



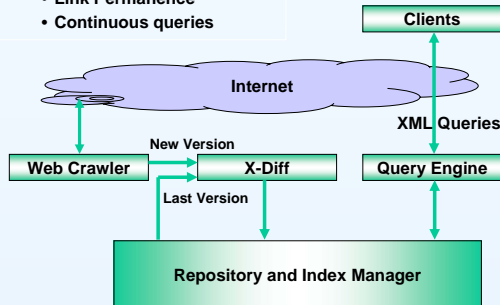
Temporal Queries in XML Document Archives and Historical WWW Warehouses

Fusheng Wang and Carlo Zaniolo Computer Science Department, UCLA

Web Information Warehouses

Applications:

- Historical queries
- Link Permanence
- Continuous queries



Technical Challenges

- Efficient document difference algorithms (x-diff)
- Efficient incremental storage and retrieval
- XML-based representation for document history
- Support for temporal queries
 - XML queries such as XPath and XQuery queries

Efficient Storage and Retrieval - Previous Work

- RCS (Revision Control System)
- SCCS (Source Code Control System)
- Our improvement
 - Usefulness-Based Copy Control (UBCC)
 - clustering pages in the new version by a threshold of usefulness
 - Reference-Based Version Modeling (RBVM)
 - keep a view for each version, and share unchanged elements among versions by references
- SPaR
 - each node is assigned SPaR numbers for document restructuring, indexing, and XPath queries

Support for Complex Queries

- Complex Queries
 - Retrieval of versioned documents
 - e.g., retrieve the 3rd version of employee document
 - Queries on changes between versions, and
 - Queries on the evolution of the document and its elements
 - e.g., find the history of the salary of employee with empno "e1"
- Can we support them in XQuery – the coming XML Query standard?

Original XML Document

```

<employees>
  <employee>
    <empno>e1</empno>
    <firstname>Bob</firstname>
    <lastname>Thompson</lastname>
    <salary>60000</salary>
    <title>Assistant Provost</title>
    <DateofBirth>1945-04-09</DateofBirth>
    <phone>3978</phone>
    <validtime start="1995-01-01" end="1995-06-01"/>
  </employee>
  <employee>
    <empno>e1</empno>
    <firstname>Bob</firstname>
    <lastname>Thompson</lastname>
    <salary>70000</salary>
    <title>Assistant Provost</title>
    <DateofBirth>1945-04-09</DateofBirth>
    <phone>3978</phone>
    <validtime start="1995-06-01" end="1995-10-01"/>
  </employee>
</employees>
  
```

Version 1

Version 2

XML Representation of a Document History

```

<employees vstart="1" vend="4">
  <employee vstart="1" vend="4">
    <empno vstart="1" vend="4">e1</empno>
    <firstname vstart="1" vend="4">Bob</firstname>
    <lastname vstart="1" vend="4">Thompson</lastname>
    <salary vstart="1" vend="1">60000</salary>
    <salary vstart="2" vend="4">70000</salary>
    <title vstart="1" vend="2">Assistant Provost</title>
    <title vstart="3" vend="4">Provost</title>
    <DateofBirth vstart="1" vend="4">1945-04-09</DateofBirth>
    <phone vstart="1" vend="3">3978</phone>
    <phone vstart="4" vend="4">4002</phone>
    <validtime vstart="1" vend="1" start="1995-01-01" end="1995-06-01"/>
    <validtime vstart="2" vend="2" start="1995-06-01" end="1995-10-01"/>
    <validtime vstart="3" vend="3" start="1995-10-01" end="1996-02-01"/>
    <validtime vstart="4" vend="4" start="1996-02-01" end="1997-01-01"/>
  </employee>
</employees>
  
```

Complex Queries in XQuery

- Snapshot queries: retrieve the 3rd version of employee document

```

for $e in document("temporalschema/delta3.xml")/employees
return snapshot( $e, 3)
  
```

- Evolutionary queries: find the history of the salary of employee with empno "e1"

```

for $salary in document("temporalschema/delta3.xml")/
  employees/employee/attr[@name='empno']='e1']/salary
return $salary
  
```

```

<quip:result>
  <salary vstart="1" vend="1">60000</salary>
  <salary vstart="2" vend="4">70000</salary>
</quip:result>
  
```

Storage and Indexing

- Each Document element is identified by:
 - its durable node number (dnn),
 - the range of its subelements (range), and
 - the version interval for its current content (vstart, vend)
- e.g., <node vstart="1" vend="6" dnn="100" range="200">My Node Value</node>
- Storage organization:
 - Usefulness-based clustering for efficient version retrieval, and
 - Multiversion B+ trees for evolution and path-expression queries