Publication of XML documents without Information Leakage with data inference

Smita Chaudhari[1], Sonali Patil[2]
1 PG Student, Dept of Computer Engineering, Alard College of engineering and management,
Sivitribai Phule Pune University, Pune
2 Professor, Dept of Computer Engineering, Alard College of engineering and management,
Sivitribai Phule Pune University, Pune

Abstract- Recent applications are using an increasing need that publishing XML documents should meet precise security requirements. In this paper, we are considering data publishing applications where the publisher specifies what information is more sensitive and should be protected from outside world user. We show that if a given document is published carelessly then users can use common knowledge to guess any information. The goal here is to protect such information in the presence of data inference with common knowledge. The most important feature of XML formatting is it allows for adding schema declarations with integrity constraints to instance data and allow composing individual pieces of data in a tree-like fashion in which a link from a parent node to a sub tree carries some ontological information about the relationship between individual pieces of data

This system work as inference problem in XML documents consists of potentially secrets and important information. Our work gives solution for this problem by providing the control mechanism for enforcing inference usability of XML document. Output of our work is again a XML document that is under their inference capabilities which neither contain nor imply any confidential information and it is indistinguishable from the actual XML document. In the proposed work it produced the weaken document which takes the consideration of inference capabilities and according to this modifies there schemas and produce inference proof documents.

Keywords:- *XML document, XML schema, DTD, XSD, Inference control, Inference rule.*

*****

I. Introduction

With the fast development of the Internet, there is an increasing amount of data published daily on the Web. Meanwhile, recent database applications see the emerging need to support data sharing and dissemination in peer-based environments. These data may contain the confidential data as well as some secure information regarding some organization which will be very delicate information for that particular organization while passing this information from one organization to another organization. There must not be leakage of such information or inference of the data associated with the document. The aim of inference control is to protect an unauthorized client from guessing this delicate information, whether directly or indirectly. Previous work has shown that to achieve this aim for an information system, for example a relational database which is necessary to control the inference channels in the information set of the system. As XML documents have more difficult structures than the relations, there may exist more fraud inference channels in XML documents. In these applications, the owner of a data source needs to publish data to others such as public users on the Web or collaborative peers. Often the data owner may have sensitive information that needs to be protected, if we publish data carelessly, users can use common knowledge to infer more information from the published data, causing leakage of sensitive information.

The aim of inference control is to prevent an unauthorized client from inferring sensitive information from xml datasets. Previous research has shown that to achieve this goal form an information system, a relational database and which is necessary to control the inference channels in the information set of the system. We propose algorithms for computing a partial document to be published without leaking of information.

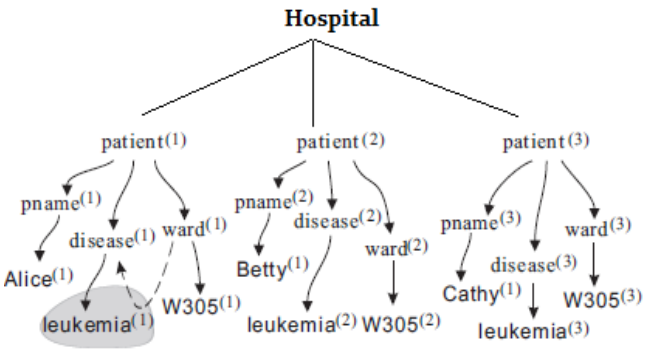

Fig: 1 Example an XML document of hospital data.

Consider example (Fig.1) a hospital at a medical school has XML documents about its patients and physicians. Such an XML document represented as a tree. Each patient has a name suffers from a disease (a disease element) and lives in a ward. Each physician has a name (phname), and treats patients identifies by their names. For instance, physician Smith is treating patient Cathy, who has leukemia and lives in ward W305.

The hospital plans to provide the data to another department at the same school to conduct related research. Some data is sensitive and should not be released. In particular, the hospital does not want the department to know the disease of patient Alice (leukemia) for some reason. One simple way is to hide the sub tree of Alice. But if it is well known that patients in the same ward have the same disease, then this common knowledge can be used by the department users to infer from the scene documents that Alice has a leukemia. Alice has a leukemia. It is because Alice and Betty live in the same ward, One solution to this information-leakage problem is that, in addition to hiding the leukemia branch of Alice, we also hide the Alice branch, so that users do not know the name of this patient.

II. Motivation

We use two stages for giving inference proof view at first stage we required such types of XML documents which is separate each contents and particular data of any organization ex. Hospital data used in above example. At second stage we required to give security view for that document while shearing it at one organization to other organization. We present in this method for generate an inference-proof view by weakening the actual XML document, eliminating confidential information and other information that could be used to infer confidential information. The inference control should make the client believe in the possibility that there is an alternative XML document that does not contain any potential secret [11]. In particular, the client should not be able to distinguish the actual document from the alternative one according to his queries.

We use weakening technique to construct the alternative XML document. The actual XML document is modified according to the weak operations and potential secrets given by organization. While previous work in this area [11] has mostly focused on the inference of Document Type Definitions (DTDs for short), we will consider the inference of XML Schema Definitions (XSDs for short).While we are using DTD the content of the elements depends on name in case of XSD depends on context itself. So the complexity of algorithm will reduce. Using XML Schema it increase their expressiveness and analyse the high complication of the algorithm The system generate effective inference control under the inference capabilities of a client.

III. Literature survey

Barceló and Abiteboul et al. [7] have worked on XML documents with incomplete information. A user might infer that an element in an XML document has an attribute with a certain value by employing integrity or semantic constraints on the XML document. Assuming this algorithm for generating an inference-proof XML document should be different from the algorithm for generating an inference-proof relation. He proved using Controlled query evaluation (CQE) an effective way to enforce inference control on information sets.

In order to protect data in XML documents, traditional access control policies, like DAC and MAC, are enforced on these documents [1], [2], [3], [4]. The smallest piece of data protected by these access control policies is an XML node. If a node is sensitive (i.e., it is not allowed to be disclosed according to the access control policy), it must be eliminated from the answers to queries.

Biskup and Wiese [9] constructed an static inference-proof database using CQE. Here they mianly worked on complete information system whereas we focus on incomplete XML documents they use lying to protect confidential information that is totally different from the weakening used in our paper. We can describe some pieces of information contained in an XML document.

Their work also focused on generating a closure of authorized data by checking the XML (inference) relations forwardly to determine [6] whether a query of a client should be refused or answered that is what different from our aim.

Meghdad Mirabi[12] Access control plays an important role to prevent unauthorized users to access private data. In traditional access control mechanisms, access authorizations are specified to the whole of entities such as tables and views.

This kind of access control is called coarse-grained access control. In coarse-grained access control mechanism, CBSAM compresses the accessibility map with minimum affect on the XML query processing when the access locality among the XML nodes is high. Accordingly, any attempt to securely handle the information contained in an XML document has to take into account the applicable schema declaration, the individual pieces of data used to compose the XML document and the link structure between them.

IV. Problem Formulation

Below figure (Fig.2) shows the Architecture for generate inference proof view. Generate inference proof view which generates weakened XML document, according to the user's desire at run time. The system uses set R of inference rules, the set P of potential secrets, and generate the inference proof XML document

We propose a formal notion of an inference-proof view of an XML document to meet the requirements of our goal of effective inference control under the inference capabilities of a client. The idea of generating the new XML Schema is to

modify or alter the some expressions of the original XML schema after every new kind of modification or changes to the content of the XML document. Consider the algorithm needs to weaken the XML document T for a path and the algorithm will modify some expressions of the original XML Schema D according to the type of the document that we have to improve to secure the content of the document. Our proposed architecture and algorithm for generating inference-proof view will eliminate all the important information and some information part that can be used to infer means (guess) some confidential information from an XML document by weakening the document step by step. We can use XSD also to produced weaken documents because schema enables many optimizations for operations on XML documents.

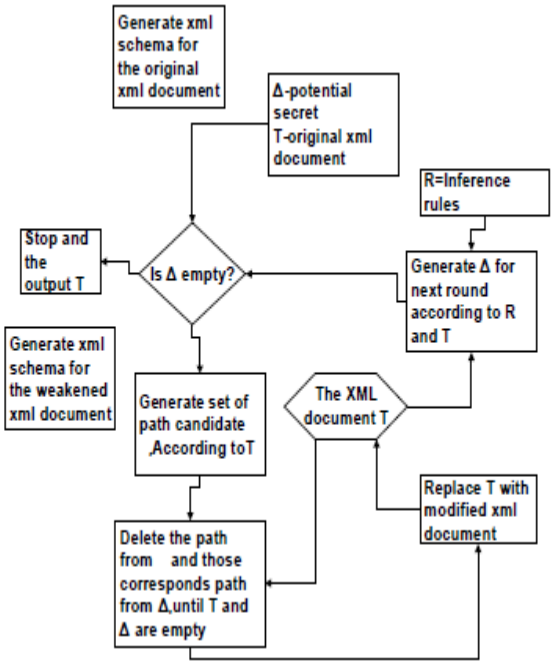

Fig. 2 Architecture for generate inference proof view

V. Methodology

Our proposed system first feature distinguishes how a client relates the information set managed by an information system to the actual situation in (what he perceives to be) the real world data. First Assuming completeness the client[8], considers the information set to contain all the information that is true in the (relevant part of the) real world depends on he holds those information not contained in the information set to be false[11]. In many applications it happens that a client will restrict himself to assume incompleteness the information set managed by the information system contains information considered to be true and nothing is postulated about information that is not contained in the information set. For example the information set given by the XML documents maintained by a hospital, like that is the hospital may not know all the diseases of all patients .here in our proposed work we focus on inference control for XML documents with incomplete information with the following properties

1. For some of its elements XML document may not include the attributes that are related to the information set given by the XML document but unknown to or hidden intentionally by the information system managing the XML document.

2. For some of its elements XML document may not include the child or descendant relations that are related to the information set given by the XML document but unknown to or hidden intentionally by the information system managing the XML document.

For some of its path instances and keeping the descendant relations of the elements in the path instances XML document may not include some intermediate elements in some path instances that are related to the information set given by the XML document but unknown to or hidden intentionally by the information system managing the XML document [5]. So any sub element of an element in an incomplete XML document might actually just be a descendant of the element in the real world.

We enhance the algorithm for generating an inference proof view by using the XML schema over Document Type Definition. The inputs include an XML document and its schema, potential secrets, and the inference capabilities The idea of generating the new XML Schema is to modify some expressions of the original XML schema after every new kind of modification to the content of the XML document. Suppose the algorithm needs to weaken the XML document T for a path and the algorithm will modify some expressions of the original XML Schema D according to the type of .

1) Restricted Inference Rules: The restrictions inference rules follow path ψ with some constraints Every variable x occurs only once in ψ , i.e., there is one and only one path node

For every path node pn ∈ ψ: PNE there is at most one path node pn' such that pn, pn' ∈ ψ. child and pn' ∈ ψ we can decrease the complexity of our modified algorithm. The set XD = x1, x2, x3, x4,…..x_n where it shows all the XML. Consider set of Inference rules where it is denoted as RD = r1, r2, r3, r4……R_n where it is the set of inference rules used to weaken the documents to protect the confidential data of the

XML document (O (p | T |) p^{T} j) to (O(p | T |) $^{p|R|}$), where | T | is the size of the document and | R | is the size of the set of inference rules. It shows that the algorithm can efficiently generate the inference-proof view

5.1 Advantages of using XML schema

1. The major advantage of schemas is their ability to more strongly type the data in XML documents. Schemas are described using XML instead of the archaic form used by

DTDs. Schemas also provide a richer approach to describing complex XML types.

2. Document structure can be described more accurately with schemas as well, using features such as minOccurs and maxOccurs to specify the number of times an element can occur within a particular context.

3. The Introduction of Algorithm Complexity Analysis to increase their expressiveness and analyze the high complexity of the algorithm because of the variables. We model the inference capabilities of a client by known information entailed by a XML Schema and the XML document, and the semantic implications between pieces of information contained in the XML document.

5.2 Confidentiality Policies

Confidentiality policies tell what is protected by inference control. In this paper, we protect the fact that an XML document contains some particular pieces of information, and we call those pieces potential secrets.

5.3 Interaction Sequences

The interactions considered in this paper contain the submissions of queries and the answers to these queries. The client could continue submitting queries over the XML documents [12], and the answers to these queries should be always controlled according to the pertinent confidentiality policy and authorization policy.

VI. User Study and Experimental Setup

6.1 Alternative XML Document:

Although the actual stored XML document may contain some potential secrets, the inference control should make the client believe in the possibility that there is an alternative XML document that does not contain any potential secret. In particular, the client should not be able to distinguish the actual document from the alternative one according to his queries. Hence, inference control aims at ensuring the existence of an alternative XML document or even at constructing such an alternative XML document as an inference-proof XML view on the actual XML document.

We have developed the alternative XML document using information weakening system using JDK 1.8 and Swing Technology. Swing provides a native look and feel and also supports a pluggable look. The database used for our system is MySql. Project Experiments were run using a machine with Intel Core i3 CPU running on Windows 7.

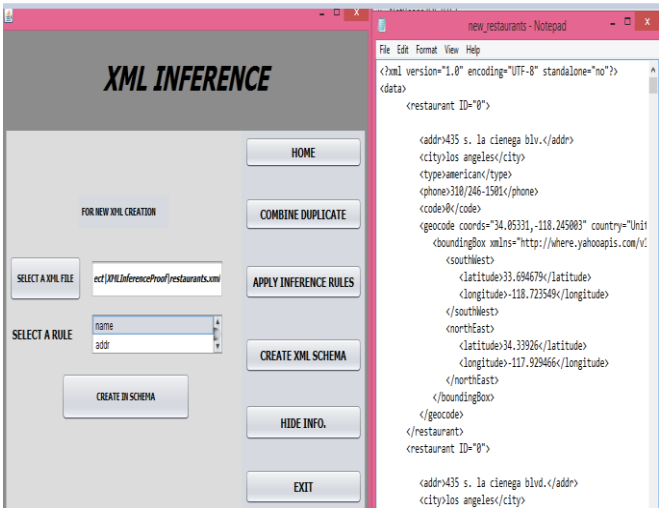

Fig.3 Result

VII. Results

The comparison table of our proposed system with existing system is shown in below table. The table shown below is the comparison of various points with our proposed work. It clearly shows that our system is better as compared to previous systems, As our system can work with secure data and also we have added more confidential policies search mechanism at runtime which improves our work.

Existing Syste	Proposed System
Inference rules algorithm	Updated inference rules algorithm.
Less inference rules applied to propose.	Improve the applied rules count.
Less Accuracy	Accuracy is high
Take more time to propose.	In less time it produces.

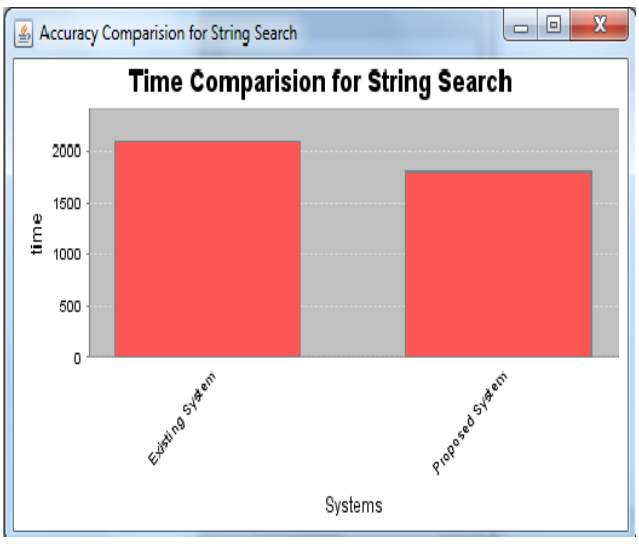

Fig.4 Comparison for existing and proposed system

In order to evaluate our system we have calculated complexity of the system and compare the results with static inference proof. The graph representing the comparison of existing and proposed system is shown in the figure below.

VIII. Conclusion and future work

We propose inference proof view of XML document, which will help us to protect the confidential data from the organization. In this system we can able to add inference rule and confidential data at the run time to provide more flexibility based on user preference. Hare we add some XML schema to provide better results and also remove duplicate or redundant data and append new unique data.

The result obtained from it should be in such a manner that it is a weakened XML document of the hospital database in which it similar to the modified schema. Thus, the result provided is not been able to infer any of the other details from the document which is provided to other organization.

This mechanism is completely different from other mechanisms for inference control, such as refusal or lying, we use weakening to generate an inference-proof view from the actual XML document, which is suitable for information set with incomplete information.

References

[1] E. Bertino, S. Castano, E. Ferrari, and M. Mesiti, “Specifying and Enforcing Access Control Policies for XML Document Sources,” 2000.

[2] E. Damiani, S.D.C. di Vimercati, S. Paraboschi, and P. Samarati, “A Fine-Grained Access Control System for XML Documents,”ACM Trans. Information Systems Security, 2002.

[3] L. Li, X. Jiang, and J. Li, “Enforce Mandatory Access Control Policy on XML Documents,” Proc. Seventh Int’l Conf. Information and Comm. Security (ICICS), S. Qing, W. Mao, J. Lopez, and

[4] B. Finance, S. Medjdoub, and P. Pucheral,The Case for Access Control on XML Relationships, Proc. 14th ACM Intl Conf. Information and Knowledge Management (CIKM), O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, eds., pp. 107-114, 2005.

[5] W. Fan, C.Y. Chan, and M.N. Garofalakis,Secure XML Querying with Security Views, Proc. ACM SIGMOD Intl Conf. Management of Data, G. Weikum, A.C. Konig, and S. Deloch, eds., pp. 587-598,2004.

[6] B. Groz, S. Staworko, A.-C. Caron, Y. Roos, and S. Tison, XML Security Views Revisited, Proc. 12th Intl Symp. Database Programming Languages (DBPL), P. Gardner and F. Geerts, eds., pp. 52-67, 2009.

[7] P. Barcelo , L. Libkin, A. Poggi, and C. Sirangelo, XML with Incomplete Information, J. ACM, vol. 58, no. 1, pp. 4.1-4.10, 2010.

[8] XML with Incomplete Information, ACM Trans. Database Systems, vol. 31, no. 1, pp. 208-254, 2006.

[9] Li and Y. Wang, An Approach for XML Inference Control Based on RDF, Proc. 17th Intl Conf. Database and Expert Systems Applications (DEXA), S. Bressan, J. Ku ng, and R. Wagner, eds., pp. 338-347, 2006.

[10] Marouane Hachicha and Jerome Darmont,”A Survey of XML Tree Patterns” Laboratoire ERIC - University Lumiere Lyon 2 5 Avenue Pierre Mendes-France 69676 Bron Cedex France 2012.

[11] Joachim Biskup and Lan Li,”On Inference-Proof View Processing of XML Documents” IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 10, NO. 2, MARCH/APRIL 2013

[12] Meghdad Mirabi, Hamidah Ibrahim, Nur Izura Udzir, Ali Mamat,”A Compact Bit String Accessibility Map for Secure XML Query Processing” International Workshop on Service Discovery and Composition in Ubiquitous and Pervasive Environment (SUPE)2011