this paper is finding spatial association rules in demographics data and number of DBD disease in Surabaya. Features that related with this problems is population density, level of prosperity, health facility, number of DBD cases and the contiguity with source of water (in this case is river and bog). For example, this software input consists of data spatial and non spatial, illustrated in Figure 2. Spatial data consists of some layers those are sub-district, health facility, and bog layer (along with related attributes), while data non-spatial consists of number of residents, population density, and number of DBD cases for every sub-district in Surabaya. Each data will be categorized into three categories those are height, low and medium.

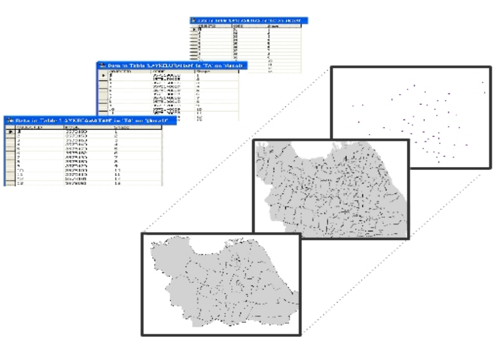
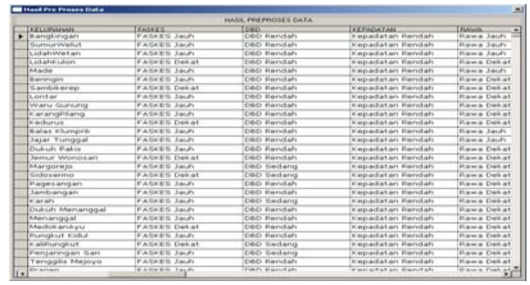


Figure 2. An example of spatial data input


Hasil Pre Proses Data

HASIL PREPROSES DATA

KELURAHAN	FASKES	DBD	KEPADATAN	RAWA
Bangkingan	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
SumurWelut	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
LidahWetan	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
LidahKulon	FASKES Dekat	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Made	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
Beringin	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Sambikerep	FASKES Dekat	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Lontar	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Waru Gunung	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
KarangPilang	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Kedurus	FASKES Dekat	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Balas Klumprik	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
Jajar Tunggal	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Jauh
Dukuh Pakis	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Jemur Wonosari	FASKES Dekat	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Margorejo	FASKES Jauh	DBD Sedang	Kepadatan Rendah	Rawa Dekat
Sidosermo	FASKES Dekat	DBD Sedang	Kepadatan Rendah	Rawa Dekat
Pagesangan	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Jambangan	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Karah	FASKES Jauh	DBD Sedang	Kepadatan Rendah	Rawa Dekat
Dukuh Menanggal	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Menanggal	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
MedokanAyu	FASKES Dekat	DBD Rendah	Kepadatan Rendah	Rawa Dekat
Rungkut Kidul	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat
KaliRungkut	FASKES Dekat	DBD Sedang	Kepadatan Rendah	Rawa Dekat
Penjaringan Sari	FASKES Jauh	DBD Sedang	Kepadatan Rendah	Rawa Dekat
Tenggilis Mejoyo	FASKES Jauh	DBD Rendah	Kepadatan Rendah	Rawa Dekat

Figure 3. Data preprocessing result

After data preprocessing, we will have data preprocessing result in the tabular form that contains spatial data transformation result (Fig. 3).

Mining process of spatial association rule used Apriori-based (1) and FP-Growth algorithms (6). The reason of using of both algorithms are widely used of those algorithms and using of two different approaches. Some interesting patterns got from mining process are:

IF Population_Density_Low
THEN DBD_Low (0.61) (0.75)

IF Health_Facility_No
AND Population_Density_ Low
THEN DBD_Low (0.47) (0.8)

IF Health_Facility_No
AND Population_Density_ Low
AND Close-to_Bog
THEN DBD_Low (0.86) (0.86) Etc

From software testing, it is indicated that data preprocessing is time consuming. This caused by spatial joint process execution that require much time.

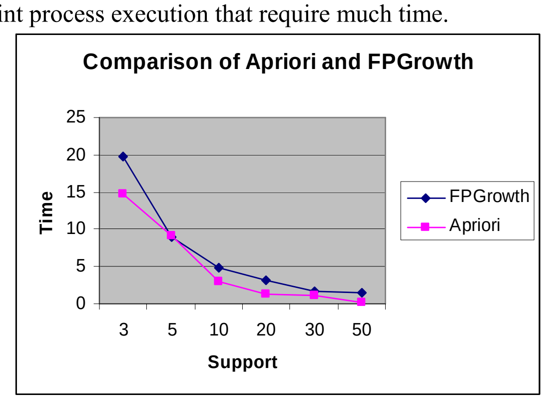


Figure 5. Comparison of Apriori and FP-Growth

Executing both association rule algorithms resulted or indicates that both algorithms generate the same patterns. Another interesting result is that Apriori algorithm is faster than FP-Growth (Fig. 5). This result may be caused by relatively few number of data (there are 163 records of sub-district in Surabaya). Another reason is the patterns that used in the case study is not a long pattern, whereas, one of several FP-Growth advantages is better to work with long pattern (6).

5. Conclusion and Future Work

In this paper, we have proposed methodology and implementation of spatial data preprocessing and then performed mining spatial association rule with conventional association algorithms. Main steps in this spatial data preprocessing is spatial and non-spatial feature selection based on parameter determined, reduction of dimension, selection and categorization of non-spatial attributes, join operation for the spatial objects based on spatial parameter spatial and transforms into form of output wanted. While finding spatial association rule used Apriori and FP-Growth algorithm.

For the near future, we plan to continue this research to accommodate temporal constraint to spatial association rule mining.

6. References

(1). Agrawal, Rakesh and Ramakrishnan Srikant, "Fast Algorithms for Mining Assosiation Rules", *Proceedings of the 20th VLDB Conference,* Santiago, Chile (1994).

(2). Bogorny, Vania, Paulo Martins Engel, and Luis Otavio Alvares, "GEOARM: an Interopera-ble Framework to Improve Geographic Data Preprocessing and Spatial Association Rule Mining", *Proceedings of SEKE06*, (2006)

(3). Clementini Clementini, Eliseo, Paolino Di Felice, Krzysztof Koperski, "Mining multiple-level spatial association rules for object with a broad boundary", *Data & Knowledge Engineering 34*, p. 251-270, Elsevier (2000).

(4). Ester, Martin, Alexander Frommelt, Hans Peter Kriegel, Jorg Sander, "Spatial Datamining: Database primitives, algorithms and Efficient DBMS Support", *Datamining and Knowledge Discovery*, 4, 193-216, Kluwer Academic Publisher (2000).

(5). Gahegan, Mark, "Data mining and Knowledge Discovery in the geographical domain", *National Academic White Paper, Intersection of Geospatial Information and Information Technology* (2001).

(6). Han, Jiawei, J. Pei, and Y. Yin, "Mining Frequent Pattern without Candidate Generation", in *Proceedings ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD 2000)*, pp. 1-12 (2000).

(7). Koperski, Krzysztof and Jiawei Han, "Discovery of Spatial Assosiation Rules in Geographic Information Databases, in *Advances in Spatial Databases*", Proc. Of 4th Symp. SSD'95, Springer Verlag, Berlin, pp. 47-66 (1995).

(8). Ladner, Roy, Kevin Shaw, and Mahdi Abdelguerfi, *Mining Spatio-temporal Information System*, Kluwer Academic Publisher (2002).

(9). Malerba, Donato and Francesca A. Lissi, "Discovering Assosiation between Spatial Object: An ILP Application", in C. Rouveirol & M. Sebag (Eds.), *Inductive Logic Programming, Lecture Notes in Artificial Intelligence, 2157*, 156-163, Springer (2001).

(10). Mennis, Jeremy and Junwei Liu, "Mining Asosiation Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change", *Transactions in GIS*, 9(1): pp. 5-17 (2005).

(11). Miller, Harvey J., "Geographic Data Mining and Knowledge Discovery" in J. P. Wilson and A. S. Fotheringham (eds.) *Handbook of Geographic Information Science* (2004).

(12). Pang Ning Tang, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining*, Addison Wesley (2006).

(13). Openshaw, Stan, "*Geographical Data Mining: Key Design Issues*", Centre for Computational Geography", School of Geography, University of Leeds, Leeds LS2 9JT UK (1999).

(14). Salleb, Ansaf and Christel Vrain, "An Application of Assosiation Rules Discovery to Geographic Information Systems", *Proceeding of Knowledge Discovery and Data Mining (PKDD) 2000*, LNAI 1910, pp. 613-618, Springer Verlag (2000).

(15). Vinnakota, Srinivas and Nina S. N. Lam, "Knowledge Discovery from Mining the Spatial Association between Cancer and Socioeconomic Characteristics", *International Journal Health Geographics*, 5(9) (2006).

(16). Wang, Lizhen, Kunqing Xie, Tao Chen, and Xiuli Ma, "Efficient Discovery of Multilevel Spatial Assiciation Rules Using Partitions", *Information and Software Technology*, vol. 47, pp. 829-840 (2005).