A Survey: Knowledge Discovery in Spatial Databases

Shishir Shrivastava\(^1\), Ajaya Kumar Akasapu\(^2\) and Lokesh Kumar Sharma\(^3\)

\(^1\)Student RCET Bhilai1CG, \(^2,3\) Reader RCET Bhilai CG

shishir.contact@gmail.com , akasapu@gmail.com and lsharmain@gmail.com

ABSTRACT

Knowledge discovery in databases is a complex process concerned with the discovery of relationships and other descriptions from data. Knowledge discovery in spatial databases represents a particular case of discovery, allowing the discovery of relationships that exist between spatial and non-spatial data, and other data characteristics that aren’t explicitly stored in spatial databases. This paper describes the conception and implementation of PADRÃO, a system for knowledge discovery in spatial databases. PADRÃO presents a new approach to this process, which is based on qualitative spatial reasoning. The spatial semantic knowledge and the principles of qualitative spatial reasoning needed for the spatial reasoning process are available in the PADRÃO’s geographic database and PADRÃO’s spatial knowledge base, allowing the integration of the geo-spatial component, associated with the analyzed non-geographic data, in the process of knowledge discovery.

This paper presented the PADRÃO system, a system for knowledge discovery in spatial databases based on qualitative spatial reasoning. PADRÃO represents a new approach to this particular case of knowledge discovery, in which the positional aspects of geographic data are provided by a spatial reference, gave by a geographic identifier (in the described case, the municipalities’ name). The PADRÃO’s geographic database and spatial knowledge base store the spatial semantic Knowledge and the qualitative spatial reasoning principles needed for the inference of new spatial relations required in the knowledge discovery process.

KEYWORDS

KDD: Knowledge Discovery in Databases
KDSD: Knowledge Discovery in Spatial Databases
DDB: Demographic Database
GDB: Geographic Database
SKB: Spatial Knowledge Base
nGDB: non-Geographic Database
PDB: Patterns Database

1. INTRODUCTION

Large amounts of operational data concerning several years’ operation are now becoming available, mainly in middle-large sized organizations. Knowledge Discovery in Databases (KDD) is the key to access the strategic valued of the organizational knowledge buried in databases, usable both for daily operation, general management and strategic planning. The process of KDD automates the discovery of relationships and other descriptions from data. Data mining is one of the steps of this process, concerned with the application of specific algorithms for extracting patterns from. The main recognized advances in the area of KDD are related with the exploration of relational databases. However, in most organizational databases exists one dimension of data, the geographic (associated with addresses or postcodes), which semantics is not used by traditional KDD systems. Knowledge Discovery in Spatial Databases (KDSD) is related with “the extraction of interesting spatial patterns and features, general relationships that exist between spatial and non-spatial data, and other data characteristics not explicitly stored in spatial databases”. Spatial database systems are normally relational databases plus a concept of spatial location and spatial. The explicit location and extension of objects define implicit relations of spatial neighborhood. The neighbor attributes of a given object may influence its behavior and therefore must be considered in the process of knowledge discovery. Knowledge discovery in relational databases doesn’t take into consideration this spatial reasoning, motivating the development of new algorithms adapted to the characteristics of spatial data. The main approaches in KDSD are characterized by the development of new algorithms that treat the objects’ position and extension through the manipulation of its co-ordinates. These algorithms are subsequently implemented, extending traditional knowledge discovery systems. In all, a quantitative spatial reasoning approach is used, although the results are presented using qualitative identifiers (like far, close, north.). This paper describes the conception of PADRÃO, a system for KDSD. PADRÃO presents a new approach to the process of KDSD based on qualitative spatial reasoning and was implemented recurring to a traditional knowledge discovery system. The spatial semantic knowledge and the principles of qualitative spatial reasoning needed for the spatial reasoning process are available in the PADRÃO’s geographic database and PADRÃO’s spatial knowledge base, allowing the integration of the data geo-spatial component in the process of knowledge discovery. The integration of a geographic database the municipality and district level, and a demographic database, in any area allowed to PADRÃO the discovery of implicit relationships existing between the analyzed geographic and demographic data. This paper is organized in several sections.
In them, qualitative spatial reasoning is defined and described how its concepts are used in the knowledge discovery process. The architecture of PADRÃO is presented, describing its main components, the several steps associated with it. The application of PADRÃO to the demographic domain is also illustrated, referring the type of discoveries that can be achieved with it.

2. QUALITATIVE SPATIAL REASONING

The positional aspects of geographic data are provided by a spatial reference, which relate the data to a given position on the Earth’s surface. Spatial references fall into two categories: based on co-ordinates or on geographic identifiers. In systems of spatial referencing using geographic identifiers (indirect referencing systems), a position is referenced to a real world location defined by a real world object. This object is termed a location, and its identifier is termed a geographic identifier. These geographic identifiers are very common in organizational databases, allowing the integration of the spatial component associated with it in the process of knowledge discovery. The adoption of an indirect geographic reference system imposes the use of qualitative spatial reasoning strategies, able to deal with the spatial semantic not explicitly associated with the adopted geographic identifiers. Spatial reasoning is the process by which information about objects in space and their relationships are gathered through measurement, observation or inference, and used to arrive to valid conclusions regarding the objects’ relationships. Qualitative spatial is based on the manipulation of qualitative spatial relations, for which composition tables facilitate reasoning, allowing the inference of new spatial knowledge. Spatial relations have been classified in several types, including direction relations (that describe order in space), distance relations (that describe proximity in space) and topological relations (that describe neighborhood and incidence). These spatial relations are briefly described in the next subsections.

3. DIRECTION RELATIONS

Direction relations describe where objects are placed relative to each other. Three elements are needed to establish an orientation: two objects and a fixed point. Cardinal directions can be expressed using numerical values specifying degrees (0°, 45°…) or using qualitative values or symbols, such as North or South, those have an associated acceptance region. The regions of acceptance for directions can be obtained by projections (also known as half-planes) or cone-shaped regions (Figure 1).

A characteristic of the cone-shaped system is that the region of acceptance increases with distance, which makes it suitable for the definition of directional relations between extended objects. Another benefit is that this system allows the definition of finer resolutions, permitting the use of eight (Figure 2) or sixteen different qualitative directions. This model uses triangular acceptance areas that are drawn from the centroid of the reference object towards the primary object (In the spatial relation A North B, B represents the reference object, while A constitutes the primary object). Since the geographic domain analyzed in this work characterizes administrative subdivisions, the cone-shaped system will be used in the identification of the direction relations existing between the several geographic subdivisions.

4. DISTANCE RELATIONS

Distances are quantitative values determined through measurements or calculated from known co-ordinates of two objects in some reference system. The most familiar definition of distance is the length of the shortest possible path between two objects, also known as Euclidean distance. Usually a metric quantity is mapped onto some qualitative indicator such as very close or far for human commonsense reasoning. Qualitative distances must correspond to a range of quantitative distances specified by an interval, and should be ordered so that comparisons are possible. Another requirement is that the length of each successive qualitative distance should be greater or equal to the length of the previous one (Figure 3).

5. TOPOLOGICAL RELATIONS

Topological relations are those relations that are invariant under continuous transformations of space such as rotation or scaling. There are eight fundamental relations that can exist between two planar regions: disjoint, contains, inside, equal, meet, covers, covered by and overlap (Figure 4). These relations can be defined considering intersections between the two regions, their boundaries and their complements.
In some exceptional cases, the geographic space can’t be characterized, in topological terms, recurring to the eight primitives presented above. One of these cases is related with application domains in which the addressed geographic regions are administrative subdivisions. Administrative subdivisions can only be related through the topological primitives disjoint, meet and contains (and its inverse inside), since they can’t have any kind of overlapping. The topological primitives used in this work are disjoint and meet, once the implemented qualitative inference process only considers regions of the same geographic level.

Integration of direction, distance and topological spatial relations
Reasoning about qualitative directions necessarily involves integrated spatial reasoning about qualitative distances and directions. Particularly in objects with extension, the size and shape of objects, and the distance between them, influence the directions. One of the ways to determine the direction and distance between regions is calculating them for its respective centroid. The extension of the geographic entities is somehow implicit in the topological primitive used to characterize its relations. Qualitative spatial reasoning requires the adoption of a set of qualitative identifiers. In the implemented approach, integrated spatial reasoning with direction, distance and topological spatial relations, the adopted set of qualitative identifiers were:

- Direction = \{N, NE, E, SE, S, SW, W, NW\}
- Distance = \{very close, close, far, very far\}
- Topological = \{disjoint, meet\}

For each of these qualitative identifiers (direction and distance, since topological relations are quantitatively by nature), is required the definition of the validity interval that limits quantitatively the region of acceptance of each one. In the case of direction relations, for the cone-shaped system with eight acceptance regions, the quantitative intervals adopted are:

\[ [337.5, 22.5), [22.5, 67.5), [67.5, 112.5), [112.5, 157.5), [157.5, 202.5), [202.5, 247.5), [247.5, 292.5), [292.5, 337.5) \]

From N to NW respectively. In the case of distances, there should exist a constant ratio (ratio-length (disti)/length (disti-1)) relationship between the lengths of two neighboring intervals. For example, for a ratio 42 the obtained intervals are:

<table>
<thead>
<tr>
<th>Ratio</th>
<th>dist1</th>
<th>dist2</th>
<th>dist3</th>
<th>dist4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>(0.1)</td>
<td>(1.5)</td>
<td>(5.21)</td>
<td>(21.85)</td>
</tr>
</tbody>
</table>

Analyzing the furthest distance that can exist between two regions of the addressed geographic domain, the ratio 4 intervals need to be magnified 3 by a factor of 10, resulting in the validity intervals: \( (0, 10], (10, 50], (50, 210] \) and \( (210, 850] \) corresponding from very close to very far respectively. The adoption of a mixed approach allowed the integration of direction, distance and topological relations, under the principles of qualitative spatial reasoning. The composition table is represented recurring to graphical symbols, like the ones presented in Figure 5. In order to exemplify the set of rules stored in the final composition table, Table 1 exhibits a subset of the rules contained in it.

The set of rules stored in the composition table (Table 1) can be used in the inference of new spatial relations needed in the knowledge discovery process. For example, knowing the facts A Northeast, very close, meet B and B East, close, disjoint C, it can be inferred that A east, close, disjoint C, as can be verified in Figure 6.

6. THE PADRÃO SYSTEM
PADRÃO is a system for KDSD based in qualitative spatial reasoning. This section presents its architecture and gives some technical details about its implementation. The PADRÃO’s architecture
The architecture of PADRÃO (Figure 7) aggregates three main components: Knowledge and Data Repository, Data Analysis and Results Visualization. The Knowledge and Data Repository component stores the data and knowledge needed in the knowledge discovery process. This process is implemented in the Data Analysis component, which allows the discovery of patterns or others relationships implicit in the analyzed geographic and non-geographic data. The discovered patterns can be visualized in a map using the Results Visualization component. These components are afterwards described.
The Knowledge and Data Repository component group three central databases:

1. A Geographic Database (GDB) constructed under the principles established by the European Committee for Normalization in the CEN TC 287 standard for Geographic Information. Following its recommendations was possible to implement a GDB in which the positional aspects of geographic data are provided by a geographic identifiers system. This system characterizes the administrative subdivisions of Portugal at the municipality and district level. Also includes a geographic gazetteer with the several geographic identifiers used and the concept hierarchies existing between them. The geographic identifiers system was integrated with a spatial schema allowing the definition of the direction, distance and topological spatial relations that exist between the adjacent regions of the municipality level.

2. A Spatial Knowledge Base (SKB) that stores all the qualitative rules needed in the inference of new spatial relations. The knowledge available in this database aggregates the composition table constructed, the set of used identifiers and the validity interval of them. This knowledge base is used in conjunction with the GDB in the inference of implicit spatial relations.

3. A non-Geographic Database (nGDB) that is integrated with the GDB and analyzed in the Data Analysis component. This procedure enables the discovery of implicit relationships that exist between the geographic and non-geographic data. The Data Analysis component is implemented through the knowledge discovery module, and is characterized by six main steps:

   1. Data Selection. This step allows the selection of the relevant non-geographic and geospatial data needed for the execution of a defined data-mining task. In this phase must be evaluated what is the minimal sub-set of data to be selected, the size of the sample needed and the period of time to be considered. The term geo-spatial is used to emphasize that geographic data has associated spatial data. Traditionally, geographic data is associated with a location, positioning an object or fact in space, while spatial data defines the characteristics of that localization, namely its geometry and topology.

2. Data Treatment. This phase is concerned with the cleaning of the selected data, allowing the corrupt data treatment and the definition of strategies for dealing with missing data fields.

3. Data Pre-Processing. This step allows the reduction of the sample set to be analyzed. Two tasks can here be affected: i) Reduction of the number of rows or, ii) Reduction of the number of columns. In the reduction of the number of rows, data can be generalized attending to the domain’s hierarchies or attributes with continuous values can be transformed into discreet values attending to the defined classes. The reduction of the number of columns attends to verify if any of the selected attributes can afterwards be omitted.

4. Geo-Spatial Information Processing. This step verifies if the geo-spatial information needed is available in the GDB. In many situations it implicitly exists, due to the properties of the spatial schema implemented. In this case, and to ensure that all geo-spatial knowledge is available for the data mining algorithms, these implicit relations are transformed into explicit relations through the inference rules stored in the SKB.

5. Data Mining. Several algorithms can be used for the execution of a given data mining task. In this step, the several available algorithms are evaluated in order to identify the most appropriate for the defined task. The selected one is applied to the relevant non-geographic and geo-spatial data, in order to find implicit relationships or other interesting patterns that exist between them.

6. Results Interpretation. The interpretation of the discovered patterns aims to evaluate their utility and importance to the application domain. In this step it may be realized that relevant attributes were ignored in the analysis, suggesting that the process should be repeated. The relevant discoveries can be stored in the database of patterns, allowing its subsequent use in further analyses (meta-rules construction) or its visualization in a map.

The Results Visualization component is responsible for the management of the discovered patterns and its visualization in a map. For that, PADRÃO uses a Geographic Information System (GIS), integrating the discovered patterns with the cartography of the analyzed region. This component aggregates two main databases:

1. The Patterns Database (PDB) that stores all relevant discoveries. In this database, each discovery is catalogued and associated with the set of rules that represents the discoveries made in a given data mining task.

2. A Cartographic Database (CDB) with the cartography of the region. It aggregates a set of points, lines and polygons with the geometry of the geographical objects.

THE KNOWLEDGE DISCOVERY PROCESS IN PADRÃO

The nGDB, of the Knowledge and Data Repository component, used in this application of PADRÃO is a Demographic Database (DDB) that stores the parish registers. This database collects attributes like birth date, birthplace, death date, death
place, occupation, number of descendants, number of marriages, etc. The several attributes related to places allow the integration of the DDB with the GDB, providing the geo-spatial data needed in the knowledge discovery module.

CONCLUSION
This paper presented the PADRÃO system, a system for knowledge discovery in spatial databases based on qualitative spatial reasoning. PADRÃO represents a new approach to this particular case of knowledge discovery, in which the positional aspects of geographic data are provided by a spatial reference, gave by a geographic identifier (in the described case, the municipalities' name). The PADRÃO’s geographic database and spatial knowledge base store the spatial semantic knowledge and the qualitative spatial reasoning principles needed for the inference of new spatial relations required in the knowledge discovery process. The analysis of a demographic database with PADRÃO allowed the discovery of implicit relationships that exist between the analyzed demographic and geo-spatial data.

The data analysis techniques described along this document constitutes a special case of Spatial Analysis and is useful in the Urban and Regional Research. This application domain analyses huge amounts of data that are related with some place on the Earth’s surface. PADRÃO has the capability to analyze databases of great dimension. These databases usually store geo-referenced data, which spatial component is included in the spatial analysis process of PADRÃO using qualitative spatial reasoning strategies.

FUTURE SCOPE
The PADRÃO’s geographic database and spatial knowledge base store the spatial semantic knowledge and the qualitative spatial reasoning principles needed for the inference of new spatial relations required in the knowledge discovery process. The analysis of a demographic database with PADRÃO allowed the discovery of implicit relationships that exist between the analyzed demographic and geo-spatial data.

REFERENCES