International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 9, September- 2014

Survey on Knowledge Observation With Spatiality Data Mining

Dr. S. Anitha Reddy
Sridevi Womens Engineering College
Hyderabad

Dr. P. Avinash
Sridevi Womens Engineering College,
Hyderabad.

***Abstract*—Huge amount of spatiality data is being collected in various applications like remote sensing, information systems, computer cryptotography, geographical information system (GIS), environmental assessment and planning, etc. Now the real challenge is holding attention in presently, and previously unknown information from this large database. This is what the objective of spatiality data mining. The real work is to extend the scope of data mining from logical and transactional database to spatiality database and apply it in the study of spatiality distribution. The paper summarizes the work that has been done so far in spatiality data mining from spatiality data generalization, mining spatiality association rule to spatiality data clustering**

Key Words: Data authentication, Data generalization, KDD, Spatiality Data Type, Spatial Data Mining, Raster Map, Vector Map…respectively

1. INTRODUCTION

The spatiality data mining in spatiality distribution of field specific database is an interdisciplinary research area that basically focuses ideas on knowledge discovery from heterogeneous format of spatiality database and homogeneous format of spatiality database. The study focuses *first* on the different format of spatiality data mining techniques and *secondly* a suitable technique to work on spatiality distribution of the database to mine knowledge from such database and get the information relating to spatiality database. The rapidly growing data creates the necessity of knowledge / information discovery from data which leads to promising emerging field, called the data mining or knowledge discovery from database (KDD). Spatiality Data Mining, Shekhar & Chawla 2003[1], Describes as a process of discovering previously unknown, but potentially useful patterns from spatiality database. The process of data mining could be the integration of many things including machine learning, database system, statistics, and information theory. There are many studies available of data mining in relational and transactional database [2, 3, 4, 5], the concept is in high demand to apply it in many other applicative area like spatiality database, temporal database, multimedia database, object-oriented database etc. Section 2 discusses various methods and research gap in between discovering interesting knowledge from spatiality data whereas section 3 discusses one of the applicative are such as applying spatiality data mining in spatiality database. Section 4 discusses the future direction of the research work.

2. SPATIALITY DATA MINING

Spatiality data are the data related to objects that occupy space. It contains topological and/or distance information and is often organized by spatiality indexing structures and accessed by spatiality access methods. The objects stored in spatiality database are the spatiality objects represented by spatiality data type and are having implicit relationship among them. The implicit relationship among the objects and the distinct feature of spatiality database poses challenge and bring opportunities for mining information from spatiality data [6].

Knowledge discovery from database refers to the extraction of implicit knowledge, spatiality relation, or other patterns not explicitly stored in spatiality database [7].

The work related to statistics [8,9,10,11], machine learning[12,13,14] and database systems[15,16] laid the foundation of knowledge discovery from database. Then after, with respect to spatiality database, the study related to computational geometry[5],spatiality data structure[17,18,19] and spatiality reasoning [20,21] paved the way for the study of spatiality data mining.

The statistical spatiality analysis [9,11] has been the most common approach for analyzing spatiality data. It handles very efficiently the numerical data which comes from the realistic model of spatiality phenomena. But the assumption of statistical independence among the spatiality distributed data causes problem as many of the spatiality data are in fact interrelated. It is because the spatiality Neighboring objects. At the same time the statistical approach cannot model non linear rules very well. Statistical methods also do not work well with incomplete or inconclusive data. Another problem related to statistical spatiality analysisis the expensive computation of the result. Tosupplement the work the machine learning techniques [12,14] and the spatiality database potential[22,23] was nicely utilized. Now to model the non linear rules out of the spatiality and non spatiality data the potential of soft computing can be used.

2.1 Spatiality Data Mining DM] Components

Following are some important attributes in the study of spatiality data mining:

Rules: With the help of combined approach of SDM techniques, various rules can be discovered such as spatiality association rule, spatiality characteristic rule, deviation and evolution rule, and discriminate rule etc.

Thematic Map : It is a map that shows a theme, which is a single spatiality distribution or a pattern, using a specific map type[2]. It presents the spatiality distribution of a single or a few attributes.
Spatiality classification is one of the techniques that analyze spatiality and non-spatiality attributes of the data objects to partition the data into a set of classes. These classes generate a map representing groups of related data objects. There are two ways to represent thematic maps: *raster map* and *vector map*. The raster image thematic maps
have pixels associated with the attribute values. In the vector form the spatiality objects are represented by its geometry i.e. boundary representation and thematic attributes.

Image Databases: These are special kind of spatiality databases which consists of images and pictures. They are stored in the form of grid array representing the image intensity in one or more spectral ranges.

2.2 SDM architecture, spatiality data structure

There are various architectures proposed for data mining. Some of the important architectures are J Han & Y. Fu's[24] architecture DBLEARN/DBMINER, M. Holsheimer and M. Kersten's [25] parallel architecture and C. J Mathu & Chan's[26] multi component architecture. These are the general data mining prototypes but they can be used or extended to handle spatiality data mining. Mathu's architecture is very general and has been used by other researchers in spatiality data mining, including M. Ester etc. [27]. The important spatiality operations like spatiality joins, map overlays, nearest neighbor queries are some important spatiality operators. Thus in order to work efficiently, the operators requires an efficient spatiality access method (SAM) and appropriate spatiality data structure. Spatiality data structure consists of points, lines, rectangles etc. and to build indices for these data , multidimensional trees have been proposed such as quad tree[46], k-d tree, R-trees, R* -tree etc.

2.3 Knowledge Discovery Methods.

The spatiality database consists of spatiality objects and non-spatiality description. The non-spatiality description of the spatiality object can be stored in the traditional relational database[22]. There are two different properties of spatiality data and they are geometric and topological.. The geometric properties could be spatiality location, area, perimeter etc. whereas opological properties can be adjacencies, inclusion etc. The figure below (Figure: 1) describes how the non-spatiality and spatiality attribute values about the states of India are mapped in a database.
The methods for discovering knowledge from spatiality database focuses on non-spatiality and/or spatiality properties of spatiality objects. Some important spatiality data mining algorithms are:

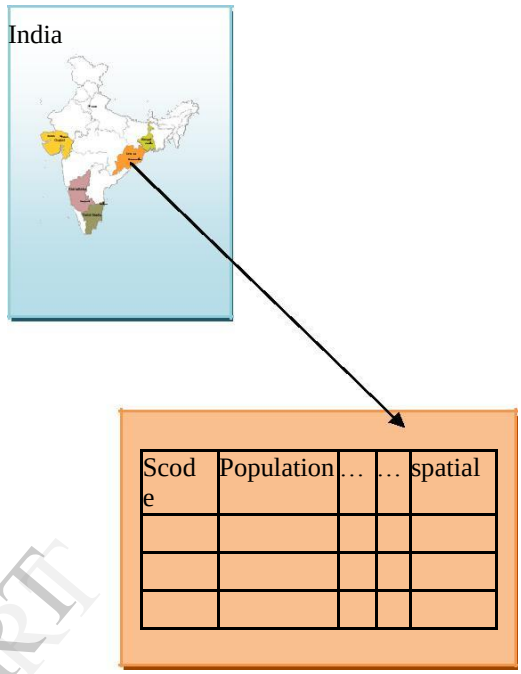


Table of data consisting of non-spatiality and spatiality attribute value.

Figure 1: Non-spatiality and spatiality attribute values of Indian states.

2.3.1. Generalization-based methods for mining spatiality characteristics and discriminate rules[4,6,28].

This is a widely used tuple-oriented technique in machine learning [13]. The method is often combined with generalization [14]. This approach cannot be used for large spatiality database because the algorithms are exponential in the number of examples and it does not handle noise and inconsistent data very well. It requires the existence of background knowledge in the form of concept hierarchies. There can be two kinds of concept hierarchies *non-spatiality* and *spatiality* and are given by the experts or as per the requirement of the analysis. Following (Figure:2) could be a concept hierarchy of epidemiology study.

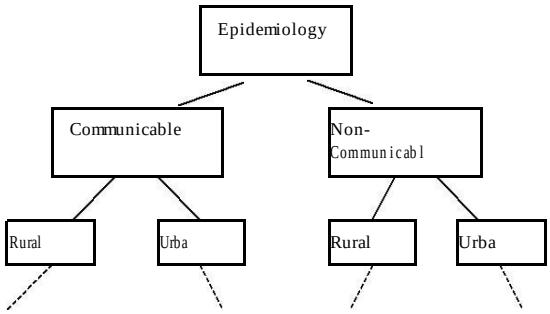
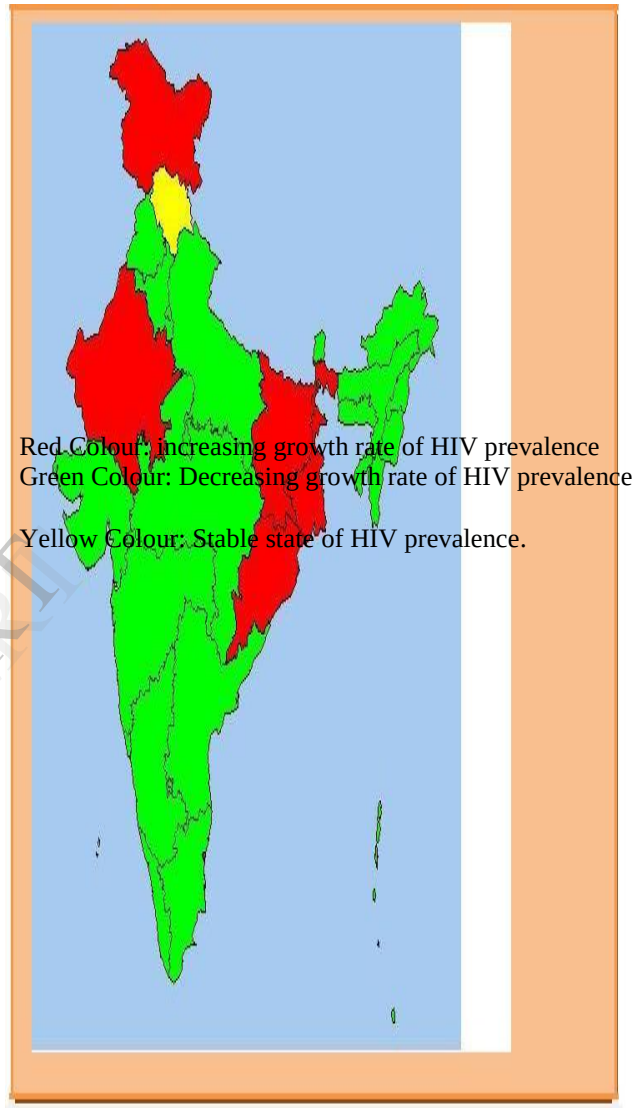


Figure 2: Concept Hierarchy of epidemiology

As we move upward in the concept hierarchy the information becomes more and more general. Similar concept hierarchy can be formed for spatiality data for example hierarchy related to region, state, district, village etc. W. Lu and J. Han[6] described two generalization based algorithms one *spatial-data-dominant* and another *non-spatial-data-dominant* generalization. In the first approach the generalization of the spatiality objects continues until the spatial in generalization threshold is reached i.e. the no of region is not bigger than a threshold value. When the spatial-oriented induction process is complete, non-spatiality data are retrieved and analyzed for each of the spatiality object using the attribute oriented induction technique. The result of the query in spatial-data-dominant algorithm could be in the form of the follows map (Figure 3). In the second approach the algorithm performs attribute oriented induction on the non-spatiality attributes, generalizing them to a higher concept level. The generalization threshold determines whether to continue or stop the generalization process. In this process the pointer to the spatiality objects are collected as a set and put with the generalized non-spatiality data. Finally the neighboring area with the same generalized attributes are merged together based on the spatiality function of adjacency (*adjacent_to*).



Red Colour: increasing growth rate of HIV prevalence
Green Colour: Decreasing growth rate of HIV prevalence

Yellow Colour: Stable state of HIV prevalence.

Figure 3 : The three categories of states having different

growth rate of HIV. Output spatiality data dominant.
For example the adjacent area having no of malaria epidemiology count ,both in male and female, more than 5% of population are merged together forming a *high-prevalence* cluster of malaria epidemiology. Similarly *low-prevalenc*e and *no-prevalence* clusters can be identified. The result of the query can be shown in the form of a map. An example of such a map (the dotted region inside the India map) is shown below Figure: 4

.

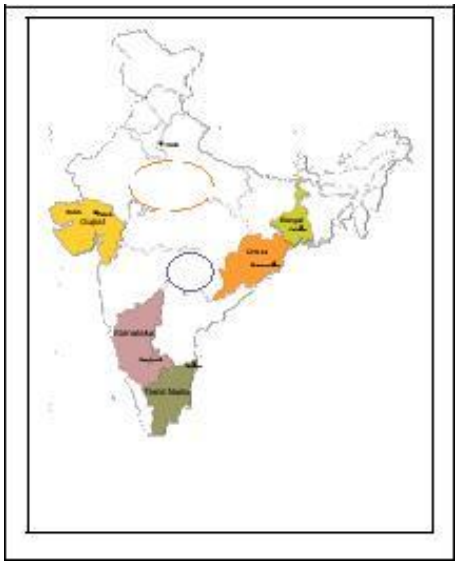


Figure 4 : Output non-spatiality data dominant

In the above described generalization based algorithms, the concept hierarchy is generated automatically. However there are cases when such hierarchy is not present *a priory*. At the same timethe mined characteristic rules is going to be much dependent on the given concept hierarchy(ies). Such concept hierarchy is given by the experts and therefore the above mentioned approaches falls under the category of *supervised classification or supervised knowledge discovery methods*.

On the other hand we have some *unsupervised techniques* of knowledge discovery methods. The method of clustering is one such approach. The conventional clustering algorithms like PAM (Partition Around Medoids) or CLARA (Clustering LARge Applications) [10], are not appropriate from computational complexity point of view. The difference between these two algorithms is that CLARA algorithm is based on sampling. CLARA can deal with large data set than PAM. Both PAM and CLARA were developed by Kaufman and Rousseeuw [10]. In PAM the cost of single iteration is $O(k(n-k)^2)$. Here n is the no of objects and k is no. of cluster. In CLARA the complexity of each iteration is $O(kS^2+k(n-k))$. Here S is the size of the sample.

Then CLARANS was developed for cluster analysis and it outperformed the previous two algorithms. This algorithm was proposed by Ng and Han [28] which tries to mix both PAM and CLARA by searching only the subset of data set and it does not confirm itself to any sample at any given time. Experimentally it has been shown that CLARANS is more efficient than PAM and CLARA. Its every iteration computational complexity is linearly proportional to number of objects [27]. Some of the drawback of CLARANS has been pointed out by Ester, Kriegel, and Xu[27]. It assumes that the objects to be clustered are stored in main memory. For a large database it is not possible and hence a disk based method would be required. This method has been shorted out by integrating CLARANS with efficient spatiality access methods, like R*-tree. But the construction of R*-tree is time consuming. Zhang, Ramakrishnan and Livny[29] presented another method BIRCH (Balanced Iterative Reducing and Clustering) for clustering of large set of points. The method is incremental one with possibility of adjustment of memory requirements to the size of memory that is available. It uses the concept called *Clustering Feature* and *CF tree*.

2.3.2. Two-step spatiality computation technique for mining spatiality association rules [34]

To minimize the number of costly spatiality computation the two-step spatiality computation technique for optimization during the search for association was introduced. Spatiality association rule is a rule that associate one or more spatiality object with other spatiality objects. Agarwal, Imielinski and Swami [30] introduced the concept of *association rules* in the study of mining large transaction database. Later Koperski and Han[7] extended this concept to spatiality database. In order to discover the useful rule the concept of *minimum support* and *minimum confidence* are used. A strong rule is a rule having large support and large confidence.

2.3.3. Aggregate proximity technique for finding characteristics of spatiality clusters [31].

An aggregate proximity is the measure of closeness of the set of points in the cluster to a feature as opposed to the distance between a cluster boundary and the boundary of a feature. Related to a cluster it would be more interesting result to know why the clusters are there. The question that would more suitable answer about the cluster is that "what are the characteristics of the clusters in terms of the feature that are close to them". For example the statement like

85% of the houses in a cluster is close to the feature F (e.g. infected by infectious disease cholera) would be more informative and interesting than statement like *one house is close to the feature F*.

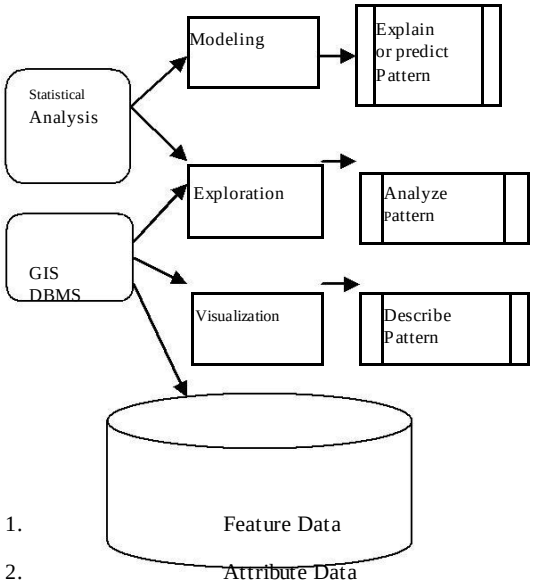
3. SPATIALITY EPIDEMIOLOGY- AN APPLICATIVE AREA

Elliott and Wartenberg [37] described "Spatiality epidemiology is the description and analysis of geographic, or spatial, variations in disease with respect to demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors". The spread of infectious disease is closely associated with the concepts of spatiality and spatio-temporal proximity, as individuals who are linked in a spatiality and temporal sense are at a high risk of getting infected [38]. Proximity to environmental risk factors is therefore important. Thus knowledge of spatiality and temporal variations of disease and characterizing its spatiality structure is essential for the epidemiologist to understand better the population's interactions with its environment [39].

Spatiality epidemiology analysis comprises of wide range of methods. Now it is a big challenge to determine which one to use[38]. The figure below (Figure: 5) is a diagrammatic representation of a spatiality analysis framework taken from Pfeiffer [38] adopted from Bailey and Bailey & Gatrell[4 In the above diagram Pfeiffer identified the following four active groups of the framework:

3.1.Data

Data is the basic need of epidemiological analysis which is conducted for description of spatiality patterns, identification of disease cluster, and explanation or 0]. prediction of disease risk [38]. Geographic data system includes georeferenced feature data and attributes, be they point and area. These data are obtained by taking field survey, remotely sensed imagery or use ofexisting data generated either by government organizations or those closely linked to government such as cadastral, meteorological or national censusstatistics and health organizations.



Source: Pfeiffer [38], Bailey and Gatrell [40]

Figure 5 : Conceptual framework of spatiality epidemiological data analysis

3.2. GIS and DBMS

Management of the data is performed using GIS and database management system(DBMS), and is of relevance throughout the various phases of spatiality data analysis. GIS provide a platform for managing these data, computing spatiality relationship such as proximity to source of infection, connectivity and directional relationships between spatiality units, and visualizing both the row data and results from spatiality analysis within a cartographic context [38].

3.3. Visualization and exploration

It covers technique that focus solely on examining the spatiality dimension of the data. Visualization tools are used resulting in maps that describe spatiality patterns and which are useful for both stimulating more complex analysis and for communicating the results of such analysis. Exploration of spatiality data involves the use of statistical methods to determine whether observed patterns are random in space. However there is some overlap

between visualization and exploration, since meaningful visual presentation will require the use of quantitative analytical methods [41].

3.4. Modeling

Modeling introduces the concept of cause-effect relationships using both spatiality and non-spatiality data sources to explain or predict spatiality patterns [38].

4. FUTURE DIRECTION

Data mining is a young field of study started during late 1980s. Spatiality data mining is an even younger. The traditional data mining researchers extended their study to work on spatiality data mining. Many spatiality data mining methods assume the presence of extended relational model for spatiality database. Some of the future directions of spatiality data mining are enlisted below.

***Data Mining in Spatiality Object-Oriented Databases*:** Many researchers have pointed out that OO database may be a better choice for handling spatiality data rather than traditional relational or extended relational models[32,33].

***Mining Under Uncertainty*:** The use of evidential reasoning [34] can be explored in the mining process for the databases where uncertainty modeling has to be done. Bell, Anand and Shapcott [35] has explained that evidential theory can model uncertainty better than traditional probabilistic models, like Bayesian methods. Fuzzy sets approach was applied to spatiality reasoning[20,36] and it can be extended to spatiality data mining.

Mining Spatiality Data Deviations and Evolution Rules: It is a more challenging and applicative work in spatiality data mining. The work would be related to spatio-temporal databases to study data deviation and evolution rules. Foe example we can find spatiality characteristic evolution rules which summarizes the general characteristics of the changing data. During the mining process we can discover the region having particular epidemiology growth rate more than the country's average
growth rate. Similarly one can make a comparison of the areas where certain epidemiology increased last year with the area where it has decreased.

These rules may be used by the government and policy makers in formulating policies and plan to curb the problem.

Multidimensional Data Analysis and Rule Visualization: Discovering rule from multidimensional data (non-spatiality and spatial) source is a challenge for the researchers. Multidimensional data analysis and visualization has been studied [42], but multidimensional rule visualization is still an immature area.

5. CONCLUSION

We have explained that spatiality data mining is a promising field of research with wide application in GIS, medical and environmental data analysis etc. We surveyed the existing methods of spatiality data mining and presented their strength and weaknesses. We have outlined one of the applicative area i.e. spatiality data mining of epidemiology database which is of great importance for the society and policy makers and we hope to give some novel and useful output from our further exploration of this field.

6. BIBLIOGRAPHY

[1] G. Say, D. Wheeler, Statistical Techniques in Geographical Analysis. London, David Fulton, 1994.
[2] R. Agarwal and R. Srikant. *Fast Algorithm for mining association rules*. In Proc. 1994 Int. Conf. VLDB,pp.487-499, Santiago,Chile, Sept. 1994.
[3] U. M. Fayyad, G Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.
[4] J. Han, Y. Cai, and N. Cercone.*Data-Driven Discovery of Quantitative Rules in Relational Databases.* IEEE Trans. Knowledge and Data Eng., 5:29-40, 1993.
[5] G. Piatetsky-Shapiro and W. J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI/MIT Press, Menlo Park, CA, 1991.
[6] W. Lu, J. Han, and B.C. Ooi. Discovery of General Knowledge in Large Spatiality Databases. In Proc. For East Workshop on Geographic Information Systems pp. 275-289, Singapore, June 1993.
[7] K. Koperski and J. Han. Discovery of Spatiality Association Rules in Geographic Information Databases. In Proc. 4th Int'l Symp. On large spatiality Databases(SSD '95), pp. 47-66, Portland, Maine, August 1995.
[8] D. K. Y. Chiu, A. K. C. Wong, and B Cheung. A Statistical technique for Extracting Classificatory Knowledge from Databases. In Piatetsky-Shapiro and Frawley [43], pp 125-141.
[9] S. Fotheringham and P. Rogerson. *Spatiality Analysis and GIS*, Taylor and Fransis, 1994.
[10] L. Kaufman and P J Rousseeuw. Finding groups in Data: an introduction to Cluster Analysis. John Wiley & Sons, 1990.
[11] S.Shekhar and S.Chawla. Spatiality Databases: A Tour. Pretice Hall (ISBN 0-7484-0064-6), 2003.
[11] D. Fisher, Improving Interface through Conceptual Clustering. In Proc. 1987 AAAI Conf., pp. 461-465, Seattle, Washington, July 1987.
[12] R. S. Michalski, J. M. Carbonnel, and T. M. Mitchell, editors. Machine Learning: An Artificial Intelligence Approach. Morgan Kaufmann, Los Altos, CA, 1983.
[13] T. M. Mitchell. Generalization and earch. In Artificial Intelligence, 18:203-226, 1982.
[14] M. Stonebraker. *Reading in Database System.* Morgan Kaufmann, 1988.
[15] M. Stonebraker. *Reading in Database System*. 2ed.. Morgan Kaufmann, 1993.
[16] R. H. Guting. *An Introduction to Spatiality Database System*. In VLDB Journal, 3(4):357-400, October 1994.
[17] R. Guttman. A dynamic index structure for spatiality searching. In Proc. ACM SIGMOD Int. Conf. on Management of Data. Bostan, MA, 1984, pp. 47-57.
[18] H. Samet. *The Design and Analysis of Spatiality Data Structure*. Addison-Wesley, 1990.
[19] S. Dutta. Qualitative Spatiality Reasoning: *A Semi Quantitative Approach Using Fuzzy Logic*. In Proc. 1st Symp. SSD'89, pp. 345-364, Santa Barbara, CA, July 1989.
[20] M. J. Egenhofer. *Reasoning about Binary Topological Relation*. In Proc. 2nd Symp. SSD'91, pp. 143-160, Zurich, Switzerland, August 1991.

[22] W. G. Aref and H. Samet . *Extending DBMS with Spatiality operation*. In Proc 2nd Symp. SSD'91, pp. 299-318, Zurich, Switzerland, August 1991.

[23] W. G. Aref and H. Samet. *Optimization Strategies for Spatiality Query Processing*. In Proc. 17th Int. Conf. VLDB, pp. 81-90, Barcelona, Spain, Sept. 1991.

[24] J. Han, and Y. Fu. *Exploration of the power of Attribute-Oriented Induction in Data Mining*. In[16]

[25] M. Holsheimer and M. Kersten. Architectural Support for Data Mining. In CWI Technical Report CS-R9429, Amsterdam, The Netherlands, 1994.

[26] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro. *Systems for Knowledge Discovery in Databases.* In IEEE Trans. Knowledge and Data Engineering, 5:903-913,1993.

[27] M. Ester, H.–P. Kriegel, and X. Xu. *Knowledge Discovery in Large Spatiality Databases: Focusing Techniques for Efficient Class Identification*. In Proc. 4th Int. Symp. On Large Spatiality Databases (SSD'95),pp.67-82, Portland, Maine, August 1995.

[28] R. Ng and J. Han. Efficient and effective clustering method for spatiality data mining. In Proc. 1994 Int. Conf. Very Large Databases, pp. 144-155, Santiago, Chile, September 1994.

[29] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an Efficient Data Clustering Method for Very Large Databases. In Proc. 1996 ACM-SIGMOD Int.Conf.Management of Data, Montreal,Canada, June 1996.

[30] R. Agarwal, T. Imielinski, and A. Swami. *Mining Association Rules Between Sets of Items in Large Databases.* In Proc. 1993 ACM-SIGMOD Int. Conf. management f Data, pp. 207-216, Washington, D.C., May 1993.

[31] E. Knorr and R. T. Ng. *Applying Computational Geometry Concepts to Discovering Spatiality Aggregate, Proximity Relationships*.InTechnical Report,University of British Columbia,1995.

[32] L. Mohan and R. L. Kashyap. *An Object-Oriented Knowledge Representation for Spatiality Information.* In IEEE Transaction on Software Engineering, 5:675-681, May 1988.

[33] J. Han, S. Nishio, and H. Kawano. Knowledge Discovery in Object-Oriented and Active Databases. In F. Fuchi and T. Yokoi(eds), Knowledge Building and Knowledge Sharing, Ohmsha/IOS Press, pp. 221-230, 1994.

[34] J. Guan and D. Bell. Evidence Theory and its Applications, vol. 1. North-Holland, 1991.

[35] D. A. Bell, S. S. Anand, and C. M. Shapcott. *Database Mining in Spatiality Databases.* International Workshop on Spati-Temporal Databases,1994.

[36] S. Dutta. Topological Constraints:

[37] *A Representational Framework for approximate Spatiality and Temporal Reasoning*. In Proc. 2nd Symp. SSD'91, pp.161-182, Zurich, Switzerland,August 1991.

[37] P. Elliott and D. Wartenberg. Spatiality epidemiology: current approaches and future challenges. Environmental health perspectives, 112(9):998, 2004.

[38] Dirk Pfeiffer. Spatiality analysis in epidemiology. Oxford University Press, GB, 2008.

[39] Frank B. Osei. *Spatiality statistics of epidemic data* : the case of cholera epidemiology in Ghana. PhD thesis, 2010.

[40] T.C. Bailey and A.C. Gatrell. Interactive spatiality data analysis. Longman Scientific & Technical Essex, 1995.

[41] A. Maroko, J.A. Maantay, and K. Grady. Using geovisualization and geospatiality analysis to explore respiratory disease and environmental health justice in New York city. Geospatiality Analysis of Environmental Health, pages 39–66, 2011.

[42] D. Keim, H. P. Kriegel, and T. Seidl. Supporting Data Mining of Large Database by Visual Feedback Queries In Proc. 10th of Int. Conf. on Data Engineering, Houston, TX, pp. 302-313, Feb. 1994.