Abstract - The organization of unstructured data is recognized as one of the major uncertain problems in the information industry and data mining paradigm. It will be in the form of computerized information that moreover, does not have a data model and there are not simply used by data mining. The task of managing unstructured data signifies possibly the major data management opportunity for our community subsequently managing relational data. The communities users such as KDD, Semantic web, AI and web, to industrial users such as Google, Yahoo, and Microsoft. This paper presents a study and analysis of the unstructured to structured distributed data mining. Several methods are available to manage unstructured data. The results of this analysis show that, there is a lack in managing unstructured data into structured data. Here, given a discussion about the impact of unstructured data in distributed data mining.

Keywords: Unstructured data, structured data, Distributed data mining, visualization, Semantic web.

The organization of unstructured data inspires vertically data separation methods for distributed environment working with structured data not an unstructured data environment. This paper aims to analyses distributed data mining for unstructured data environment and to attain a device of new structured data models from NETMARK. It benefits data integration of unstructured data; protect the Meta data by use hiding technique.

INTRODUCTION

Distributed data mining is the extraction of appropriate knowledge of the huge amount of data, while defending at the same time complex information or individually identifiable information in the unstructured distributed data environment. Terrovitis [2] have modified group based methods such as k-anonymity to "unstructured “data by discussing text data as a variety of variable length database record, or set of un-typed ethics, with the assumption that the sensitive value to defend is deterministically confined in this set. Clifton [3]
information that conforms to this specification. The Example of structured data is a relational database system. Figure 2 illustrates an Entity Relationship diagram (ER diagram), its concrete on tables within an RDBMS [relational database management system].

Figure 2 Example of Structured data

An analyst is an adept who collects data from legitimate sources in order to deliver meaningful, usable and relevant information used in decision making. This paper aim to analyze a distributed data mining for unstructured data environment and to accomplish a method of the new structured data model from NETMARK. It benefits data integration for unstructured data and protect Meta data by routine hiding technique.

II UNSTRUCTURED TO STRUCTURED DATA ENVIRONMENT IN DISTRIBUTED DATA MINING SURVEY

1. Data generation and Exploitation model [DGE model]

AnHai Doan, et al., [6] analysis to manage unstructured data effectively, a clear data generation and exploitation model [or DGE model as short] will have to emerge. The model describes the interaction between the data, the system and the users. The DGE model explains who are the users, how they define the needs, what their material needs are, how the data is created in the system, and how user interact with the system to satisfy these needs. The Jeffrey F. Naughton, et al., [6] generates the data by using a schema, populates it with conforming data, and perhaps modifies the data from update transactions. The DGE model mainly focuses on unstructured documents, in the relational context; its essence handles only sophisticated, SQL-knowing developers. The user interacts with the database (to create and query the data) simply by appealing canned SQL commands and queries (written by some developers) via relatively simple form interfaces. For data generation the AnHai Doan, et al.,[6] proposed to generate new data by extracting unstructured data to structured data, where in its simple form this structured data is attribute value pairs, such as city names, person names from Wikipedia, temperatures. Hence the nature of unstructured data, the mined structured data will often be semantically heterogeneous.

For example, of DGE model, the two different names “David Smith” and “D. Smith” extracted from Wikipedia may in fact mention to the same person, or characteristics the location and address mined from two Wikipedia info boxes may in fact equal. Accordingly, we will often have to accomplish an information integration step to resolve the semantic heterogeneity and unify the extracted structured data. Then automatic IE and II (i.e., information extraction and integration) often will not be 100% accurate. Accordingly, the DGE model [5,6] should agree the structured data to be generated in an incremental, best-effort fashion, should the application decide on to do so. Another good example of Data exploitation by our author is Q = “find the average temperature of madison” in the Wikipedia. Supposing we have mined the monthly temperatures, then a refined user can direct formulate Q as a structured query (e.g., in SQL), and acquire an answer from the system. By using our model DGE should permit users to start in whatever data-exploitation mode they deem comfortable (e.g., structured querying, browsing, visualization, keyword search) then help them move effortlessly into the mode that is finally appropriate for their information need.
The DGE model for unstructured data should use a grouping of IE, II, and HI to generate structured data from the initially unstructured data, in a potentially mass collaboration, best-effort fashion. The model should allow a wide range of data exploitation modes as keyword search, structured querying, browsing, visualization, monitoring as well as continuous transition from one mode to another, in an iterative fashion through interaction with the user. In that, found some difficulties as well as some limitations. By using the automatic information extraction and integration the DGE model will not be 100% accurate. Finally, to improve the accuracy of the many applications may want to generate structured data incrementally, in a best determination and the user deems necessary (instead of generating all of them in one shot).

2. Semantics and NLP technologies

Harish Jadhao, et al., [7] analyze by using the approach of semantic and NLP technologies to build a mechanism for semantic from un-structured data to structuring data. It explores a methodology for processing domain specific data and delivers appropriate, compact size information to the analyst. The author approaches another method to visualize extracted structured data through a graph visualization technique called spring growth. To develop semantic methods for analyzing unstructured data, it is essential to build systematic information extraction framework. This automatically misappropriate entity and relations between them from legacy data created on domain ontologies, and signify this abstract structure into an RDF (Resource Description Framework) graphs. These can be additional stored in RDF knowledge base and queried using SPARQL query language.

2.1 Different Methods for Extracting Information from Unstructured Data

Ontology [8] learning and population from unstructured data are two dynamic research fields. Harish Jadhao, et al., most of the work had been carried out in a similar direction within Artificial Intelligence, Machine Learning and NLP. The main aims at mining domain terms, concepts, individuals, concept attributes and relations from textual data. The different approaches for text relies in Ontology Learning is 1) Linguistic methods to determine information mining rules by inspection of a corpus. 2) Combination of linguistic and machine learning methods. 3) Machine Learning method and statistical method to learn rules from annotated corpora. The key benefit of the linguistic method above statistical method is it does not require huge amounts of training corpus. This is often exclusive to obtain. On the other side 1) domain adaptation may involve significant reconfiguration 2) rule identification process is tedious and laborious. Now statistical approach, domain adaptation is relatively improved due to automatic rule induction. It shows results for entity annotation, such as identifying gene name in system biology [9]. This method is not active in case of related documentation due to lack of annotation text corpora. Likewise, training data may be hard to obtain. Significant change in domain arrangement requires re-annotation of complete training data. This needs repeating in the preparation process for every new domain in order to be accurate.

Maedche, Staab and Volz [10] present an active approach for creation, reuse and maintenance of ontologies from domain text by relating statistical approach (Text-to-Onto tool). The combinations of two approaches are: 1) statistical techniques to simplify and accelerate ontology construction, 2) linguistic techniques. The general system includes filtering and preprocessing textual data. This is carried out by processing resources module and NLP system, and then modeling properties are recognized by applying machine learning algorithm. This application also delivers ontology refinement, pruning. It focuses on learning the significance of indefinite words over the time. On to Learn [11], OntoLT [12], OntoGen [13] tools established to support the user in creating ontologies from a text data. As a result, there has been an extensive focus on rule based approach, even though lacking a lot of manual work; proves to be more active and clear in capturing the semantic criteria.

2.2 Different Frameworks for Information Visualizations
Dr. Jagannath Aghav, Anil Vegirajuet al., [7] point of view graph visualizations have been developed for demonstrating ontology extraction in a graphical format. But these visualizations are effected in the conventional graph representation and no existing work is being carried out in the perspective of a spring embedded graph generation. The spring graph model is based on the web based search engine is provided an aggregation of information archived from different sources like audio and video content, news and magazine articles, blog entries and many other sources. The result in a graphical representation beside with charts provides a 360-degree search and permits for unpredictable the relevancy of a topic as per user requirement.

Another implementation creates visualizations of graphs using a spring embedded algorithm that eliminates edge crossings [14]. This implementation creates three dimensional graph visualizations that provide a graph with quality and attractiveness. Another paper discusses the usage of graph theory for social network analysis [15]. This paper suggests that by defining the conceptual distance between people and groups, information about the type of communication in an association can be inferred. it generally focuses on exploration of military groups and visualization of the link between people in the organization through diagrams and graphs. The theoretic techniques and statistical approach are making use of graph visualization spring embedding algorithm is used to visualize the social network, anywhere the spring distance relates to the actual distance between two nodes (i.e. The link distance).

The Author analyzes with another paper intends an extension of the spring embedding algorithm to the three dimensional realm, it is known as GEM-3D (Graph Embedder3D) [16]. It provides graphs representation at communicating speed and improved display quality. The graph provides examples as accommodating hundreds of nodes and makes use of visual clues for easier readability. This algorithm can be realistic to both directed graph and undirected graphs. The implementation routines spring graph to visualize the information mined from a digital library after semantic analysis [17]. It signifies a meaningful relationship between the documents and provides an effective way of extracting the documents. The spring embedding algorithm signifies the semantic relationship between the entities, with minor spring distance representing extra similarity between the nodes, and vice versa. Existing systems basically focus on the knowledge acquisition, aggregation and manipulation of data through graph representation with efforts made towards giving the data to a user.

Table 1: show the comparison of unstructured and structured content.

<table>
<thead>
<tr>
<th></th>
<th>Unstructured data</th>
<th>Structured data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Character and binary data</td>
<td>Relational database tables</td>
</tr>
<tr>
<td>Transaction management</td>
<td>No transaction management and no concurrency</td>
<td>Matured transaction management, various concurrency techniques</td>
</tr>
<tr>
<td>Version management</td>
<td>Versioned as a whole</td>
<td>Versioning over tuples, rows, tables, etc.</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Very flexible, absence of schema</td>
<td>Schema-dependent, rigorous schema</td>
</tr>
<tr>
<td>Scalability</td>
<td>Very scalable</td>
<td>Scaling DB schema is different</td>
</tr>
<tr>
<td>Robustness</td>
<td>-</td>
<td>Very robust, enhancements since 30 years</td>
</tr>
</tbody>
</table>

Table: 1: The comparison of unstructured and structured content

3. Randomization method and the k- Anonymity model

In the year of 2012, Thavavel, et al., [18] distributed data mining is the extraction of related knowledge from the huge amount of data, while defending at the
same time sensitive information or personally identifiable information in the unstructured distributed data environment. Terrovitis [2] have modified group based methods such as k-anonymity to "unstructured"data by considering text data as a type of variable length database record, or set of un-typed values, with the statement that the sensitive value to defend is deterministically contained in this set. The author Clifton [2] point of view inspires vertically data separation techniques for distributed environment and functioning with structured data not a unstructured data environment. Miller [21] has aimed a system to simplify the interaction of structured and unstructured data. The main aim of the view mechanism, particularly as they relate to textual documents are presented in the paper. The approach allows the interface between the data occupied from unstructured (example: Text) data source and, semi structured (object oriented), structured (example: Relational). The author described in extensible point of view system that support the combination of data from heterogeneous structured and unstructured data sources in moreover the multi database or data warehouse environment. Make use of global schema if the first approach; it normally makes use of common data models and a global language. The Second Approach is used to define the common language in multi database that is how the data sources are combined, transferred and presented. Third Approach is the local site works thoroughly with a fixed of interrelated sites to set up the partial global schema.

David A.Maluf [22] deal with NETMARK was NASA Ames Research center aimed and developed a data management and integration system. NET MARK that accomplishes data integration across several structured and unstructured data. A source which is highly scalable and cost efficient manner. Querying and integration of originally unstructured data such as several formatted reports in Excel Spreadsheets and power point presentations, Microsoft word, is a key focus, Adobe portable Document Format(PDF),given that the substance of creativity data is indeed unstructured.

It compact with converting the structured data from unstructured data, the unstructured data are transformed to Xml, node representation and relational storage with Meta data. The Meta data, perform text handling to find words or attribute in documents and testing parallel measure using normal distribution method. It extends to achieve and analyzing data for distributed data mining in unstructured data like e-mail messages, complicated information, appearances, voicemail, images, and video.

III DISCUSSIONS

Existing methods are related to unstructured to structured data environment. The results of the above motioned paper are discussed here. AnHai Doan, et al., the analyses DGE model describes the interface between the data, the system, and the users. The model is explained in the form of relational data, to generate data, a user defines a schema with conforming data and feasibly modifies the data from update transactions. To manage different kind of data efficiently AnHai Doan, et al., identifies problems that may 5-10 years ahead of the industry, thus situating us in a position to lead instead of responding. By analyzing on structured data extracted from unstructured documents, the model for relational data, keyword search that has been proposed for the DB+IR environment [6], are not suitable.

Harish Jadhao, et al., found that the existing approach by using of NLP technologies and semantic mechanism for semantically un-structured data to structured data. Hence the inherent experiments of natural language processing; the existing methods for
finding a particular piece of information from text incline to be domain specific. The author also analyses the approach to visualize mined structured data done a graph visualization technique is called spring growth. To analyze unstructured data semantic tool are used develop to build the systematic information. Further information is stored in Resource Description Framework [RDF] knowledge base and queried using SPARQL query language. By using a Semantic Validation rule to validate relation semantic is formulated as nsubj - verb - dobj/pub that simply finds paths in the chunk dependence tree that for a start-point (in general as the NEntitySubject) to an endpoint (in general as the NEntityObject). For these conditions, we construct (NEntitySubject-Predi-NEntityObject) to the set of validated constructs, as equation 1 [7] and equation 2 [7]

\[
\text{Predi= } \{\text{Relation} \mid \text{Node with two outgoing edges with labels “nsubj” and “dobj”}\} \quad \rightarrow \quad \text{equation 1}
\]

\[
\text{NEntitySubject= } \{\text{Entity} \mid \text{Node containing named entity; which is connected to the predi by edge with label “nsubj”}\} \quad \rightarrow \quad \text{equation 2}
\]

Thavavel, et al., [18] introduced methods by using VB.NET application. Unstructured data environment Unstructured data environment changed to the structured data environment is designed by using VB.NET application. The implementation structure considered the text documents in (.txt file extension) only. The two text file document size as 42.4KB (43,497 bytes) and 19.7KB (20,178 bytes) was used to create the resultant shown below. In example dataset (2 examples, 4 special attributes, 2199 regular attributes), where two examples are stored in m1.txt and m2.txt. Extraordinary attributes are Labels (me9, me10 are type is binomial), the Metadata-file, metadata-path (type is polynomial) and Meta-data-date (type is date time), regular attributes of both text files in words or attribute names.

**IV Conclusions**

Unstructured data (information that lies external of databases where business intelligence is regularly stored) represent the largest and fastest growing basis of information accessible to businesses and governments worldwide. In this Paper consists of study and analysis of the structured data from an unstructured data environment with a distributed mechanism. Several methods are available to find the structured data from unstructured data environment. Many researchers said that, unstructured data communities such as semantic web, AI, IR, KDD, and Web to industrial users such as Microsoft, Google, and Yahoo. In Distributed data mining the unstructured data like voice mail, e-mail messages, still images, complicated reports, video and presentations. Most of the existing methods contain some bias. So our goal is to focus on the structured data environment from unstructured data for industry players.

**V References**

7. Dr. Jagannath Aghav, Anil Vegiraju ,Harish Jadhao, “Semantic Tool for Analysing Unstructured data”.
14. Weimin Wu, Yongfeng Cao, Qing Su, Yonghe Zhang, Guangdong University of Technology, Guangzhou, “Visualization of Graph based on the Three-dimensional Spring Model”.
16. Ingo Brub and Arne Frick, Universitat Karlsruhe, Fakultat fur Informatik, D-76128 Karlsruhe, Germany, Fast Interactive 3-D Graph Visualization.
18. V. Thavavel And S. Sivakumar, “A generalized Framework of Privacy Preservation in Distributed Data mining for


Author Profile

Padmapriya G received the first degree in B.C.A from Bharathiyar University in 2004, Tamilnadu, India. She obtained her master degree in Computer Communication from Bharathiyar university and She obtained her master degree in Computer Applications from Bharathiyar University in 2012, Tamilnadu, India. She is currently pursuing her Ph.D. degree Under the guidance of Dr. M. Hemalatha, Head, Dept of Software Systems, Karpagam University, Tamilnadu, India.

Dr. M. Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women’s University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science in Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.