A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment

V.THAVAVEL and S.SIVAKUMAR*

Department of Computer Applications, Karunya University, Coimbatore, Tamilnadu, India-641 114.

*Corresponding Author

Abstract

The management of unstructured data is recognized as one of the major unsolved problems in the information industry and data mining paradigm. Unstructured data in computerized information that either does not have a data model and there are not easily usable by data mining. This paper proposes a solution to this problem by managing unstructured data in to structured data using legacy system and distributed data partitioned method for gives distributed data for mining multi text documents. This frame work gives the testing of the similarities among text documents and privacy preserving meta data hiding technique, which are explored in text mining.

Keywords: *Unstructured data, Privacy preserving data mining, Distributed data mining, Testing Similarity.*

1 Introduction

Privacy preserving distributed data mining is the extraction of relevant knowledge from large amount of data, while protecting at the same time sensitive information or personally identifiable information in the unstructured distributed data environment. Terrovitis[14] have adapted group-based methods such as k-anonymity to "unstructured" data by treating text data as a sort of variable length database record, or set of un-typed values, with the assumption that the sensitive value to protect is deterministically contained in this set. Chris Clifton[3] address the problem of data is vertically partitioned and privacy means preventing others from learning the value of private attribute values for each entity. The need for privacy preservation is privacy of source because unauthorized user interact to damage the data of misuse of information and to support heterogeneity of source. Clifton[15] motivate vertically data separation techniques for distributed environment and working with structured data not a unstructured data environment. This paper aim to develop a privacy preserving distributed data mining frame work for unstructured data environment and to achieve a device of new structured data model from NETMARK. It helps data integration for unstructured data; protect Meta data by use hiding technique. This paper attempt to distribute heterogeneous data to the network by using vertically data separation method and it enables text mining to test the similarity measure among text documents.

This paper has proposed and designed a new generalized frame work for privacy preservation in distributed data mining for unstructured data environment and implementing the testing of similarity among free form of text using testing of hypothesis (Inferential Statistics). In this paper, section 2 describes the literature work, section 3 provides designing of a new frame work for privacy preservation in distributed data mining for unstructured data environment, section 4 discusses implementation of a frame work and testing the similarity of the given documents and finally, section 5 gives the conclusion of the work.

2 Literature Review

2.1 Privacy preserving data mining

A number of approaches and techniques such as randomization, and k-anonymity have been developed in order to carry out privacy-preserving data mining task.

2.1.1 Randomization method

In randomization method for privacy-preserving data mining, the noise is added to the data set in order to mask the attribute values of sensitive record fields[1][2]. The amount of noise added is large enough to smear original values, so individual records values cannot be recovered techniques are designed to derive aggregate distributions from the perturbed records. Subsequently data mining methods can be developed in order to work with these aggregate distributions. The key advantage of the randomization method lies in its simplicity as method does not require knowledge of the distribution of other records in the data, unlike other methods such as k-anonymity which require the knowledge of other records in the data. Therefore, the randomization can be applied at data collection time without the use of a trusted server containing all original records.

Kargupta et.al[7] challenges perturbation and randomization –based approaches. They claim that such approaches may lose information as well as not provide privacy by introducing random noise to the data by using random matrix properties, Kargupta et al. successfully separates the data from the random noise and subsequently discloses the original data.

2.1.2 The k- Anonymity model

The K-anonymity model [12] was proposed to deal with the possibility of indirect identification of records form public databases. Since combinations of records attributes can be used to exactly identify individual records. In k-anonymity the granularity of data representation is reduced by employing techniques such as generalization and suppression. The granularity is reduced to such a level that any given record maps onto a least K other records in the dataset. The k-anonymity method was first proposed by Samarati [11]. The approach uses domain generalization hierarchies of the quasi-identifiers in order to build K-anonymous tables. The concept of K-Minimal generalization has been proposed Samarati [11] in order to limit to level of possible for a given level of anonymity.

2.1.3 Secure multi-party computation

Secure multi-party computation (SMC) deals with general problem of functions secure computation with distributed inputs. In privacy preserving data mining the solutions that posses the rigor of work in SMC settings and typically make use of cryptographic techniques are known as SMC solutions.

Yao first formulated the two-party comparison problem (Yao's Millionaire protocol) and presented a provably secure solution [13]. It was extended to multiparty computations by Goldreich et al.[6]. They developed a frame work for secure multiparty computations, and in [5] proved that computing a function privately is equivalent to computing if securely. A semi honest party(also known as honest but curious) follows the rules of the protocol using its correct input, but is free to later use what it sees during execution of the protocol to compromise security.

2.2 privacy preserving Distributed Data mining

The primary role of distributed methods for privacy preserving data mining is to enable computation of useful aggregate statistics over the joint databases without compromising the privacy of the individual datasets within the different participants. So, the participants may wish to collaborate in obtaining aggregate results, may not fully trust each other in terms of the distribution of their own databases Lindell and Pinkas[7] first introduced this technique to the data mining community. Their method enabled two parties to jointly contract a decision tree without either party gaining any knowledge about each other's data except what might be revealed through the final decision tree. Specifically they targeted ID3 Algorithm with horizontally partitioned data. For the purpose of privacy, the datasets may be partitioned either horizontally or vertically. In case of horizontally partitioned datasets, the individual records are spread out across multiple entities, each have the same set of attributes. In vertical partitioning the individual entities may have different attributes (or view) of the same record sets. Chris Clifton[3] deals with finding to address the problem of association rule discovery, where data is vertically partitioned, and privacy means preventing others from learning the value of private attribute values for each entity. The problem of distributed privacy preserving data mining closely resembles field of cryptography for determining secure multi-party computations and share common techniques [10].

2.3 Unstructured Data Environment

Miller[8] have designed a system to facilitate the interaction of structured and unstructured data. The main features of the view mechanism, especially as they relate to textual documents are presented in the paper. It also looks at how the views approach allows the interaction between the data taken from structured (example: Relational) semi structured (object oriented) and unstructured (example: Text) data source. Data mining and how the system will operate in the complete environment. This paper, describe in extensible view system that support integration of both data from heterogeneous structured and unstructured data sources in either the multi database or data ware house environment. First Approach is makes use of a global schema; it typically makes use of common data model and a global languages. Second Approach is multi database language approach common language to define how the data sources are integrated, transferred and presented. Third Approach is the local site works closely with a set of inter-related sites to setup the partial global schema.

David A.Maluf [4] deal with NETMARK was NASA Ames Research centre designed and developed a data management and integration system. NET MARK that achieves data integration across multiple structured and unstructured data source in a highly scalable and cost efficient manner.

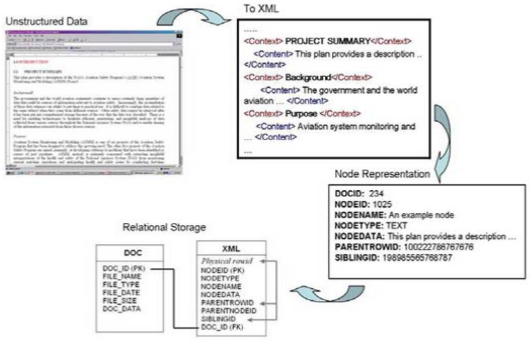
Querying and integration of originally unstructured data such as various formatted report in micro soft word, Adobe portable Document Format(PDF), Excel Spread sheets and power point presentations, is a key focus , given that the bulk of enterprise data is indeed un structured

3. Proposed Method

The objective of this paper is proposed a generalize framework for the privacy preservation distributed data mining for unstructured environment.

3.1 Unstructured data into Structured Data Environment

It deal with converting the unstructured in to structured data, the unstructured data is converted to Xml, node representation and relational storage with Meta data.



```
......
<Context> PROJECT SUMMARY</Context>
   <Content> This plan provides a description ...
</Content>
<Context> Background</Context>
     <Content> The government and the world
aviation ... </Content>
<Context> Purpose </Context>
    <Content> Aviation system monitoring and
... </Content>
...
```

```
DOCID:  234
NODEID: 1025
NODENAME: An example node
NODETYPE: TEXT
NODEDATA: This plan provides a description ...
PARENTROWID: 100222786767676
SIBLINGID: 198985565768787
```

Figure 1 Unstructured data into Structured Data

3.2 Distributed Mechanism

Distributed mechanism was designed for real storage of metadata with text. Distribute data as method of vertically separation techniques for heterogeneous data.

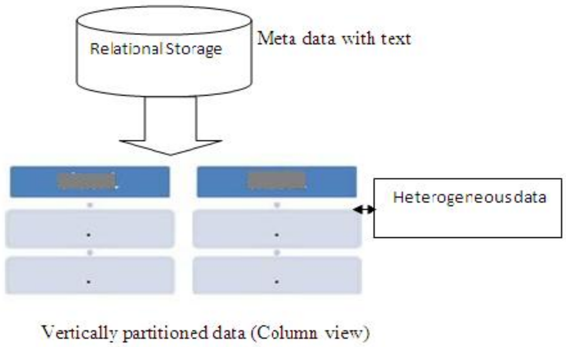


Figure 2 Distributed Mechanisms

3.3 Security Mechanism

It deals with to apply in Meta data with hiding techniques for privacy policy.

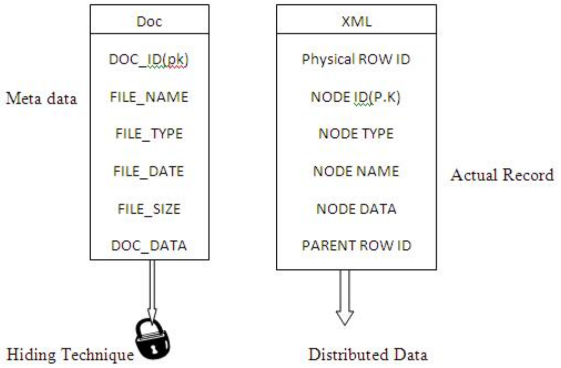


Figure 3 Security Mechanism with hiding Meta data for privacy preservation

3.4 Text Mining

It perform text processing to find words or attributes in documents occurrences in word list process document from files

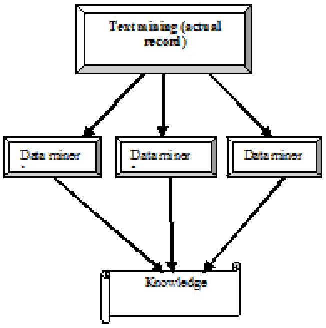


Figure 4 Text process is distributed to data miner for Text mining

3.5. Privacy Preservation in Distributed Data mining for Unstructured Data Environment

Design and develop a data model (structured data) from unstructured data, managing unstructured data are converted into XML and then create data table contain Meta data.

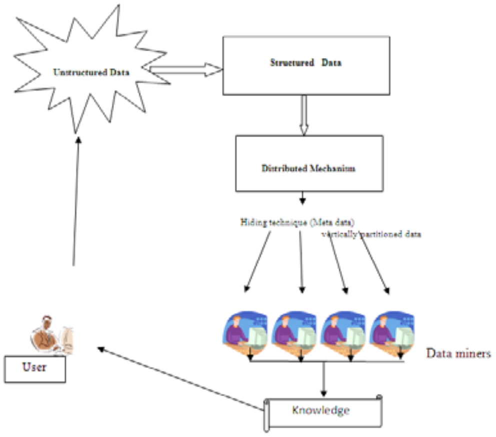
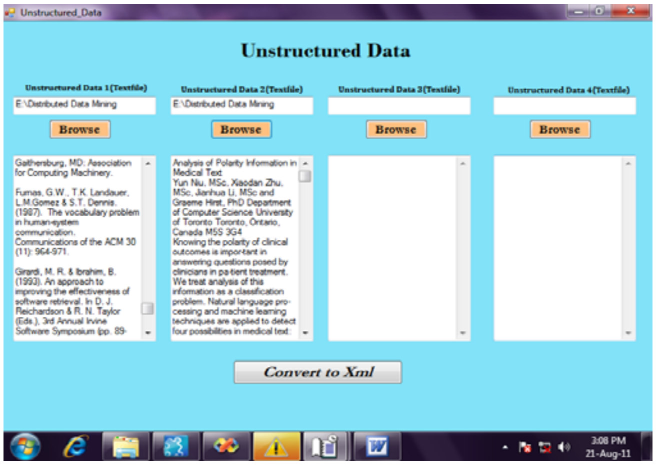


Figure 5 Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment

Meta data like data about data in textual information, For example file name, Date of creation, file type, file size, Author of document etc., Because of this conversion is used to well efficient of distributed data mining of textual area and secure of personally identifiable information. Data table contains Meta data and xml data from various text files is stored in relational Storage. Relational storage data (heterogeneous) is distributed along network path with more securely. Privacy preserving distributed data mining distribute as datasets like horizontally or vertically partitioned. In this frame work distribute only xml data (Real storage data with column view) by using vertically partitioned method. Privacy policy is applied to Meta data with information hiding technique for security purpose. Data miner only knows xml text information not a personally identifiable information and also intruder does not interact with personal data without authentication. Xml data is distributed to data miners; data miners perform information extraction and measure text document similarity. In this frame work using text mining is extract knowledge from heterogeneous data into related data groups. (i.e. here clustering method is extract similar/related data group from heterogeneous data).Distributed data mining paradigm have more one than data miner to perform knowledge extraction process, at the same time interaction among one miner to another for sharing data mining intermediate results not a source data. In this frame work perform automatic discovery of new, previously unknown, information from unstructured textual data. Finally develop a generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment.

4. IMPLEMENTATION

Unstructured data environment converted structured data environment and privacy preservation of Meta- data designed by using VB.NET application. The implementation setup considered the text documents (.txt file extension) only. In this paper take two text file documents both size as 42.4kb (43,497 bytes) and 19.7kb (20,178 bytes) was used to produce the resultant shown below. Actual text processing as like document occurrences and total occurrences of words in text document was designed by using Rapid miner tool. Example dataset (Two examples, 4 special attributes, 2199 regular attributes), where two examples are m1.txt and m2.txt.Special attributes are Labels (me9, me10 are type is binominal), Metadata-file, metadata-path (type is polynomial) and Meta data-date (type is date-time). Remaining are regular attributes of both text files in words or attribute names. The following screenshot shows the working environment of this frame work


Unstructured_Data
Unstructured Data
Unstructured Data 1(Textfile)
Unstructured Data 2(Textfile)
Unstructured Data 3(Textfile)
Unstructured Data 4(Textfile)
Browse
Browse
Browse
Browse
Convert to Xml
3:08 PM
21-Aug-11

E:\Distributed Data Mining

E:\Distributed Data Mining

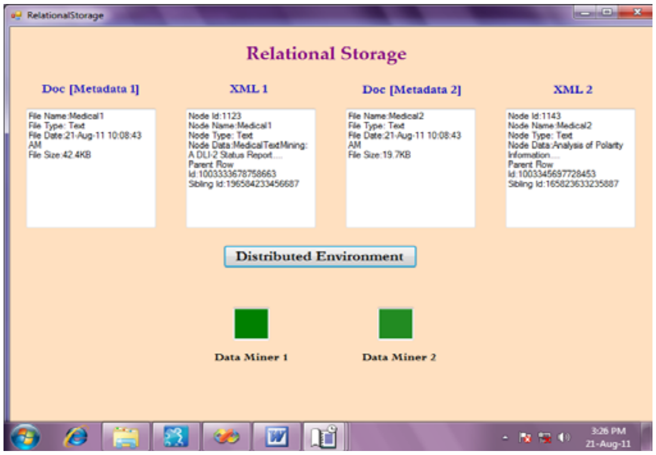
Gaithersburg, MD: Association for Computing Machinery.

Furnas, G.W., T.K. Landauer, L.M.Gomez & S.T. Dennis. (1987). The vocabulary problem in human-system communication. Communications of the ACM 30 (11): 964-971.

Girardi, M. R. & Ibrahim, B. (1993). An approach to improving the effectiveness of software retrieval. In D. J. Reichardson & R. N. Taylor (Eds.), 3rd Annual Irvine Software Symposium (pp. 89-

Analysis of Polarity Information in Medical Text
Yun Niu, MSc, Xiaodan Zhu, MSc, Jianhua Li, MSc and Graeme Hirst, PhD Department of Computer Science University of Toronto Toronto, Ontario, Canada M5S 3G4
Knowing the polarity of clinical outcomes is important in answering questions posed by clinicians in pa-tient treatment. We treat analysis of this information as a classification problem. Natural language pro-cessing and machine learning techniques are applied to detect four possibilities in medical text:

Figure.6.Unstructured data (Text file) browse from node and apply operation convert to XML and Meta data format



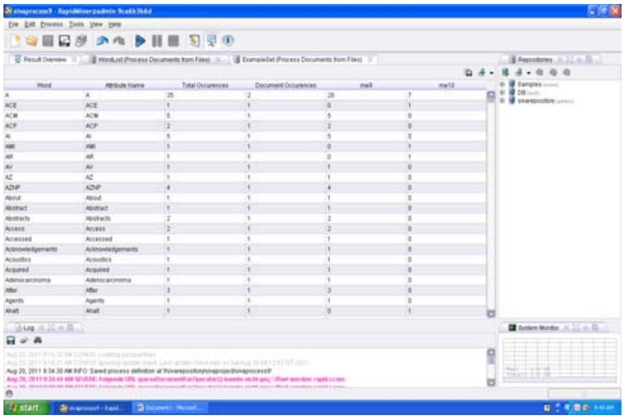
File Name:Medical1
File Type: Text
File Date:21-Aug-11 10:08:43 AM
File Size:42.4KB

Node Id:1123
Node Name:Medical1
Node Type: Text
Node Data:MedicalTextMining: A DLI-2 Status Report....
Parent Row Id:1003333678758663
Sibling Id:1965842334566887

File Name:Medical2
File Type: Text
File Date:21-Aug-11 10:08:43 AM
File Size:19.7KB

Node Id:1143
Node Name:Medical2
Node Type: Text
Node Data:Analysis of Polarity Information....
Parent Row Id:1003345697728453
Sibling Id:1658236332358887

Figure.7. XML are distribute through distributed environment into data miners instant of hiding Meta data



Word	Attribute Name	Total Occurences	Document Occurences	me9	me10
A	A	35	2	28	7
ACE	ACE	1	1	0	1
ACM	ACM	5	1	5	0
ACP	ACP	2	1	2	0
AI	AI	5	1	5	0
AMI	AMI	1	1	0	1
AR	AR	1	1	0	1
AV	AV	1	1	1	0
AZ	AZ	1	1	1	0
AZNP	AZNP	4	1	4	0
About	About	1	1	1	0
Abstract	Abstract	1	1	1	0
Abstracts	Abstracts	2	1	2	0
Access	Access	2	1	2	0
Accessed	Accessed	1	1	1	0
Acknowledgements	Acknowledgements	1	1	1	0
Acoustics	Acoustics	1	1	1	0
Acquired	Acquired	1	1	1	0
Adenocarcinoma	Adenocarcinoma	1	1	1	0
After	After	3	1	3	0
Agents	Agents	1	1	1	0
Ahalt	Ahalt	1	1	0	1

```
Aug 20, 2011 9:34:38 AM INFO: Saved process definition at //sivarepository/sivaproject/sivaprocess9
```

Figure.8. Text data processed by text processing using rapid miner tool find words or attributes in documents occurrences in word list process documents from files

4.1 Result and Discussion

For testing the similarity of given documents using the method of testing of hypothesis (Inferential Statistics). This paper deals the testing of two documents, namely m1 and m2. Words occurrences of the above said documents are 2199 regular attributes.

$$Z_c = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}}$$

where Z_c is the calculated value of standard normal variate. $\overline{x}$ and $\overline{y}$ are the mean values of the words occurrences in documents m1 and m2 respectively. σ_1^2 and σ_2^2 are the variances of the words occurrences in documents m1 and m2 respectively. n$_1$ and n$_2$ are the number of regular attributes of documents m1 and m2 respectively. Z_t = 1.96 is the table value of standard normal variate at 5% level of significance. Let us assume that, there is no similarity in documents m1 and m2 (null hypothesis). Here Z_c = 2.936789 , $\overline{y}$ = 1.398818, σ_1^2 =163.3763, σ_2^2 =38.69847 and n$_1$=n$_2$= 2199 and Z_c = 5.0735, therefore $Z_c > Z_t$, so rejects null hypothesis. Hence the documents m1 and m2 are Similar.

4. Conclusion and Future work

` This paper provides frame work for privacy preservation of Meta data using hiding technique in unstructured data environment with a distributed mechanism. The proposed system is also perform text processing to find words or attributes occurrences in documents and testing similarity measure using normal distribution method which was discussed. In future generalized frame work is extend to manage and analyzing privacy preservation for distributed data mining in unstructured data like e-mail messages, complicated reports, presentations, voice mail, still images, and video.

5 References:

[1] Agrawal. D and Aggarwal .C.C, *"On the Design and Quantification of Privacy Preserving Data Mining Algorithms":ACM PODS Conference*,(2002).

[2] Agrawal.R and Srikant.R, *"Privacy-Preserving Data Mining";ACM SIGMOD Conferene,pp*.439-450,Dallas,TX,May 14-19,(2000).

[3] Chris Clifton, *Privacy Preserving Distributed Data mining*, Department of Computer Science, Nov 9,2001*www.cs.purdue.edu/homes/**clifton**/.../**Clifton**DDM.pdf*

[4] David A.Maluf and Peter B.Tran, *Managing Unstructured Data with Structured Legacy Systems*, 2008 IEEE.

[5] Goldreich.O, *The Foundations of Cryptography*, volume 2,chapter General Crytographic Protocols. Cambridge University Press,2004.

[6] Goldreich.O, Micali.S and A.Wigderson, *How to play any mental game- a completeness theorem for protocols with honest majority*, In 19th ACM Symposium on the Theory of Computing, pages 218-229,1987.

[7].Kargupta.H, Datta.S. Q.Wang and K.Sivakumar, *"On the privacy preserving properties of random data perturbation techniques"*,IEEE ICDM,2003.

[8].Lindell.Y and B.Pinkas. Privacy preserving data mining. Journal of Cryptology,15(3):177-206,2002.

[9].Park.B and H.Kargupta, "Distributed data mining": Algorithm,System,and applications.N.Ye,editor,The Hand book of Data mining,Pages 341-358.Lawrence Erlbaum Associates, Mahwah,N.J.,(2003).

[10] Pinkas. B, *Cryptographic Techniques for Privacy-Preserving Data Mining*,ACM SIGKDD Explorations,4(2),2002.

[11] Samarati.P, *"Protecting Respondents Identities in Microdata Release*, IEEE Trans.Knowl.Data Eng.13(6):1010-1027(2001).

[12] Samarati.P and Sweeney.L,*"Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression"*, IEEE Symposium on Security and Privacy,(1998).

[13] Yao.A.C, *"How to generate and exchange secrets"*, In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, Pages 162-167.IEEE, 1986.

[14] Terrovitis. M, Mamoulis. N and Kalnis. P, *"Privacy-Preserving Anonymization of Set-Valued Data". In Proc. of VLDB Endowment,* 1(1), 2008.

[15] Chris Clifton, Murat kartarcioglu, Jaideep vaidya,Xiaodong Lin and Michael Y.Zhu, *Tools for Privacy Preserving Distributed Data mining*, SIGKDD Eplor. Vol 4,Issue 2,2002.