An Agent Based Burgeoning Framework for Privacy Preserving Information Harvesting Systems

Gitanajali.J [1], Shaik Nusrath Banu [1],Geetha Mary.A[1] , Indumathi.J *[2],Dr.G.V.Uma [2]
***Corresponding author**

jgitanjali_99@yahoo.co.in, nusrath_fatima97@yahoo.com, erpgeetha@gmail.com ,

[1]Department of Computer Science, Vellore Institute of Technolog,Vellore-632006;
[2]Department of Computer Science and Engineering, Anna University,Chennai – 600 025.
Tamilnadu,India

"All truths are easy to understand once they are discovered; the point is to discover them".
Galileo Galilei.

Summary

Call it by any epithet, nectar of life or life-giver, the precious mined knowledge nugget from raw data means life to all business applications. As the technology soars to an all time high in this information era, the people are in the grip of insecurity panic originating from diverse loopholes owing to the oscillation between privacy & utility loss, spanning from the internet to local networks. We are enforced to scrutinize the skies for the elusive rain- bearing clouds of Privacy Preserving Information Harvesting Techniques and measures. The desiccated users & miners look for the small mercy from the heavens in the form of a framework for Privacy Preserving Information Harvesting Techniques.

In modern times, a ground-breaking consortium of Information Harvesting methods, branded as Privacy Preserving Information Harvesting Techniques (PPIHT), erstwhile developed performs the central objective of protecting sensitive information held in a database from being infringed by a generic database user. The venture of these techniques is the withdrawal of relevant information from colossal amount of data, whilst protecting at the same time sensitive information. A number of Information Harvesting techniques, incorporating privacy protection mechanisms, have been developed that allow one to masquerade sensitive item sets or patterns, at the vanguard of the carrying out of the Information Harvesting process.

In this paper, we present a framework to shore up in the established scrutiny of the database deduction problem. We have specified the burgeon conceptual framework in order to compare and contrast each and every one of the techniques in a general podium which will be the basis for ascertaining the suitable technique for a given type of application. Having studied and examined the existing frameworks, we proposed a Burgeon Framework for Agent Based Privacy Preserving Information Harvesting Systems (ABBPPIHS)model for privacy preserving cooperative mining.We hope the proposed solution will tarmac way for investigation track and toil well according to the evaluation metrics including hiding effects, data utility, and time performance.

***Key Words*:**

Agent Based, Burgeon, fuddle, PPDM Ontology,Privacy Preserving Information Harvesting Techniques,Privacy, Privacy-Boosting Technologies,Privacy-Preserving Data Transformation (PPDT)Selector.

1. Introduction

Information Harvesting or data archaeology or data dredging data mining in databases is defined as the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. Information Harvesting is useful to support both decision-making processes and to promote social goals. The partaking of data has raised a number of ethical issues as those of privacy, data security, and intellectual property rights. These datasets enclose tantalizing personal information, which inexorably gets naked to diverse parties. As a

Manuscript received November 5, 2007
Manuscript revised November 20, 2007

result secrecy issues are persistently under the glare of publicity and the public discontent may well pressurize the employment of Information Harvesting and all its profits. Hence there is a great implication to develop drivable security techniques for defensive secrecy of individual values used for Information Harvesting.

Privacy

Former work [21] recognized that privacy is a acuity which varies from person to person, changes over time and emerges from a society's communication practices[18,15].Additionally, people sustain these privacy perceptions by restraining their accessibility to others [6,11] and these efforts broaden to exercising power over the information that describes them [20]. Cooley defined privacy as simply "the right to be let alone" and later Warren & Brandeis used prominent British cases to show that the common law supported this right [3]. This paper established an understanding of privacy as relevant to individual people and proposed a Privacy-Boosting architecture which incorporates user-defined privacy preferences into operational databases in such a way that the privacy protection it offers is extended to primary and secondary data processing applications[20].Future work was to attempt to demonstrate the viability of the architecture.

The exhort for a Privacy-Boosting Architecture

The impetus for the architecture anticipated here emerges through a deliberation of the role that privacy plays in individual peoples' lives and a epigrammatic survey of both the research and the state of practice with regard to Privacy-Boosting Technologies .The proposed architecture is thus constructed on the privacy legislation in addition to an acknowledgement of individual citizens' privacy preferences and any supplementary privacy constraints necessary by organisations or duty-bound by regulatory bodies.

In this paper we put forward a Privacy-Boosting architecture that enables an approach to protecting the privacy of data. Commentators squabble for stability amid innovation and directive [5, 15, 19, and 22].With respect to secondary data processing applications, the equilibrium to be established is a means of exploiting the data resource while simultaneously complying with applicable legislative requirements and safeguarding the data privacy rights of individuals.

2. Literature Survey

2.1. Privacy Preserving Information Harvesting Techniques (PPIHT)

Innumerable approaches to Privacy Preserving Information Harvesting (PPIH) tools have recently emerged for boosting privacy protection. The PPIH techniques are classified into three categories based upon the approach adopted: heuristics, cryptography or reconstruction.

Classification, Association Rule Discovery and Clustering are all Information Harvesting techniques for which heuristic privacy preservation approaches exist[4].PPIH approaches using cryptography [9, 11, 13, and 16] endow with superior privacy of data chattels in any Information Harvesting perspective where data is at risk of exposure, providing Secure Multi-Party Computation (SMC). Finally, reconstruction approaches [2] apply perturbation followed by aggregation to provide privacy protection.

Allied to privacy preserving in Information Harvesting, but in another direction, Eviffmievski et al. [14] proposed a framework for mining association rules from transactions consisting of categorical items in which the data has been randomized to preserve privacy of individual transactions. Infringe is analyzed with an association network, which consists of the probabilistic dependency structure, the taxonomy structure and the similarity measure. This provides a unified framework for database infringe analysis [27].

The proposed framework in [9] regarded as "knowledge sanitization" approach, first performing sanitization on an itemset lattice called a knowledge base from which association rules can be derived. In [23] the research focuses on further investigating effective knowledge sanitization, data reconstruction based techniques for association rule hiding. Particularly, they have proposed a FP-tree-based method for inverse frequent set mining [12] which can be used in their proposed reconstruction based framework.

The remainder of this paper is organized as follows: Section 2 offers an overview of the related works in framework for Privacy Preserving Information Harvesting, the different problems in Information Harvesting, the existing solutions, and our solution to the problem.Section 3 discusses the problem statement, assumptions, notations used etc., for accomplishing our work..Section 4 presents the block diagram, architectural diagram and the work flow architecture. Section 5 discusses the system architecture design, datasets used, user interface design, and subsystem

architecture. Section 6 discusses about the implementation of the system. Section 7 analyses the results and discusses the results. Section 8 concludes this paper with a brief summary and outlines the future research directions to be carried out.

3. Problem Description

With the intent of enforcing Privacy Preservation we portray a conceptual Agent Based Burgeoning Framework for Privacy Preserving Information Harvesting Systems. We prove that any kind of Information Harvesting can be done securely with this architecture without sacrificing accuracy. Thus, our framework attempts to unearth poise amid privacy and revelation of information by attempting to minimize the impact on the sanitized transactions.

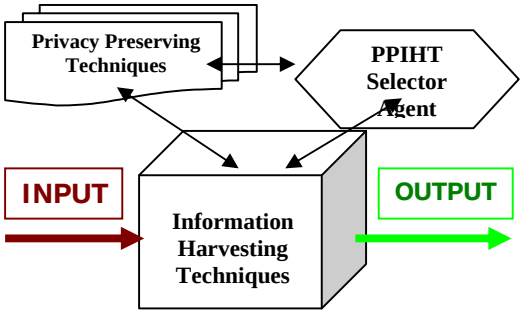

Figure 3.1: An Agent Based Burgeoning Framework for Privacy Preserving Information Harvesting Systems: High-Level

3.1. Problem Statement

The goals of An Agent Based Burgeoning Framework for Privacy Preserving Information Harvesting Systems were to design, develop and implement functionalities like privacy preservation, User friendly framework, Reusability, Portability secure protocol for preserving private data's and knowledge. Specification of a burgeoning framework in order to compare and contrast each and every one of the techniques in a general podium which will be the basis for ascertaining the suitable technique for a given type of application.

3.2. Problem Description

We have to develop mechanism for modifying the unique facts by some means, with the intention that the private data and private knowledge linger private even subsequent to the mining process. There are many mechanisms which have been adopted for privacy preserving data mining. We have the techniques stored in the PPDM Ontology module. PPIHT Selector selects the most apt PPDT apt technique and preserves the privacy of the sanitized data and outputs the fuddled data.

4. Architecture of the Proposed Work

We bring out a diagrammatic schematic representation of the blocks as shown in figure 3. 1 involved in the proposed architecture.

4.1. Block Diagram

We have specified a conceptual framework which can be used to compare and contrast each and every one of the techniques on a general platform which will be the basis for ascertaining the suitable technique for a given type of application.

4.1.1. Original database as shown in figure 4. 1 may be a depiction of a database server or data warehouse server. These datasets and rules may be owned either by a single party or by various parties who are in all probability forbidden from partaking, or not agreeable to dole, their datasets.

4.1.2.PPIHT Framework : Owing to the versatility of the Information Harvesting tasks, a family of privacy-preserving data transformation (PPDT) methods for protecting privacy before data are shared can be used to the address privacy preservation in Information Harvesting. The input to this block is unpreserved data whereas its output is privacy preserved and fuddled data. This is given as an input and is subject to any of the Information Harvesting techniques.

4.1.3. Data Preprocessing - is done in the real world as these data are dirty viz., incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), inconsistent: containing discrepancies in codes or names.

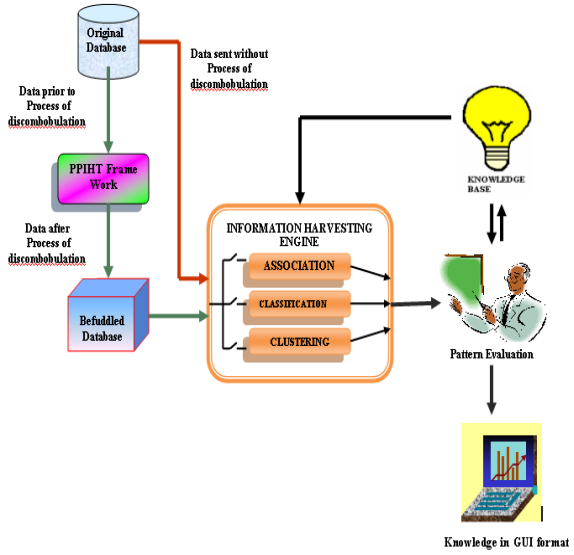

Figure 4.1. Block Diagram of the proposed architecture

4.1.4. Information Harvesting Engine

(i) **Feature extraction:** obtaining only the interesting attributes of the data.
(ii) **Pattern extraction and discovery with the help of knowledge bases:** In this block we extract the interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.

4.1.5.Visualization of the data : The principal graphical techniques used for the visually representing the mined knowledge for further analysis of the information viz.,Box plot,Histogram,MultiVari chart,Run chart,Pareto chart,Scatter plot,Stem-and-leaf plot, tree diagram , bar chart ,pie chart ,function graph,scatter plot, Euler diagram, Venn diagram, existential graph etc.,

4.1.6. Evaluation of results: Based on the visual representations of data we can do decision-making.

5. System Architecture Design of PPIHT

5.1. Data structure design and datasets used

For testing this framework we have utilised datasets of patients which contain sensitive information. We have tested the system on real-time data sets garnered from hospitals in Vellore and Chennai.

5.2. Subsystem Architecture of the Agent Based Burgeoning Framework for PPIH Systems

The architecture of the proposed system has two major subsystems called Privacy Preserving Framework, Information Harvesting Subsystems. Agent Based PPIHT Selector Module of PPDT Methods intelligently decides the best suited technique as it is the routine that waits in the background and performs an action when a specified event occurs.

5.2.1.Privacy Preserving Framework Subsystem

The Privacy Preserving Framework system has eight components, as shown in figure 4. 2 namely, an Analyzer cum filter module, metrics module, Library of algorithms module, retrieval facilities module, technique selector module, user level arbiter module, PPDM ontology module and Strategy module. The Privacy Preserving Framework categorically decides the Privacy preservation mode of selection (manual or automatic or interactive) of technique from the technique selector module, with the help of Library of algorithms module. The metrics module is used to quantify the work. The retrieval facilities module is used to check the usefulness of the data.

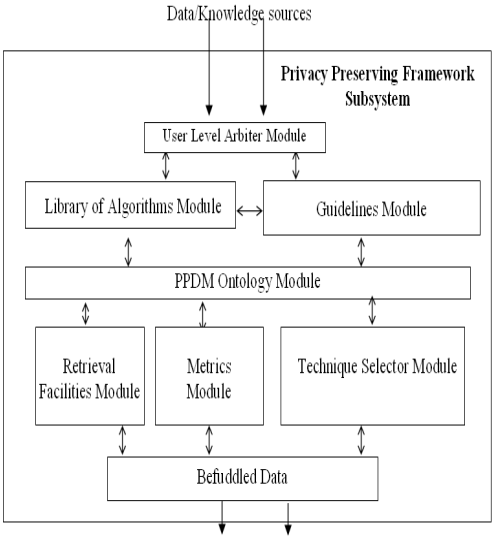

Figure 4.2. Privacy Preserving Framework Subsytem

5.2.2.Agent Based PPIHT Selector Module of PPDT Methods

Agent Based PPIHT Selector Module of PPDT Methods intelligently decides the best suited technique as it is the routine that waits in the background and performs an action when a specified event occurs. According to the level of privacy requires this agent comes to a decision on the potential track of feat to acquire in Privacy preservation of data or Privacy preservation knowledge. The PPDM Technique which is best suited as shown in figure 4.3. for fuddling the data of the particular domain application is suggested to the person.If it is acceptable then they proceed.This is usually prefered by Abceredian users.Advanced user /expert is a person with a high degree of skill in or knowledge of the PPDM selection subject have the highest grade that can be achieved in marksmanship of the subject they know which PPDM technique is suitable for a particular application based on their experience. This can be made manual for them. Middle level users require an interactive system which asks them to answer a set of questions and based on the answer inputs the techniques are semi-automatically chosen.

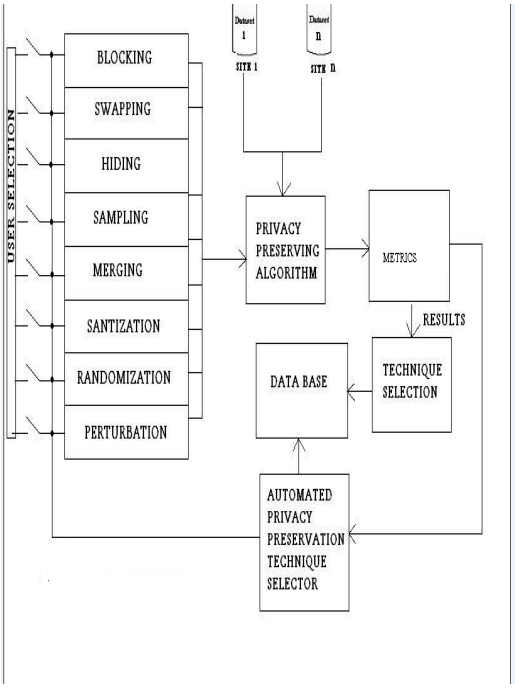

Figure 4.3. Technique Selector

The Technique selector module framework consists of the following procedure:

Step 1: Analyze Input of PPDM;
Step 2: Go to last step if no input data;
(Based on the knowledge level of the inputting person decide to jump to the concerned step or else go to last step

Step 3: If an expertise user of PPDM technique, Choose manual selection path and jump to step 6;
Step 4: If an Abecedarian user of PPDM technique, Choose automatic selection path and jump to step 11;
Step 5: If an intermediate user of PPDM technique, Choose intermediate selection path and answer interactive questionnaire jump to step 12;
Step 6: Choose the best suited PPDM technique manually, if an expertise in PPDM technique for the desired data;
Step 7: Identify the best PPDM technique for the given data;
Step 8: Execute the best PPDM technique on the data to be preserved;
Step 9: Output the befuddled data;
Step 10: Exit the technique selector module; go to step 14;
Step11: Answer the abecedarian questionnaire and say that you need the systems assistance in choosing the best suited PPDM technique for the desired data;
Step12: Perform the selection of the best suited PPDM technique for the application in hand based on the selection decision of the program; go to step 7;
Step13: Repeat step7 - step 10 until all privacy preservation is accomplished;
Step 14: Generate the privacy preserved data as output.

5.2.3. Information Harvesting Subsystem as shown in figure 4.4. has five **c**omponents, namely, Database Server Module, Data Warehouse Server Module, Visualization Module, Data Mining Module and Data Preprocessing Module. Data Mining Module has been confined only to three components, namely, association, classification, clustering.

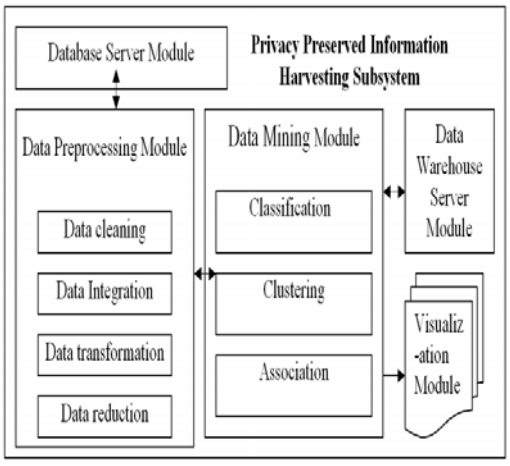

Figure 4.4. Privacy Preserving Framework Subsystem.

6. Implementation

Association Rule Hiding is based on the concept of distance flanked by the original database and its sanitized version, where all sensitive rules have been hidden. By quantifying distance, knowledge is gained with minimum modification that needs to be made in the original dataset in order to hide sensitive, while austerely affecting nonsensitive, itemsets. In this paper, we have endeavored to enhance an existing concealment technique[1] in order to make safe susceptible knowledge from being uncovered in pattern mining. By hiding the sensitive frequent itemsets that lead to the production of the association rules, we are able to secure the sensitive knowledge and minimize the side effect on the quality of the sanitized database so that non-sensitive knowledge can still be mined. They have used the Apriori algorithm to compute the large itemsets, which is less efficient. In this paper, we have used to harness the advantages of Frequent Pattern Growth Method which mines the complete set of frequent itemsets without candidate generation. The investigational appraisal shows that this modus operandi can yield good results on real world datasets, demonstrating its effectiveness towards solving the problem with good data utility, privacy and performance.

Figure 6.2.1 shows the screenshot which prompts the user to fill the values of expected maximum item set and threshold value. After the data file and product file are loaded by the user it displays the association rules. From this selected screen the users are allowed to select the sensitive item sets which form the association rule and to add sensitive itemset which are not listed in the list box.

Figure 6.2.2 displays the association rules of the original dataset and association rules of modified dataset. It helps the user easily, visualize and compare the association rules of the original and modified dataset.

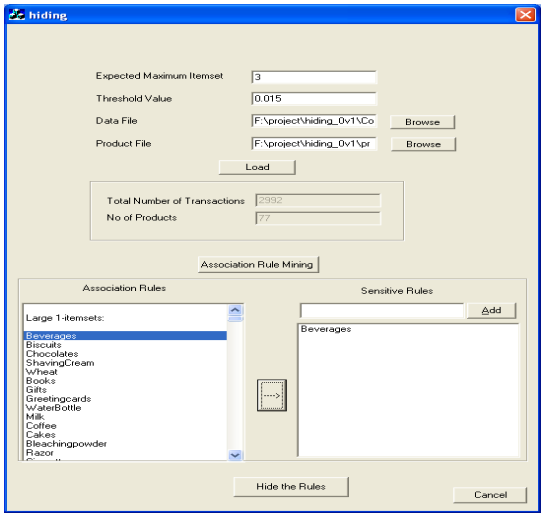

Figure 6.2.1. Association rule and sensitive itemsets

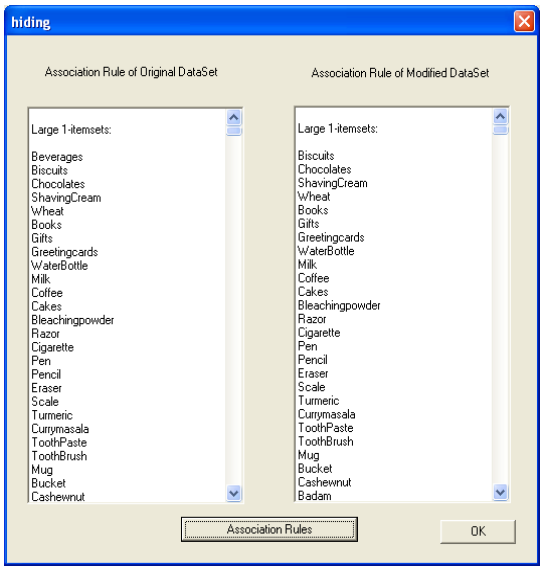

Figure 6.2.2. Final output

7. Results and Analysis

Data Utility is the percentage of similarity between the data mined results from original data and concealed data.

7.1. Privacy Analysis

Figure 7.1 shows the degree of privacy that can be achieved using this algorithm. As seen from the figure we can note that the degree of privacy can be increased as we increase the minimum support value. Based on

the support count value, the number of transaction to be modified is decided.

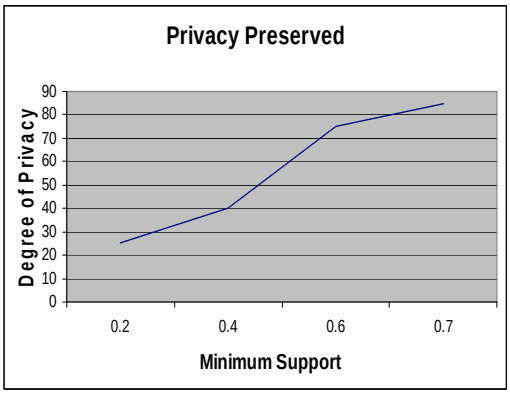

Figure 7.1 Degree of Privacy

7.2.Error Analysis

The graph as in Figure 7.2 shows the expected error percentage in comparison with the existing privacy preserving algorithms. In this graph we can see that the error percentage during rule mining using the integer programming technique is much less than the other privacy preserving algorithms, which is an improvement.

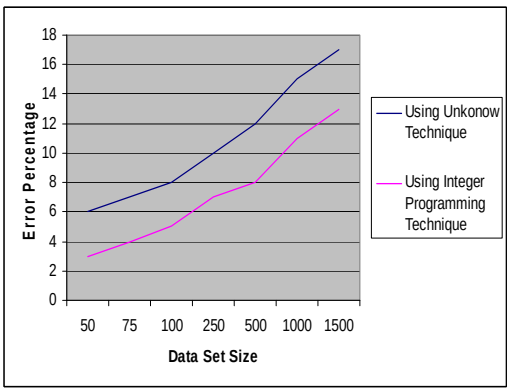

Figure: 7.2 Error Percentage

7.3. Data Utility Analysis

The graph as in Figure 7.3 shows the percentage of accuracy that can be achieved using the proposed method compared to the existing method. Here we can see that the percentage of accuracy that can be achieved using the proposed using the integer programming technique is higher than the other methods, which is an enhancement.

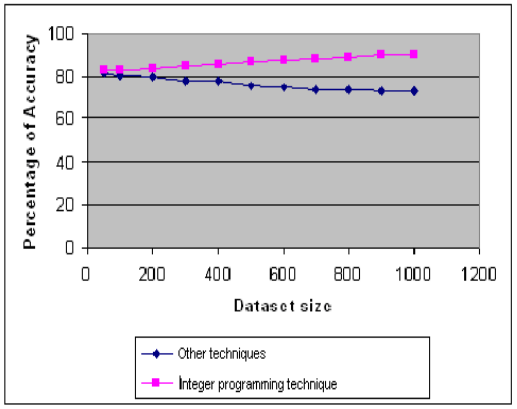

Figure: 7.3 Percentage of Accuracy.

7.4. Performance Analysis

The above graph as in Figure 7.4 shows the expected run time of the clustering algorithm increases with an increase in the size of the data set. The proposed algorithm takes relatively extra run time while comparing with other hiding techniques, which requires our attention.

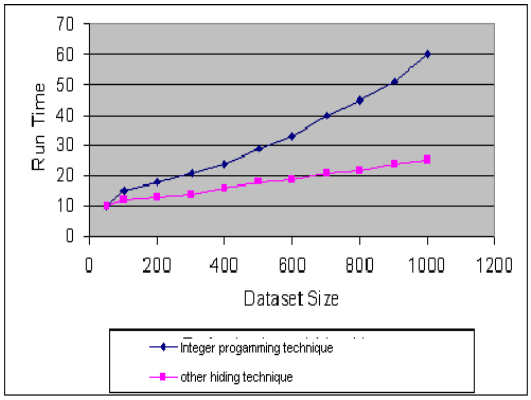

Figure. 7.4 Run time

8. Conclusion and Future Work

Privacy and Efficiency are the two eyes of a person, equally imperative for safe Information Harvesting. Compromising on both is not sensible. Hence we have suggested a framework for Privacy Preserving Information Harvesting. We need to implement and evaluate true efficiency, after including improvements such as sampling.

Future work will attempt to demonstrate the viability of the architecture through a proof-of-concept prototype. We demonstrate how the masking-integer programming technique can be effectively done using this architecture. In the future, we hope to perk up the efficiency of this approach. As a first direction, we sketch to investigate firmly generating diplomat samples from the database. This would be an orthogonal technique for applications not requiring perfect accuracy, but very high security. We hope the proposed solution will get hold of new frameworks, techniques, paving way for research track and work well according to the evaluation metrics including hiding effects, data utility, and time performance.

9.References

[1].A.Aris Gkoulalas-Divanis and Vassilios S. Verykios ,"An Integer Programming Approach for Frequent Itemset Hiding ", CIKM'06 , November 2006.

[2].Agrawal, R. & Srikant, R. (2000): Privacy-preserving Information Harvesting. Proc. of the ACM SIGMOD Conference on Management of Data. Dallas, Texas, US, 439-450, ACM Press.

[3].Akdeniz, Y., Clarke, O., Kelman, A. & Oram, A. (1997): Cryptography and Liberty: Can the Trusted Third Parties be Trusted? A Critique of the Recent UK Proposals. The Journal of Information, Law and Technology.

[4].Atallah, M. J., Bertino, E., Elmagarmid, A. K., Ibrahim, M. & Verykois, V. S. (1999): Disclosure limitation of sensitive rules. Proc. of the IEEE Knowledge and Data Engineering Workshop. Chicago, Illinois, US, 45-52,IEEE Computer Society Press.

[5].Baase, S. (1997): A gift of fire: social, legal, and ethical issues in computing, New Jersey, Prentice-Hall.

[6].Cavoukian, A.: Information Harvesting: staking a claim on your privacy. http://www.ipc.on.ca/. Accessed 27 Apr 2000.

[7].Chen, X., Orlowska, M., and Li, X. A new framework for privacy preserving data sharing. In: Proc. of the 4Pth P IEEE ICDM Workshop: Privacy and Security Aspects of Information Harvesting. IEEE Computer Society, 2004. 47-56.

[8].Clifton, C., Kantarcioglou, M., Lin, X. & Zhu, M. Y. (2002): Tools for privacy preserving distributed Information Harvesting. SIGKDD Explorations 4(2).

[9].Du, W. & Atallah, M. J. (2001), Secure Multi-Party Computation Problems and their Applications:A Review and Open Problems, in `10th ACM/SIGSAC New Security Paradigms Workshop',Cloudcroft, New Mexico, pp. 13-22.

[10].Evfimievski, A., Srikant, R., Agrawal, R. & Gehrke, J. (2002), Privacy Preserving Mining of Association Rules, in `8th ACM SIGKDD International Conference on Knowledge Discovery and Information Harvesting', Edmonton, AB, Canada, pp. 217-228.

[11].Gavison, R. (1984): Privacy and the limits of law. Yale law journal 89: 421-71.

[12].Guo, Y.H., Tong, Y.H., Tang, S.W., and Yang, D.Q. A FPtree-based method for inverse frequent set mining. In: Proc. of the 23P rd P British National Conf. on Databases (BNCOD'06). LNCS 4042, Springer-Verlag. 2006. 152-163

[13].Ioannidis, I., Grama, A. & Atallah, M. (2002): A secure protocol for computing dot-products in clustered and distributed environments. Proc. of the International Conference on Parallel Processing. Vancouver, British Columbia, Canada, 379-384, IEEE Computer Society Press.

[14].Kirsten Wahlstrom & Gerald Quirchmayr the Australasian Information Security Workshop: Privacy Enhancing Technologies (AISW), Ballarat, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 68. The motivation and proposition of a Privacy-Boostingarchitecture for operational databases

[15].Kizza, J. M. (1998): Ethical and social issues in the information age, Springer-Verlag Inc, New York, US.

[16].Lindell, Y. & Pinkas, B. (2002): Privacy Preserving Information Harvesting . Journal of Cryptology 15(3):177-206.

[17].LiWu Chang and Ira S. Moskowitz, 2001An Integrated Framework for Database Privacy Protection.

[18].Rachels, J. (1975): Why privacy is important. Philosophy and public affairs 4(4):323-333

[19].Registratiekamer (Netherlands) & Information and Privacy Commissioner (Ontario, Canada): Privacy-BoostingTechnologies: the path to anonymity, vol 1.http://www.ipc.on.ca/. Accessed 19 February 2002

[20].Tavani, H. (1996); Computer matching and personal privacy: Can they be compatible? In Proc of the Symposium on Computers and the Quality of Life. Philadelphia, Pennsylvania, US, 97-101, ACM Press.

[21].Wahlstrom, K. & Roddick, J. (2000): On the impact of Knowledge Discovery and Information Harvesting. Conferences in research and practice in information technology:Second Australian Institute of Computer Ethics Conference (AICE2000). 1:22-27

[22].Weckert, J. & Adeney, D. (1997): Computer and information ethics, Greenwood Press, Connecticut, US.

[23].Yuhong Guo , In Reconstruction-Based Association Rule Hiding Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative DatabaseResearch 2007(IDAR2007), June 10, 2007, Beijing, China.

ACKNOWLEDGMENTS

For my mother Ms.A.Indirani who pinned her ears back ever so patiently as I discussed my work though she rarely understood it. For my guide Dr.G.V.Uma, for challenging me, supporting me, and removing my training wheels.

AUTHORS PROFILE

Gitanjali.J received her B.Com and M.C.A. from Sri Venkateswara University, Andhra Pradesh, India in year 2002 and 2005 respectively. At present she is a student in Information Technology at Vellore Institute of Technology, Vellore, India. She is currently doing her M.Tech from Vellore Institute of Technology, Vellore. Her field of interest spans Mobile Computing, Data Mining, and Networking.

Shaik Nusrath Banu received her B.Tech. from JNTU, Andhra Pradesh, India in year 2006. At present she is a student in Information Technology at Vellore Institute of Technology, Vellore, India. She is currently doing her M.Tech from Vellore Institute of Technology, Vellore. Her field of interest spans Mobile Computing, Data Mining, and Networking.

Geetha Mary A received BE degree in computer science from University of Madras in 2004 and Studies MTech Computer science and Engineering at Vellore Institute of Technology (VIT). Her research interests include networks, security, mobile computing and data mining.

Indumathi.J received her M.E. from Anna University, Chennai, India in year 1992 and M.B.A from Madurai Kamaraj University, Madurai, India in 1994. She is working for Anna University as a Senior Lecturer. She is currently doing her Ph.D from Anna University, Chennai. Her field of interest spans and is not limited to Computer Science and Financial Management. Her research interests include Security for Data Mining, Databases, Networks, Computers, Software Engineering, Software Testing, Project Management, Biomedical Engineering, Genetic Privacy and Ontology.

G.V.Uma, a Polymath received her M.E. from Bharathidasan University, India in year 1995 and Ph.D. from Anna University, Chennai, India in 2002. She is working for Anna University as a Assistant Professor. Her research interests include Software Engineering, Genetic Privacy, Ontology.Knowledge Engineering & Management, and Natural Language Processing.