ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 4 April 2015, Page No. 11434-11436

A New Approach for Distributed Data Mining based on SOA

K.R.Shankar[1], K.Manikandan[2]
[1]PG Student, Department of MCA , SRM UNIVERSITY,CHENNAI
[2]Assistant Professor, Department of MCA , SRM UNIVERSITY ,CHENNAI

Abstract

Distributed Data mining is a data mining where computation and data is dispersed over multiple independent sites. Service-oriented architecture (SOA) provides integration of many services which coordinate and communicate to one another for their respective goals. Web service based SOAs are used for on-demand computing as well as for developing more interoperable intra or inter organizational systems. Generally data needs be securely protected for privacy purposes in the distributed environment and services provided on time. To serve both the purpose we use Distributed Data Mining (DDM) integrated on web service Business Process Execution Language (BPEL). Use of local abstraction technique used for privacy purpose

Keywords: Distributed data, Service oriented , data abstraction

1. Introduction

Data Mining is a process of transformation of data to an information, and then information to an action and action to a value or benefit [1]. In other terms Data Mining is process of extracting the useful information from different database for mining the patterns. Distributed computing plays a vital role since it requires large volumes which helps in storage and the time. To make more productive in a way, distribution of load dispersed in several sites. Here it involves some security issues since it is dispersed in different sites. Hence data mining requires use of heterogeneous resources.

DDM is the process of mining distributed data sets using distributed resources. Distribution of data at different sites and location of data is addressed in DDM process [2]. To store the privacy information and impose limits on the centralized server impose challenging issues. Mining information needs to be private[3] . Also the information should be provided at the right time is really important. This real time conditions along with this privacy requirements needs to be addressed. Hence it involves to be delivered just in time as well as in a secured manner .

Hence we implement this DDM with the use of service oriented option with the help of BPEL along with the web services. SOA has advantage of using some messaging protocol and extension of the service and their implementation in their business applications. BPEL is both executable is the use of full implementation logic and message exchange between process. SOA helps in building an infrastructure for DDM for the integration of distributed and heterogeneous environment with a complex connection. Here mainly this paper focus on the privacy, reduce the computational complexity and use of adaptive mining process. Hence we use the use of clustering and visualization process.

2. Data mining as a distributed service:

Before the use of distributed mining the service is the use of parallel method. Parallel clustering is compared to the distributed since distributed over many sites. Parallel process maintains the process very locally. But in the case of distributed it involves the real time measure. DDM is a method that is used for the purpose of distributed. So here it involves the use of the mining algorithm. Here in the distributed case the dataset is partitioned and distributed in the different sites. Data is replicated and replicas are placed at the different sites. This helps in the failure of replicas at a site that can be replaced from the other sites. It essentially helps in the case of the fault tolerance and flexible. But due to this heterogeneity that exists in local data sources that reflects the data privacy and mining accuracy [4].

3. SOA based DDM framework:

SOA is the popular web based services [5]. Web services form the basis of complex mining environment. Here it identifies the services based on the specification [6] and by using this highly composite data mining functionality is defined. This is implemented by use of the high level mining algorithms. Also SOA provides integration of services that coordinate and communicate for their goal. This result in emerge of web services that include the use of WSDL, UDDI, SOAP and BPEL.

WSDL is the web service definition language which is the combination of both the SOAP and XML for the web services over the internet. Network service involves operations and messages with a confined protocol in a defined message format. UDDI is Universal Description, Discovery and Integration by which the services as be listed in the internet to register and locate the services. SOAP is the simple object access protocol is messaging protocol that operate for different location. BPEL is the business process execution language that is a language for the business process and the services for the web. This is de facto standard for specifying business process behaviors. Interactions can be sequential ,concurrent or conditional

4. Privacy in data mining:

Storing data and retrieving plays a very vital role in all part of business. It that made process oriented then retrieving will be made in an easier manner. That is done by the data mining technique. This was recently very popular due to the field of research. One of the most important role is the privacy. We use data mining in number of applications. So the data must be sensitive and private. Protecting the data plays an important role in the data mining[7]

Without leaking individual information the data needs to be mined. We have two privacy models one is interactive other is non interactive. In interactive noise is added for privacy**.** Privacy preserving in data mining is very important for the usage and storage of personal data and with the algorithms. Many algorithms put forward for this preserving technique.

Privacy is important due to

- Data publishing
- Changing the data mining results
- Auditing
- Cryptographic methods
- Theoretical challenges

Most of this privacy issue is related in the database not an issue with the data mining.

Reputation and ability to protect privacy result in a greater willingness to provide accuracy in personal data[8]. Privacy-preserving data mining technology used to protect sensitive data that is not concerned with individual privacy. This is commonly done using a trusted broker to manage information. The goal of research on privacy-preserving data mining techniques needs to go beyond developing the basic techniques. Amount of data dealt is also considered as an important factor.

Best approach is to work with the risk of re-identification. Here addresses this issue by bounding the probability of a given person being in a private dataset so that risk of identification can be controlled[9]

5. Local abstraction Technique:

This local abstraction technique is the discovery of data patterns, reveals only the local data details of abstractions to the outside world. Privacy policies used by the local data owner control data's granularity level which determine how much privacy protection is necessary. Then we use global data analysis for data abstraction. Resampling will not be ideally suite for this and also much expensive. Here in this method it discards the similar data contributes equally for the estimation of global parameter. Then we use local abreaction technique to enhance the efficiency. This forms the basis of next-generation knowledge discovery in data systems, which must address challenges associated with supporting multi-dimensional. This involves process level integration. Discovering based on temporal knowledge is a challenging task. Temporal abstraction is based on temporal reasoning, which provides an intelligent interpretation and summary of large amounts of data or looking the important co relation or patterns for large dataset. This is one of the static learning techniques. Decision Trees induction algorithm is one to measure of the temporal dimension[10]. One advantage of temporal decision trees is that the output of the induction algorithm is a tree that can immediately used for pattern recognition purposes. However, the method can only be applied to time points, not to time intervals. This will help in the purpose of diagnosis and monitoring. Match the expets knowledge from the background. In order to respresent the data in means of distributed for scalability and privacy preservation we use GMM(Gaussian mixture model). One advantage is it provided a summary of whole data set in local and other is granularity, one referred by other.

Depending on the privacy preservation of local source each local data observed from outside is only the specific GMM abstraction from the abstraction hierarchy. We can also apply clustering technique to achieve this hierarchical local data abstraction. Each cluster is computed at each level of hierarchy in iterative manner for computation of the abstraction.

Two key characteristics of learning from abstraction
1. Analysis level degrades with quality as level in detail of abstraction decreases.
2. Determines scalablity , degree of privacy protection and accuracy

We have this application incorporated in the gaming environment. Here the learning from abstraction is very active in exploring the right level of details for cost purpose.

6.Conclusion:

Privacy is a vast area and real time also plays a vital role. So there must be more research that needs to be done on the privacy issue along with the new algorithms. So that should meet the need of privacy as well as the timing issue. Challenges are there in the both of the issues and most of them were covered. This proposed work from abstraction methodology, together with the SOA services, provides the essential for enabling privacy concerns in DDM.

References:

[1]Wang, K., Zhou, S., & Han, J. (2002). Profit mining: From patterns to actions. In: Proceedings of EDBT'02 (pp. 70–87).

[2] Subramonian R., Parthasarathy, An Architecture for Distributed Data Mining,
Fourth International Conference on Knowledge Discovery and Data Mining, New York, 1998, pp. 44-59

[3]H. Kargupta et al., "Collective Data Mining: A New Perspective towards Distributed Data Mining," *Advances in Distributed and Parallel Knowledge Discovery*, H. Kargupta and P. Chan, eds., MIT/AAAI Press, 2000, pp. 133–184.

[4] S. Merugu and J. Ghosh, "Privacy-Preserving Distributed Clustering Using Generative Models," *Proc. 3rd*

IEEE Int'l Conf. Data Mining (ICDM 03), IEEE CS Press, 2003, pp.
211–218.

[5] Kantardzic, M., Data Mining: Methods, Tools and Techniques, IEEE Press and John Wiley, 2002, Pages 380.

[6] Kumar, A., Kantardzic, M., "Web Application Protocols and Services for Distributed Data Mining", The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - Workshop on Data Mining Standards, Services and Platforms (DM-SSP 03), Washington, DC, August 2003.

[7] Zhang N. , Ming L. and Wenjing L. 2011 Distributed Data Mining with Differential Privacy. IEEE.

[8] A. Kobsa. Privacy-enhanced personalization. Communications of the ACM, 50:24–33, Aug. 2007

[9] M. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, June 11-14 2007.

[10] X. Zhang and W.K. Cheung, "Learning Global Models Based on Local Data Abstractions," *Proc. Int'l Joint Conf. Artificial Intelligence* (IJCAI 05), Morgan Kaufmann, 2005, pp. 1645–1646.