Design of XML data coding based on fuzzy data

Yan Jiang[1, a *], Xin Li[2,b], Xin Jin[3,c]

[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] Shenyang Hexing Testing Equipment Co., Ltd, Shenyang 110180, China

[a]panjiang1@163.com, [b]644876180@qq.com, [c]sealovesoft@163.com

Keywords: fuzzy data, XML Schema, coding scheme, lower level.

Abstract. To solve the change in the content and structure resulting from fuzzy data, we should complete the definition of fuzzy data in the parser and propose reasonable coding scheme. The semantic of XML document specification was settled by using the corresponding elements to define them in XML Schema. For these uncertain elements, this paper proposed a new coding scheme by using a quad, including the document number, the serial number of preorder traverse, fuzziness, and the remark in group, to accomplish the coding of XML document tree with fuzzy information. Taking experiments by documents with different levels and numbers, the level is less than 15, and the node number is within 150000. Finally, the document with the most nodes does not need the most time and coding bits, but the lower level document has less coding time. It is suitable for the lower level document.

Introduction

Currently, in order to adapt the increasing requirements to the Web, people take the concept called fuzzy data [1, 2, 3] to XML (eXtensible Markup Language) database. While users could judge the structure of each node in the XML document based on the data coding. Nevertheless, the most existing encoding schemes [4, 5, 6, 7, 8] are for accurate data, they can not reflect the fuzziness of data even when the encoding scheme has reserved space, and the present coding can only satisfy the various cases of fuzzy data storage. What is more, there are two kinds of XML parser, including XML DTD (Document Type Definition) and XML Schema. At the same time, only when finish the definition of fuzzy data in XML parser first, can we accomplish the data specification in a fuzzy XML document.

In recent years, a lot of research on XML is committed to the definition of fuzzy data in XML DTD parser, and experts have proposed adding two elements to use to standardize the fuzzy data in an XML document. And yet, it does not mean to solve the fuzzy data of the document. As XML DTD [9, 10] have a large number of disadvantages, including not supporting namespace, inheritance, the definition of data type and so on. For instance, dialing code for the city is different, some of which is three, and some of which is four, but XML DTD is unable to meet the phenomenon. In addition, each XML document can have only one XML DTD, and DTD stipulates the corresponding language information in XML document by its own grammar rules. Therefore, we should complete the definition of fuzzy data in XML Schema. XML Schema is in itself a well-formed XML document, so it is easier than DTD to manage.

After accomplishing specification of the fuzzy data in language parser, we would like to understand its encoding, as the structure relationship of any two nodes can be determined according to the encoding in an XML document. The current related coding mainly focus on accurate data coding scheme which reserved space. It roughly divided into two kinds, that is based on path and the interval, no matter what kind of coding scheme cannot better complete fuzzy data encoding in the

XML document. Consequently, in order to solve the deficiency, this paper proposes a coding scheme on the basis of fuzzy set theory [11], which can support directly to fuzzy data, what is more, this coding scheme can be have a more comprehensive expression on fuzzy node in an XML document, not only to judge the structural relationship of each node, but also to know the information such as fuzziness, level of each element and so on. The method has a higher efficiency for lower level document, more transparency to users, and more in line with user requirements, so that improve the ability of the interaction between the user and the system.

Definition of Fuzzy Data in XML Schema

There are two forms of fuzzy data in an XML document, including fuzziness of set and fuzziness of values of the attributes. For the fuzziness of set, the information of a student is a case in point. For instance, the information of Kara is part of Class 2 or not. For the fuzziness of values of the attributes, it can be further divided into fuzzy disjunction and conjunction [12]. The fuzzy disjunction as the age of people, it can take an arbitrary value in the set {20, 21, 25, 29, 30}, and fuzzy conjunction as the email of people, but it can take multiple values, such as Kara@163.com, Kara@yahoo.com, Kara@qq.com, and so on.

The fuzziness of set says degree of membership that any set belonged to the corresponding class instance. However, fuzziness of values of the attributes is the uncertainty of values of fuzzy data. The fuzzy disjunction can only take one of the possible values, but fuzzy conjunction can take more possible values. XML Schema is a document with well format, which can use XML editor to edit. Its function is to normalize the language of XML document, and to determine the elements, the attributes, the quantity of same elements, sequence and so on. In order to adapt to the introduction of fuzzy information, the definition of fuzzy data must be completed in XML Schema, so we can predicate the structural relationship of nodes in all document then.

The method is to add corresponding elements. For the elements with fuzziness of set, we would like to add element ‘chance’ when defining it in XML Schema. The added data type is float, and the range of values is [0, 1]. It means to the membership of related element. Specific syntax is as follows:

```
<xs:element ref=”chance”/>
<xs:element name=”chance” type=”xs:float”/>
```

For the elements with fuzziness of values of the attributes, we should add element ‘Forms’ and ‘Fuzzytype’ when defining the element in XML Schema. Forms element is used to guide a variety of possible values of attribute. It is a complex data type. Fuzzytype element is used to mark fuzzy disjunction or fuzzy conjunction, and its type is Boolean type. Specific syntax is as follows:

```
<xs:element ref=”Forms”/>
<xs:element ref=”Fuzzytype”/>
<xs:element          name=”Fuzzytype” type=”xs:Boolean”/>
```

Each element ‘Forms’ contains more elements ‘Table’. As follows:

```
<xs:element name=”Forms”>
<xs:element ref=”Table”/>
</xs:element>
```

Element ‘Table’ is a bipropellant. It is made up of attributes ‘P’ and ‘Value’. Furthermore, ‘P’ is each of the possible values for membership, and ‘Value’ is the corresponding multiple possible values.

```
<xs:element name=”Table”>
<xs:element ref=”P”/>
<xs:element ref=”Value”/>
</xs:element>
<xs:element name=”P” type=”xs:float”/>
<xs:element name=”Value” type=”xs:string”/>
```

Coding of the XML Document with Fuzzy Data

This paper proposes a coding scheme supporting fuzzy data in order to adapt to the introduction of fuzzy data.

The Rules of Coding

When to encode XML documents containing fuzzy elements, we will establish a four-tuple for each element, that is (docld,leftpos:rightpos,fuzzynumber,x).The meanings of four parameters are as follows:

The docld is the sign of the document, which is used to distinguish different XML document, and the XML document with the same identifier must be the same document, so it is unique.

The leftpos and the rightpos are the start and the end serial number of preorder traversal of XML document. The leftpos of the root node is 1. The rightpos of the root node is the largest rightpos all the nodes in XML document.

The fuzzynumber presents the fuzzy number of the XML document. Its value depends on its level and its own fuzziness. The level determines its figures. The level of the root node is 1, so the fuzzynumber of the root node is a single digit. In addition, if the root node is a fuzzy data, then its fuzzynumber is 1, otherwise, it is 0. The fuzzynumber has the characteristics of the inheritance. If the fuzzynumber of the parent node is 01, then the fuzzynumber of the child node will be 010 or 011.

The x means the group ID, which is to distinguish the sibling nodes, if they are sibling nodes, their x will remain consistent. Its value can be a, b, ..., z, aa, ab, …, az, …, za, zb, …, zz, …

The Structural Relationships of Nodes

Firstly, we can define two functions, which to be called h and f. The function h is to obtain the sum of 0 and 1 in fuzzynumber parameter, to put it another way, its function is to obtain the current element's level. And the function f is to acquire the last number of its fuzzynumber, we can know that the current element is fuzzy or not by f.

Definition 1: (ancestor-descendant) Assume that there are two nodes, u ,v∈N, N is the set of nodes, and their coding are C(u) and C(v) respectively, if C(u).docld=C(v).docld, C(u).leftpos<C(v).leftpos, C(u).rightpos>C(v).rightpos, ∀k∈(u+1,v-1), ∃C(u+1).f ∧C(k).f ∧C(v-1). f=0, then node u is an ancestor of node v.

Definition 2: (parent-child) Assume that there are two nodes, u ,v∈N, N is the set of nodes, and their coding are C(u) and C(v) respectively, if C(u).docld=C(v).docld, C(u).leftpos<C(v).leftpos, C(u).rightpos> C(v).rightpos, C(u).h-C(v).h=1, then node u is the parent of node v.

Definition 3: (sibling nodes) Assume that there are two nodes, u ,v∈N, N is the set of nodes, and their coding are C(u) and C(v) respectively, if C(u).docld=C(v).docld, C(u).h=C(v).h, C(u).x=C(v).x, then node u is a sibling node of node v.

The Character of Coding

This coding scheme meets many properties.

For ∀u∈N, if C(u).leftpos=C (u).rightpos, then node u is a leaf node.

For ∀u∈N, if C(u).h=A, then node u is in level A. One other thing to note is that the level of root node is 1, is not 0.

For ∀u∈N, if the number of 1in the parameter fuzzynumber is B, so there are B fuzzy elements on the path from root node to the node u.

For ∀u∈N, if the last number of the parameter fuzzynumber is 1, the node u is a fuzzy element, otherwise, it is precise.

The Instance of Coding

Fig. 1 is an original XML document tree, which is used to store the information of students. Fig. 2 is the detailed coding of the fuzzy data for Fig. 1 original XML document tree. The university has many classes. The fuzziness of set exists between university and class. There are a lot of students in each class. We discuss the students with four attributes, including id, age, email, address. Among them, age and email are fuzzy disjunction and conjunction respectively. For instance, the value of age is {21, 22, 23}, and the possibility of the corresponding values is {0.3, 0.5, 0.6}. The same is true in email.

Among this XML document tree, the node A_1 and A_3 are two fuzzy children of the root node R, the set fuzziness exist between them, and the membership degree is 0.7 and 0.5. That is to say, the possibilities of node A_1 and node A_3 belonging to root node R are 0.7 and 0.5. In addition, as the document tree has shown, root node R is a fuzzy node. The relationship between node A_2 and root node R is accurate. Therefore, the membership degree between them is 1. As default is 1. What is more, node A_2 is an ancestor of node D_2, D_3, D_4, … The node C_1 is the accurate parent of node D_1, however, the node C_2 is the fuzzy parent of node D_2, the node D_2, D_3, D_4 are sibling nodes. Table 1 is the detailed coding scheme of the XML document in Fig. 2.

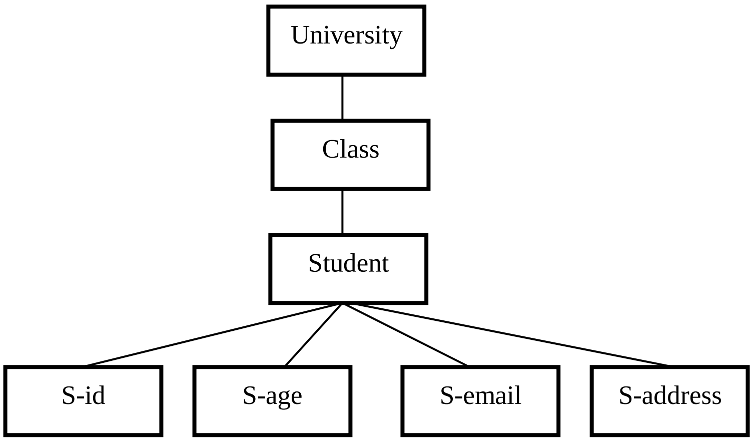

Fig. 1 The structure tree stored student information

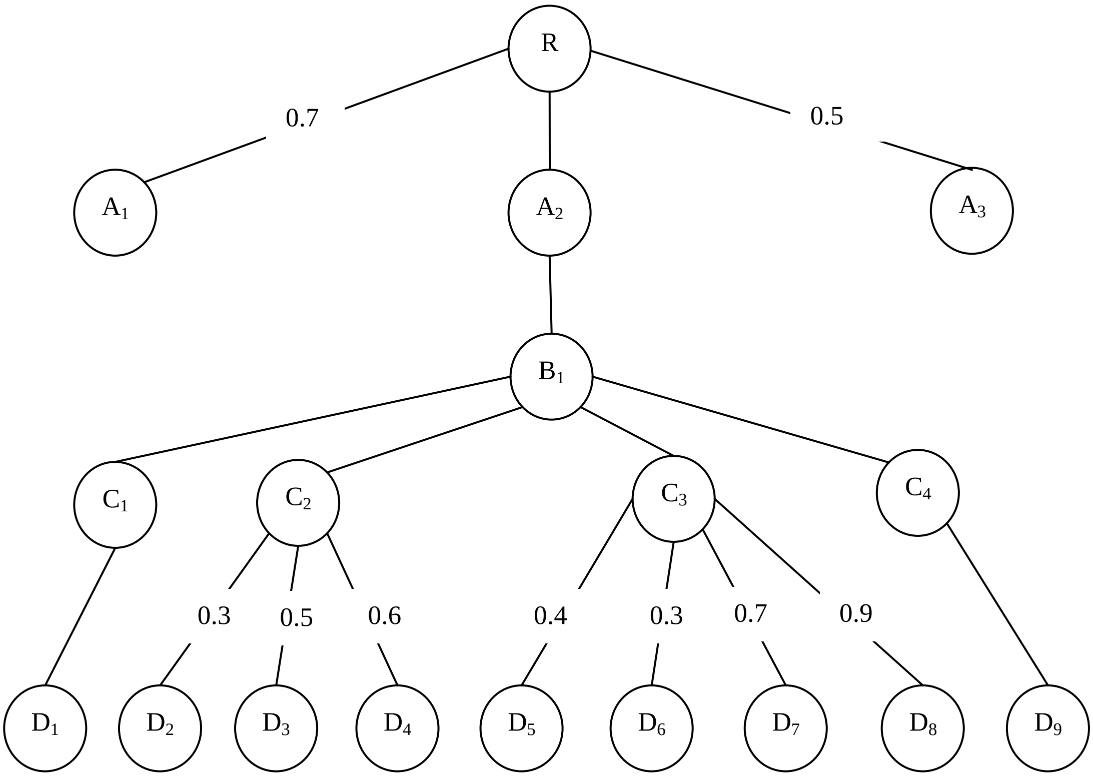

Fig. 2 The coding scheme of XML document with fuzzy data

Results

After proposing this coding scheme, the corresponding experiment would be carried out to demonstrate the effectiveness of this kind of coding scheme for fuzzy XML document. This experimental platform is 2.53 GHz Intel dual-core processor, and the memory is 4GB. What is more, this operating system is Windows 7 and uses Visual C + + 6.0. The selection of test datasets is generated automatically by the tool XMark of XML. The information, including the number of nodes and the depth of this document tree, are shown as in Table 2.

Fig. 3 and Fig. 4 are the comparison of total node and the comparison of coding length in datasets. Fig. 5 is the comparison of coding time in datasets. Finally, we discuss the distribution graph of average time and level according the results, as shown in Fig. 6.

From this experiment, we can conclude that the dataset of the largest coding length and the longest time of coding is not the one which has the most number of nodes. It is obvious that the coding efficiency is related to the depth of the document.

Table 1 The detailed coding scheme

Node	The coding
R	[1,1:25,1,x]
A_1	[1,2:2,10,a]
A_2	[1,3:23,10,a]
A_3	[1,24:24,10,a]
B_1	[1,4:22,100,a]
C_1	[1,5:7,1000,a]
C_2	[1,8:12,1001,a]
C_3	[1,13:18,1001,a]
C_4	[1,19:21,1000,a]
D_1	[1,6:6,10000,a]
D_2	[1,9:9,10010,b]
D_3	[1,10:10,10010,b]
D_4	[1,11:11,10010,b]
D_5	[1,14:14,10010,c]
D_6	[1,15:15,10010,c]
D_7	[1,16:16,10010,c]
D_8	[1,17:17,10010,c]
D_9	[1,20:20,10000,d]

Table 2 The number and depth of XML testing datasets

The dataset	The number of nodes	The maximum depth	The average depth
Xmark_1	467420	6	4.92
Xmark_2	556435	4	3.51
Xmark_3	746340	12	9.25
Xmark_4	928534	10	7.86
Xmark_5	1204783	9	8.04

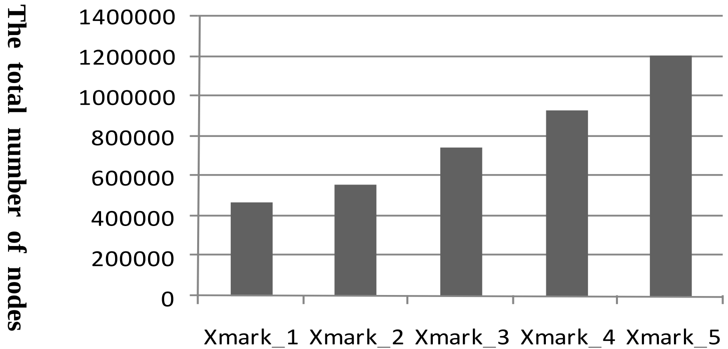

Fig. 3 The comparison of total node in datasets

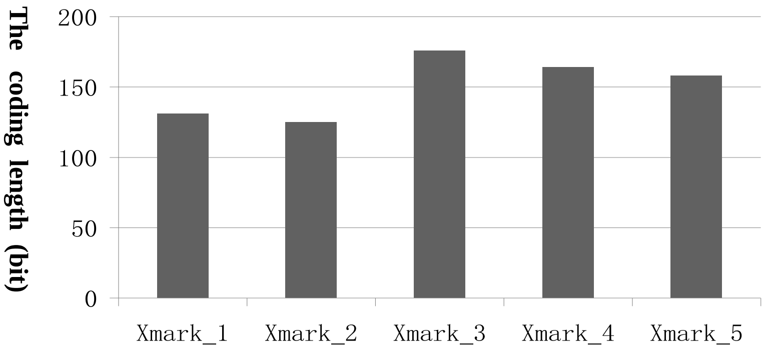

Fig. 4 The comparison of coding length in datasets

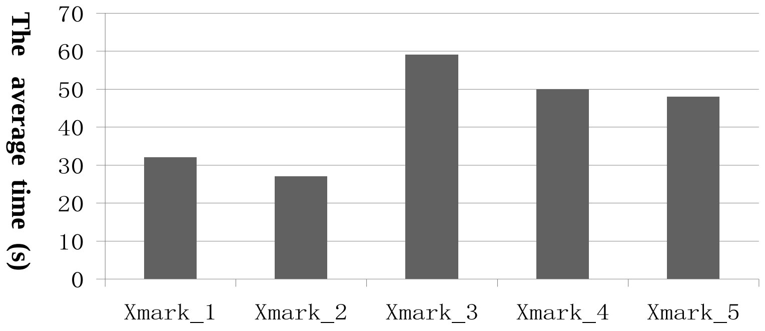

Fig. 5 The comparison of average time of coding in datasets

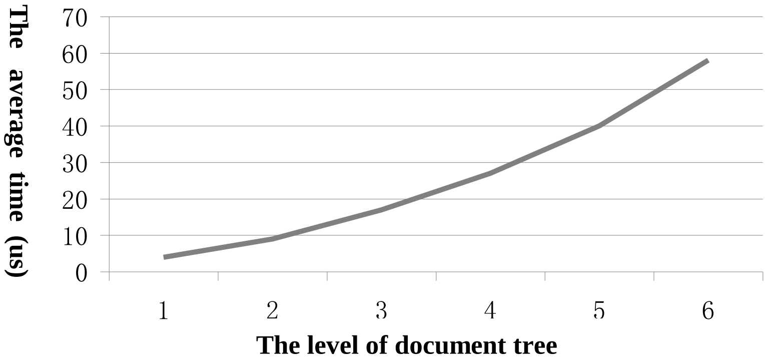

Fig. 6 The distribution graph of average time and level

Conclusion

This paper completed the definition of fuzzy data in XML Schema, and also designed a new coding scheme for XML document containing fuzzy data, which could reflect the fuzziness of XML. At the same time, multiple sets of data were taken to validate the feasibility of this coding scheme under limited space. The lower levels the documents own, the higher efficiency they gains. Finally, the next step to research is to make a kind of coding scheme for longer level of document.

Acknowledgements

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

References

[1] Xingtong Zhu, Bo XU, A Fuzzy Association Rules Algorithm for XML Document.11 (26) 5467-5470. (In Chinese)

[2] Chris Tseng*, Wafa Khamisy, Toan Vu, University fuzzy system representation with XML.28 (2005) 218-230.

[3] Elisabetta Binaghi*, Ignazio Gallo, CristinaGhiselli, An integrated fuzzy logic and web-based framework for active protocol support.77 (2008) 256-271.

[4] Huanan Wen, Xianfeng Liu, XML coding scheme for efficient query processing.30 (3) 831-834. (In Chinese)

[5] Baofeng Yao, Cheng Ma, Na Xie, Dynamic prefix XML encoding scheme based on fraction.30(3) 71-74. (In Chinese)

[6] Shaorong Feng, Tianshuo Chen, Research of Dynamic XML Coding Method Based on Vector.38 (13) 64-66. (In Chinese)

[7] Xianfeng Liu, Zhou Zhou, Ping Liu, Fraction and Prefix XML Encoding Scheme.38 (12) 29-31. (In Chinese)

[8] Lihong Guo, Jian Wang, He Du, An XML Encoding Scheme Based on Concentric Circular Cutting.39 (6) 52-54. (In Chinese)

[9] Li Yan, Jian Liu, Conceptual Design Methodology for Fuzzy XML Model.38 (12) 157-158. (In Chinese)

[10] Xiangfu Meng, Xiayan Zhang, Zongmin Ma, An XML fuzzy query answering approach based on domain knowledge. 7 (6) 527-528. (In Chinese)

[11] Zadeh L A, Fuzzy sets. 8 (3) 338-353.

[12] Li Yan, Zongmin Ma, Jian Liu, XML Modeling of Fuzzy Data with Relational Databases.34 (2) 292-300. (In Chinese)