The Three-dimensional Coding Based on the Cone for XML Under Weaving Multi-documents

Jiang Yan[1,*], Wang Yu-xuan[2] and Gong Yi-shan[3]

[1]*School of Software, Shenyang University of Technology, Shenyang, 110023, China*

[2]*School of Information Science and Engineering, Shenyang University of Technology, Shenyang, 110023, China*

[3]*School of Foreign Language, Shenyang University of Technology, Shenyang, 110023, China*

Abstract: In view of the deficiency of the XML document definitions extend the XML elements have to modify the original elements, under the multi-document analyzes the problem of low efficiency of the coding and the query when expand the node, this paper proposes the three-dimensional coding scheme applied to multi-document environment. The XML tree will be regarded as a cone, combining height, radian and marks to encode. Using the corresponding symbol to represent and using mathematical model to describe several new added marker elements, the coding method avoids the recoding of the node due to expansion of the XML document, improving the reusability of the document. The results show that, the method of definition improves the modularity and portability of the document definitions, compared with the existing coding methods, the three-dimensional coding has better performance and feasibility.

Keywords: Coding scheme, multi-document, reusability, XML.

1. INTRODUCTION

One file format to be described in text form, XML (eXtended Markup Language) is becoming a mainstream form of data. The current research focus is on how to achieve efficient query XML data. Most XML related query technology [1] based on modeling ability of XML DTD and XML Schema and the coding method to the XML tree and the corresponding structural join algorithm.

Traditional static coding schemes such as prefix coding [2, 3], interval coding [4], prime coding [5] can support the judgment of positional and structural relationships [6] of nodes, but when the update occurs, the entire XML tree need to recode, resulting in low coding efficiency, higher system cost issues, cannot effectively support XML document update. The researchers have proposed a number of dynamic coding [7] method in relation to the issues, including floating-point number interval, CDBS and QED and so on. These methods are relatively static coding scheme, support the document updates, but larger time and space overhead, reduce the query performance [8, 9], in case the document is not updated is particularly evident, therefore, develop the coding scheme has a good performance in all the circumstances under the document updates or not is particularly important. Based on this meaning, this paper proposes a new dynamic coding scheme — the three-dimensional XML coding based on cone, on the basis of the static coding scheme is extended, can effectively support the XML documentupdate [10, 11],at the same time under the condition of the document is not updated with good performance.

2. MATERIALS AND METHODOLOGY

XML documents are usually represented by a tree structure, for a tree of n layer, if makes the root node of the tree as the vertex of a cone, the nodes of each layer of the tree evenly occupy the bottom of different radius of the cone, the child nodes evenly cut up the sector region of the parent node occupies the bottom, resulting the XML document tree similar to a cone of the root node as the vertex, based on the thought, this paper proposes a three-dimensional XML coding scheme based on cone.

2.1. Coding of the Original Document

2.1.1. The Rules of Coding

Definition 1: XML document node is a four-tuple (docID,H,rad,X). The docID represents the XML document number, which is used to distinguish the original document node and the weaving document node, the docID of the original document node and the weaving document node is different. The H represents the height of the node in the cone, also represents the layer of the node in the nodes tree at the same time, layer by layer increase from the vertex of the cone, the H of the root node is 0, the H of its child nodes is 1, layer by layer increase with 1 step size. The rad represents the radian of the node occupies the cone, any child node evenly cuts up the region occupied by the parent node, choosing the starting radian value to represent the node. The X represents the group marking of the node, which is used to

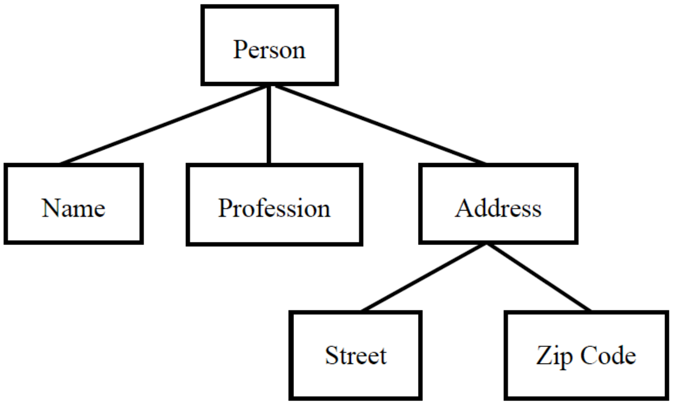

Fig. (1). The original XML document tree.

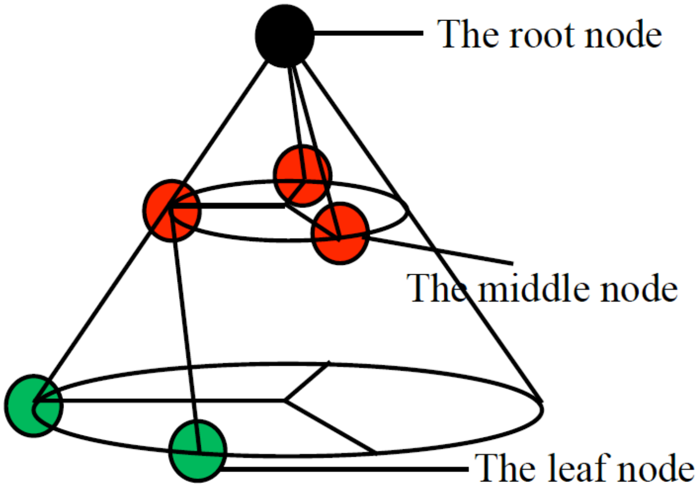

Fig. (2). Cone coding scheme.

represent brother nodes, brother nodes are in the same group and with the same group marking, marking are represented by letters, followed by a, b, c, … , z, aa, ba, … , za, ab, bb, … , zb, … .

Fig. (**1**). is an original XML document tree, Fig. (**2**). is the detailed scheme of the cone coding for Fig. (**1**). original XML document tree, among them, the black node represents the root node, the red node represents the middle node, the green node represents the leaf node. Table **1** is the detailed cone coding, among them, the docID of the original document node defaults to 1.

Table 1. Cone coding.

The Name of the Node	The Cone Coding
Person	(1,0,0,X)
Name	(1,1,0,a)
Profession	(1,1,120,a)
Address	(1,1,240,a)
Street	(1,2,240,a)
Zip Code	(1,2,300,a)

2.2. Coding of the Weaving Document

Usually, weaving the document has the following three methods:

2.2.1. The Root Node of the Original Document is Woven

Fig. (**3**) is the weaving operation for the root node of the original XML document in the Fig. (**1**), the root node of "people" is expanded, as shown by the red dashed line; after weaving, besides the structural relationship between the original document nodes need to judge, may also need to judge the structural relationship between the node of the weaving documents and the root node of the original document, as shown by the blue dashed line. Fig. (**4**) is the detailed coding scheme for the weave method 1 in Fig. (**3**), when the root node of the original document is the woven node, need to make the root node as the vertex of the cone again to form another cone upward according to the way of definition 1, among them, the yellow node represents the root node of the weaving document, the blue node represents the node of the weaving document. Table **2** is the detailed cone coding of the weaving document.

Definition 2: The weaving document node is a four-tuple (docID,H,rad,X). The docID represents the XML document number, which is different from the docID of the original

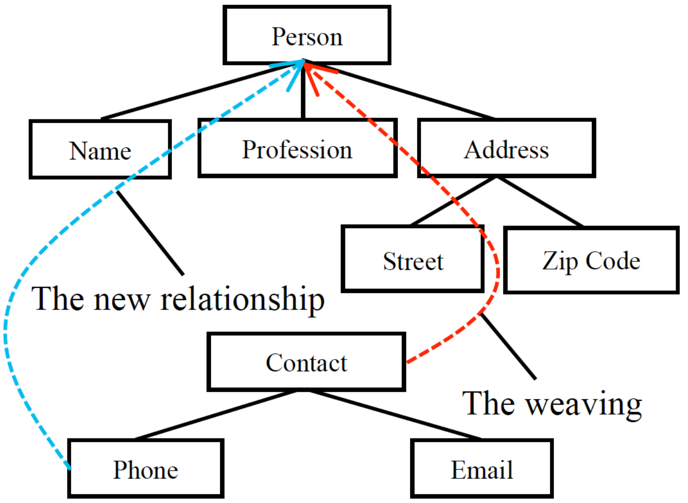

Fig. (3). Weave method 1.

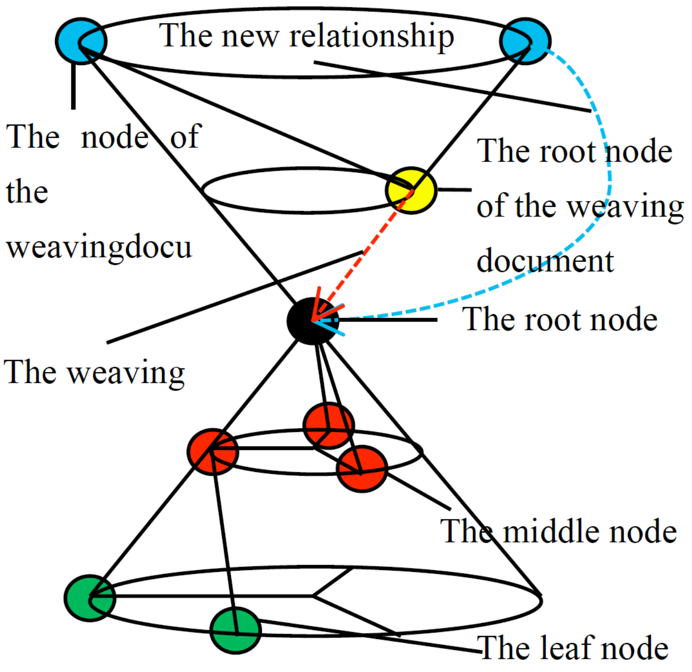

Fig. (4). Weave 1 coding scheme.

XML document. The H represents the height of the node in the cone, layer by layer decrease from the vertex of the cone, the H of the root node is 0, and because of the root node of the weaving document as its child node, then the H is -1, layer by layer decrease with 1 step size.

Table 2. Weave 1 cone coding.

The Name of the Node	The Cone Coding
Contact	(2,-1,0,a)
Phone	(2,-2,0,a)
Email	(2,-2,180,a)

2.2.2. The Middle Node of the Original Document is Woven

Fig. (**5**) is the weaving operation for the middle node of the original XML document in the Fig. (**1**), the middle node of “address” is expanded, as shown by the red dashed line; after weaving, besides the structural relationship between the original document nodes need to judge, may also need to judge the structural relationship between the node of the weaving documents and the root node of the original document, as shown by the blue dashed line. Fig. (**6**) is the detailed coding scheme for the weave method 2 in Fig. (**5**), when the middle node of the original document is the woven node, need to make the woven middle node as the vertex of the cone to form another cone outward according to the way of definition 1, among them, the yellow node represents the

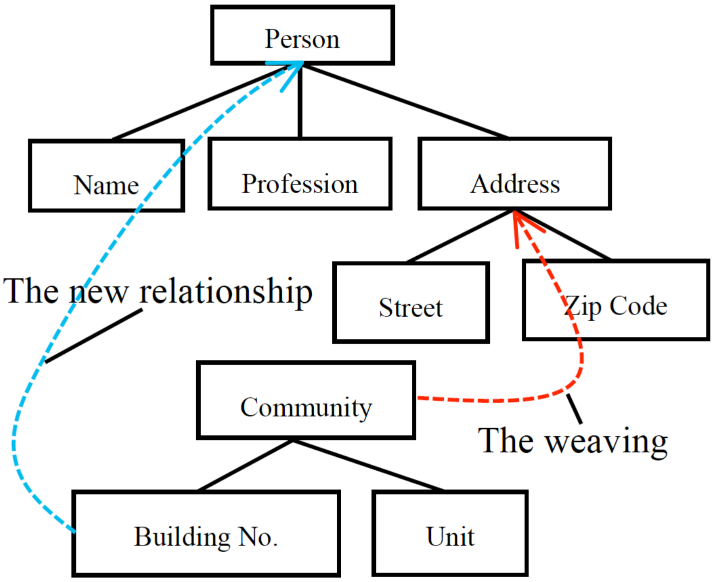

Fig. (5). Weave method 2.

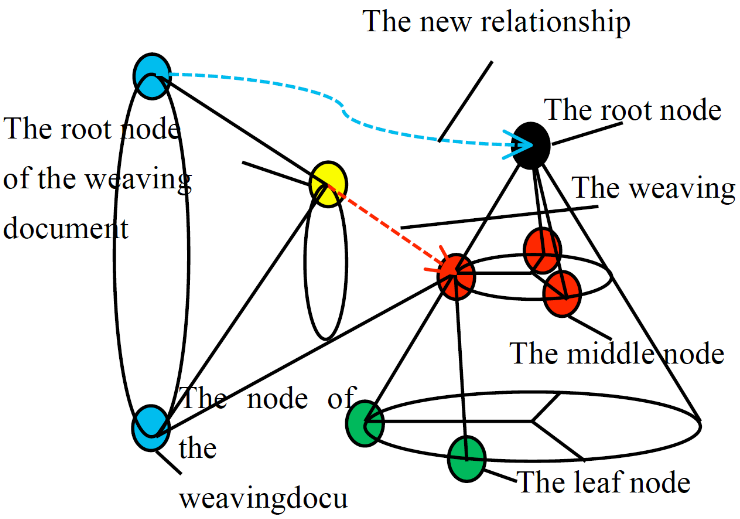

Fig. (6). Weave 2 coding scheme.

root node of the weaving document, the blue node represents the node of the weaving document Table **3** is the detailed cone coding of the weaving document and the independent coding of the woven middle node of the original document.

Definition 3: The weaving document node is a four-tuple (docID,H,rad,X). The H represents the height of the node in the cone, layer by layer increase from the vertex of the cone, if the H of the woven middle node of the original document is h, and because of the root node of the weaving document as its child node, then the H is h+1, layer by layer increase with 1 step size. The woven middle node of the original document need to be coded independently at this time, its rad is returned to 0.

2.2.3. The Leaf Node of the Original Document is Woven

Fig. (**7**) is the weaving operation for the leaf node of the original XML document in the Fig. (**1**), the leaf node of "profession" is expanded, as shown by the red dashed line; after weaving, besides the structural relationship between the original document nodes need to judge, may also need to judge the structural relationship between the node of the weaving documents and the root node of the original document, as shown by the blue dashed line. Fig. (**8**). is the detailed coding scheme for the weave method 3 in Fig. (**7**), when the leaf node of the original document is the woven node, need to make the woven leaf node as the parent node

Table 3. Weave 2 cone coding.

The Name of the Node	The Cone Coding
Address	(1,1,0,a)
Community	(2,2,0,a)
Building No.	(2,3,0,a)
Unit	(2,3,180,a)

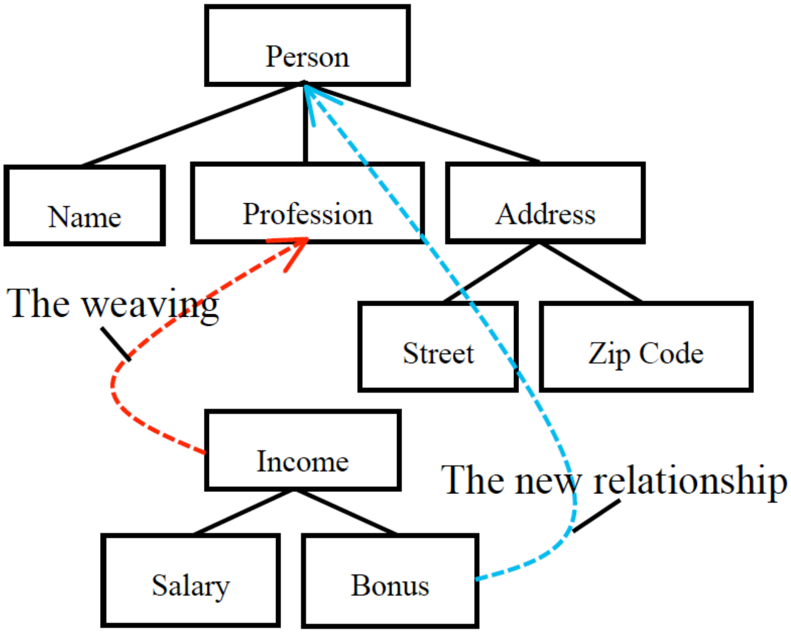

Fig. (7). Weave method 3.

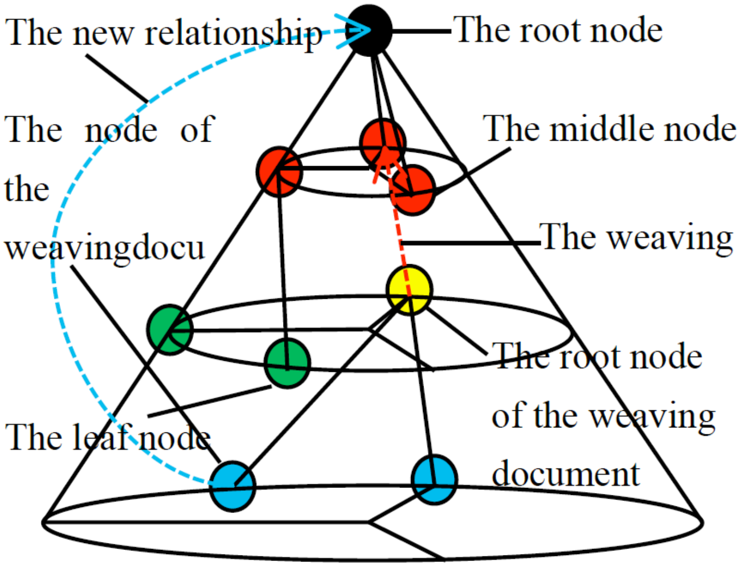

Fig. (8). Weave 3 coding scheme.

for the root node of the weaving document and continue to form the next layer of the cone according to the way of definition 1, among them, the yellow node represents the root node of the weaving document, the blue node represents the node of the weaving document. Table **4** is the detailed cone coding of the weaving document.

Table 4. Weave 3 cone coding.

The Name of the Node	The Cone Coding
Income	(2,2,120,b)
Salary	(2,3,120,a)
Bonus	(2,3,180,a)

Definition 4: the weaving document node is a four-tuple (docID,H,rad,X). The H represents the height of the node in the cone, layer by layer increase from the vertex of the cone, if the H of the woven leaf node of the original document is h, and because of the root node of the weaving document as its child node, then the H is h+1, layer by layer increase with 1 step size.

3. RESULTS

In order to test the effectiveness, rationality and practicability of the coding scheme of this paper, the corresponding experiment is designed. Comparing the cone coding scheme (CCS for short) of this paper with CDBS and QED of region coding. In this paper, the experimental platform is 2.53 GHz

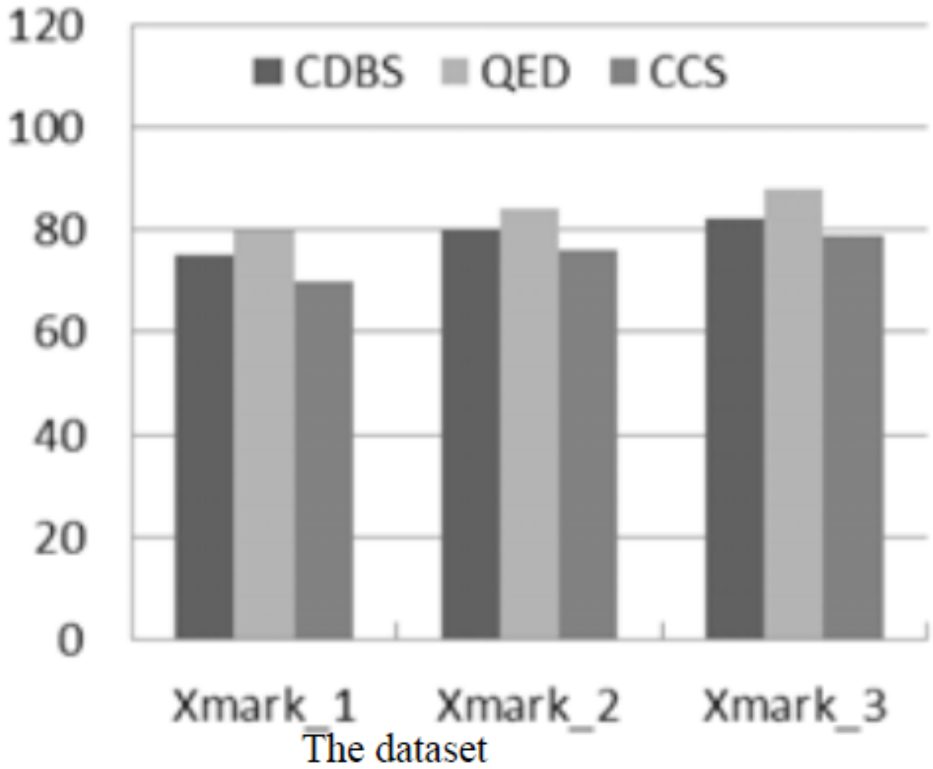

Fig. (9). The comparison of the coding bit length.

Intel core dual-core processor, memory is 4 GB, operating system is Windows 7, using the Visual c++ 6.0, based on the DOM programming technology to realize. The chosen test datasets is generated by XMark of the XML automatic generation tool, the information such as the depth and number of nodes of the document tree, as shown in Table **5**.

Table 5. XML test datasets.

The Dataset	The Total Number of Nodes	The Maximum Depth	The Average Depth
Xmark_1	475844	8	3.12
Xmark_2	1034867	4	2.96
Xmark_3	2254776	32	7.68

3.1. The Static Performance Analysis

In the experiment, the three datasets are coded by CDBS, QED and CCS respectively, and test the static performance.

Fig. (**9**) is the comparison of the average coding bit length of the three coding scheme, because the coding of QED based on quaternary, and the “0” of quaternary only be used for the coding end mark, and the “0” does not appear in the effective bit, therefore, the coding storage proportion is lower. While the fixed length storage efficiency of CCS is higher than the variable length storage efficiency of CDBS, therefore, it can be seen from the figure, compared with CDBS, the coding bit length of CCS is shorter.

Fig. (**10**) is the comparison of the coding time of the three coding scheme, because QED and CDBS need to recursively generate the sequence of the code value, then use the generated sequence of the code value to encode, while CCS traverses only once, mark is simple, therefore, it can be seen from the figure, CCS has the best time performance.

3.2. The Dynamic Performance Analysis

Testing the dynamic performance under the condition of evenly inserting 2^n-1(n=1, 2, 3, … , 20) nodes.

Fig. (**11**) is the comparison of the actual coding bit length of the three coding scheme when inserting the nodes. The coding of CDBS based on binary, the storage proportion is higher, while the effective bit of QED coding only have the “1”, “2”, “3” of quaternary, the storage proportion is lower, it can be seen from the figure, the bit length of CDBS is slowest-growing, while QED is fastest-growing, then CCS is between CDBS and QED.

Fig. (**12**) is the comparison of the actual coding time of the three coding scheme when inserting the nodes. When calculating the coding of the new node, CDBS and QED need to rescan the existing coding, compared with QED, the length of CDBS coding is shorter, so CDBS costs less time, while the time cost of CCS is least, this is mainly due to CCS don’t need to rescan the existing coding, saved the time.

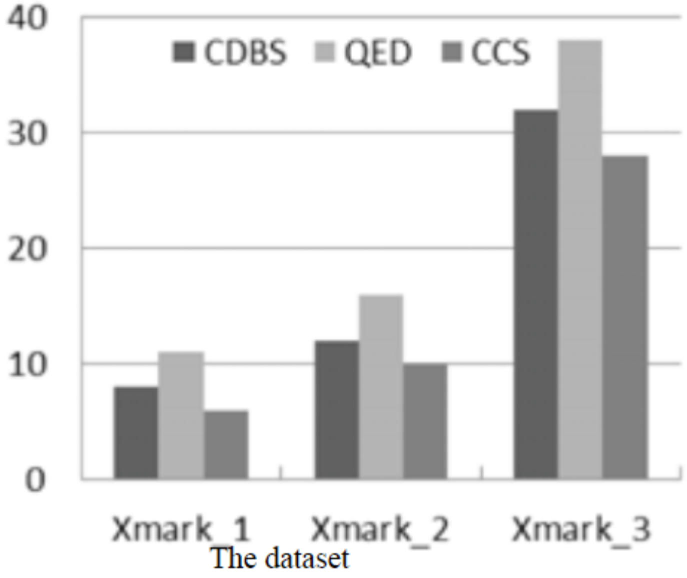

Fig. (10). The comparison of the coding time.

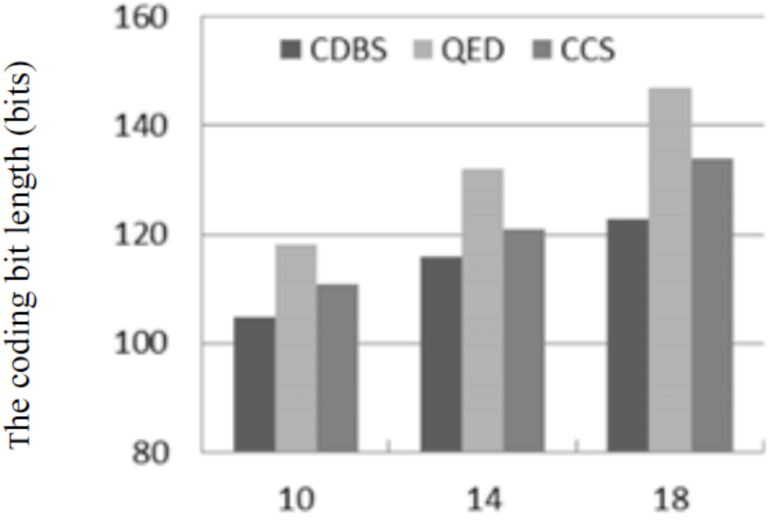

Fig. (11). The comparison of the coding bit length.

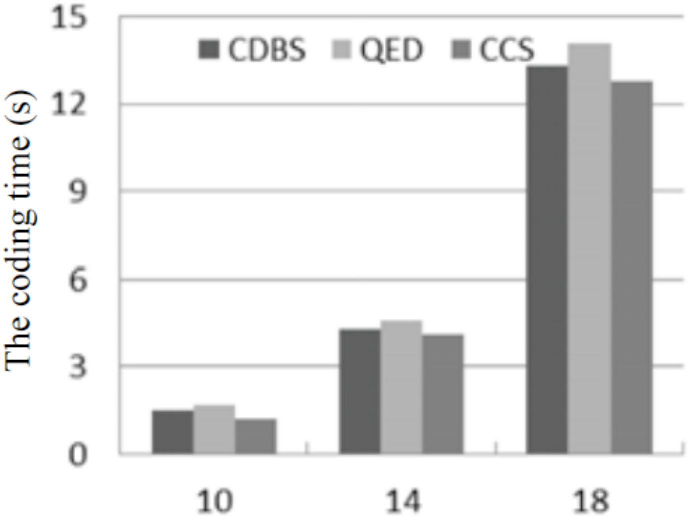

Fig. (12). The comparison of the coding time.

CONCLUSION

This paper proposes a new efficient XML tree dynamic coding scheme — the three-dimensional XML coding based on cone, which has a good static performance, at the same time, supports the document update and update efficiency is higher. Through theoretical analysis and related experiments show that, in this paper, the coding scheme is effective and feasible. In addition, in view of this coding scheme, improves the existing structural join algorithm, realizes the structural join algorithm under the multi-document, and outputs the judgment of related relationships of the node, which is the next step research direction.

CONFLICT OF INTEREST

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "The Three-dimensional Coding based on the Cone for XML under Weaving Multi-documents".

ACKNOWLEDGEMENTS

The authors would like to thank for the support of Natural Science Foundation of Liaoning Province under Grant 2013020032. The authors also thank for the support of the Education Department of Liaoning Province under Grant L2011021.

REFERENCES

[1] L. J. Chen, Y. Papakonstantinou, "Supporting top-K keyword search in XML databases", *Proceedings of the 26th International Conference on Data Engineering* (ICDE2010), 2010.

[2] D. An, S. Park, "Efficient access control labeling scheme for secure XML query processing", *Computer Standards & Interfaces*, vol. 33, no. 5, pp. 439-447, 2011.

[3] L. Xu, T.W. Ling, H. Wu, Z.F. Bao, "DDE: From Dewey to a fully dynamic XML labeling scheme", *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data* (SIGMOD 2009), 2009.

[4] N. Na, D. Guoqing, "A New Labeling Scheme for XML Trees Based on Mesh Partition", *Proceedings of the 2nd International Conference on Future Computer and Communication,* New York, USA: [s. n.], 2010.

[5] L. Jiaheng, M. Xiaofeng, W. L. Tok, "Indexing and query xml using extended dewey labeling scheme", *Data & Knowledge Engineering*, vol. 70, no. 1, pp. 35-59, 2011.

[6] A. Termehchy, M. Winslett, "Using structural information in XML keyword search effectively", *ACM Transactions on Database Systems* (TODS), vol. 36, no. 1, 2011.

[7] A. Li, T. W. Ling, M. Hu, "Efficient updates in dynamic XML data: from binary string to quaternary string", *The VLDB Journal,* vol. 3, pp. 573-601, 2008.

[8] J. Li, J. Wang, "Effectively inferring the search-for node type in XML keyword search", *Proceedings of the 15th International Conference on Database Systems for Advanced Applications*, 2010.

[9] J. Li, C. Liu, R. Zhou, W. Wang, "Suggestion of promising result types for XML keyword search", *Proceedings of the 13th International Conference on Extending Database Technology*, 2010.

[10] W. Li, X. Li, Y. Zhao, "XML documents clustering research based on weighted cosine measure", *Proceedings of the 5thInternational Conference on Frontier of Computer Science and Technology* (FCST''10), 2010.

[11] W. Li, X. Li, R. Te, "Cluster dynamic XML documents based on frequently changing structures", *Advances in Information Sciences and Service Sciences*, vol. 4, no. 6, p. 70, 2012.

Received: November 21, 2014 Revised: January 08, 2015 Accepted: January 20, 2015