Proposed Model for E-commerce Distributed Data Mining based on SOAP and Ontology

Ibrahim S. Alwatban[1] **and Dr. Abdulrahman A. Mirza**[2]
[1] Computer and Information Technology Department, College of Telecommunication and Information, Technical and Vocational Training Corporation , Riyadh, Saudi Arabia
[2] Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Abstract- *The increasing amount of online customer data in many e-commerce applications has resulted in interesting opportunities for data driven knowledge discovery and data mining techniques. The aim of this paper is to propose an e-commerce distributed data mining model (ECDDM model) to get the benefits from this opportunity. The proposed ECDDM model consists of different components that interact with each other to overcome the expected difficulties which are related to these kinds of data. Distributed nature, heterogeneous data, privacy preserving data mining, data learning, and increasing interoperability within secured communications and performance issues are considered by the proposed ECDDM model.*

Keywords*:* E-Commerce, Distributed Data Mining, Ontology, SOAP, PPDM.

1 Introduction

Several important issues need to be addressed to design the ECDDM Model. The first one is the distributed nature and semantically disparate e-commerce customer data. Several approaches have been implemented for distributed data mining (DDM) [1]. One of these approaches is moving all distributed customer data into a central location and then performing mining tasks on the integrated data. Another approach is to perform mining tasks on each set of distributed customer data (for each location) to build local models. Then, local models will be moved to a central location and combined as a global model [2].

The final years of the past decade have seen the rapid development of learning classifiers from a semantically homogeneous relational database in the machine learning literature [3] [4]. In recent years, there has been an increasing interest to extend such approaches for learning classifiers from multiple semantically disparate, geographically distributed, relational data sources on the Semantic Web [5].

Privacy Preserving Data Mining (PPDM) is another important issue that needs to be considered during DDM model design. Most e-commerce providers may not be willing to share their data but they would like to take the benefits from DDM applications. According to [6], PDDM research is still in its infancy and there is no a practical system or development framework for PDDM. In recent years, there has been an increasing interest in data mining privacy methods [7]. Examples of these methods are: sanitation, data distortion, and Secure Multi-party Computation. Sanitation method aims to modify or remove sensitive data from data sources. Removing or modifying process may give a negative impact in the data mining results [8][9]. Distortion method provides privacy for e-commerce data by modifying the original data [2]. Another different method uses Secure Multi-party Computation (SMC) that uses cryptographic techniques to ensure almost optimal privacy [6].

The data learning process itself is a critical issue for this kind of model due to the relational nature of e-customer data. Relational nature of e-commerce customer data basically violates two assumptions made by traditional data mining techniques as stated by [4]: "The instances in relational data are not recorded in a homogeneous structure and are not independent and identical distributed".

In the ECDDM proposed model, we have used the approach that performs mining tasks for each distributed customer data to build local models and then combining them as a global model on the user side (central side). Statistical Relational learning is used as a classification process, and hierarchical ontologies are used to solve the problem of semantically heterogeneous data. There are privacy-based and performance-based components in both distributed customer data and user data. Simple Object Access Protocol (SOAP) is an XML-based protocol that will be used as a communication protocol to enable the user side to communicate with heterogeneous E-commerce applications. In section 2 the proposed model is discussed in more details.

2 ECDDM Proposed Model Structure

2.1 Overall Structure for the Proposed Model

There are two sides for the proposed model. User side as central side and distributed competitor side. Various components in both sides work together as follows: Mining request is generated from user architectural components at the user side, decomposed to many SOAP requests for each competitor side and SOAP results will be composed to generate integrated results at the user side. Figure 1 shows the overall structure of this DDM model. SOAP adopts SSL encryption to encrypt the information, so it is secure to transfer information [10].

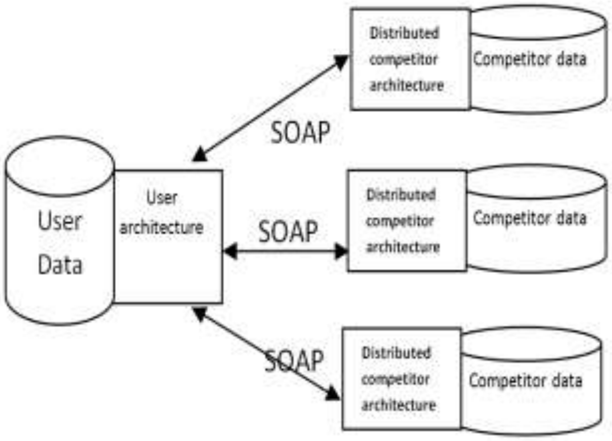
Distributed competitor architecture
Competitor data
SOAP
User Data
User architecture
SOAP
Distributed competitor architecture
Competitor data
SOAP
Distributed competitor architecture
Competitor data

Figure 1: Overall structure for the proposed model

2.2 Distributed Competitor-Side Architectural Components

As indicated in figure 2, there are three types of competitor side components. Physical component, coordination processes, and memory buffers. Physical component includes a local database or any other data source formats and algorithms library. Algorithms library contains all available algorithms for DDM and PPDM.

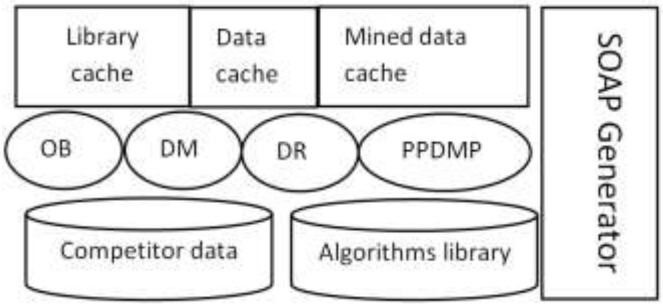

Figure 2: Competitor-Side Architectural Components**.**

Coordination processes are interfaces between physical components and memory buffers. Coordination processes includes SOAP Generator, ontology builder (OB), data miner (DM), data reader (DR) and Privacy Preserving Data Mining Process (PPDMP). The functions of these processes are as follows:

- SOAP Generator function: Receive "request SOAP message" and then convert it to competitor side environment. Convert mined result as "Response SOAP message" and send back to user side.
- Ontology builder (OB) function: It is responsible for Building hierarchical ontology in competitor side to be at the same abstraction level with user ontology. Hierarchical ontology will be constructed according to predefined mapping constraints between user side ontology and competitor side ontology.
- Data miner (DM) function: It provides the required algorithms from the algorithms library to perform data mining tasks and send the result to PPDMP to perform privacy process. After that, DM will place the result inside the data mined cache.
- Data Reader (DR) reads the required data from Competitor data and places them inside data buffers.
- Privacy Preserving Data Mining Process (PPDMP) provides the required algorithms from the algorithms library to perform privacy tasks and then send protected results to Data Miner.

Memory buffers component includes *library cache*, *data cache* and *data mined cache*. The aim of these buffers is to improve the performance of the proposed model. The library cache stores the most recently used DDM and PPDM algorithms, data cache stores the most recently used data, and data mined cache stores the most recent results. These buffers reduce the amount of physical reads (I/O reads) through the following steps:

a- Competitor side receives "request Soap message from user side.

b- Search for identical request and result inside these buffers. If they are available, the competitor side will directly send the result to the requester (user side) and there is no need to repeat DDM and PPDM processes again.

2.3 User-Side Architectural Components

As indicated in figure3, component types are similar to distributed competitor side. In user side, *mapping library* in physical components includes interoperation constraints, and the associated set of mappings from the user ontology to the competitor-side ontologies.

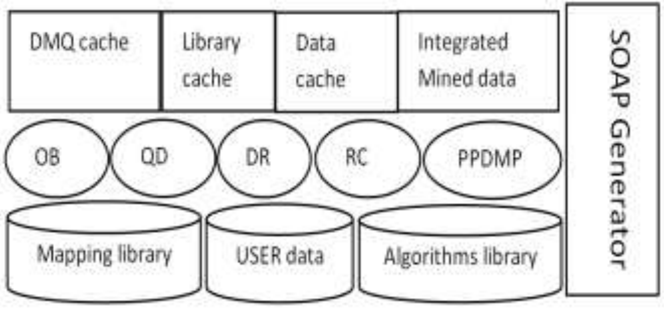

Figure 3: User-Side Architectural Components

In coordination processes, the new processes are as follows: *Query decomposer (QD*) maps and decomposes user side query into sub-quires for each competitor' side. *Result composer (RC)* integrates all competitor side results as one result according to the used data mining algorithm. Integrated result will be placed by RC inside integrated mined data. New Memory buffers components in user side are as follows: *Data mining query cache (DMQ cache)* that stores the most recently used composed/decomposed queries and *Integrand mined data cache* that stores the most recently used integrated results.

3 Conclusions

Data mining in e-commerce environments is an application of distributed data mining (DDM). To design a suitable model for Data mining in e-commerce environments, you need to consider several issues such as: heterogeneous data, privacy, data learning, increasing interoperability within secured communication, and performance. Our ECDDM model contains components which are dedicated to solve these issues. We have used memory buffers to overcome performance issues. We have used hierarchical ontologies to add semantics on data levels. We have proposed background processes for Privacy Preserving Data Mining, composition, and decomposition of data mining quires. We have used SOAP technology to deal with heterogeneous e-commerce applications, increase interoperability and secure communication messages.

4 References

[1] K. Hammouda and M. Kamel. “Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization”, IEEE Transactions on Knowledge and Data Engineering, , 2009, Vol. 21(5), pp.681-698.

[2] J. da Silva, C. Giannella, R. Bhargava, H. Kargupta, and M. Klusch, “Distributed Data Mining and Agents,” Eng. Applications of Artificial Intelligence, 2005,vol.18(7) , pp. 791-807.

[3] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer, “Learning probabilistic relational models”, In S.Dzeroski and Eds. N. Lavrac, editors, Relational Data Mining. Springer-Verlag, 2001.

[4] J. Neville, D. Jensen, and B. Gallagher,“Simple estimators for relational Bayesian classifiers”. In ICDM, pages 609–612. IEEE Computer Society, 2003.

[5] D. Caragea, J. Bao and V. Honavar. “Learning Relational Bayesian Classifiers on the Semantic Web”, Data Mining and Bioinformatics Laboratory, Department of Computing and Information Sciences, 2006.

[6] J.Secretan. “An Architecture for High-Performance Privacy-Preserving and Distributed Data Mining”,P.hD dissertation, College of Engineering and Computer Science at the University of Central Florida, Orlando, Florid, 2009.

[7] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the art in privacy preserving data mining”, In SIGMOD Record, 33(1):50–57, March 2004.

[8] Y. Saygin, V. Verykios, and C. Clifton, “Using unknowns to prevent discovery of association rules”, ACM SIGMOD Record, 30:45–54, December 2001.

[9] Y. Saygin, V. Verykios, and A. Elmagarmid, “Privacy preserving association rule mining”, In Research Issues in Data Engineering (RIDE), 2002.

[10] J. Tom, “SOAP: Cleans up Interoperability Problems on the Web”, Los Angeles: IT Professional, 2001.