Lecture 17

Wednesday, May 29, 2002

XML Storage

Final Review

XML Storage in a Relational DB

- Use generic schema
  - [Florescu, Kossman 1999]
- Use DTD to derive schema
  - [Shanmugasundaram, et al. 1999]
- Use data mining to derive schema
  - [Deutsch, Fernandez, Suciu 1999]
- Use the Path table
  - [T.Amagasa, T.Shimura, S.Uemura 2001]

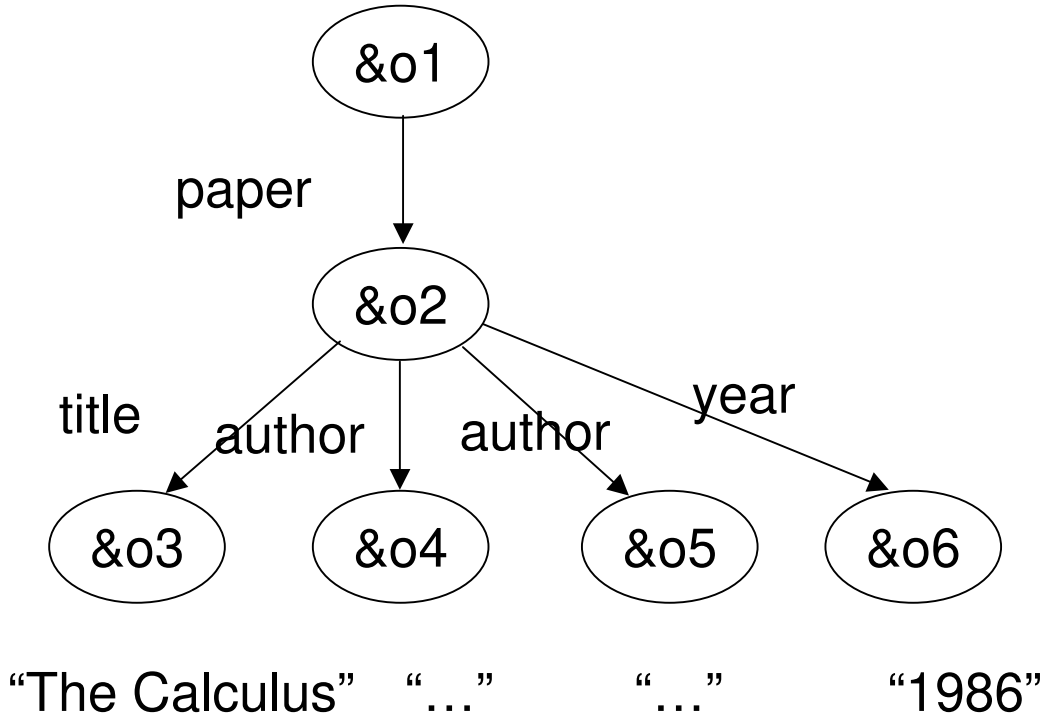
XML Stoarge: Ternary Relation

- [Florescu, Kossman 1999]
- Use generic relational schema (independent on the XML schema):

Ref(source,label,dest)

Val(node,value)

XML Stoarge: Ternary Relation



Ref

Source	Label	Dest
&o1	paper	&o2
&o2	title	&o3
&o2	author	&o4
&o2	author	&o5
&o2	year	&o6

Val

Node	Value
&o3	The Calculus
&o4	…
&o5	…
&o6	1986

[Florescu, Kossman 1999]

XML Stoarge: Ternary Relation

- Xpath to SQL translation:
- Xpath:

```
/paper[year="1986"]/author
```

- SQL:

```
Select . . . . . .

From . . . . . .

Where . . . . . .
```

XML Stoarge: Ternary Relation

- In practice may need more table:

RefTag1(source,dest)

RefTag2(source,dest)

…

IntVal(node,intVal)

RealVal(node,realVal)

…

XML Storage: DTD to Schema

[Christophides, Abiteboul, Cluet, Scholl 1994]

[Shanmugasundaram, Tufte, He, Zhang, DeWitt, Naughton 1999]

- Idea: use the XML schema to derive the relational schema

XML Storage: DTD to Schema

- DTD:

```
<!ELEMENT paper (title, author*, year?)>
<!ELEMENT author    (firstName, lastName)>
```

- Relational schema:

Paper(<u>pid</u>, title, year)
Author(<u>aid</u>, pid, firstName, lastName)

XML Storage: DTD to Schema

- Xpath to SQL translation:

```
/paper[year="1986"]/author
```

- Xpath:
- SQL:

```
Select . . . . . .

From . . . . . .

Where . . . . . .
```

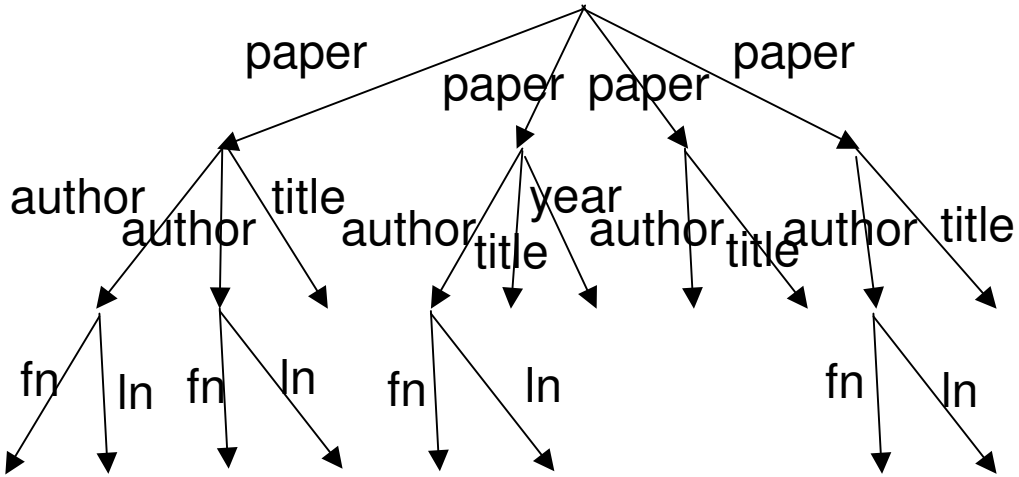
XML Storage: Data Mining to Schema

[Deutsch, Fernandez, Suciu 1999]

- Given:
  - One large XML data instance
  - No schema/DTD
  - Query workload
- Problem: find a “good” relational schema for it

- Notice: even when a DTD is present, it may be imprecise:
  - E.g. when a person may have 1-3 phones: phone*

XML Storage: Data Mining to Schema



Paper1

fn1	ln1	fn2	ln2	title	year
X	X	X	X	X	-
X	X	-	-	X	X
X	X	-	-	X	-

Paper2

author	title
X	X

[Deutsch, Fernandez, Suciu 1999]

XML Storage: Data Mining to Schema

- Xpath to SQL translation:

```
/paper[year="1986"]/author
```

- Xpath:

- SQL:

XML Storage: the Path Relation Method

- [T.Amagasa, T.Shimura, S.Uemura 2001]

- Store paths as strings
- Xpath expressions become the SQL **like** operator
- Additional information for parent/child, ancestor/descendant relationship

XML Storage: the Path Relation Method

Path

pathID	Pathexpr
1	#/bib
2	#/bib#/paper
3	#/bib#/paper#/author
4	#/bib#/paper#/title
5	#/bib#/paper#/year
6	#/bib#/book#/author
7	#/bib#/book#/title
8	#/bib#/book#/publisher

One entry for every *path* in the database
Relatively small

XML Storage: the Path Relation Method

Element

NodeID	pathID	Start	End	ParentID
1	1	0	1000	-
2	2	5	200	1
3	3	8	20	2
4	3	21	30	2
5	3	31	100	2
6	3	101	150	2
7	4	151	180	2
8	2	300	500	1
. . .				

One entry for every *element* in the database
Relatively large

XML Storage: the Path Relation Method

Val

NodeID	Val
3	Smith
4	Vance
5	Tim
6	Wallace
7	The Best Cooking Book Ever
6	3
7	4
8	2
. . .	

One entry for every *leaf* in the database
Relatively large

XML Storage: the Path Relation Method

- Xpath to SQL translation:

```
/bib/paper[year="1986"]//figure
```

- Xpath:
- SQL:

```
Select . . . . . .

From . . . . . .

Where . . . . . .
```

The Project

What to do:

- A website.
- A short printed description.
  - Could be a printout of the website.
- A presentation (this Friday).

Due dates:

- soft deadline is Friday, 5/31 (for most of the project)
- hard deadline is Friday, 6/7 (for selected remaining experiments)

The Project

What to address:

- What problem you are trying to solve ? Why is it interesting ?
- How did you approach it ?
- What did you achieve ? What did you implement, evaluate, learn ?
- Who did what in the project ?

The Project

The Presentations:

- Friday, 1:30-2:20, Low 105
- Following order:
  1.
  2.
  3.
  4.

The Final

- Monday, June 10, 2:30-4:30
- Lowe 102 (this room)
- Open book exam !

The Final

1. SQL
2. XPath/XQuery
3. Theory
4. Database implementation
5. XML processing

1. SQL

- Select-from-where
- Group-by, having
- Insert, delete, modify tables
- Create tables
- Need to understand E/R diagrams

Excluded: constraints, triggers

2. XQuery

- Basic FLWR expressions
- Nested queries
- Joins
- Aggregates
- Please use correct syntax (slides often don't do that)
  - see XQuery's use cases, www.w3.org/TR/xmlquery-use-cases

Should be simpler than SQL

3. Theory

- First Order Logic
- Domain independence
- Expressive power
- Query complexity
- Conjunctive queries
  - Containment
  - Semijoin reduction

4. Database Implementation

- Data storage
- Indexing
  - B^+ trees
  - Hash tables
- Execution
  - Various algorithms and their complexity
- Optimization
  - Know basic algebraic laws
  - Dynamic programming

5. XML Processing

- Basic syntax (well-formed XML documents):
  - Elements, attributes
- XML and semistructured data
- Schemas (DTDs)
- Publishing
  - Define XML view in Xquery
  - Translate XQuery to SQL
- Storing XML in relational databases

Grading

Breakdown:

- Homework: 35%
- Project: 35%
- Final: 25%
- Intangibles: 5%

Compared to the syllabus: more weight on the project, less on the final

...and finally !

- Enjoy taking the final !
- I enjoyed teaching this class ☺