

The Study of XML Functional Dependency and Multi-Valued Dependency and Inference Rules Set

Liu Xiyin¹, Cao Lijun² and Zhongping Zhang³

^{1,2}*Hebei Normal University of Science & Technology, China*

³*College of Information Science and Engineering Yanshan University, China*

Email: ¹misscao6666@163.com, ²liuxiyin2003@sina.com, zpzhang@ysu.edu.cn

Abstract

This paper discusses and defines XML functional dependency and XML multi-valued dependency, making the formal definition of XML functional dependency and XML multi-valued dependency and further defines the XML-trivial functional dependency and multi-valued dependency. It defines the logical implication, base closures and minimal dependence and gives the set of inference rules that effectiveness and completeness are proved when functional dependency and multi-valued dependency exist simultaneously.

Keywords: *functional dependency; multi-valued dependency; key; inference rules*

1. Introduction

In the wake of the range XML application being wider and wider, there search of XML constraints becomes one of the most important tasks. More and more data are stored permanently in the form of XML, XML data-dependency is a very important field of research, such as how to design XML database efficiently to avoid redundancy and update anomalies. Data dependency plays an important role in database design, data management, data security, data mining and other fields. Data dependency can express the semantic information between the data, and more importantly the purpose of it is to design the core part of good data patterns, also the basis of standardized design. With XML being introduced by W3C as a standard for data exchange on the internet, but because of the limitations in the semantic relationship expression between the data, the study of it is becoming increasingly important. In this essay, related data-dependency theory systems from relational database are drawn, and the formal definition of XML functional dependency and XML multi-valued dependency is proposed, so is the inference rules based on XML functional dependency and XML multi-valued dependency.

2. XML Mode

XML plays an important role in the data on the internet, XML mode design has become a hot topic of research. XML mode is a language to describe a class of XML documents which related constraints. The constituent part of the mode include: data type, element part, text type, entity type and its notation. Through a general document vocabulary can explain the meaning, function and usage of each part, and the relationship of various components by the structure constraints. Currently there are numerous XML modes being widely used, such as XML-Data, XML Schema, DTD (document type definition) and *etc.*, while DTD is favored by the researchers because of its good consistency, scalability, simple use, compatibility and other features as the application of XML schema. DTD schema can guarantee the document format is correct. By judging whether the XML document meets the DTD schema, the effectiveness of the document and the correctness of the element of the document and the use of properties is obviously knowable. Therefore, DTD schema of XML document is used the most frequently and maturely.

Before XML functional dependency and multi-valued dependency formal definition given, an example should be presented.

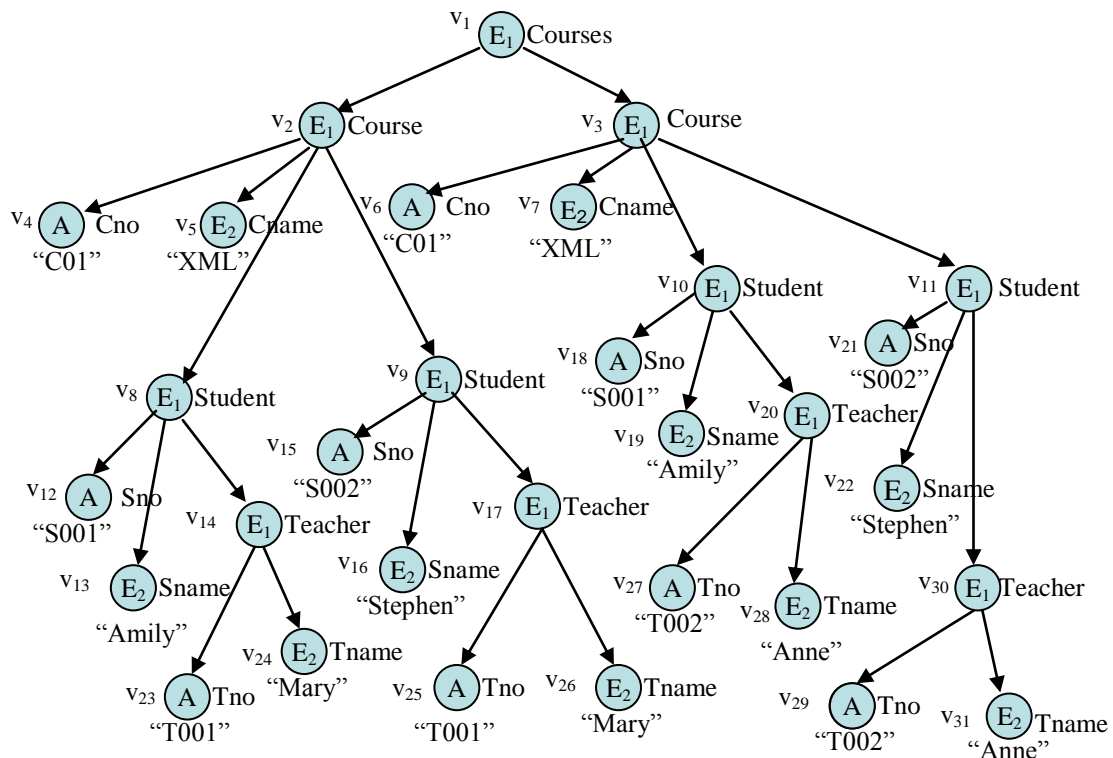


Figure 1. An XML DTD D1 which Meets Fully Documented Instances of T1

As can be seen from the document tree T1, the data dependency between the data of this XML document exists in the document tree structure. In the information of the course, its number is the unique identifier which determines its name accordingly. In the information of the student, the number of the student determines the name of the student, similarly, teachers' job number determines the name of the teacher, the modification of which bears similarities with the key concept of the relational database. In addition, there exist constraints: a course is taught by several teachers and a student can be guided by several teachers. These are more complex situations.

According to the description of the constraints, it can be seen that the data redundancy exists in the XML document. From the above example of the XML document tree, the duplication of the information of courses, students and teachers stored exists. Therefore the XML document is not a well-designed DTD document. In the relational schema, mentioning redundancy implies some dependencies existing between the data. XML document is similar to relational database, and the interdependency relation between the data in the XML document leads to the presence of the data redundancy. Reference relationship database represents the concept of the data dependency relation, which defined as functional dependency and multi-valued dependency in the XML document. Because relational model is complicate while the XML document is hierarchical, description of the functional dependency and multi-valued dependency in the XML document is different from the description of the functional dependency and multi-valued dependency in the relational model. XML document's hierarchy is described by the path expressions.

Example1: the relational table that expresses the case in Figure.1 is presented as follows:

Table 1. A Case Based on Mode R

| Cno | Cname | Sno | Sname | Tno | Tname |
|-----|-------|------|---------|------|-------|
| C01 | XML | S001 | Amily | T001 | Mary |
| C01 | XML | S002 | Stephen | T001 | Mary |
| C01 | XML | S001 | Amily | T002 | Anne |
| C01 | XML | S002 | Stephen | T002 | Anne |

From the analysis of the Figure 1 it is knowable that the redundancy exists in the relational model R. Two students (S001 and S002) have selected Course C01, and there are two teachers (T001 and T002) valid. Within each course, all the students and teachers have their corresponding name. Namely, during students learning each course, the relevant information will be stored repeatedly according to the number of the teachers who teach this course.

The one-to-one relation among the properties in mode R can be represented by the functional dependency in the model R as follows: $Cno \rightarrow Cname$, $Sno \rightarrow Sname$, $Tno \rightarrow Tname$

In relational schema R, each course has a corresponding name, so does the student and teacher. This one-to-one relationship between properties is defined as functional dependency.

The one-to-one relation among the properties in mode R can be represented by the functional dependency in the model R as follows: $Cno \twoheadrightarrow Sno \mid Tno$

In relational schema, there are many students selecting the same course, (one-to-many relationship), which is similar to the relation between teachers and courses (one-to-many), but there is no direct relationship between students and teachers. This one-to-many relationship among properties is defined as multi-valued dependency.

In Figure 1 the XML document instance tree is used to represent the relationship of chart 1, it can be seen from the Figure 1 that instance tree T1 meets the existing of data redundancy in DTD D1, because in the XML document needs each course stored the information of students and teachers many times. If not, data inconsistencies and operational anomalies will arise.

(1) Data redundancy. If a course is taught by several teachers, this course is required to store repeatedly.

(2) Operational anomalies. Due to the presence of the redundancy in the data, all kinds of anomalies during operation will arise.

(a) Update anomalies. If the information of a course needs to be updated, accordingly, all the information about the relevant students who have selected the course needs to be updated. Otherwise the data will be inconsistent.

(b) Insertion anomalies. If a new teacher needs to be inserted in a course, his information needs to be inserted into the information files of all the students who have selected this course. Otherwise the data will be inconsistent.

(c) Deletion anomalies. If a teacher needs to be deleted, his information in the information files of all the students who have selected this course needs to be deleted. Otherwise the data will be inconsistent.

From the above analysis, it's know able that the DTD D1mode is clearly not ideal, since there exists a direct relation both between the courses and students, and the courses and teachers. However, due to the connection mentioned above, D1 nests the students and the teacher which bear no direct relation, resulting in data redundancy and abnormal operation

3. The Functional Dependence and the Multiple-Valued Dependence of XML

3.1. XML Functional Dependence

It has relations among node sets in the document of XML. For instance, in the instance tree of Course.xml, each course has only one name, the same as teachers and students. This type of relation is named as the Functional Dependence of XML. Its formalized definition is as follow:

Definition 1(Functional Dependence of XML, XFD) DTD D given, T is a complete tree that satisfies D. Provided that $P=\{p_1, \dots, p_k\}$ and $Q=\{q_1, \dots, q_m\}$ are the route sets of D, $k \geq 1, m \geq 1, (P \cup Q) \subset Paths(D)$. To any two groups of the route instance in Paths (Q): $V=\{v_1^1, v_2^1, \dots, v_{i-1}^1, v_i^1, \dots, v_1^m, v_2^m, \dots, v_{i-1}^m, v_n^m\}$ And $W=\{w_1^1, w_2^1, \dots, w_{i-1}^1, w_i^1, \dots, w_1^m, w_2^m, \dots, w_{i-1}^m, w_n^m\}$, $n \geq 1$, if the following conditions are satisfied, it is called that the Functional Dependence of XML($p_1, \dots, p_k \rightarrow q_1, \dots, q_m$) is presented in the mode D.

By definition 1, route $p_i \cap q_j$ is a prefix of q_j . Hence, there is only one node that satisfies both in the set of $\{v_i^1, \dots, v_i^m\}$ and the set of $N(p_i \cap q_j)$. Namely, there is only one node included in x_{ij} . Similarly, there is only one node included in y_{ij} .

Case 2: In view of the presentation of Courses. Course. from the above analysis, it's obvious(knowable)that the DTD D1mode is clearly not ideal. since there exists a direct relation both between the course and students, and the course the teacher. however, due to the connection mentioned above, D1 nests the students and the teacher which bear no direct relation, resulting in data redundancy and abnormal operation $Cno \rightarrow Courses. Course.Cname$, ($k=m=1$),the route living Cases of root $v_1.v_2.v_5$ and $v_1.v_3.v_7$ taken, and to make $v_1^1, v_2^1, \dots, v_{i-1}^1, v_i^1$ as $v_1.v_2.v_5, w_1^1, w_2^1, \dots, w_{i-1}^1, w_i^1$ As $v_1.v_3.v_7$, it is knowable that $x_{11}=v_2, y_{11}=v_3$.

Because last (Courses. Course. Cno) $\notin E_1, Nodes(v_2, Courses.Course.Cno)=v_4, Nodes(v_3, Courses.Course.Cno)=v_6, val(v_4)=val(v_6)$. Namely, $val(Nodes(v_2, Courses.Course.Cno)) \cap val(v_3 Courses.Course.Cno) \neq \emptyset$, then it has $val(v_5)=val(v_7)$. Therefore, XFD; $Courses.Course.Cno \rightarrow Courses.Course.Cname$ is presented. In a similar way, $Courses.Course.Student.Sno \rightarrow Courses. Course. Student.Sname$ $Courses.Course.Student.Teacher.Tno \rightarrow Courses.Course.Student.Teacher. Tname$ is presented as well.

Based on definition of the Functional dependence of XML, this paper will give a conception of the trivial functional dependence of XML.

Definition 2(trivial functional dependence of XML) Given DTD D, T is a complete tree that satisfies D. Provided $P=\{p_1, \dots, p_k\}$ and $Q=\{q_1, \dots, q_m\}$ are the route sets of D, $k \geq 1, m \geq 1$. If $P \subset Q$, then functional dependence of XML($p_1, \dots, p_k \rightarrow q_1, \dots, q_m$) is presented in mode D. It is called trivial functional dependence of XML.

3.2. Multiple-Valued Dependence of XML

Definition 3 (Multiple-valued dependence of XML, XMVD) Given DTD D, T is a complete tree that satisfies D. $P=\{p_1, \dots, p_k\}, Q=\{q_1, \dots, q_m\}$ and $R=\{r_1, \dots, r_s\}$ are all the routes of Paths(D), and $k \geq 1, m \geq 1, s \geq 1, \{p_1, \dots, p_k\} \cup \{q_1, \dots, q_m\} \cup \{r_1, \dots, r_s\} \subset Paths(D)$. To any two different route Cases: $V=\{v_1^1, v_2^1, \dots, v_{i-1}^1, v_i^1, \dots, v_1^m, v_2^m, \dots, v_{i-1}^m, v_n^m\}$ And $W=\{w_1^1, w_2^1, \dots, w_{i-1}^1, w_i^1, \dots, w_1^m, w_2^m, \dots, w_{i-1}^m, w_n^m\}$ ($1 \leq i \leq m, 1 \leq t$), if the following conditions are satisfied, the multi-valued dependence of XML ($p_1, \dots, p_k \twoheadrightarrow q_1, \dots, q_m | r_1, \dots, r_s$) is presented in mode D.

$$Val(v_i^1) \neq val(w_i^1);$$

$$(1) val(v_i^1) \neq val(w_i^1);$$

(2) There exists a route $r_j \in R$ ($1 \leq j \leq s$), which has two nodes z_1, z_2 . Then it has conditions that $z_1 \in \text{Nodes}(x_{ij}, r_j)$ and $z_2 \in \text{Nodes}(y_{ij}, r_j)$ which will make $\text{val}(z_1) \neq \text{val}(z_2)$;

(3) For all the routes $p_h \in P$ ($1 \leq h \leq k$), there exist two nodes z_3 and z_4 , within which, $z_3 \in \text{Nodes}(x_{ijh}, p_h)$ and $z_4 \in \text{Nodes}(y_{ijh}, p_h)$ which will make $\text{val}(z_3) = \text{val}(z_4)$;

(4) There exists a route $v_1^i, v_2^i, \dots, v_{t-1}^i, v_t^i$ in $\text{Paths}(q_i)$, which will make $\text{val}(v_i^i) = \text{val}(v_i^i)$. And there exists a node z'_1 at $\text{Nodes}(x'_{ij}, r_j)$, and it will make $\text{val}(z'_1) = \text{val}(z_2)$; there exists a node z'_3 at $\text{Nodes}(x'_{ijh}, p_h)$, and it will make $\text{val}(z'_3) = \text{val}(z_3)$;

(5) There exists a route $w_1^i, w_2^i, \dots, w_{t-1}^i, w_t^i$ In $\text{Paths}(q_i)$, which will make $\text{val}(w_i^i) = \text{val}(w_i^i)$. And there exists a node z'_2 at $\text{Nodes}(y'_{ij}, r_j)$, and it will make $\text{val}(z'_2) = \text{val}(z_1)$; there exists a node z'_4 at $\text{Nodes}(y'_{ijh}, p_h)$, and it will make $\text{val}(z'_4) = \text{val}(z_4)$.

Among which, $x_{ij} = \{v | v \in \{v_1^i, \dots, v_t^i\} \text{ and } v \in N(r_j \cap q_i)\}$, $y_{ij} = \{v | v \in \{w_1^i, \dots, w_t^i\} \text{ and } v \in N(r_j \cap q_i)\}$, $x_{ijh} = \{v | v \in \{v_1^i, \dots, v_t^i\} \text{ and } v \in N(p_h \cap r_j \cap q_i)\}$, $y_{ijh} = \{v | v \in \{w_1^i, \dots, w_t^i\} \text{ and } v \in N(p_h \cap r_j \cap q_i)\}$; $x'_{ij} = \{v | v \in \{v_1^i, \dots, v_t^i\} \text{ and } v \in N(r_j \cap q_i)\}$, $y'_{ij} = \{v | v \in \{w_1^i, \dots, w_t^i\} \text{ and } v \in N(r_j \cap q_i)\}$, $x'_{ijh} = \{v | v \in \{v_1^i, \dots, v_t^i\} \text{ and } v \in N(p_h \cap r_j \cap q_i)\}$, $y'_{ijh} = \{v | v \in \{w_1^i, \dots, w_t^i\} \text{ and } v \in N(p_h \cap r_j \cap q_i)\}$.

From Definition 2, it can be known that route $r_j \cap q_i$ is the prefix of q_i . Therefore, there is only one node which is both in set $\{v_1^i, \dots, v_t^i\}$ and set $N(r_j \cap q_i)$. Then it can be concluded that only one node is included in x_{ij} . In a similar way, there is only one node included in $y_{ij}, x_{ijh}, y_{ijh}, x'_{ij}, y'_{ij}, x'_{ijh}$ and y'_{ijh} . It can be also known that XMVD is symmetrical, namely, $\text{XMVD}; p_1, \dots, p_k \rightarrow q_1, \dots, q_m | r_1, \dots, r_s$ is presented; if and only if $\text{XMVD}; p_1, \dots, p_k \rightarrow r_1, \dots, r_s | q_1, \dots, q_m$ is presented.

Case 3: In view of document tree T1's satisfying the multiple-valued dependence of XML in

Figure.1:

$\text{Courses.Course.Cno} \rightarrow \text{Courses.Course.Student.Sno} | \text{Courses.Course.Student.Teacher.Tno}$.

Here $k=m=s=1$, provided $v_1^1, v_2^1, \dots, v_{n-1}^1, v_n^1$ be the case v_1, v_2, v_8, v_{12} , and $w_1^1, w_2^1, \dots, w_{n-1}^1, w_n^1$ the case

v_1, v_3, v_{11}, v_{21} . $v_1, v_2, v_8, v_{12} \in \text{Paths}(\text{Courses.Course.Student.Sno})$ and

$v_1, v_3, v_{11}, v_{21} \in \text{Paths}(\text{Courses.Course.Student.Sno})$, $\text{val}(v_{12}) \neq \text{val}(v_{21})$, satisfying condition (1)

in definition 3.2: since $x_{11} = v_8, y_{11} = v_{11}, x_{111} = v_2, y_{111} = v_3$,

$\text{Nodes}(x_{11}, \text{Courses.Course.Student.Teacher.Tno}) = \{v_{23}\}$,

$\text{Nodes}(y_{11}, \text{Courses.Course.Student.Teacher.Tno}) = \{v_{29}\}$. Provided $z_1 = v_{23}, z_2 = v_{29}$, within which

$$z_1 \in$$

$\text{Nodes}(x_{11}, \text{Courses.Course.Student.Teacher.Tno}), z_2 \in \text{Nodes}(y_{11}, \text{Courses.Course.Student.Teacher.Tno})$ and

$\text{Nodes}(x_{111}, \text{Courses.Course.Cno}) = \{v_2\}, \text{Nodes}(y_{111}, \text{Courses.Course.Cno}) = \{v_3\}$, and

$\text{val}(z_3) = \text{val}(z_4)$, satisfying condition (3) in definition 2; provided $v_1^1, v_2^1, \dots, v_{n-1}^1, v_n^1$ be

the case v_1, v_3, v_{10}, v_{18} , then $\text{val}(v_n^1) = \text{val}(v_n^1)$, and $x'_{11} = v_{10}, x'_{111} = v_3$. Provided $z'_1 =$

$v_{27}, z'_1 \in \text{Nodes}(x'_{11}, \text{Courses.Course.Student.Teacher.Tno})$ and

$\text{val}(z'_1) = \text{val}(z_2)$. Provided $z'_3 = v_2, z'_3 \in \text{Nodes}(x'_{111}, \text{Courses.Course.Cno})$, and

$\text{val}(z'_3) = \text{val}(z_3)$, satisfying condition (4) in definition 2. Provided $w_1^1, w_2^1, \dots, w_{n-1}^1, w_n^1$ be

v_1, v_2, v_9, v_{15} , then $\text{val}(w_n^1) = \text{val}(w_n^1)$. $y'_{11} = v_9, y'_{111} = v_2$. Provided $z'_2 =$

$v_{25}, z'_2 \in \text{Nodes}(y'_{11}, \text{Courses.Course.Student.Teacher.Tno})$, and $\text{val}(z'_2) = \text{val}(z_1)$. Provided

$z'_4 = v_2, z'_4 \in \text{Nodes}(y'_{111}, \text{Courses.Course.Cno})$, and $\text{val}(z'_4) = \text{val}(z_4)$, satisfying condition

(5) in definition 2. Therefore, Tree T1 satisfies $\text{XMVD}; \text{Courses.Course.Cno} \rightarrow \text{Courses.Course.Student.Sno} | \text{Courses.Course.Student.Teacher.Tno}$.

Definition 4 (trivial multiple-valued functional dependence of XML): Given DTD D, T is a complete tree that satisfies D. Provided $P = \{p_1, \dots, p_k\}$ and $Q = \{q_1, \dots, q_m\}$ are route sets of D

($k \geq 1, m \geq 1, (P \cup Q) \subseteq \text{Paths}(D)$). If $P \subseteq Q$ or $V\text{Paths}(D) \subseteq (P \cup Q)$, the trivial multiple-valued functional dependence: $p_1, \dots, p_k \rightarrow q_1, \dots, q_m$, is presented in mode D , which is called trivial multiple-valued functional dependence.

Based on the definitions of XFD and XMVD above, XFD will be proved as a particular case of XMVD.

Lemma 1: Given DTD D , T is a complete tree that satisfies D . Provided $p_h, r_j, q_i \in D$, $1 \leq h \leq k, 1 \leq i \leq m$ and $1 \leq j \leq s$. If tree T satisfies XFD which satisfies $p_1, \dots, p_k \rightarrow q_1, \dots, q_m$, then T will satisfy XMVD which satisfies $p_1, \dots, p_k \rightarrow q_1, \dots, q_m | r_1, \dots, r_s$ ($R = \{r_1, \dots, r_s\} \in D$).

Prove: Suppose that there exist two different route Cases in $\text{Paths}(q_i)$: $v_1^i, v_2^i, \dots, v_{t-1}^i, v_t^i$ and $w_1^i, w_2^i, \dots, w_{t-1}^i, w_t^i$ ($1 \leq t$), which satisfies $\text{val}(v_t^i) \neq \text{val}(w_t^i)$. According to the definition of XFD, it can be known that $\text{val}(\text{Nodes}(x_{ij}, p_i)) \cap \text{val}(\text{Nodes}(y_{ij}, p_i)) = \emptyset$, then it can be known p_i is key. Then, it can be concluded that it doesn't satisfies the condition 3 of the definition of XMVD. Namely, there is no existence to all p_h ($1 \leq h \leq k$) which will have two nodes z_3 and z_4 ($z_3 \in \text{Nodes}(x_{ijh}, p_h), z_4 \in \text{Nodes}(y_{ijh}, p_h)$) to make $\text{val}(z_3) = \text{val}(z_4)$. Therefore, T satisfies the multiple-valued dependence: $p_1, \dots, p_k \rightarrow q_1, \dots, q_m | r_1, \dots, r_s$.

Definition 5 (logical entailment) Given DTD D , and $\text{Paths}(D)$ is a route way in D , and Σ is a data dependence set in D . If each complete tree T which satisfies Σ also satisfies $P \rightarrow Q | R$ or $P \rightarrow Q$, then Σ logically entails $P \rightarrow Q | R$ or $P \rightarrow Q$, marked as $\Sigma \models P \rightarrow Q | R$ or $\Sigma \models P \rightarrow Q$.

Definition 6 (Closure) Given DTD, $\text{Paths}(D)$ is a set of all routes in D . Suppose that Σ is a mixed set of XFD and XMVD, and the set of all XFD and XMVD which is logically entailed by Σ is called the closure of XFD and XMVD, which is marked as Σ^+ . Namely $\Sigma^+ = \{ P \rightarrow Q | R \text{ or } P \rightarrow Q | \Sigma \models P \rightarrow Q | R \text{ or } \Sigma \models P \rightarrow Q \}$; and the route set $P \subseteq \text{Paths}(D)$. Hence, the closure of route set P which has relative to Σ is marked as P^+ . It's a set of q which has all route that satisfies $P \rightarrow q$ by means of using rule of inference set from set Σ : $P^+ = \{ \text{route } q | P \rightarrow q \text{ in } \Sigma^+ \}$.

Definition 7 (set equivalence) Given DTD D , T is a complete tree that satisfies D . Definite $V\text{Paths}(D) = \{ p | \text{last}(p) \in (E_2 \cup A) \}$. Order $P = \{ p_1, \dots, p_k \}$ and $Q = \{ q_1, \dots, q_k \}$, and P and Q are route set of T , $k \geq 1$. If it satisfies $\text{last}(p_i) = \text{last}(q_i)$ and the values of the end nodes of p_i and q_i are equal, the values of P and Q will be equal, marked as $P_{\text{val}} = Q$.

4. Inference Rules Set

The method of solving the decision problem of logical implication is to derive some more dependencies established from a group of known dependency sets. While the primary method when solving the decision of logical implication, in general, is to introduce inference rules. Correctness of inference rules means "the dependencies derived by using inference rules set from dependency set Σ are certainly in Σ^+ ", completeness means "all dependencies in Σ^+ can be derived by using inference rules set in Σ ". The correctness ensures that all dependencies derived are correct, and the completeness ensures that all implicated dependencies can be derived. Which ensure the validity and reliability of derivation. Here is a group of inference rules sets under the coexistence of XFD and XMVD.

(Reflexivity of XFD) If $Q \subseteq P$, then $P \rightarrow Q$

(Augmentation of XFD) If $P \rightarrow Q$ and $S \subseteq R \subseteq \text{Paths}(D)$, then $PR \rightarrow Q$

(Transitivity of XFD) If $P \rightarrow Q, Q \rightarrow R$, then $P \rightarrow R$

(Supplement rule of XMVD) If $P \rightarrow \rightarrow Q$, then $P \rightarrow \rightarrow R, R = \text{Paths}(D) - P - Q$

(Reflexive rule of XMVD) If $Q \subseteq P$, then $P \rightarrow \rightarrow Q$; This rule can also be defined as trivial multivalued dependency.

(Augmentation rule of XMVD) If $P \rightarrow \rightarrow Q$, and $V \subseteq W \subseteq \text{Paths}(D)$, then $PW \rightarrow \rightarrow QV$

- (Transmission rule of XMVD) If $P \rightarrow \rightarrow Q$, $Q \rightarrow \rightarrow V$, then $P \rightarrow \rightarrow V-Q$
(Plyability) If $P \rightarrow Q$, then $P \rightarrow \rightarrow Q$
(Connectivity) If $P \rightarrow \rightarrow Q$, $R \subseteq Q$, and there $W \subseteq \text{Paths}(D)$ meet $W \cap Q = \emptyset$ and $W \rightarrow R$, then $P \rightarrow R$
(Complement rule of XMVD) If $P \rightarrow \rightarrow Q$, $P \rightarrow \rightarrow R$, then $P \rightarrow \rightarrow QR$
(Residual rule of XMVD) If $P \rightarrow \rightarrow Q$, $P \rightarrow \rightarrow R$, then $P \rightarrow \rightarrow Q-R$, $P \rightarrow \rightarrow R-Q$
(Intersection rule of XMVD) If $P \rightarrow \rightarrow Q$, $P \rightarrow \rightarrow R$, then $P \rightarrow \rightarrow Q \cap R$
(Pseudotransitivity rule of XMVD) If $P \rightarrow \rightarrow Q$, $WQ \rightarrow \rightarrow R$, then $WX \rightarrow \rightarrow Z-WY$
(Mixed pseudotransitivity rule) If $X \rightarrow \rightarrow Y$, $X \rightarrow \rightarrow YZ$, then $X \rightarrow \rightarrow Z-Y$

4.1. Correctness of Inference Rules Set

The followings will prove that the Rule (1) to (14) is valid for the derivation of logical implication relationship between XFD and XMVD.

To prove conveniently, it is assumable all XFD and XMVD in the following proofs are dependencies based on the document tree T.

(1) Condition (1) in definition1 shows that $Q \subseteq P$ meets the definition of XFD. Therefore the rule is established.

(2) Provided that XFD: $P \rightarrow Q$ is established, while $PR \rightarrow QS$ is not. Since $P \rightarrow Q$ is established, when path q_i meets the Condition(1), Definition1, it still meets Condition(1), Definition1, when the left path extends to $\{p_1, \dots, p_k, r_1, \dots, r_g\}$, ② when path q_i meets Condition(2), Definition3.1, if $r_i = q_i$ meets the Condition(1), Definition1, XFD $PR \rightarrow Q$ is established when the left path set extends to $\{p_1, \dots, p_k, r_1, \dots, r_g\}$, and it meets the Condition(1), Definition1; $\forall t$, if r meets Condition(2), Definition1. Namely, $\text{Last}(r_t) \notin E_1$ and $\text{val}(\text{Nodes}(x_{it}, r_t)) \cap \text{val}(\text{Nodes}(y_{it}, r_t)) \neq \emptyset$ or $\text{Last}(r_t) \in E_1$ and $x_{it} = y_{it}$, then XFD $PR \rightarrow Q$ is established; And because $S \subseteq R \subseteq \text{Paths}(D)$, then $PR \rightarrow QS$ is established according to Definition1 as for path s_i meets $r_i = s_i$. If $\exists t$, making r_t fail to meet Condition(2), Definition1, it shows that all the instance path sets which correspond to the left path sets are different, and in this case, that $PR \rightarrow QS$ is established can be certainly derived. $P = \{p_1, \dots, p_k\}, 1 \leq j \leq k$; $Q = \{q_1, \dots, q_m\}, 1 \leq i \leq m$; $R = \{r_1, \dots, r_g\}, 1 \leq t \leq g$; $S = \{s_1, \dots, s_h\}, 1 \leq d \leq h$ are included.

(3) Provided that XFD $P \rightarrow Q$ and $Q \rightarrow R$ are established, according to Definition1, as to two arbitrary path instance sets P_1 and P_2 , if $P_{1 \text{ val}} = P_2$, $Q_{1 \text{ val}} = Q_2$ must be corresponding. And since XFD $Q \rightarrow R$ is established, as to two path instance sets Q_1 and Q_2 , and $Q_{1 \text{ val}} = Q_2$, $R_{1 \text{ val}} = R_2$ must be corresponding. Namely, to two arbitrary path instance sets P_1 and $P_{2,i}$ if $P_{1 \text{ val}} = P_2$, then $R_{1 \text{ val}} = R_2$ must be corresponding. Hence, XFD $P \rightarrow R$ is established.

(4) The rule of XMVD can be derived directly by the definition of XMVD.

(5) The rule of XMVD can be derived directly by the definition of XMVD.

(6) Using proof by contradiction: Namely, assuming that $P \rightarrow \rightarrow Q$ is established while $PW \rightarrow \rightarrow QV$ is not. Since $P \rightarrow \rightarrow Q$ is established, definition3 shows that there is a path $v_1^i, v_2^i, \dots, v_{i-1}^i, v_t^i$ in $\text{Paths}(q_i)$, which makes $\text{val}(v_1^i) = \text{val}(v_t^i)$. Meanwhile, there is a crunode z'_1 in $\text{Nodes}(x_{ij}^i, r_j)$, which makes $\text{val}(z'_1) = \text{val}(z_2)$, and there is a crunode z'_3 , making $\text{val}(z'_3) = \text{val}(z_3)$; There is a path $w_1^i, w_2^i, \dots, w_{i-1}^i, w_t^i$, which makes $\text{val}(w_1^i) = \text{val}(w_t^i)$, while, there is a crunode z'_2 in $\text{Nodes}(y'_{ij}, r_j)$, making $\text{val}(z'_2) = \text{val}(z_1)$, and there is a crunode z'_4 in $\text{Nodes}(y'_{ijh}, p_h)$, making $\text{val}(z'_4) = \text{val}(z_4)$. According to reflexive rule, $W \rightarrow \rightarrow V$ is established.

To sum up, for all paths u within PW , that there exists a path $v_1^i, v_2^i, \dots, v_{i-1}^i, v_t^i$ in $\text{Paths}(u_i)$ making $\text{val}(v_1^i) = \text{val}(v_t^i)$. Meanwhile, there is a crunode z'_1 in $\text{Nodes}(x_{ij}^i, r_j)$, which makes $\text{val}(z'_1) = \text{val}(z_2)$, and there is crunode z'_3 in $\text{Nodes}(x'_{ijh}, p_h)$, which makes $\text{val}(z'_3) = \text{val}(z_3)$. Consequently, it meets the Condition (4), Definition3. As for the rest of paths $P' = \text{Path}(D) - \{PQWV\}$, they are subsets both of $\text{Paths}(D) - PQ$ and $\text{Path}(D) - WV$, which shows that arbitrary

paths in P' meet Condition(5) ,Definition3. Consequently, $PW \twoheadrightarrow QV$ is established, and thus conflicts with the assumption.

(7) Assuming that $XMVD: Q \twoheadrightarrow V$ is established:According to Definition3, there is $Q \twoheadrightarrow V-Q$. Similarly, according to Definition3, when $Q \twoheadrightarrow V-Q$ is established, $Q \twoheadrightarrow V$ is also established. Therefore, it is quite considerable that the condition where $Q \twoheadrightarrow V$ does not have a common path between the left and the right part.

Each $q \in Q_T(p)$ meets $V_T(p)=V_T(pq)$, which p,q are path instances of $P,Q.V_T(pq) \subset V_T(p)$ is obvious. If there is a path instance $v \in V_T(p)-V_T(pq)$,and v belongs to $V-Q$, then path instances p and q consist in some tree tuples, but not in tree tuples t , within which path instance sets p,q,v are all in tree tuples. It can be further deduced that $Q_T(pv)$ certainly exclude q .

Since $Q \twoheadrightarrow V$, according to Definition3, V is only relevant to path set Q , but irrelevant to the other paths. Consequently $V(pq)=V(q)$. Since $V_T(p)=V_T(pq)$ is established, $q \in Q_T(p), V_T(p)=V_T(q)$ are thus established. It can be proved that $V_T(pw)=V_T(p)$ be established only under the condition that $q \in Q(p)$, in which $w \in W, W=Paths(D)-PV$. Since multi-valued dependency $Q \twoheadrightarrow V$ fails to maintain a common path between the left and the right part. Consequently it shows $Q \subset PW$. There is a map q' consists of Q against path instance set pw . According to the Definition of $Q_T(p)$, q' belongs to $Q_T(p)$;And because $Q \twoheadrightarrow V$ is established, so is $V_T(pw)=V_T(q')$. Consequently, $P \twoheadrightarrow V$ is established.

(8) The inference rules can be derived by Lemma 1.

(9) Assuming that $P \twoheadrightarrow Q$ and $W \cap Q = \Phi$ are established, according to Definition3, $Q_T(p) = Q_T(pw)$. Since $R \subset Q$, then $R_T(p) = R_T(pw)$.According to Definition1,it shows $WP \twoheadrightarrow W$, and because $W \twoheadrightarrow R$, it is reasonable that $WP \twoheadrightarrow R$ by transmission rule of XFD. Accordingly $R_T(pw)$ only contains one element, namely, a path set value of R . And because $R_T(p) = R_T(pw)$. $R_T(p)$ also only contains a path set value of R , which shows $P \twoheadrightarrow R$ according to the definition or XFD.

(10) Given that $P \twoheadrightarrow Q$ and $P \twoheadrightarrow R$ are established, $P \twoheadrightarrow P \cup Q$ and $P \cup Q \twoheadrightarrow R \cup Q$ are established according to the augmentation rule of XMVD. According to the reflexivity of XMVD,it shows that $P \cup Q \twoheadrightarrow Paths(D)-P \cup Q \cup R$ is established. According to transmission rule of XMVD, it is obvious that $P \twoheadrightarrow Paths(D)-P \cup Q \cup R$ is established. According to reflexivity of XMVD, it shows that $P \twoheadrightarrow Q \cup R$ is established.

(11) Given that $P \twoheadrightarrow Q$ and $P \twoheadrightarrow R$ are established, and according to complement rule of XMVD, it shows that $P \twoheadrightarrow QR$ is established. According to reflexivity of XMVD, it show that $P \twoheadrightarrow Paths(D)-P \cup Q \cup R$ is established. According to $P \twoheadrightarrow R$ and $P \twoheadrightarrow Paths(D)-P \cup Q \cup R$ and complement rule of XMVD, $P \twoheadrightarrow (Paths(D)-P \cup Q \cup R) \cup R$ is established. By reflexivity of XMVD, it is knowable that $P \twoheadrightarrow Paths(D)-P \cup (Paths(D)-P \cup Q \cup R) \cup R$ is established, that is $P \twoheadrightarrow (Q-R)-P$. According to reflexivity of XMVD, $P \twoheadrightarrow Paths(D)-P \cup ((Q-R)-P)$,namely, $P \twoheadrightarrow Paths(D)-(Q-R)$,is established. Again, by reference to reflexivity of XMVD, it shows that $P \twoheadrightarrow Q-R$ is established. The same goes for $P \twoheadrightarrow R-Q$.

(12) Given that $P \twoheadrightarrow Q$ and $P \twoheadrightarrow R$ are established, and according to complement rule of XMVD, then $P \twoheadrightarrow Q \cup R$ is established. According to reflexivity of XMVD, $P \twoheadrightarrow Paths(D)-P \cup Q \cup R$ established. According to residual rule of XMVD and complement rule of XMVD, $P \twoheadrightarrow (Paths(D)-P \cup Q \cup R) \cup (Q-R)$ established. By reference to reflexivity of XMVD, $P \twoheadrightarrow Paths(D)-P \cup (Paths(D)-P \cup Q \cup R) \cup (Q-R)$,that is $P \twoheadrightarrow (Q \cap R)-P$ established. And again, by reference to reflexivity of XMVD $P \twoheadrightarrow Q \cap R$ established.

According to augmentation of XMVD, $W \cup P \twoheadrightarrow W \cup Q$, and according to transitivity of XMVD, it shows $WP \twoheadrightarrow R-(W \cup Q)$.

(13) According to augmentation of XMVD, $P \twoheadrightarrow PQ$; according to reflexivity of XMVD, $W \twoheadrightarrow W-PQ$; and because of $PQ \twoheadrightarrow W$, according to transitivity of XFD, $PQ \twoheadrightarrow W-PQ$; and according to transitivity of XMVD,, $P \twoheadrightarrow W-PQ$. According to Joint inference rules applied

to $P \rightarrow W-PQ$ and $PQ \rightarrow W-PQ$, $P \rightarrow W-P \cup Q$. According to augmentation of XFD, $P \rightarrow P(W-PQ)$. Hence, $P \rightarrow P(W-Q)$ is established. According to reflexivity of XFD, $P \rightarrow W-Q$ established.

4.2. The Completeness of Inference Rules Set

The concept of path dependency basis should be presented before proving the completeness of inference rules set.

Definition9(minimal base set) gives the complete instance document tree T which meets DTD D and the set of path set $P=\{P_1, \dots, P_k\}$. Supposing $Paths(D)=P_1 \cup \dots \cup P_k$, and the minimal base set(mdb) of P is a partition of path(D), S_1, \dots, S_q , then it makes :

(1) Each P_i is the set of S_j which operated by Complement rule.

(2) There is no partition which meets Condition (1) and is less than q; Meanwhile, $1 \leq i \leq k, 1 \leq j \leq q$.

Definition10 (dependency basis) makes Σ is a set of XFD and XMVD that on the complete instance document tree T which meets DTD D.

Demonstration: the proof of completeness is to prove “XMVD and FD in Σ^+ can be derived by using inference rule in Σ ” is established

provided that Σ is the set of multi-valued dependency and function dependency, making $P=\{p_1, \dots, p_k\}$, then $P \rightarrow P^+, P^+ \rightarrow P$, and P_1, \dots, P_m is the dependency set of $Paths(D)-P^+$. Consequently P^+, P_1, \dots, P_m is a partition based on $Paths(D)$ Hence it can be proved by T construction Constructing an instance document tree T constituted by tree tuples in the number of 2^m , each tree tuple values 1 on all paths of P^+ , and values $a_i, a_j(0,1)$, on instance path of P_i . The document tree has the following features. Property1: for each XMVD whose right path set is P_i is established in T; Property2: if the right of one dependency is the non-empty subset of P_i , then the dependency is established in T, and only if the left of the dependency is non-empty intersection with $P_i, i=1, \dots, m$.

The followings will demonstrate the aforesaid two properties.

(1) Since $P \rightarrow P_i$ is established in T, based on extension rule, for arbitrary path instance set Y, $Y \rightarrow P_i$ is also established in T. Property1 is proved.

(2) When the dependency is function dependency, if $Y \cap P_i$, since the value of each paths is equal in P_i , further, the values of all paths equal the value of path in $Y \cap P_i$, then $Y \rightarrow P_i$. For each $Z \subseteq P_i$ and $Z \subseteq P_i$, $Y \rightarrow Z$ is established; if $Y \cap P_i = \emptyset$, $Z \subseteq P_i$ and $Z \subseteq P_i$, according to construction T, $Z_T(y)$ contains two values 0 and 1, according to Definition 1, $Y \rightarrow Z$ is not established.

When the dependency is multi-valued dependency, if $Y \cap P_i$, since the value of each path is equal in P_i , further, the values of all paths equal the value of path in $Y \cap P_i$, then $Y \rightarrow P_i$. For each $Z \subseteq P_i$ and $Z \subseteq P_i$, $Y \rightarrow Z$ is established; therefore $Y \rightarrow Z$ is established on T. If $Y \cap P_i = \emptyset$, $Z \subseteq P_i$ and $Z \subseteq P_i$, Supposing r is a path in P_i-Z , according to constitution T, as to arbitrary tree elementu, there are two values 0 and 1 of $Z_T(y)$, but because Z and r must be assigned the same value, the value of $Z_T(y)$, $Z_T(yr)$ is only one, therefore $Z_T(y)Z_T(yr)$, and $Y \rightarrow Z$ is not established on T.

To sum up, Property 2 is proved.

The followings will discuss when meets the conditions of completeness assumption in T.

First of all, supposing f denotes a functional dependency in Σ^+ to prove f is established in T. According to Definition1, f can be formalized as $Q \rightarrow p(Q \rightarrow PW)$ can be indicated as $Q \rightarrow P$ and $Q \rightarrow W$, p denotes a path, if $p \subseteq P^+$, since there is only one value of P^+ in T, according to Definition1, f is established in T; If $p \subseteq P^+$, that is, $p \subseteq P$ and $i=1, \dots, m$, if the intersection of Q and P_i is empty, according to Definition 10, there is $P \rightarrow P_i$. According to connectivity rule, $P \rightarrow p$ is established derived by $P \rightarrow P_i$ it is conflict with $p \subseteq P^+$. Therefore, Q must have intersection with P_i . According to Property 2, $Q \rightarrow p$ is established in T.

Thensupposing $g; Q \twoheadrightarrow V$ denotes a multi-valued dependency in Σ^+ to prove g is established in T . Known from the analysis of the above, function dependency $Q \rightarrow V \cap P^+$ is established in T , according to recombination, $Q \twoheadrightarrow V \cap P_i$ is also established in T . Then, the followings will prove multi-valued dependency $Q \twoheadrightarrow V \cap P_i$ is established in T . As to some i , $V \cap P_i =$ or $V \cap P_i = P_i$, according to Property 1, $Q \rightarrow P_i$ is established in T , and $Q \twoheadrightarrow$ is also established in T ; As to some i , $V \cap P_i \neq P_i$, if $Q \cap P_i =$ according to extension rules, $Paths(D)-P_i \twoheadrightarrow V$ is established. According to Definition 10, $P \twoheadrightarrow P_i$ is established, known from symmetry of multi-valued dependency, $P \twoheadrightarrow Paths(D)-P_i$ is established, and derived by transitivity of multi-valued dependency, $P \twoheadrightarrow V-(Paths(D)-P_i)$ is established, further, $P \twoheadrightarrow V \cap P_i$ is established, and that is contradictory with $P_i \in DEP(P)$. Therefore, the intersection of and is not empty. According to Property 2, $Q \twoheadrightarrow V \cap P_i$ is established in T . Then as to arbitrary i $Q \twoheadrightarrow V \cap P_i$ is established, according to $Q \rightarrow V \rightarrow P^+$ and complement rule, $Q \twoheadrightarrow V$ is established in T .

Last but not least, considering the dependency whose left is P is not in Σ^+ . If it is a functional dependency, supposing f' whose left is P is a function dependency not in Σ^+ , it is formalized as $P \rightarrow Q$, then there is $Q \in P^+$, according to Definition 3.10, Q must have intersection with some P_i . According to Definition 1, if $P \rightarrow Q$ established, then $P \rightarrow Q \cap P_i$ established, according to Definition 10, P^+ and P_i do not have intersection, hence, is contradictory with Property 2. Hence, $P \rightarrow Q$ is not established in T ; If the dependency is multi-valued dependency, supposing g' whose left is P is a function dependency not in Σ^+ , it formalized as $P \twoheadrightarrow Q$, according to Definition 10, $P \twoheadrightarrow Q \cap P^+$ and $P \twoheadrightarrow Q \cap P_i$ both established in T , then $Q \cap P_i \neq P_i$ (otherwise, since $Q \rightarrow V \cap P^+$ is in Σ^+ , and as to arbitrary i , $Q \twoheadrightarrow V \cap P_i$ is in Σ^+ , then it contradictory with the assumption). According to Property 2, $P \twoheadrightarrow Q \cap P_i$ is not established in T . Since $P \twoheadrightarrow P_i$ is established in T , if $P \twoheadrightarrow Q$ is also established in T , known from intersection rule of XMVD, $P \twoheadrightarrow Q \cap P_i$ is established in T , contradicted with the assumption. Hence, $P \twoheadrightarrow Q$ is not established in T .

To sum up, the inference rules is complete.

5. Conclusions

The text, in the first place, analyzes the requirements presented by function dependency and multi-valued dependency of XML through a number of examples, proposed the formalized definition of XML function dependency and XML multi-valued dependency on this basis in detail. Then it further defined XML trivial function dependency and XML trivial multi-valued dependency. It described the definition of logical implication and coverage in the case of XML function dependency and XML multi-valued dependency coexist, gave the concept of path dependency set and member set, gave inference rules set in the case of XML function dependency and multi-valued dependency coexist, And proved the validity and completeness of the rule set.

Acknowledgment

This work has been financially supported by the Project of Major Reform to the Network Engineering Programmed in Hebei Normal University of Science and Technology, which was assigned by Hebei Education Department in 2012.

References

- [1] J. J. Lu and S. Renjen, "Normalizing XML Schemas through Relational Database", //Proceedings of the 43rd Annual Southeast Regional Conference, Kennesaw, Georgia, Alabama, USA, (2005), pp. 220-221.
- [2] C. Yu and H. V. Jagadish, "XML schema refinement through redundancy detection and normalization", The VLDB Journal, vol. 17, no. 2, (2008), pp. 203-223.

- [3] F. Ferrarotti, S. Hartmann and H. Köhler, “The Boyce-Codd-Heath Normal Form for SQ”, Proceedings of 18th International Workshop on Logic, Language, Information and Computation, Philadelphia, PA, USA, (2011), pp. 110-122.
- [4] F. Ferrarotti, S. Hartmann and H. Köhler, “Foundations for a Fourth Normal Form over SQL-Like Databases”, Conceptual Modeling and Its Theoretical Foundations. Essays Dedicated to Bernhard Thalheim on the Occasion of His 60th Birthday. London, USA, (2012), pp. 85-100.
- [5] Md. S. Shahriar and J. Liu, “Towards the Preservation of Keys in XML Data Transformation for Integration”, Proceedings of the 14th International Conference on Management of Data, IIT Bombay, Mumbai, India, (2008), pp. 116-126.
- [6] F. Ferrarotti, S. Hartmann and S. Link, “Performance Analysis of Algorithms to Reason about XML Keys”, Proceedings of 23rd International, Conference of Database and Expert Systems Applications, Vienna, Austria, (2012), pp. 101-115.
- [7] G. Li, B. C. Ooi, J. Feng, J. Wang and L. Zhou, “EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data”, Proceedings of the ACM SIGMOD international conference on Management of data ,SIGMOD 2008, Vancouver, BC, Canada, (2008), pp. 903-914.
- [8] R. Cyganiak and D. Wood, “RDF 1.1 Concepts and Abstract Syntax [EB/OL]”, (2011-08-30)[2012-10-15].<http://www.w3.org/TR/2011/WD-rdf11-concepts-20110830>.
- [9] F. Ferrarotti, S. Hartmann and S. Link, “A Precious Class of Cardinality Constraints for Flexible XML Data Processing”, Proceedings of 30th International Conference on Conceptual Modeling, Brussels, Belgium, (2011), pp. 175-188.
- [10] M. W. Vincent, J. Liu and M. Mohania, “The implication problem for ‘closest node’ functional dependencies incomplete XML documents”, Journal of Computer and System Sciences, vol. 78, no. 4, (2012), pp. 1045-1098.
- [11] J. Liu, J. Li, C. Liu and Y. Chen, “Discover Dependencies from Data—A Review”, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 2, (2012), pp. 251-264.
- [12] S. Hartmann, S. Link and T. Trinh, “Solving the Implication Problem for XML Functional Dependencies with Properties”, Proceedings of the 28th international Workshop on Logic, Language, information and Computation, Brasilia, Brazil, (2010), pp. 161-175.

