Asian Journal of Information Technology 15 (20): 4113-4119, 2016
ISSN: 1682-3915
© Medwell Journals, 2016

Secure and Efficient Labeling Scheme for XML Content Dissemination

S. Sankari and S. Bose
Department of Computer Science and Engineering, College of Engineering, Guindy (CEG),
Anna University, 600025 Chennai, Tamil Nadu, India

Abstract: XML has become the default standard followed for information exchange across organizations using the web. Any XML document can be viewed as an XML tree. A labeling scheme is necessary to uniquely identify every node of an XML document tree. Secure labeling scheme is preferred in XML content dissemination of publish/subscribe system. In this study, we propose a secure labeling scheme called Enhanced Dewey Coding (EDC). EDC labeling scheme is efficient in terms of storage space required to store labels and time needed for labeling. EDC labeling scheme also preserves the basic hierarchical structural relationships that exists among the nodes of an XML tree. In addition, EDC provides security without leaking the actual structure of the XML document. We implement EDC labeling scheme for various real-time XML documents that varies in document structure and size. Our experimental results show that the performance of EDC in memory space and labeling time is better than the existing method. We also identified optimal random value range for EDC by evaluating the results of EDC label size with various random value ranges.

Key words: Publish/subscribe, security, XML, XML content dissemination, XML labeling scheme

INTRODUCTION

XML emerged as a de facto standard to store and exchange data over the Internet. XML (World Wide Web Consortium is widely followed because of its variable structure and can be designed by the user. In publish/subscribe model (Sankari and Bose, 2013), XML document is widely used for secure content dissemination over the web. Producer owns an XML document that needs to be securely labeled, encoded and encrypted. Selective contents of securely labeled XML document with secure labels are disseminated to the subscribed consumers with the help of publisher also called as message broker. Secure XML labeling scheme plays a vital role in secure content dissemination. The advent of XML and its flexible structure for content representation and dissemination necessitates a secure XML labeling scheme with efficient label size and labeling time. To label the XML document the notion of XML Document Object Model (DOM) is utilized. XML DOM helps to view an XML document as XML tree with nodes. XML labels are used to uniquely identify every node of an XML tree by preserving the hierarchical structural relationships exists between the nodes. From XML labels, consumer can estimate the overall structure of the XML document which leads to structural information leakage. To prevent this, secure XML labeling scheme is preferred. Secure XML labeling scheme hides the actual structural information of an XML document that can be inferred from the secure XML labels but preserves the hierarchical structural relationship.

Secure XML labels are additional information sent along with the actual XML content that needs to be disseminated by the producer. Therefore, an efficient secure XML labeling scheme is required with the following properties:

- Memory: Memory required for storing secure XML labels should be less
- Time: Time required to generate secure XML labels should be minimal

Hence, the efficiency of a secure XML labeling scheme is measured in terms of label size and labeling time. The main objective is to develop a secure labeling scheme that requires minimal memory for XML secure labels and efficient labeling time.

For efficient XML content dissemination, the most challenging issue is to design a secure XML labeling scheme with efficient label size and labeling time. The secure labeling scheme must prevent additional information leakage about the structure of entire XML document. In this study, to achieve secure XML labeling scheme with efficient label size and labeling time for secure and effective content dissemination, we accomplish the following:

Corresponding Author: S. Sankari, Department of Computer Science and Engineering, College of Engineering, Guindy (CEG), Anna University, 600025 Chennai, Tamil Nadu, India

- Propose a secure XML labeling scheme called Enhanced Dewey Coding (EDC) that follows level order
- Implement the secure XML labeling scheme EDC using different real-time datasets
- Evaluate the performance of EDC on label size and labeling time by measuring and comparing the results with the existing method
- Identify optimal random value range by assessing the storage space required by EDC for different random value ranges

Need for secure XML labeling scheme: In publish/subscribe model, producer sends selective XML content to the consumer. XML document queries are usually represented in using Xpath or XQuery format. For a consumer, selective XML contents are specified by an access control policy represented as set of XPaths. The result of XPath may be single text content or a content that includes set of XML tags with text content in a hierarchical way. To uniquely identify the XML content and to maintain a hierarchical relationship between the contents, an identifier is required. Therefore, producer considers an XML document as an XML DOM tree. An XML document forms a XML tree where every element/tag, text content in an XML document denotes a node encompassing the hierarchical structural relationship between them. Thus, XML labels are assigned to all the nodes of an XML tree to identify every node distinctly and also to preserve the structural relationship. XML labeling scheme is a mechanism used to label the XML tree. Producer labels the XML document. XML labels of subscribed XML content are sent to the consumer at the time of subscription. After decrypting the received disseminated content, consumer uses the XML labels to identify the structural relationship between the received content. From the XML labels, consumer can infer the structural information like total number of nodes, number of nodes existing between the received nodes, etc. This leakage in structural information may lead the consumer to perform any attack. To avoid these structural information leakages a secure XML labeling scheme is essential for an XML document. Hence, a secure XML labeling scheme must uniquely identify every node and preserve structural relationship exists between them without revealing actual structural information. Also, these secure XML labels used by the producer and consumer are additional information apart from the actual XML content to be dissemination. Therefore, secure XML labels must be efficient in terms of label size and labeling tme to minimize the additional cost. XML secure labeling scheme are helpful in identifying structural relationships that influences the performance of XML query (Lu *et al*., 2005; Xu and Papakonstantinou, 2008; Li, 2010) processing.

Literature review: In selective dissemination of XML documents is discussed and used an access control system for disseminating the selective contents to various set of users (Bertino and Ferrari, 2002; Kundu and Bertino, 2006). Carzaniga *et al.* (2004) suggested a routing approach for content-based publish/subscribe systems (Datta *et al*., 2003) but security threats and measures for secure content dissemination were not considered. (Bertino *et al*., 2004) explores the difficulty exists in assuring the integrity of XML data and used Merkle hash. However, these methods does not prevent from sending additional data that leads to information leakage and also fails to support scalability.

Several labeling schemes for XML document have been proposed. Labeling schemes helps to identify every node of XML content uniquely. A numbering approach called Dewey coding was presented in (Tatarinov *et al*., 2002; Gou and Chirkova, 2007). (Kundu and Bertino, 2006) utilizes DOM properties of XML document to solve the security related issues. (Ko, 2010) discusses M-IBSL (Modified Improved Binary String Labeling) that splits the XML DOM nodes into public and sensitive nodes. Content of sensitive nodes are encrypted and disseminated to the consumer along with the content of public nodes without avoiding information leakage. (Kundu and Bertino, 2008) follows Post Order Numbering (PON) that is obtained by traversing the XML tree nodes in post order traversal (Dietz, 1982) and labels the nodes by calculating a Structural Identifier (SID) from Encrypted Post Order Number (EPON) (Kundu and Bertino, 2006). SID acts as a unique label for XML tree nodes and is determined using EPON. Though SID provides security, but it does not protect from structural information leak. Also, the label size and label generation time of SID including EPON is high. Further, it does not preserve the major structural relationships exists among the nodes of an XML tree. Therefore, the major drawbacks in existing secure labeling schemes are structural information leak with high label size and labeling time. Our proposed secure labeling scheme EDC prevents additional information leak along with reduced label size and labeling time. We introduce a preliminary result for EDC secure labeling scheme (Sankari and Bose, 2013). In this study, we extend our work in (Sankari and Bose, 2013) by evaluating our secure labeling scheme EDC over numerous real-time XML document datasets and also identifying optimal random value range for EDC labeling scheme without compromising security.

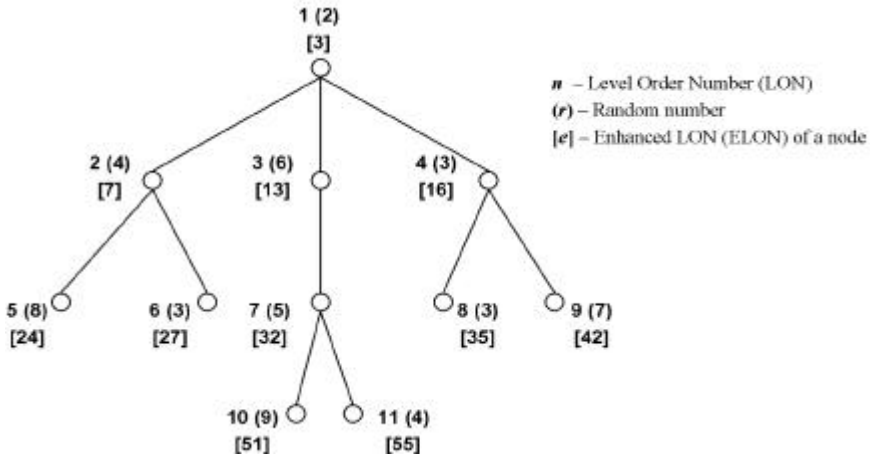

Fig. 1: XML Document Numbered using LON and ELON

MATERIALS AND METHODS

Level order: Level order is the order of traversing an XML tree nodes starting from the root node level to bottom level nodes with every node in a level from left to right.

Level Order Number (LON): Level order number is a unique number assigned to every node of an XML tree while visiting the nodes in level order.

Figure 1 shows an XML tree with LON assigned to all nodes. A LON (Sankari and Bose, 2013) uniquely identifies every node in an XML tree. Also, LON of a parent node is always lesser than its child nodes and descendant nodes. LON has the following properties that conversely results as drawbacks since they reveal the actual structure of an XML document:

- Lowest value 1 denotes root node and highest value of LON denotes the total number of nodes in an XML tree
- At any level, the number nodes exists in a level can be determined from the starting and ending nodes LON value

These are the structural informations that are leaked about an XML document. To avoid this, an enhanced version of LON is designed.

Encrypted Level Order Number (ELON) is calculated using LON. ELON is intended to surpass the drawbacks of LON with the help of random numbers. Any random number is selected and used with LON to provide security by hiding the structural information revealed by LON.

Enhanced Level Order Number (ELON): ELON of a node is a unique secure number that can be calculated using LON with a random number and is represented in Eq. 1 where i is any node whose ELON has to be calculated, j is a preceding node of node i in level order and r is a random number generated for each node whose value ranges from Eq. 2-9:

$$ELON_i = \begin{cases} LON_i + r, & i \text{ is root} \\ LON_j + r, & i \text{ is non-root} \end{cases} \quad (1)$$

XML tree nodes assigned with ELON is shown in Fig. 1. Therefore, ELON calculated for a node avoids structural information leak by preserving the ordering relationships. From Fig. 1, it can be seen that ELON has the following advantages:

- Uniquely identifies every node in an XML document tree.
- Maintains the level order among the nodes of an XML tree.
- Prevents information leakage occurred using LON by concealing the actual number of nodes exists between any two nodes and also the total number of nodes occur before the current node in level order

Though, ELON has various advantages over LON it could not evolve as a complete secure label for a node. Since ELON could not preserve hierarchical relationships like Parent-Child (PC), Ancestor-Descendant (AD),

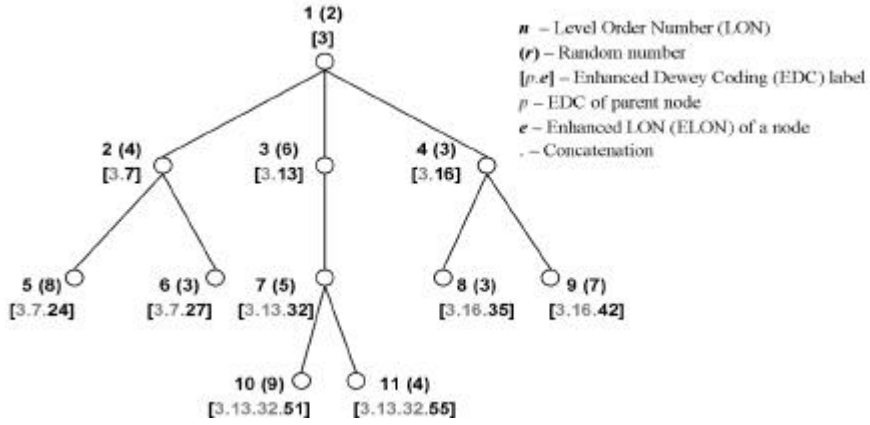

Fig. 2: XML document labeled using EDC

Siblings and these are the major relationships for a hierarchy based XML document tree. Hence, a secure label is preferred that should support both hierarchical and ordering relationships among the nodes without leaking the actual structure of an XML document tree.

Enhanced Dewey Coding (EDC) is a secure label that can be assigned to every node of an XML document tree. EDC (Sankari and Bose, 2013) label is calculated based on the ELON of a node and its parent node. EDC of a node preserves both hierarchical and ordering relationships and at the same time provides security thereby not revealing the actual structure of the whole XML tree through the label.

Enhanced Dewey Coding (EDC): EDC of a node is calculated by concatenating ELON of a node with EDC of its parent node and is denote in Eq. 2 where i is any node whose EDC has to be calculated, j is a parent node of node i and ◦ denotes concatenation of values:

$$ELON_i = \begin{cases} LON_i & i \text{ is root} \\ LON_j.ELON_i, & i \text{ is non-root} \end{cases} \quad (2)$$

Hence, EDC label uniquely identifies every node of an XML document tree, maintains ordering and structural relationships and provides securtity without leaking the actual structure of the whole XML document. Figure 2 shows the XML tree with EDC secure label assigned for all the nodes.

RESULTS AND DISCUSSION

All the experiments of the proposed secure labeling scheme EDC are implemented in a system with configuration of 1.83 GHz Core2 Duo CPU with 3 GB RAM using Java. The datasets utilized for implementation includes numerous XML documents with different sizes from real-time applications like SigmodRecord, Partsupp, Uwm, Wsu, Orders ("UW XML Repository", n.d.) and XMark XML Benchmark project dataset (Schmidt *et al*., 2002; "Xmark-An XML Benchmark Project", n.d.). Table 1 shows the details of several XML document datasets used.

Table 1: XML dataset

XML document	Total nodes count	File size (Bytes)
Sigmod record	11,526	467 K
X mark	17,132	1.12 M
Partsupp	48,001	2.13 M
Uwm	66,729	2.22 M
Wsu	74,557	1.57 M
Orders	150,001	5.12 M

Label size: Label size is measured by calculating the storage memory space required to store labels. The performance of the proposed labeling scheme EDC can be analyzed by comparing the label size of EDC with the existing method SID. Label sizes of SID and EDC for a XML document is obtained by labeling the XML tree using SID and EDC labeling schemes, respectively. Figure 3 shows label sizes of numerous XML documents using SID and EDC labeling schemes. From the graph, it is evident that the label size of EDC labeling scheme is very less compared to SID. Using, EDC label size of Partsupp XML document is reduced to maximum of 53% and overall 39% of label size is reduced for all the XML document datasets. Hence, EDC has greater performance in label size compared to the existing method.

Labeling time: Labeling time for a labeling scheme is measured by calculating the time required to label the

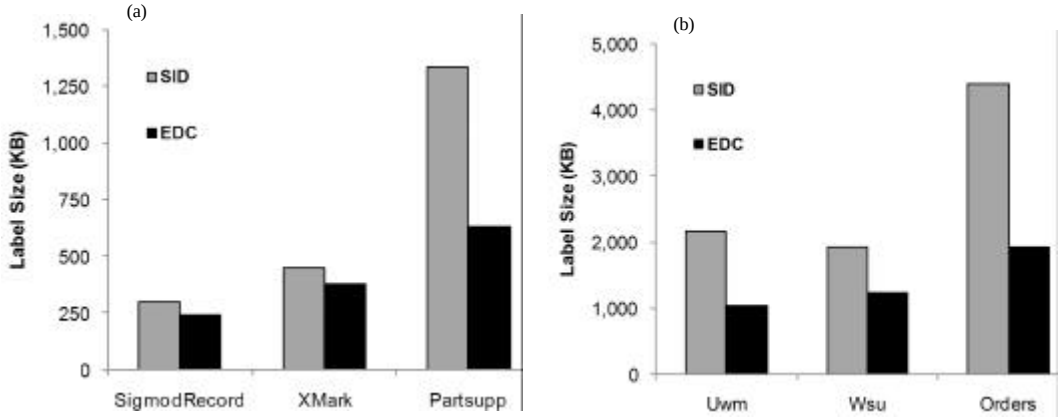

Fig. 3: XML Label size of SID and EDC

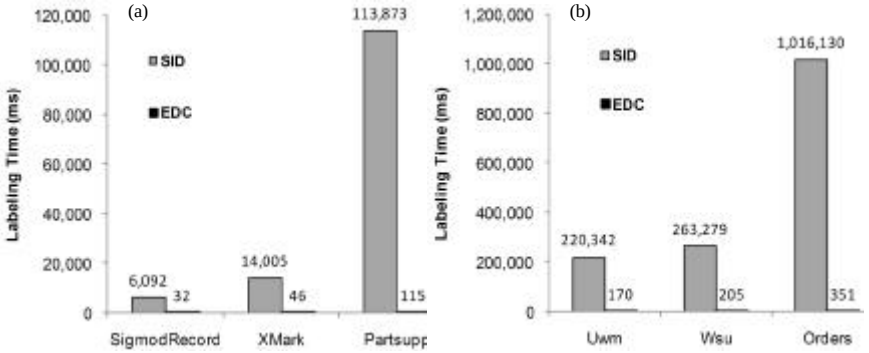

Fig. 4: XML labeling time of SID and EDC

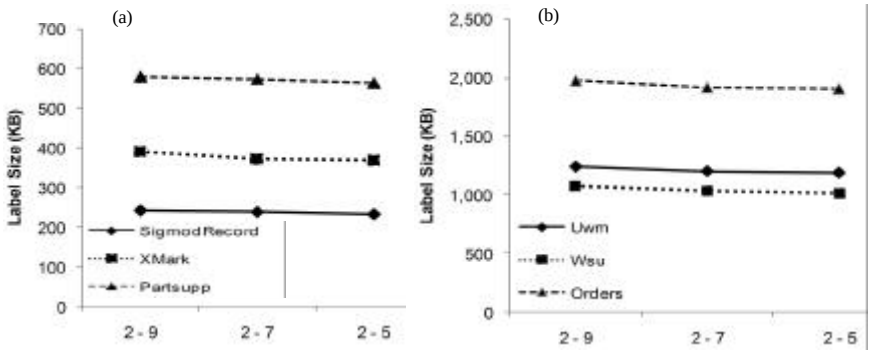

Fig. 5: EDC label size for various random value ranges

nodes of an XML tree. XML documents in Table 1 are implemented using SID and EDC labeling schemes and their corresponding labeling time is determined. The performance of SID and EDC are evaluated based on their labeling time. Figure 4 shows the labeling time of SID and EDC labeling schemes. From the graph, it is clear that the performance of EDC in labeling time is better than SID. Overall, the proposed labeling scheme EDC decreased the labeling time to 99%. Thus, the proposed EDC is better than the existing method (Fig. 5).

Random range values: To identify the optimum random value ranges, the proposed EDC labeling scheme is implemented using different random value ranges such as 2-9, 2-7 and 2-5 and their label sizes are computed for different XML documents. Figure 5 shows the EDC label size of various XML documents with a variety of random value ranges. From the graph, it is obvious that random value ranging from 2-5 has smaller label size compared to other ranges without compromising security. Therefore, random value range 2-5 can be used in EDC labeling scheme for enhancing the efficient EDC results.

CONCLUSION

In this study, we proposed a secure labeling scheme called EDC for XML document whose contents needs to be disseminated. The label size and labeling time of EDC is efficient. EDC label preserves the structural relationships that exist between the nodes of an XML tree, without revealing the actual structure of the XML document. EDC labeling scheme is implemented for various real-time XML documents. The experimental results showed that the performance of EDC labeling scheme in label size and labeling time is better than the existing method. We also implemented EDC with different random value ranges. From the implementation results, optimal random value range is identified based on the label size of EDC. EDC labeling has efficient storage space for label size and labeling time and hence, EDC can be preferred over other existing methods for secure XML labeling and efficient XML content dissemination.

ACKNOWLEDGEMENTS

This research is supported by Anna Centenary Research Fellowship from Anna University, Chennai, India. Any recommendations, conclusions, etc., stated in this proposed work belongs to the authors and do not necessarily reflect those of Anna University.

REFERENCES

Bertino, E. and E. Ferrari, 2002. Secure and selective dissemination of XML documents. ACM Trans. Inform. Syst. Sec., 5: 290-331.

Bertino, E., B. Carminati, E. Ferrari, B. Thuraisingham and A. Gupta, 2004. Selective and authentic third-party distribution of XML documents. IEEE. Trans. Knowl. Data Eng., 16: 1263-1278.

Carzaniga, A., M.J. Rutherford and A.L. Wolf, 2004. A routing scheme for content-based networking. Proceedings of 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Volume 2, March 7-11, 2004, Hong Kong, China, pp: 918-928.

Datta, A.K., M. Gradinariu, M. Raynal and G. Simon, 2003. Anonymous publish subscribe in p2p networks. Proceedings of the International Symposium on Parallel and Distributed Processing, April 22-26, 2003, IEEE, Nice, France, ISBN: 0-7695-1926-1, pp: 8-8.

Dietz, P.F., 1982. Maintaining order in a linked list. Acta Inform., 21: 122-127.

Gou, G. and R. Chirkova, 2007. Efficiently querying large XML data repositories: A survey. IEEE. Trans. Knowl. Data Eng., 19: 1381-1403.

Ko, H.K., 2010. Preventing information leakage in selective dissemenation of web contents. Proceedings of the 2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE), January 9-13, 2010, IEEE, Las Vegas, USA, ISBN: 978-1-4244-4314-7, pp: 327-328.

Kundu, A. and B. Elisa, 2006. Secure dissemination of XML content using structure-based routing. Proceedings of the 2006 10th IEEE International Conference on Enterprise Distributed Object Computing (EDOC'06), October 16-20, 2006, IEEE, West Lafayette, Indiana, ISBN: 0-7695-2558-X, pp: 153-164.

Kundu, A. and E. Bertino, 2008. A new model for secure dissemination of xml content. IEEE. Trans. Syst. Man Cybern. Appl. Rev., 38: 292-301.

Li, C., 2010. Advanced Applications and Structures in XML Processing: Label Streams, Semantics Utilization and Data Query Technologies. IGI Global, Pennsylvania, USA., ISBN: 978-0-61520-727-5, Pages: 467.

Lu, J., T.W. Ling, C.Y. Chan and T. Chen, 2005. From region encoding to extended dewey: On efficient processing of XML twig pattern matching. Proceedings of the 31st International Conference on Very Large Data Bases, October 04-06, 2005, VLDB Endowment, Trento, Italy, ISBN: 1-59593-154-6, pp: 193-204.

Sankari, S. and S. Bose, 2013. Efficient encoding of XML document for secure dissemination. Proceedings of the 2013 Fifth International Conference on Advanced Computing (ICoAC), December 18-20, 2013, IEEE, Guindy, India, ISBN: 978-1-4799-3448-5, pp: 565-570.

Schmidt, A., F. Waas, M. Kersten, M.J. Carey, I. Manolescu and R. Busse, 2002. Xmark: A benchmark for XML data management. Proceedings of the 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China, pp: 974-985.

Tatarinov, I., S. Viglas, K.S. Beyer, J. Shanmugasundaram, E.J. Shekita and C. Zhang, 2002. Storing and querying ordered XML using a relational database system. Proceedings of the International Conference on Management of Data, June 3-6, 2002, Madison, Wisconsin, pp: 204-215.

Xu, Y. and Y. Papakonstantinou, 2008. Efficient LCA based keyword search in XML data. Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology, March 25-30, 2008, ACM, Nantes, France, ISBN: 978-1-59593-926-5, pp: 535-546.