ISSN: 2321-7782 (Online)
Volume 3, Issue 11, November 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Data Integration with Spatial Data Mining and Security Model in Cloud Computing

CH V V Narasimha Raju
Assistant Professor,
Marri Laxman Reddy Institute of Technology & Management,
Hyderabad, Telangana

Abstract: Cloud computing has is a popular design in managing world to back up large volumetric details using cluster of commodity computer systems. In spatial data mining, we have to deal with uncertainties in data and mining process. The nature of uncertainties can be, for example, fuzziness and randomness. With the cloud computing time arrival, spatial data storage and management technology based on cloud computing are getting more extensive attention and application. But under the cloud environment, how to ensure that the data stored in the cloud security will be a serious challenge. This paper introduces the meaning characteristics and development present situation of cloud computing, and gives the analysis about the advantage of using cloud computing technology to spatial data management. In cloud model context, spatial data preprocessing pays more attention to data cleaning, transform between qualitative concepts and quantitative data, data reduction and data discretization. Spatial knowledge is represented with qualitative concepts from large amounts of data and also the cloud model. The effectiveness and efficiency of the proposed approach was evaluated by using an analytical cost model and an extensive experimental study on a geographic database.

Keywords: cloud computing; spatial data; spatial knowledge.

I. INTRODUCTION

The computerization of many business and government transactions and the advances in scientific data collection tools provide us with a huge and continuously increasing amount of data. This explosive growth of databases has far outpaced the human ability to interpret this data, creating an urgent need for new techniques and tools that support the human in transforming the data into use-ful information and knowledge. *Knowledge discovery in databases* (*KDD*) has been defined as the non-trivial process of discovering valid, novel, and potentially useful, and ultimately understandable patterns from data [FPS 96]. The process of KDD is interactive and iterative, involving several steps such as the following ones:

Selection: selecting a subset of all attributes and a subset of all data from which the knowledge should be discovered.

Data reduction: using dimensionality reduction or transformation techniques to reduce the effective number of attributes to be considered.

Data mining: the application of appropriate algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

***Evaluation*:** interpreting and evaluating the discovered patterns with respect to their usefulness in the given application.

1.1.The Summary of Cloud Computing

So far the cloud computing is not accepted exact definition. Along with continuous research in cloud computing, the definition of which is in the dynamic change.Baidu encyclopedia [1] gives the definition of cloud computing at present: it is a web based method, the sharing of software and hardware resources and information can provide the need for computer and

other equipment in this way. The whole operation mode is just like the grid. Cloud Computing-China [2] defined "cloud" as: Cloud Computing is the development of Distributed Computing, Parallel Computing and Grid Computing, or is the commercial realization of these scientific concepts.

1.1.1.Characteristics of Cloud computing

Cloud computing is accepted by the enterprise and the technology IT passionately because of its enormous advantages. Its specific characteristics can be summed up in the following aspects

- ***The large scale of server:*** "Cloud" has a certain scale, Google who is the earliest practicer of cloud computing, already has millions of servers. And the Amazon, IBM, Microsoft, Yahoo, have more than 50 ten thousand servers in the "cloud". So "cloud" can give users super computing power.
- ***Resources virtualization:.*** Cloud computing can provide service in different geographical locations and all sorts of terminals. The requested resource is a dynamic and invisible. When the applications run in "cloud" somewhere, the users need not consider what the exact location it is. Only there is a laptop or a mobile phone, you can realize our need through network service, even super computer such task.
- ***High reliability:***. "Cloud" takes advantage of the multiple fault-tolerant measures to ensure the high reliability of the service.
- ***Strong generality****:* In the support of "cloud", a variety of applications can be constructed. Computing clouds aim at the changes of the application. Yet different applications operation can be supported by the same "clouds", which can save a lot of network resources.
- ***Expansibility:*** Although application and user scale are in the unceasing growth, "cloud" scale can also use the dynamic expansion and satisfy these needs.
- ***Service according to the needs:***The cloud is just like running water, electricity, coal gas that billing. Therefore, users can purchase what they need. And it's also more convenient for enterprise and business network resources management.
- ***Low price:*** "Cloud" can be made by very cheap node, so "cloud" has no burden of more and more high data center management cost. When users enjoy the low cost of "cloud", the traditional system of resources utilization has improved at the same time.
- ***High popular degree:*** Based on the network platform and computer technologies, more and more users start to understand and enjoy the service from cloud computing, which is convenience in daily life with all kinds of practical application.

II. The Advantage of Spatial Data under the Cloud Environment

The Advantage of Spatial Data under the Cloud Environment. One of the most famous data management technology based on cloud computing is Google's Big Table [4] data management system. At the same time, Hadoop team is developing a similar open source data management module like Big Table. The way to manage massive spatial data by cloud computing has advantages in the following aspects:

- ***Rich Data Resource:*** Resource sharing is the important target for us in information era, especially when some spatial data source is much scarce. The arrival of cloud computing brings the good solution to this problem. Through the computer network, we can get any spatial data, which stored on the network and satisfy the work and study requirements on-demand at anytime and anywhere. Cloud computing provides a very good development environment and application environment.

» ***Low Cost:*** Based on cloud computing the low cost distributed parallel computing environment is implemented, so the cost of data processing greatly reduces, especially for complex topology relation in GIS, space and time data.

» ***Updates Timely:*** Usually, as the operator of the GIS spatial data processing, the spatial data that we can get may be collected probably a few days, months and even years ago. So the update of the data becomes the main concern by a lot of enterprise and the users. The spatial data based on cloud computing is able to update timely. In this way, the information which is reflected by the GIS spatial data and analysis can play a greater role.

» ***Shields the Ground Floor:*** In the parallelism conditions, cloud computing will be able to use the original equipment to improve the large-scale data processing power and speed, not only ensure the fault tolerance ,but also increases nodes.

III. ALGORITHMS FOR SPATIAL DATA MINING

To support our claim that the expressivity of our spatial data mining is adequate, we demonstrate how typical spatial data mining algorithms can be integrated with spatial data knowledge as follows.

3.1 Spatial Clustering

Clustering is the task of grouping the objects of a database into meaningful subclasses (that is, clusters) so that the members of a cluster are as similar as possible whereas the members of differ-ent clusters differ as much as possible from each other. Applications of clustering in spatial data-bases are, e.g., the detection of seismic faults by grouping the entries of an earthquake catalog or the creation of thematic maps in geographic information systems by clustering features spaces.

Different types of spatial clustering algorithms have been proposed, e.g. *k*-medoid clustering algorithms such as CLARANS. This is an example of a *global* clustering algorithm (where a change of a single database object may influence all clusters) which cannot make use of our database primitives in a natural way. On the other hand, the basic idea of a *single scan algorithm* is to group neighboring objects of the database into clusters based on a *local* cluster condition per-forming only one scan through the database. Single scan clustering algorithms are efficient if the retrieval of the neighborhood of an object can be efficiently performed by the SDBS. Note that lo-cal cluster conditions are well supported by our database primitives, in particular by the neighbors operation on an appropriate neighborhood graph. The algorithmic schema of single scan clustering is depicted below.

Different cluster conditions yield different notions of a cluster and different clustering algorithms'. For example, *GDBSCAN* (*Generalized Density Based Spatial Clustering of Applications with Noise*) relies on a density-based notion of clusters. The key idea of a density-based cluster is that for each point of a cluster its *Eps*-neighborhood for some given *Eps* > 0 has to contain at least a minimum number of points, i.e. the "density" in the *Eps*-neighborhood of points has to exceed some threshold. This idea of "density-based clusters" can be generalized in two important ways. First, any notion of a neighborhood can be used instead of an *Eps*-neighborhood if the definition of the neighborhood is based on a binary predicate which is symmetric and reflexive. Second, instead of simply counting the objects in a neighborhood of an object other measures to

SingleScanClustering(Database db; NRelation rel) set *Graph* to create_NGraph(db,rel); initialize a set *CurrentObjects* as empty;

for each node *O* in *g* **do**

if *O* is not yet member of some cluster **then** create a new cluster *C*;

insert *O* into *CurrentObjects*; **while** *CurrentObjects* **not** empty **do**

remove the first element of *CurrentObjects* as *O*;

set *Neighbors* to neighbors(*Graph*, *O*, TRUE); **if** *Neighbors* satisfy the cluster condition **do**

add *O* to cluster *C*;

add *Neighbors* to *CurrentObjects*;

end SingleScanClustering;

Algorithm1: Schema of single scan clustering algorithms

define the "cardinality" of that neighborhood can be used as well. Whereas a distance-based neighborhood is a natural notion of a neighborhood for point objects, it may be more appropriate to use topological relations such as *intersects* or *meets* to cluster spatially extended objects such as a set of polygons of largely differing sizes. See for a detailed discussion of suitable neighborhood relations for different applications.

3.2 Spatial Characterization

The task of *characterization* is to find a compact description for a selected subset of the data-base. In this section, we discuss the task of characterization in the context of spatial databases and review two relevant methods.

Extending the general concept of association rules, introduces *spatial association rules* which describe associations between objects based on spatial neighborhood relations. For instance, a user may want to discover the spatial associations of towns in British Columbia with roads, waters, mines or boundaries having some specified support and confidence. Then, the following spatial association rule may be discovered:

$\forall X \in DB \; \exists Y \in DB$: is-a(X,town) $\rightarrow$ close-to(X,Y) $\wedge$ is-a(Y,water) (80%)

This rule states that 80% of the selected towns are close to water, i.e. the rule characterizes towns in British Columbia as generally being close to some lake, river etc.

The algorithm presented in to find spatial association rules consists of 5 steps. Step 2 (coarse spatial computation) and step 4 (refined spatial computation) involve spatial aspects of the objects and are briefly examined in the following. Step 2 computes spatial joins of the object type to be characterized (such as town) with each of the other specified object types (such as water, road, boundary or mine) using a neighborhood relation (such as close-to). For each of the candidates obtained from step 2 (and which passed an additional filter-step 3), the exact spatial relation, for ex-ample *overlap*, is determined in step 4. Finally, a relation such as the one depicted in figure 3 results which is the input for the final step of rule generation.

Town	Water	Road	Boundary
Saanich	<meet, J.FucaStrait>	<overlap,highway1>, <close-to,highway17>	<close-to,US>
PrinceGeorge		<overlap, highway97>	
Petincton	<meet,OkanaganLake>	<overlap, highway97>	<close-to,US>
. . .	. . .	. . .	. . .

Figure 1. Input for the step of rule generation.

[EFKS 98] introduces the following definition of spatial characterization with respect to a data-base and a set of target objects which is a subset of the given database. A *spatial characterization* is a description of the spatial and non-spatial properties which are typical for the target objects but not for the whole database. The relative frequencies of the non-spatial attribute values and the rel-ative frequencies of the different object types are used as the interesting properties. For instance, different object types in a geographic database are communities, mountains, lakes, highways, rail-roads etc. To obtain a *spatial* characterization, not only the properties of the target objects, but also the properties of their neighbors (up to a given maximum number of edges in the relevant neigh-borhood graph) are considered.

A spatial characterization rule of the form $target \Rightarrow p_1\ (n_1, freq\text{-}fac_1) \wedge \ldots \wedge p_k\ (n_k, freq\text{-}fac_k)$ means that for the set of all targets extended by n_i neighbors, the property p_i is $freq\text{-}fac_i$ times more (or less) frequent than in the database. The characterization algorithm usually starts with a small set of target objects, selected for instance by a condition on some non-spatial attribute(s) such as "rate of retired people = HIGH" (see figure 2, left). Then, the algorithms expands regions around the target objects, simultaneously selecting those attributes of the regions for which the distribution of values differs significantly from the distribution in the whole database (figure 2, right).

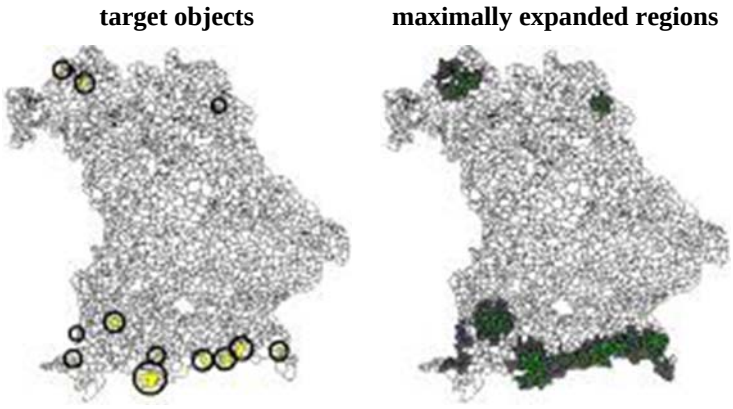

Figure 2. Characterizing wrt. High rate of retired people

In the last step of the algorithm, the following characterization rule is generated describing the target regions. Note that this rule lists not only some non-spatial attributes but also the neighbor-hood of mountains (after three extensions) as significant for the characterization of the target re-gions:

community has high rate of retired people ⇒
apartments per building = very low (0, 9.1) ∧
rate of foreigners = very low (0, 8.9) ∧
rate of academics = medium (0, 6.3) ∧
average size of enterprises = very low (0, 5.8) ∧
object type = mountain (3, 4.1)

Obviously, this algorithm is well suited for support by the proposed database primitives.

3.3 Spatial Classification

The task of *classification* is to assign an object to a class from a given set of classes based on the attribute values of this object. In *spatial classification* the attribute values of neighboring ob-jects are also considered.

The algorithm presented in [KHS 98] works as follows: The relevant attributes are extracted by comparing the attribute values of the target objects with the attribute values of their nearest neigh-bors. The determination of relevant attributes is based on the concepts of the *nearest hit* (the nearest neighbor belonging to the same class) and the *nearest miss* (the nearest neighbor belonging to a different class). In the construction of the decision tree, the neighbors of target objects are not con-sidered individually. Instead, so-called *buffers* are created around the target objects and the non-spatial attribute values are aggregated over all objects contained in the buffer. For instance, in the case of shopping malls a buffer may represent the area where its customers live or work. The size of the buffer yielding the maximum information gain is chosen and this size is applied to compute the aggregates for all relevant attributes. Figure 3 depicts an example of a spatial decision tree classifying stores as having a high or low profit.

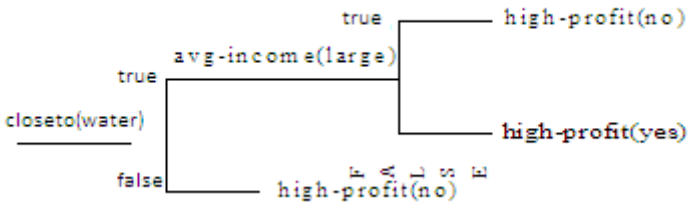

Figure 3. Spatial decision tree [KHS 98]

Whereas the nearest neighbor cannot be directly expressed by our neighborhood relations, it would be possible to extend our set of neighborhood relation accordingly. The proposed database primitives are, however, sufficient to express the creation of buffers for spatial classification.

3.4 Spatial Trend Detection

A spatial trend has been defined as a regular change of one or more non-spatial attributes when moving away from a given start object o [EFKS 98]. Neighborhood paths starting from o are used to model the movement and a regression analysis is performed on the respective attribute values for the objects of a neighborhood path to describe the regularity of change. For the regression, the distance from o is the independent variable and the difference of the attribute values are the depen-dent variable(s) for the regression. The correlation of the observed attribute values with the values predicted by the regression function yields a measure of confidence for the discovered trend.

Global as well as local trends are possible. The existence of a global trend for a start object o indicates that if considering all objects on all paths starting from o the values for the specified at-tribute(s) in general tend to increase (decrease) with increasing distance. Figure 4 (left) depicts the result of algorithm global-trend for the attribute “average rent” and the city of Regensburg as a start object. Algorithm local-trends detect single paths starting from an object o and having a certain trend. The paths starting from o may show different pattern of change, e.g., some trends may be positive while the others may be negative. Figure 4 (right) illustrates this case for the attribute “average rent” and the city of Regensburg as a start object.

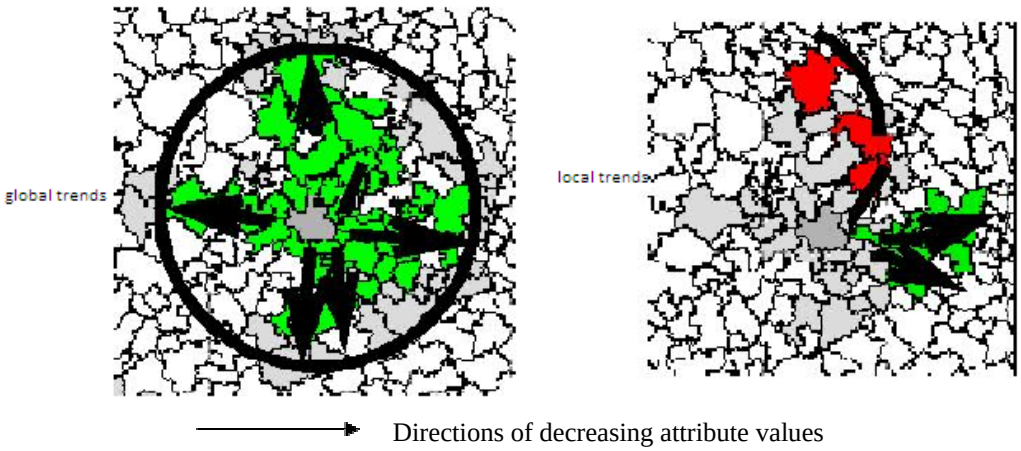

Figure 4. Spatial trends of the“average rent” starting from the city of Regensburg

IV. SPATIAL DATA SECURITY MODEL AND TECHNOLOGIES

For many of the spatial data security issues, many experts at home and abroad have made a lot of experiments and researches. Relational database model which achieves space safety management is the effective method.The model mainly has the information collection and access validation two parts. Information collecting module is used to collect user identity information, and will transfer the effective information to access validation module at the same time. After access validation module receives information, it ensures user identity uniqueness through the database of cloud.

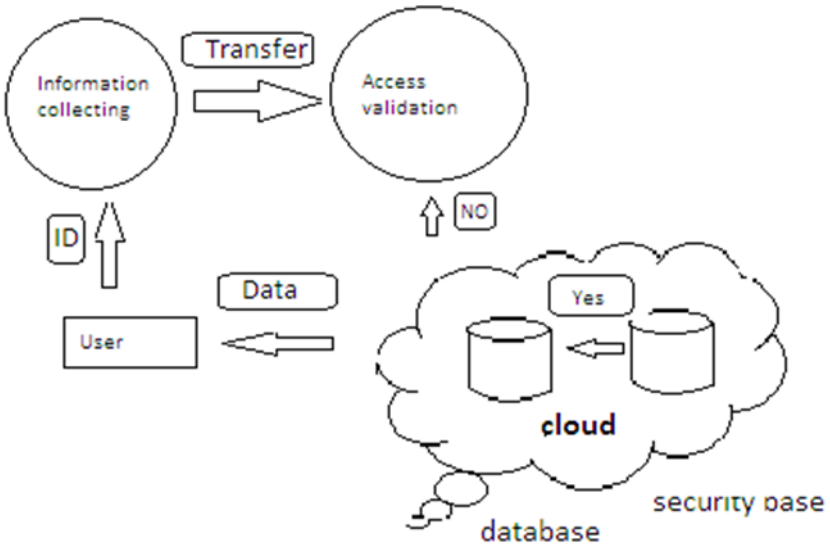

Fig. 1 Spatial Data Security Model

4.1.Spatial Data Security Technologies.

Here are some main technologies for spatial data security.

» ***Save the spatial data for encryption***: Encryption technology can be used in the data encryption, only the correct password can be declassified. Encryption can protect your spatial data, including the data sent to the data center.

» ***The E -mail encryption****:* E-mail can visit your inbox by the peeper style. In order to ensure that the E-mail security, you can use Hushmaii or Mutemail to encrypt all E-mails automatically.

» ***Use credit good service****:* The reputable service is a good choice for spatial data users. They are unlikely to take risks with their famous brand. So the data won't be allowed leak occurs, also won't be shared with others.

» ***Business model in use****:* Charge of Internet application service is safer than the free service. When you are choosing spatial data storage environment, paying storage is your first consider.

» ***Reading privacy statement****:* When you store the data in the cloud computing environment, you must be sure to read the privacy statement. Because there are many leaks in privacy policies about Internet application, so some key data must be shared in some cases with your permission. So you can determine what spatial data should be stored in the cloud and what data should be stored in your computer

V. CONCLUSION

We showed that spatial data mining algorithms such as spatial clustering, characterization, classification and trend detection are well supported by the proposed operations. Although cloud computing in spatial data storage and management, which is favor to many IT enterprise and user, has a lot of advantages. Generally speaking, cloud computing technology is still at the beginning stage. How to use cloud computing and make it gradually regularized, commercial and popular, needs a long process. The future is bright while the road ahead is tortuous; we have a long way to go!

References

1. Data mining concepts and techniques J Han, M Kamber, J Pei - 2011 - books.google.com
2. ZHOU Yan, SANG Shu-juan,Data Mining Technology Based on Cloud Computing, Computing knowledge and Technology, Vol.6, No.34, December 2010, pp.9681-9683.
3. Fay Chang, Jeffrey Dean, Sanjay Ghemawat et al. BigTable: a distributed storage system for structured data [A]. Operating Systems Design and Implementation, 2006.

4. [AIS 93] Agrawal R., Imielinski T., Swami A.: “Database Mining: A Performance Perspective” , IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, 1993, pp. 914-925.

5. [BF 91] Bill, Fritsch: “ Fundamentals of Geographical Information Systems: Hardware, Software and Data” (in German), Wichmann Publishing, Heidelberg, Germany, 1991.

6. [Ege 91] Egenhofer M. J.: “Reasoning about Binary Topological Relations” , Proc. 2nd Int. Symp. on Large Spatial Databases, Zurich, Switzerland, 1991, pp. 143-160.

7. [EKS 97] Ester M., Kriegel H.-P., Sander J.: “Spatial Data Mining: A Database Approach” , Proc. 5th Int. Symp. on Large Spatial Databases, Berlin, Germany, 1997, pp. 47-66.

8. Big Data: Techniques and Technologies in Geoinformatics, Edited by Hassan A. Karimi. CRC Press, 2014.

9. R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar, “ Spatiotemporal Data Mining in the Era of Big Spat ial Data: Algorithms and Applications,” in Proceedings of ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, 2012, pp. 1–10.

10. M. Zaki, “Parallel and Distributed Association Mining: A Survey,” Concurrency, IEEE, vol. 7, no. 4, pp. 14–25, 1999.

11. “Apache Hadoop,” http://hadoop.apache.org/.

12. [Gue 94] Gueting R. H.: “An Introduction to Spatial Database Systems” , Special Issue on SpatialDatabase Systems of the VLDB Journal, Vol. 3, No. 4, October 1994.

13. [Gut 84] Guttman A.: “R-trees: A Dynamic Index Structure for Spatial Searching“, Proc. ACMSIGMOD Int. Conf. on Management of Data, 1984, pp. 47-54.

