



Keyword: Indexing technique, XML, Query processing

1.

XML 가 가

가

[1]. XML XML , XML XML

XML , XML 가 가

가 .

(semistructured) (regular path expression)

XML

XML 가 XML 가

. XML

가 가

XML 가

XML XML

(simple path expression)

XML 가 가

가 가

가 가

XML XML  
 ,  
 가 .  
 1 2 , 3 XML  
 XML . 4 XML  
 5 가 , 6 .

2.

XML [2-9]  
 , (semistructured) [10-18]  
 .  
 XML 가 . [4]  
 XML 가  
 like [3], XML SQL- (join)  
 [2] 가  
 가 Oracle XML DB[7] SQL 2000 [8] XML  
 가 SQL  
 XML  
 가 DB2 XML Extender[9]  
 contains

(regular path expression) [2,11,12,15,16,17]  
 . XML  
 / XML 가 XML  
 가 XML 가  
 [10,12,18] XML 가

XML XPath[19] ,  
XML / , / ,  
, XML  
, /  
XML 가 (family line) XML  
가 가  
XML / ,  
/ , , 가 , ,

### 3. XML

XML  
XML  
3.1 XML  
XML  
OEM (edge labeled graph)  
XML  
XML 가 가  
(directed graph)  
3.1 XML  $G = (L, V, E, A, R, O)$  6가  
가 (directed graph)  
• L



```

</account>
<customer customer_id="C100" accounts="A-401">
  <customer_name >Elton John</customer_name >
  <customer_addr>
    <customer_street>Monroe</customer_street>
    <customer_city>Madison</customer_city>
  </customer_addr>
</customer>
<customer customer_id="C102" accounts="A-401 A-402">
  <customer_name >John Lennon</customer_name >
  <customer_addr>
    <customer_street>Austin</customer_street>
    <customer_city>Texas</customer_city>
  </customer_addr>
</customer>
</bank>
<bank id="3">
...
</banks>

```

3-1

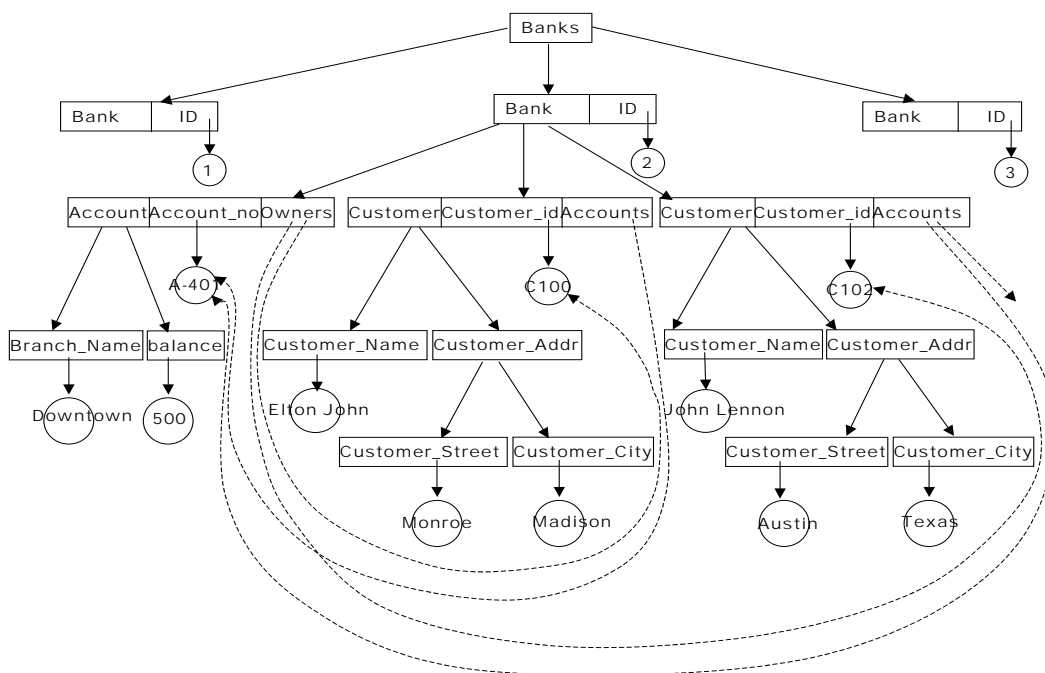
XML

3.1

3-2

가

가



3-2 XML

XML

IDREF IDREFS

4.

XML

W3C

XQuery[21]

가

4.1

XML

가

가

가

XML

FMFT (First order monadic theory of finite binary trees)

[20]

XML

가

FMFT+

4.1

$t = (\{0, \dots, 9\}^*, \dots, d, >, (P_{a1}, P_{b1}) \dots, (P_{an},$

$P_{bn}))$

$\{0, \dots, 9\}^*$

$(P_a, P_b)$

$P_a, P_b$

가 0 9

$(P_a,$

$P_b)$

$P_a, P_b$

$\{0, \dots, 9\}$

,  $P_a$

,

,  $P_b$

$\alpha$

$P_a, P_b$

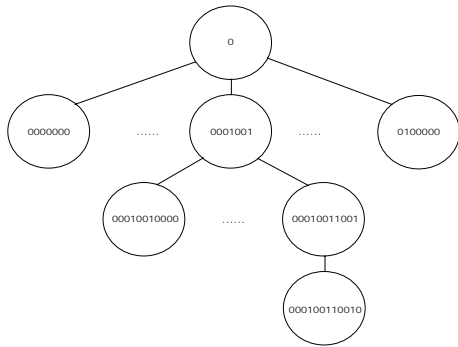
(concatenation)

$P_a + P_b$ 가

$\beta$

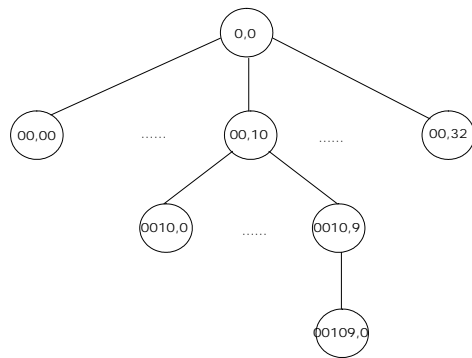
$P_a$ 가  $P_b$   $\beta$  .  $P_a$   $P_b$   
 . n  
 가  
 .  
 • 가 w 가 / v ,  
 • d 가 w 가 / v 가  
 • > 가 w 가 / v ,  
 / .

4.1 4-1 FMFT



4-1 FMFT

4-2 FMFT+



4-2 FMFT+

가 ,

FMFT+

. , FMFT+

가 .

, d, >

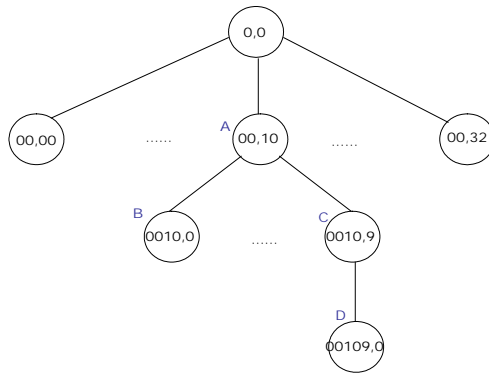
FMFT+

4.1

XML 가 가

가 .





4-3

**4.2**  $(P_{a1}, P_{b1})$   $(P_{a2}, P_{b2})$   
 $P_{a1} + P_{b1}(\dots) = P_{a2}$  /  
 ( d) . , 4-3 B :  $(P_{a2} = 0010, P_{b2} = 0)$ , C  
 :  $(P_{a3} = 0010, P_{b3} = 9)$  A :  $(P_{a1} = 00, P_{b1} = 10)$   $(P_{a1} + P_{b1})$   $P_{a2}, P_{a3}$  가  
 B,C A A B 가 .

**4.3**  $(P_{a1}, P_{b1})$   $(P_{a2}, P_{b2})$   
 $P_{a1} + P_{b1}(\dots)$  가  $P_{a2}$  (prefix)가 /  
 ( ) . , 4-3 B :  $(P_{a2} = 0010, P_{b2} = 0)$ , C  
 :  $(P_{a3} = 0010, P_{b3} = 9)$ , D :  $(P_{a4} = 00109, P_{b4} = 0)$  A :  $(P_{a1} = 00, P_{b1} = 10)$   
 $(P_{a1} + P_{b1})$   $P_{a2} = 0010, P_{a3} = 0010, P_{a4} = 00109$  (prefix) 가  
 B, C, D A A . ( d)  
 ( ) .

**4.4**  $(P_{a1}, P_{b1})$   
 $(P_{a2}, P_{b2})$   $P_{a1} = P_{a2}$  ( $>$ )  $P_{b1} > P_{b2}$   $P_{b1}$   $P_{b2}$   
 $P_{b2}$   $P_{b1}$  . , 4-3 B :  $(P_{a2} = 0010, P_{b2} = 0)$ , C :  $(P_{a3} =$   
 $0010, P_{b3} = 9)$   $P_{a2} = P_{a3}$   $P_{b2} < P_{b3}$  B C .

**4.5** A B가 C 가 .  
 A B / A가 B  
 A B C 가 . , C 가 가  
 B . , 4-3 D :  $(P_{a4} = 00109, P_{b4} = 0)$  A :  $(P_{a1} = 00,$

$P_{b1} = 10$ ) C : ( $P_{a3} = 0010, P_{b3} = 9$ ) D  $P_a$   $P_a + P_b$  가  
 D . C  $P_a$  가 A  $P_a + P_b$   
 A B C 가 .  
 가 가  
 가 .

FMFT+  
 (genealogy-based index) XML 가 가

가

#### 4.2 XML

$NI(\text{Name Index}) = \text{Name}_1 \rightarrow ((P^l_{a1}, P^l_{b1}), \dots, (P^l_{ak}, P^l_{bk})), \dots, \text{Name}_n \rightarrow ((P^n_{a1}, P^n_{b1}), \dots, (P^n_{am}, P^n_{bm}))$  .  
 $(P_{a1}, P_{b1}), \dots, (P_{an}, P_{bn})$  FMFT+  
 $(\quad) = k + \dots + m$  .

#### 4.3 XML

RNI(attribute Name

Index for Reference) =  $\text{Name}_1 \rightarrow (((P^l_{a1}, P^l_{b1}), (RP^l_{a1}, RP^l_{b1}), \dots, (RP^l_{an}, RP^l_{bn})), \dots, ((P^l_{ak}, P^l_{bk}), (RP^l_{ak1}, RP^l_{bk1}), \dots, (RP^l_{akn}, RP^l_{bkn}))), \dots, \text{Name}_n \rightarrow (((P^n_{a1}, P^n_{b1}), (RP^n_{a1}, RP^n_{b1}), \dots, (RP^n_{ap}, RP^n_{bp})), \dots, ((P^n_{am}, P^n_{bm}), (RP^n_{am1}, RP^n_{bm1}), \dots, (RP^n_{amq}, RP^n_{bmq})))$  .  $(P_a, P_b)$  FMFT+  
 $(RP_a, RP_b)$   
 .  $\text{Name}_1$  가 k  
 n .

#### 4.4 XML

$VI(\text{Value Index}) = (\text{Keyword}_i, \text{frequency}_i) \rightarrow ((P^l_{a1}, P^l_{b1}), \text{wp}^l_{i1}), \dots, ((P^l_{am}, P^l_{bm}), \text{wp}^l_{im}), \dots, (\text{Keyword}_n, \text{frequency}_n) \rightarrow ((P^n_{a1}, P^n_{b1}), \text{wp}^n_{i1}), \dots, ((P^n_{ak}, P^n_{bk}), \text{wp}^n_{ik})$  .  $\text{Keyword}_i$   
 $\text{frequency}_i$   $\text{frequency}_i = m$  .  
 $\text{wp}(\text{word position})$  keyword가

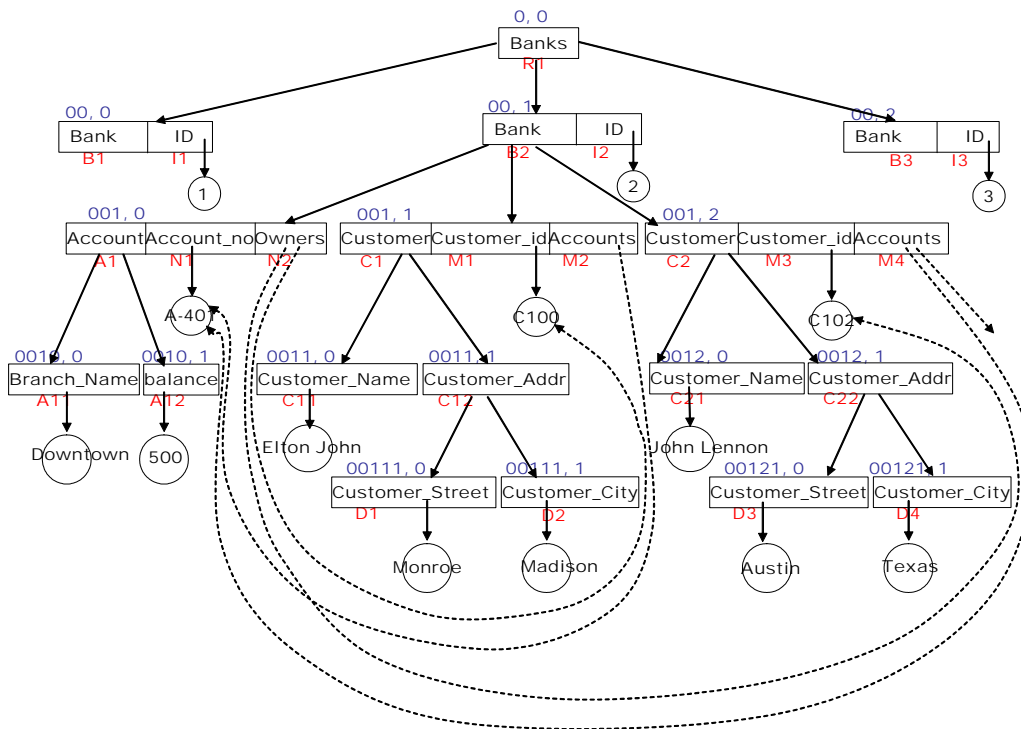
4.2

4.6 3 3-2  
4-4

(Documents Identifier) 가

가

- Banks->(0,0)
- Bank->(00,0), (00,1), (00,2)
- Customer->(001,1), (001,2)
- Customer\_name->(0011,0), (0012,0)
- ...
- account\_no->(001,0)
- customer\_id->(001,1), (001,2)
- ...
- owner->((001,0), (001,1), (001,2))
- accounts->((001,1), (001,0)), ((001,2), (001,0))
- ...



4-4 FMFT+

owner accounts IDREFS 4.3

가

XQuery

4.7 customer\_street bank

FOR \$b in bank  
WHERE \$b/customer\_street  
RETURN \$b

		bank		customer_street
	가	bank	, bank	B1, B2, B3가
customer_street		bank	B2가	customer_street
	00111 00121		(00,1)	가

4.8 customer\_city가 'Texas' customer\_addr

FOR \$c in customer\_addr  
WHERE \$c/customer\_city= "Texas"  
RETURN \$c

customer_city가 'Texas'	D4	customer_addr	D4
00121	가	customer_addr	C22가
			XML

**4.9** customer\_street가 'Austin' customer\_city가 'Texas' customer  
customer\_street customer\_city

**FOR** \$c in customer  
**WHERE** \$c/(customer\_street="Austin" AND customer\_city="Texas")  
**RETURN** \$c

customer\_street가 'Austin' D3  
customer\_city D4가 D3  
D4 00121 가 customer  
C2 가 가

**4.10**

bank/customer[first]

bank customer 가

가

customer[1TO10]/customer\_name

customer

customer\_name

가

**4.11** account customer

Account/@owner=>customer/customer\_name

C1, C2                      owner                      ((001,1), (001,2))                      customer                      가  
 C11, C21                      .                      가                      customer\_name

5. 가

가                      5-1

5-1 가

				/		
( )				<5	1	
				>=5	1	
				>3	1	
	( >3)				<5	1
					>=5	1
					<5	1
( >3)				>=5	1	

XML

IDREF(S)

가

5-2

가  
 가 5  
 가  
 가

	=		Q1
	>		Q2
1=	2		Q3
BETWEEN 1 AND 2			Q4
IN ( )			Q5
1= 1 OR	2= 2		Q6
1= 1 AND	2= 2		Q7
NOT ( = )			Q8
( / )			Q9
( / )			Q10
			Q11
			Q12

가

가

5-2

가

P.Griffiths Selinger “Access Path Selection in a Relational Database Management System”[22] (Selectivity Factor) XML

가

가

가

[23]

5-2

$$Cost = \sum (T * N_{pt}) + P + \sum (N_{pv})$$

- T:
- $N_{pt}$ : 가
- $N_{pv}$ : (predicate)
- P:

[19]

8가

가

가

(attribute table with inlining)

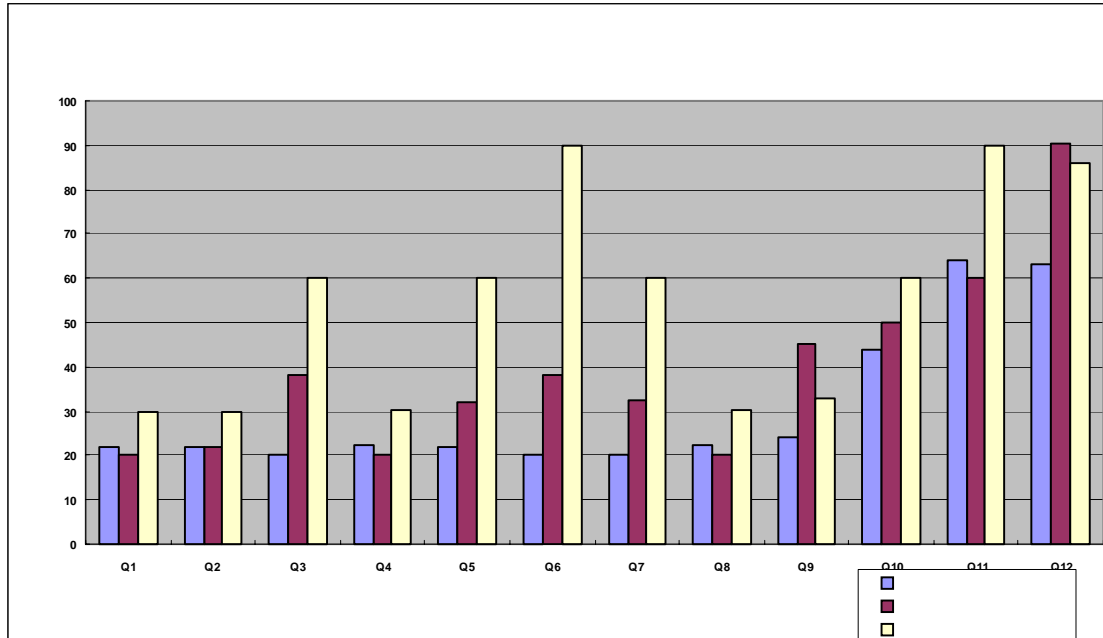
[21]

가

(hybrid inlining)

가 ,

3가



5-1 가

가

가

, XML

가

가

“ 1= 1 OR

2= 2”

, Q6

가

가

6.

XML 가

XML

가

, XML

XML

/ ,



/ ,  
가

XML

XML

가

- [1] W3C Consortium, XML1.0 (Second Edition), W3C Recommendation, 6 Oct. 2000, available at <http://www.w3.org/TR/2000/WD-xml-2e-20000814>
- [2] Q. Li and B.Moon, Indexing and Querying XML data for Regular Path Expressions, VLDB, 2001
- [3] C. Zhang, J. Naughton, D. DeWitt, Q. Luo, G. Lohman, On Supporting Containment Queries in Relational database Management Systems, SIGMOD, 2001
- [4] D. Florescu and D. Kossmann, Storing and Querying XML data using an RDBMS, IEEE Data Engineering Bulletin, 22(3): 27-34, 1999
- [5] J. Shanmugasundaram, K. Tuffe, G. He, C. Zhang, D. DeWitt, and J. Naughton, Relational databases for Querying XML Documents: Limitations and Opportunities, VLDB, 1999
- [6] C. Zhang, Q. Luo, D. DeWitt, J. Naughton, and F. Tian, On the Use of a Relational database Management System for XML Information Retrieval, 2000.
- [7] S. Banerjee, Oracle XML DB, Oracle Corporation Technical White Paper Release 9.2, Jan. 2002
- [8] S. Howlett and D. Jennings, SQL Server 2000 and XML: Developing XML-Enabled data Solutions for the Web, MSDN magazine, Jan. 2002, available at <http://msdn.microsoft.com/library/default.asp?url=/msdnmag/issues/0800/sql2000/toc.asp>
- [9] IBM Corporation, DB2 XML Extender, IBM Corporation, 2000, available at <http://www-4.ibm.com/>
- [10] V.Aguilera, S.Cluet, P.Veltri, D.Vodislav, and F.Wattez, Querying XML Documents in Xyleme, SIGIR, 2000
- [11] C. Chung, J. Min, and K. Shim, APEX: An Adaptive Path Index for XML data, SIGMOD, 2002
- [12] F. Rizzolo and A. Mendlzon, Indexing XML data with ToXin, 4<sup>th</sup> Int. Workshop on the Web and database, 2001
- [13] B. F. Cooper, N. Sample, M. J. Franklin, G. R. Hialtason, and M. Shadmon, A Fast Index for semistructured data, VLDB, 2001
- [14] B. F.Cooper, N. Sample, and M. Shadmon, A Parallel Index for semistructured data, SAC, 2002

- [15] M. Fernandez and D. Suciu, Optimizing Regular Path Expressions Using Graph Schemas, ICDE, 1998
- [16] R. Goldman and J. Widom, Dataguides: Enabling Query Formulation and Optimization in semistructured databases, VLDB, 1997
- [17] T. Milo and D. Suciu, Index Structures for Path Expressions, ICDT, 1997
- [18] J. McHugh, J. Widom, S. Abitboul, Q. Luo, and A. Rajaraman, Indexing semistructured data, In Stanford Technical Report, Jan. 1998
- [19] W3C Consortium, XML Path language (XPath) Version 1.0, W3C Recommendation 16 Nov. 1999, available at <http://www.w3.org/TR/xpath.html>
- [20] W. Thomas, Automata on Infinite Objects, Handbook of Theoretical Computer Science, Vol. B, 135-191, 1990
- [21] W3C Consortium, Xquery1.0 Formal Semantics, W3C Working Draft 16 Aug. 2002, available at <http://www.w3.org/TR/query-semantics>
- [22] P. Selinger, M. Astrahan, D. Chamberlin, R. Lorie, and T. Price, Access Path Selection in a Relational Database Management System, Proceedings of the ACM SIGMOD International Conference on Management of Data, Boston, MA, 1979
- [23] J. Kim, W. Lee, K. Lee. "The Cost Model for XML Documents", In Proc. of ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, Jun. 2001.