Asian Journal of Applied Science and Technology (AJAST)
(Open Access Quarterly International Journal) Volume 2, Issue 1, Pages 28-33, 2018

Design and Implementation of Data Refinement Technique Using Medical Support System in Cloud Architecture

S.Acharimonica[1], M.Jeevarathinam[2], M.Gayathri[3] and Dr.M.Ramesh Kumar[4]

[1,2,3]*UG Students, Department of Computer Science and Engineering, VSB College of Engineering Technical Campus, Coimbatore, Tamilnadu, India.*
[4]*Associate Professor, Department of Computer Science and Engineering, VSB College of Engineering Technical Campus, Coimbatore, Tamilnadu, India.*

Article Received: 21 September 2017 *Article Accepted: 23 December2017* *Article Published: 07 January 2018*

ABSTRACT

The emergence of the Cloud has represented a fundamental change in the way information technology services are designed and deployed in business and governments and there is a growing trend of using cloud environments for storage and data processing needs .However, this environment represents a serious threat for data privacy, since document containing confidential information might be made available for unauthorized parties. Although measures to automatically remove or hide sensitive information that may disclose identities of referred entities or reveal their confidential data of publicly available document have been purposed. But there is no big implementation regarding security using refinement method was implemented for data stored in cloud server. We are developing this Project for Medical Purpose .Here we use the Cloud Server as a main Server, where all the Data from the Users are Stored. We design this system using Registered Doctors, Paid and unpaid users. Here document is sanitizing dynamically so that different users get different view of same document. Data Refinements achieved by Three Processes. 1. Entity Generalization-Preserving the Privacy data with its semantics. 2. Entity Swapping is used to reduce the Document Size. 3. Noise Addition: an entity substituted by another similar one extracted from another repository.

Keywords: Natural Language Processing, Patient Data Privacy, Data Refinement and Sensitive information.

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.1 Data Refinement Techniques

1.1.1 Technique: NULL'ing Out

Simply deleting a column of data by replacing it with NULL values is an effective way of ensuring that it is not inappropriately visible in test environments. Unfortunately it is also one of the least desirable options from a test database standpoint. Usually the test teams need to work on the data or at least a realistic approximation of it. For example, it is very hard to write and test customer account maintenance forms if the customer name, address and contact details are all NULL values.

1.1.2 Technique: Masking Data

Masking data means replacing certain fields with a Mask character (such as an X). This effectively disguises the data content while preserving the same formatting on front end screens and reports.

1.1.3 Technique: Substitution

This technique consists of randomly replacing the contents of a column of data with information that looks similar but is completely unrelated to the real details. For example, the surnames in a customer database could be sanitized by replacing the real last names with surnames drawn from a largish random list.

1.1.4 Technique: Shuffling Records

Shuffling is similar to substitution except that the substitution data is derived from the column itself. Essentially the data in a column is randomly moved between rows until there is no longer any reasonable correlation with the remaining information in the row.

1.1.5 Technique: Number Variance

The Number Variance technique is useful on numeric data. Simply put, the algorithm involves modifying each number value in a column by some random percentage of its real value. This technique has the nice advantage of providing a reasonable disguise for the numeric data while still keeping the range and distribution of values in the column within viable limits.

1.1.6 Technique: Gibberish Generation

In general, when sanitizing data, one must take great care to remove all imbedded references to the real data. For example, it is pointless to carefully remove real customer names and addresses while still leaving intact in stored copies of correspondence in another table. This is especially true if the original record can be determined via a simple join on a unique key.

2. PROPOSED SYSTEM

To overcome this drawback, we are developing this Project for Medical Purpose. Here we use the Cloud Server as a main Server, where all the Data from the Users are Stored. We design this system using Registered Doctors, Paid and unpaid users. Here document is sanitize dynamically so that different users get different view of same document. Data Refinement is achieved by Three Process that's hide the sensitive entities instead of removing it as to preserve the utility of the document. 1. Entity Generalization-Preserving the Privacy data with its semantics. 2. Entity Swapping is used to reduce the Document Size. 3. Noise Addition: an entity substituted by another similar one extracted from another repository.

3. ARCHITECTURE DIAGRAM

The paid users and the doctor are register to access the cloud server where the information is stored. When the paid user requests some information to the doctor, the sanitized information will be available to the users. The refinement is performed by the following three refinement techniques. These techniques hide the sensitive terms instead of removing them. So, it preserves the utility of the information. Moreover, it allows the user to configure

the level of refinement applied to the document being more flexible than methods based on fixed refinement policies.

3.1 Entity Generalization: entities can be generalized to achieve some degree of privacy while preserving some of their semantics.

3.2 Entity swapping: entities of different documents of the same set or within the same document can be swapped depending on the concrete case

3.3 Entity noise addition: an entity can be substituted by another similar one extracted from another repository.

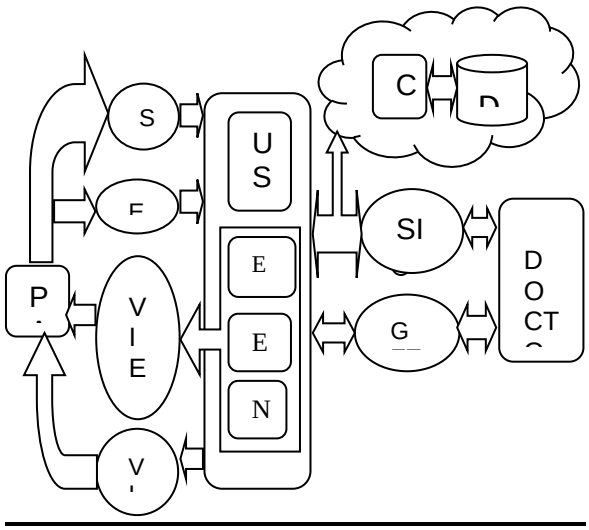

Fig 3.1 System Architecture

4. MODULES

4.1 Users & Doctors

Users and doctors are persons who are going to access the information Stored in the Cloud Server. To implement this module we are going to implement the User Interface frame to send the User's and Doctors request. The Doctors are the persons who are going to provide the suggestions for the User requested queries. So that they have to register their details in the Cloud Server. Once registered the data of the both Users and Doctors are Stored in the Cloud Server's Database.

4.2 Cloud Server

The Cloud Server will maintain the Data of the User and Doctors information stored in the Cloud Server. At the mean while the Cloud Server will monitor the access information of the Cloud Network. So that they can protect the Network from the Unwanted Users. Also the Cloud Server will retrieve the data for the User requested Query.

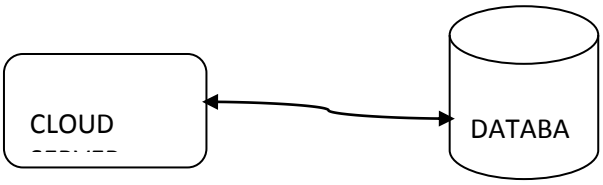

Fig 4.1: Cloud Server

4.3 Entity Generation

In this phase of the project, we are implementing an mechanism to get the similar results for the User Entered Query based on the Semantics. Semantics is nothing but similar information regarding the User entered queries. So that the User can get more information about the query that they are surfing for.

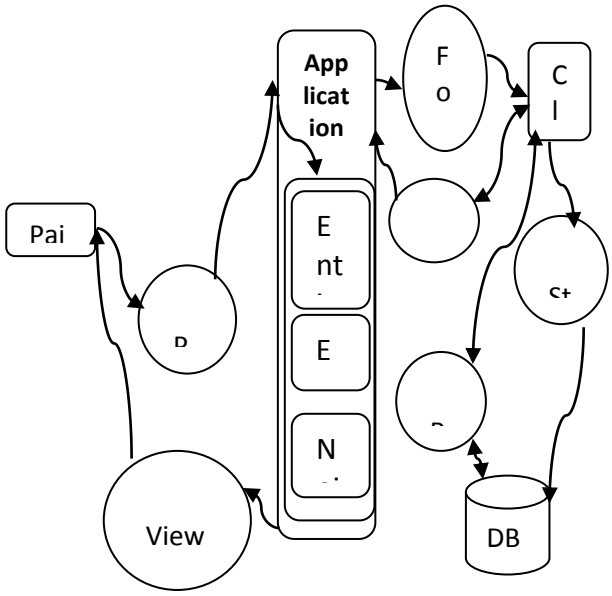

Fig 4.2: Refinement Technique (Entity Generation, Entity Swapping and Noise Addition)

4.4 Entity Swapping

In this module, we are implementing Swapping mechanism to reduce to document size. So that the User can get the documents with precise information. Swapping will replace document. Entities of different documents of the same set or within the same document can be swapped depending on the concrete case.

4.5 Noise Addition

In this module an entity can be substituted by another similar one extracted from another repository. So that the User can get more information about their entered query. In the Case of noise addition we are able to retrieve more information regarding for the User's requested.

4.6 Authentication and Data Retrieval

Once the registered User send the query to the Cloud Server, they Server will transfer the query to the registered Doctors only after they are authenticated by the Cloud Service Provider. So that any unauthorized Users are not allowed to provide any information via the Cloud Server. This will increase the Security Level in this Cloud Network.

5. CONCLUSION

Since cloud computing is the vast developing technology, security is the major in the cloud environment. To overcome this drawback many existing approaches has been introduced but they have not fulfilled the security issue. At this situation storing medical records in the cloud environment is the major issue. Because the data will be hacked (corrupted) modified by the unauthorized person in the network. So we need new mechanism need to be implemented in the cloud environment for storing and access the medical records stored in the cloud servers. By implementing this project we can allow the authorized doctors to enter into the network and respond to the queries submitted by the registered patient in the cloud network using sanitation mechanism so that we get more result about the query that the user is enter. In future we allow the registered doctors, nurse and pharmacist based on the attributed based encryption scheme to view the records so that we can increase the security level and we can also try to improve the way of detection of sensitive information.

REFERENCES

[1] L. Sweeney, “Replacing personally-identifying information in medical records, the scrub system,” in Proc. 1996 American Medical Informatics Association Ann. Symp., 1996, pp. 333–337.

[2] L. Sweeney, Computational Disclosure Control: A primer on data privacy protection. Ph.D. Thesis, Massachusetts Institute of Technology, 2001.

[3] L. Sweeney, “K-anonymity: A model for protecting privacy,” Int. J. Uncertainty, Fuzziness and Knowledge-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[4] A. Tveit, O. Edsberg, T. B. Rost, A. Faxvaag, O. Nytro, M. T. Nordgard, M. T. Ranang, and A. Grimsmo, “Anonymization of general practioner medical records,” in Proc. Second HelsIT Conf., Trondheim, Norway, 2004.

[5] Nat. Security Agency, RedactingWith Confidence: How to Safely Publish Sanitized Reports Converted From Word to pdf, Tech. Rep. I333-015R-2005, 2005.

[6] M. M. Douglass, G. D. Cliffford, A. Reisner, W. J. Long, G. B.Moody, and R. G. Mark, “De-identification algorithm for free-text nursing notes,” Proc. Computers in Cardiology’05, pp. 331–334, 2005.

[7] D. A. Dorr, W. F. Phillips, S. Phansalkar, S. A. Sims, and J. F. Hurdle, "Assessing the difficulty and time cost of de-identification in clinical narratives," Methods Inf. Medicine, vol. 45, no. 3, pp. 246–252, 2006.

[8] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania, "Efficient techniques for document sanitization," in Proc. ACM Conf. Information and Knowledge Management'08, 2008, pp. 843–852.

[9] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: A review of recent research," BMC Med. Res. Methodol., vol. 10, pp. 70–86, 2010.

[10] D. Sánchez, M. Batet, A. Valls, and K. Gibert, "Ontology-driven web-based semantic similarity," J. Intell. Inf. Syst., vol. 35, no. 3, pp.383–413, 2010.

[11] S. K. Dash, R. Mishra, D. P. Mishra, and A. Tripathy, "A privacy preserving repository for securing data across the cloud," in Proc. 3rd Int.Conf. Electronics Computer Technology, 2011, vol. 5, pp. 6–10.

[12] S. Marston, Z. Li, S. Bandyopadhyay, A. Ghalsasi, and J. Zhang, "Cloud computing the business perspective," Decision Support Syst., vol. 51, no. 1, pp. 176–189, 2011.

[13] D. Abril, G. Navarro-Arribas, and V. Torra, "On the declassification of confidential documents," in Proc. Modeling Decisions for Artificial Intelligence'11, 2011, pp. 235–246.

[14] C. Cumby and R. Ghan, "A machine learning based system for semiautomatically redacting documents," in Proc. 23rd Innovative Applications of Artificial Intelligence Conf., 2011, pp. 1628–1635.

[15] National Security Agency, Redaction of pdf Files Using Adobe Acrobat Professional X 2011 [Online]. Available: http://www.nsa.gov/ia/ files/vtechrep/I73 025R_2011.pdf

[16] B. Anandan and C. Clifton, "Significance of term relationships on anonymization," in Proc. Web Intelligence/IAT Workshops'11, Lyon, France, 2011, pp. 253–256.

[17] U.S. Department of Justice, U.S. Freedom of Information Act (FOIA) 2012 [Online]. Available: http://www.foia.gov/

[18] S. Martínez, D. Sánchez, A. Valls, and M. Batet, "Privacy protection of textual attributes through a semantic-based masking method," Inf. Fusion, vol. 13, no. 4, pp. 304–314, 2012.