2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Exploring Spatial Relationships for Knowledge Discovery in Spatial Data

Norazwin Buang [+], Abdullah Mohd Zin and Mohamad Shanudin Zakaria

Faculty of Information Science and Technology (FTSM)
Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, Malaysia

Abstract. The process of exploring spatial relationships is the most expensive task for knowledge discovery in spatial data. This task is crucial to reduce the data pre processing time in discovering interesting patterns in spatial data. However, this task has received little attention in the literature. With the increased amount of spatial data, an efficient technique for exploring spatial relationships is necessary. This paper presents a preliminary study of approaches for exploring spatial relationships for knowledge discovery in spatial database system.

Keywords: knowledge discovery, spatial relations, spatial data, data mining

1. Introduction

Exploring spatial relationship refers to a process of computing topological, distance and direction relationship among spatial objects. Usually, spatial relations are not explicitly stored in spatial databases and are computed when necessary. The process of exploring spatial relationships involves complex computational geometry algorithm and hence computationally expensive. The problem is more prevalent when the spatial objects are complex or consists of multi lines or multi polygons or the combination of both. With the proliferation of spatial data collected nowadays, there is a need for a far more efficient approach in exploring spatial relationships.

The process of exploring spatial relationships is one of the steps in knowledge discovery in spatial databases system. Knowledge discovery in spatial data (KDSD) or spatial data mining (SDM) refers to the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases [8]. Knowledge discovery in spatial data differs from knowledge discovery in relational data because the analysis process use spatial predicates that relate two spatial objects as one of the attributes of spatial data.

KDSD involves data selection, pre processing, exploring spatial relationship, data transformation, data analysis and result interpretation in discovering various types of knowledge for numerous purposes. Among them are spatial characteristic rules [4], spatial classification rules [8, 13], spatial association rules [8, 9], spatial associative classification rules [2, 6] and spatial trend detection [4]. Such rules using the spatial predicates representation are:

is_a (x, hospital) ^ *contains* (x, religious_site) → *close_to* (x, park) [24%, 80%]

is_a (y, store) ^ *close_to*(y, school) → profit (y, high)

The first rule is a spatial association rule where 80% of hospitals with a religious site onsite are also *close to* a park with 24% support. The second rule classifies that the profits of any store that *close to* a school is high.

[+] Corresponding author. Tel.: +60389216087; fax: +60389256732.
E-mail address: azwin@ftsm.ukm.my.

Many traditional data mining techniques such as decision trees and apriori algorithms have been applied to discover spatial rules. However, we focus only on the approaches to explore spatial relationship instead of techniques to analyze and mine spatial patterns.

2. Spatial Relations

Spatial relations can be classified into three categories [3]. They are topological, direction and distance. Topological relations are the spatial relations that are preserved under transformations such as rotation and scaling. It also refers to the relation that describes neighbourhood and incidence. Egenhofer and Franzosa (1991) pointed out that there are eight basic topological relations (*disjoint, overlap, inside, covers, coveredBy, contains, equal and meet*) within the topological category.

The second category of spatial relation is direction, which refers to the relation that describes order in space. Two formalization models for representing direction relations introduced by Frank (1996) are projection-based and cone-based direction model [5]. The last category of spatial relation is distance, the description of proximity in spaces [7].

3. Current Approaches in Exploring Spatial Relationships

The process of exploring spatial relations depends on the complexity of spatial data types and spatial relations. Ceci et al (2004) presented more than 700,000 spatial relations in discovering classification rules for mortality index [2]. The number of spatial relations is high because the process involves calculation for all pairs of spatial objects. For example, to determine if a district *contains* or *disjoint* to a university, it may need to compute spatial relations from the district to all the universities.

Several approaches have been proposed in the last decade to reduce the processing time for exploring spatial relationships in KDSD. Ester et al (1997) introduced neighborhood graph to modeling geographic neighbours [4]. Others optimize the computational geometry algorithm by introducing progressive-refinement approach [8], join-indexing [13], clustering and join-indexing [6] and rule-based qualitative spatial reasoning strategy [9, 10]. Another approach utilizes the knowledge constraints to reduce the number of spatial relationships to be computed [1]. We present such approaches for exploring spatial relationship together with its strengths in Table 1.

TABLE I. Description of Exploring Spatial Relationships Approach

Approach	Description	Indexing	Reference
Progressive-Refinement	optimize the computation using MBR		Koperski et al (1995)
Neighborhood Graphs	simplify spatial relation into 'neighbour' only	√	Ester et al (1997)
Spatial Join Indexing	using join index	√	Zeitouni et al (2000)
Micro Cluster Join Indexing	integrate clustering and indexing	√	He (2001)
Rule-based Qualitative Spatial Reasoning	implement qualitative spatial reasoning approach using rule induction algorithm		Santos et al (2000, 2004), Wang et al (2006)
Elimination of Well-Known Dependencies	require domain experts to eliminate well-known spatial relations		Bogorny (2006)

Koperski and Han (1995) proposed a progressive refinement approach to optimize the process of exploring spatial relationships for KDSD system [8]. Fig. 1 illustrates the process of KDSD using the progressive refinement approach. The process involves six steps. The first step is data query followed by coarse predicate computation. This step computes coarse spatial relations (such as *course generalized close to, coarse_g_close_to*) for all pairs of spatial objects. The spatial relation is at coarse resolution because it computes between two minimum bounding rectangles (MBR). MBR is used as an approximation of the spatial object when exploring spatial relations because it can reduce the computation time. The next step is data reduction or relevance analysis. The output of this step will generate a candidate set for fine predicate computation. Here, the number of spatial relationships has been reduced and a more precise spatial relationship (such as *intersects*, *contains*) is computed. Finally, patterrns are extracted using techniques such as data mining and are visualized to the users.

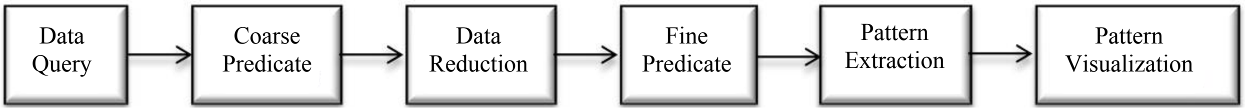

Fig. 1: The Progressive Refinement Approach

Although the progressive refinement approach can optimize the process of exploring spatial relationships, it has been integrated within the apriori-based algorithm. Therefore, the approach has also inherited some of the downsides of apriori-based algorithm. The approach becomes very costly if there are a large number of frequent patterns, very long patterns and patterns with low minimum support thresholds since it requires repeatedly scanning the database and checking for a large set of candidates and very sensitive to the user defined support threshold. Furthermore, it is also difficult to make the approach offline because it depends on the user specified minimum support thresholds.

Ester et al (1997) working independently from Koperski and Han (1995) introduced graph data structure to store spatial objects and neighborhood relations [4]. He defines a neighbourhood graph G for spatial relation “neighbour” as a graph (N, E) with the set of nodes, N and the set of edges, E. Each node corresponds to an object of the database and two nodes n_1 and n_2 are connected via some edge if and only if the relation neighbour (object (n_1), object (n_2)) holds.

There are a few identified disadvantages of this approach. First, this approach is limited to modeling only one spatial relationship, called ‘*neighbour’*. According to Koperski and Han (1995), one neighbour relation is insufficient to discriminate and segment the data correctly. Furthermore, in real applications, each object could have many neighbours and the spatial relations can be of various types. Another limitation of this approach is it is computationally expensive to compute all relationships between all objects in the database to obtain a neighbourhood graph and neighborhood paths for real database. Finally, although this approach is supported by neighbourhood index, it is expensive to construct indices for ‘*neighbour’* relationship because of possibly infinite number of relations and different distance values.

The third approach is indexing. Indexing has been implemented for KDSD by Zeitouni et al (2000) where spatial relations are pre-computed and stored in the form of spatial join index [13]. However, similar to neighborhood graph approach, the infinite number of relationships makes it hard to identify the relations to be pre-computed. Furthermore, it is still expensive to pre-compute spatial relations and storing them in join index when there are a large number of training objects and reference objects in the analysis of spatial data [6]. He (2001) reported that the progressive refinement outperformed spatial join index approach because the progressive refinement performs fine spatial computation on those candidates found by coarse computation while join index performs on all training reference object pairs [6].

Since spatial join index is still costly compared to progressive refinement approach, He (2001) has extended a spatial join indexing to a micro cluster join indexing to reduce the size and the computational cost for building a spatial join index. This approach involves two steps. First, it groups the training objects into small clusters and second, it computes spatial relations between each reference objects and the relevant micro-cluster and stores them in spatial join index. In order to identify spatial relations between reference objects and the relevant micro-cluster, spatial relations are evaluated between reference object and each objects in the micro-cluster. A spatial relation is inserted into the join index *if and only if* all the objects in the cluster satisfy the spatial relations *p* and nothing is inserted otherwise. In cases where some of the objects in a cluster satisfy spatial relations while some are not, a spatial relation has to be evaluated at the individual object level. The spatial relation is inserted into the join index only for pair of spatial objects that satisfy the spatial relations. Therefore in the spatial join index, there are two types of spatial relations; between an object and a cluster, and between two objects.

The micro-cluster join indexing approach has contributed to a far more efficient spatial computation and less space for storing materialized features. However, this approach can become complex when clustering data like lines and polygons. This approach looks appropriate when most of the data are in points form. The work of He (2001) is the only work that extensively evaluate the performances of exploring spatial relations approaches for KDSD. He (2001) demonstrates that micro-clustering outperformed progressive refinement approach because micro clustering shares the expensive process of spatial computation.

The approaches presented require computational geometry algorithm to compute all necessary spatial relationships. The coordinates of the spatial objects are initially extracted before the computation process can begin. However, the computational geometry algorithm cannot be completed in the absence of the coordinates. There is a case when a KDSD does not have function to extract and process the coordinates, and only some of the pre-computed spatial relations can be extracted. In this case, qualitative spatial reasoning (QSR) approach can be used to infer unknown spatial relations. Introduced by Sharma (1996), QSR refers to the process by which information about aspects in space and their relationships are gathered through measurement, observation or inference and used to arrive at valid conclusions regarding the relationships of the objects [11].

Using this approach, given

A *contains* B and

B *contains* C,

one can infer that

A *contains* C

However, the inference of new spatial relations can only be achieved based on the composition table, i.e a table contains the pre-defined qualitative rules.

Several attempts have been made to utilize this approach in exploring spatial relationships for KDSD [9, 10, 12]. Santos and Amaral (2000 and 2004) have proposed this approach in the PADRAO's system [9, 10]. The architecture of PADRAO consists of three main components; Knowledge and Data Repository, Data Analysis and Results Visualization. The first component is used to store data and knowledge needed in the knowledge discovery process. It is also required for spatial reasoning to complete the process of exploring spatial relationship. This component can be divided into two sub-components; Knowledge Repository and Data Repository. PADRAO's Data Repository stores the spatial semantic knowledge while PADRAO's Knowledge Repository stores the principle of qualitative spatial reasoning or composition tables. Data Analysis and Results Visualization components are used to discover patterns or other relationships implicit in the geographic and non-geographic data and visualize the discovered patterns respectively.

One of the advantages of QSR is it can infer different types of spatial relationships given two computed spatial relations. For instance, by computing that object A is *North_of* B and object B is *North_of* C, one can infer that A and C are *disjoint*. It is also known as integrated spatial reasoning [11]. In the past decade, there are many attempts to integrate topological and direction, direction and distance and topological and distance. Although this approach can support integrated spatial reasoning, the reasoning process requires several composition tables as a look up which will then, complicate the inference process.

Another approach to reduce the process of computing spatial relations is by eliminating well known dependencies between spatial objects [1]. Introduced by Bogorny et al (2006), this approach reduces the number of spatial relations for KDSD using geo-ontology and semantic spatial integrity constraints. In this approach, mandatory and prohibited topological relationships are identified by experts and then eliminated. Two algorithms, AprioriKC and AprioriKC+ have been developed to apply this approach in KDSD. The former eliminates well known dependencies while the latter eliminate pairs with either well known dependencies or pairs with two predicates containing the same feature type but different topological relationships (e.g *contains* (university) and *touches* (university)).

At high granularity level, a spatial object might be an association of several spatial objects. Nowadays, with the higher resolution of satellite and remote sensing images, spatial objects can be derived at low granularity level. For instance, the state of Terengganu in Malaysia might be seen as one polygon at high granularity level but multi polygons at low granularity level because the state has many islands. This results in more complex spatial objects and hence the process of exploring spatial relationships will become more computationally expensive. Therefore, the approaches of exploring spatial relations are worth to be studied further and evaluated in more details.

4. Conclusions

In this paper, we present a preliminary study on exploring spatial relationship approaches introduced in the past decades for KDSD. This work can contribute to improve a KDSD system because the process of exploring spatial relationships is a bottleneck of the KDSD. Furthermore, this issue is rarely being discussed extensively in the past researches.

5. References

[1] V. Bogorny, P. Engel, L. O. Alvares. Towards the reduction of spatial joins for knowledge discovery in geographic databases using geo-ontologies and spatial integrity constraints. *ECML/PKDD Second Workshop on Knowledge Discovery and Ontologies (KDO'05).* Porto, Portugal. 2005, pp. 51-58.

[2] M. Ceci, A. Appice, D. Malerba. Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach. In *Knowledge Discovery in Databases: PKDD*. 2004, pp. 99-111.

[3] M. J. Egenhofer and R. D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Science*. 1991, (**5**):161-174.

[4] M. Ester, H.-P. Kriegel, J. Sander. Spatial data mining: A database approach. *Proc. of the Fifth International Symposium on Large Spatial Databases*. Germany. 1997, pp. 47-68.

[5] A. U. Frank. Qualitative spatial reasoning: cardinal directions as an example. *International Journal of Geographical Information Science*. 1996, (**10**):269-290.

[6] J. He. *SPARC - An Association-Rule-Based Classification Algorithm for Spatial Data Mining*. In School of Computing Sciences. Master of Sciences: Simon Fraser University. 2001, p. 71.

[7] D. Hernández, E. Clementini, P. Di Felice. *Qualitative Distances*: Inst. für Informatik. 1995.

[8] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. *Proceedings of the 4th International Symposium on Advances in Spatial Databases*. 1995, pp. 47-66.

[9] M. Y. Santos and L. Amaral. Knowledge discovery in spatial databases: the PADRÃO's qualitative approach. *Cities and Regions, GIS Special Issue*. 2000, pp. 33-49.

[10] M. Y. Santos and L. Amaral. Mining geo-referenced data with qualitative spatial reasoning strategies. *Computers and graphics*. 2004, (**28**):371-379.

[11] J. Sharma. *Integrated Spatial Reasoning in Geographic Information Systems: Combining Topology and Direction*. University of Maine, PhD Thesis. 1996.

[12] S.s Wang, D.y. Liu, X.y. Wang, J. Liu. Spatial Reasoning Based Spatial Data Mining for Precision Agriculture. *Advanced Web and Network Technologies, and Applications*. 2006, pp. 506-510.

[13] K. Zeitouni, L. Yeh, M. A. Aufaure. Join indices as a tool for spatial data mining. *International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence*. 2000, pp. 102-114.