ARTICLE IN PRESS

Journal of King Saud University – Computer and Information Sciences (2017) **xxx**, xxx–xxx

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Incremental topological spatial association rule mining and clustering from geographical datasets using probabilistic approach

Y. Jayababu [a,*], **G.P.S. Varma** [b], **A. Govardhan** [c]

[a] *Department of CSE, Pragati Engineering College, Surampalem, Andhra Pradesh 533437, India*
[b] *S R K R Engineering College, Andhra Pradesh, India*
[c] *University College of Engineering, Jawaharlal Nehru Technological University Hyderabad, India*

Received 26 March 2016; revised 20 December 2016; accepted 21 December 2016

KEYWORDS

Spatial database;
Association rule mining;
Topological support;
Probabilistic approach;
Dynamic database

Abstract Due to the dynamic updating of real time spatial databases, the preservation of spatial association rules for dynamic database is a vital issue because the updates may not only invalidate some existing rules but also make other rules relevant. Consequently, the dynamic updating of spatial rules was handled by many researchers through the incremental association rule mining algorithm. Accordingly, in this paper we have developed an incremental topological association rule mining of geographical datasets using probabilistic approach. Initially, the spatial database is read out and it is passed through probability-based incremental association rule discovery algorithm to mine the topological spatial association rules. Once the rules are mined from the spatial database, the assumption here is that the database is dynamically updating for every time interval. In order to handle this dynamic nature, the proposed incremental topological association rule mining process is used in this paper. Here, the candidate topological rule generation is done from the spatial association rules using the topological relations such as, nearby, disjoint, intersects and inside/outside and the topological support is calculated using the proposed probabilistic topological support model. Finally, the spatial clustering is performed based on the mined spatial rules. From the experimentation, we proved that the maximum accuracy reached by the proposed method is 83.14% which is higher than the existing methods, which is defined as the ratio of the occurred rules and total number of topological data objects.

1. Introduction

The collection of a large amount of spatial data is gathered by various developing fields such as remote sensing, e-commerce, and other data collection tools. Due to the huge amount of spatial data, the extraction of information from the spatial database (Ding et al., 2006) is one the most

* Corresponding author.
E-mail addresses: jayababuy2015@gmail.com (Y. Jayababu), gpsvarma@yahoo.com (G.P.S. Varma), govardhan_cse@yahoo.co.in (A. Govardhan).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.12.006

challenging task. In order to overcome the above challenge, the automated information is discovered from the spatial data, which leads to favorable fields of data mining or knowledge discovery in databases (KDD) Mukhopadhyay et al., 2014; Wu et al., 2014; Jiang et al., 2015. With the help of KDD database useful information can be retrieved from the data such as previously unknown values, hidden information and uncertain values (Clementini et al., 2000; Frawley et al., 1991). The classification of spatial data mining (Shyu et al., 2006; Laube et al., 2008; Guo et al., 2015; Ding et al., December 2008; Dao and Thill, 2012) is performed based on the types of rules, which have been located in the spatial database. Basically, the spatial association rule can be represented as $X \rightarrow Y$, in which the representation of X and Y shows the predicate set. A few of the aforementioned predicate sets contain spatial data. Basically, the large database contains various association relationships but some association relationships are not occurring regularly based on the concepts of minimum support and minimum confidence. In a set of spatial objects S, the support pattern of A is the probability that a member of spatial object S satisfies the pattern A.

Then, the confidence measure of pattern A and B is the probability that the pattern B occurs when the pattern A occurs. The threshold value is given by the user to determine the strong spatial association rules (Koperski and Han, 1995; Han and Fu, 1995; Dong et al., 2012). Basically, the spatial database is used to accumulate and control the spatial objects. The spatial object consists of two components such as descriptive component and spatial component. The components based on the spatial data mainly include their geometry, which is based on the type of point, line, surface, etc. Based on the topological relationships, the spatial objects are related to each other. The topological relationship (Pascucci et al., 2011; Fang et al., 2010; Doraiswamy et al., 2014) is based on the representation of a spatial extent by a set of points and composition of three subsets such as boundary, interior and exterior. In this paper we have taken the topological relations, which have been used as spatial predicates for complex objects. The spatial predicates are named as adjacent, within, close and overlap. A set of topological relations for geographical datasets using probabilistic approach is used in this paper. In order to reduce the computational overhead the topological relations are used in the complex objects. The generated spatial database is passed through the topological relations to mine the topological spatial association rules. Basically, the spatial database is dynamically updated for every time interval due to nature. Here, a probability-based dynamic discovery of rules is performed for the newly added database and the preservation of the important spatial rules are computed based on the probability (Mohamed and Refaat, 2011) of occurrence in the existing and new database.

The organization of the paper is as follows: Literature review is presented in Section 2. The problem definition and contributions of the paper are presented in Section 3. Proposed methodology: Incremental topological association rule mining of geographical datasets using probabilistic approach is presented in Section 4. In Section 5, the experimental results and the Performance analysis of topological spatial rule mining are presented. Finally, we conclude this paper in Section 6.

2. Literature review

Literature presents various techniques for spatial association rules' mining and clustering. Here, we present the review of different works. Koperski and Han (1995) have proposed an efficient method for mining strong spatial association rules in geographic information databases. A spatial association rule is a rule indicating certain association relationship among a set of spatial and possibly some nonspatial predicates. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships. Several optimization techniques were explored, including a two-step spatial computation technique (approximate computation on large sets, and refined computations on small promising patterns), shared processing in the derivation of large predicates at multiple concept levels, etc. This work faces issues when we incorporate the multiple concepts into the mining algorithm without much computational overhead. Clementini et al. (2000) have used objects with broad boundaries, the concept that absorbs all the uncertainty by which spatial data were commonly affected and allows computations in the presence of uncertainty without rough simplifications of the reality. The topological relations between objects with a broad boundary can be organized into a three-level concept hierarchy. The progressive refinement approach was used for the optimization of the mining process. Even though the rule mining process utilizes the optimization algorithm, the mining for accurate spatial rules are completely missed due to the random initialization.

Shyu et al. (2006) have customized the data mining algorithms using visual content and potential objects extracted from geospatial image databases with other relevant information, such as text-based annotations. Queries utilizing the mining results were also discussed in this paper. These mining and query processing algorithms play an important role in GeoIRIS-Geospatial Information Retrieval and Indexing System. The query processing for the multiple concept and topological relations pose manual preparation for the rule mining processes. Laube et al. (2008) have investigated the support and confidence measures for spatial and spatio-temporal data mining. Using fixed thresholds to determine how many times a rule that uses proximity is satisfied seems too limited. It allowed the traditional definitions of support and confidence, but does not allow to make the support stronger if the situation is "really close", as compared to "fairly close". The traditional measure of support and confidence are not suitable to mine the spatial rules if they considered the topological relations.

Qin et al. (2008) have proposed an efficient approach to derive association rules from spatial data using Peano count tree (P-tree) structure. P-tree structure provided a lossless and compressed representation of spatial data. Based on P-trees, an efficient association rule mining algorithm PARM with fast support calculation and significant pruning techniques was introduced to improve the efficiency of the rule mining process. The P-tree based association rule mining (PARM) algorithm was implemented and compared with FP-growth and Apriori algorithms. Even though the tree-based mining algorithm are effective than Apriori, the memory requirement to store the tree structure is high as compared with the candidate-based methods. Dao and Thill (2012) have proposed a comprehensive framework and library of algorithms of spatial analysis and