SCHEMA EVOLUTION IN WIKIPEDIA:

toward a Web Information System Benchmark

Carlo A. Curino
(carlo.curino@polimi.it)

Hyun J. Moon
Letizia Tanca
Carlo Zaniolo
Motivations

- **Information Systems (IS)** are subject of *continuous evolution*
  - requirements change to adapt to an evolving environment
  - waterfall development methodologies are *inadequate* (especially on the WEB)

- **Evolution** in IS is an extremely difficult and *expensive* task
  - software evolution and maintenance represents 90% of the *costs*
  - Legacy Applications cannot be modified, but must be supported

- **The data management core:**
  - is one of the *most difficult* portion of a system to evolve
  - data evolves: need for *historical archiving* due to accountability
  - schema evolves: with *dramatic impact* on queries and applications
• **Why Wikipedia**: DB-centric Web Information System (WIS), very popular, open-source license of the software platform (MediaWiki) and data, rich schema history

• **Tools**: We developed a tool-suite to analyze WIS DB backends

• **The analysis**: We collect and dissect MediaWiki schema history: 170+ schema versions in 4.5 years

• **The benchmark**: We publicly released tools, results and datasets as a first step: “towards a Benchmark for Schema Evolution”
Scalability issues:

- DB size exceeds 700 Gb (excluding multimedia content)
- Wikipedia receives an average of 29k requests/sec (peak 85k) each producing several DB queries
- Several Layers of caching
- DBMS performance is the bottleneck: poor partitioning?
• **Tables** can be grouped in:
  - article and content (6 tables)
  - links and structure (9 tables)
  - users and permissions (5 tables)
  - performance and caching (7 tables)
  - statistics and special features (3 tables)
  - history and archival (4 tables)

• **NOTE**: History management represents about **30% of the schema** (4 tables + several attributes in the other tables)
• **Schema Evolution:**
  - 170+ versions in 4.5 years
  - almost 250% increase

• **WIS evolve faster** than Traditional IS
  - 38% w.r.t. [Sjoberg93]
  - 539% w.r.t. [Marche93]

• **Note:** Collaborative WISs embrace information sharing (better data to study!)
• More frequent schema changes far away from releases

• Schema Elements Lifetime:
  • a group of stable relations
  • young tables and columns
## Type of Changes

<table>
<thead>
<tr>
<th>Type of Change</th>
<th># of evolution steps</th>
<th>% of evolution steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Schema</td>
<td>94</td>
<td>54.9%</td>
</tr>
<tr>
<td>Index/Key</td>
<td>69</td>
<td>40.3%</td>
</tr>
<tr>
<td>Data Type</td>
<td>22</td>
<td>12.8%</td>
</tr>
<tr>
<td>Syntax Fix</td>
<td>20</td>
<td>11.7%</td>
</tr>
<tr>
<td>Rollback</td>
<td>15</td>
<td>8.8%</td>
</tr>
<tr>
<td>Doc Only</td>
<td>13</td>
<td>7.6%</td>
</tr>
<tr>
<td>Engine</td>
<td>6</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

- **NOTE**: sum exceeds 100%, since several changes might coexist in a single evolution step
- Total lack of integrity constraints (except for primary keys)!!
## Changes: by Schema Modification Operators

<table>
<thead>
<tr>
<th>SMO type</th>
<th># of usages</th>
<th>% of usage</th>
<th>% per version</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREATE TABLE</td>
<td>24</td>
<td>8.9%</td>
<td>14%</td>
</tr>
<tr>
<td>DROP TABLE</td>
<td>9</td>
<td>3.3%</td>
<td>5.2%</td>
</tr>
<tr>
<td>RENAME TABLE</td>
<td>3</td>
<td>1.1%</td>
<td>1.75%</td>
</tr>
<tr>
<td>DISTRIBUTED TABLE</td>
<td>0</td>
<td>0.0%</td>
<td>0%</td>
</tr>
<tr>
<td>MERGE TABLE</td>
<td>4</td>
<td>1.5%</td>
<td>2.33%</td>
</tr>
<tr>
<td>COPY TABLE</td>
<td>6</td>
<td>2.2%</td>
<td>3.5%</td>
</tr>
<tr>
<td>ADD COLUMN</td>
<td>104</td>
<td>38.7%</td>
<td>60.4%</td>
</tr>
<tr>
<td>DROP COLUMN</td>
<td>71</td>
<td>26.4%</td>
<td>41.5%</td>
</tr>
<tr>
<td>RENAME COLUMN</td>
<td>43</td>
<td>16.0%</td>
<td>25.1%</td>
</tr>
<tr>
<td>MOVE COLUMN</td>
<td>1</td>
<td>0.4%</td>
<td>0.58%</td>
</tr>
<tr>
<td>COPY COLUMN</td>
<td>4</td>
<td>1.5%</td>
<td>2.33%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>269</strong></td>
<td><strong>100%</strong></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** This operators have been used in PRISM [vdlb2008a] and PRIMA [vdlb2008b] systems.
• NOTE: version 41-42 represents a MAJOR evolution step where article versioning management is heavily modified!!
**NOTE**: over ~4000 queries (version 28) from which we extract 75 templates
Real Wikipedia Queries

- NOTE: 500 most common templates out of 2k extracted from over 780 millions query instances from the On-Line Wikipedia Profiler http://noc.wikimedia.org/cgi-bin/report.py
Conclusions (1/2)

• **Strong evidence:** need for better Schema Evolution and Data Archiving

• Schema history, analysis data (raw and aggregated), queries and tool-suite are available at:

Goal: create a benchmark for schema evolution (and in general a standard relational DB dataset).

- Extend the analysis to several other Open-Source WIS (Joomla!, TikiWiki, Slashcode, Zen-Cart, Wordpress)
- Extend the analysis towards Public Scientific DB (Genome, HGVS)

This work is part of the bigger project "Panta Rhei" tackling:

- Schema Evolution: PRISM [vldb2008a]
- Transaction Time DB under schema evolution: PRIMA [VLDB-2008b]
- History Metadata Management [STSM-2008], [ECDM-2008]

for more information visit:

http://yellowstone.cs.ucla.edu/schema-evolution/

