An XML-IR-DB Sandwich

Is it better with an Algebra in Between?

Vojkan Mihajlovic
Djoerd Hiemstra
Henk Ernst Blok
Peter M. G. Apers
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
XML-IR and Relational DBs

- XPath and XQuery:
  - Navigation in XML structure

- Relational Databases:
  - Rules for relational table manipulation

- Missing:
  - Sound specification of IR tasks
  - Rules for score propagation and correlation
  - Connection between the two
Our Approach

Three level DBMS:
- Conceptual level:
  - XPath+IR (NEXI)
- Logical level:
  - extended region algebra
- Physical level:
  - relational model
Intermediate level

- Algebraic approach
  - XML navigation is supported
  - Ranking is a part of the algebra

- Opportunities
  - Query rewriting and optimization
  - ... also for IR-like queries
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
XML example

<article lang="en" date="10/02/04">
  <title>Region algebra</title>
  <bdy>
    <sec>
      <sec>
        <p>Structured documents ...</p>
        <p>Text search ...</p>
      </sec>
    </sec>
  </bdy>
</article>
XML example - Index (step 1)

<article lang="en" date="10/02/04">
  <title>Region algebra</title>
  <bdy>
    <sec>
      <p>Structured documents ...</p>
      <p>Text search ...</p>
    </sec>
    ...
  </bdy>
</article>

WIRD'04
## Indexed XML example (step 2)

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10034</td>
<td>article</td>
<td>node</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>lang</td>
<td>attr_name</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>en</td>
<td>attr_value</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>date</td>
<td>attr_name</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>10/02/04</td>
<td>attr_value</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>title</td>
<td>node</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>-</td>
<td>text</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>region</td>
<td>term</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>algebra</td>
<td>term</td>
</tr>
<tr>
<td>9</td>
<td>9876</td>
<td>bdy</td>
<td>node</td>
</tr>
<tr>
<td>10</td>
<td>576</td>
<td>sec</td>
<td>node</td>
</tr>
<tr>
<td>11</td>
<td>54</td>
<td>p</td>
<td>node</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
<td>-</td>
<td>text</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>structured</td>
<td>term</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>documents</td>
<td>Term</td>
</tr>
</tbody>
</table>

... ... ...
The Storage of XML

- Fragmentations
  - Horizontal
    - XML node type
  - Vertical
    - name and type of XML elements
- Path-based
- Not unified
Queries & Relational Algebra

Bottleneck

- Descendant/ancestor step
- Join and projection combination

\[ R = \pi_{\text{start}_2, \text{end}_2, \text{name}_2} (R_2 \bowtie_{\text{start}_2 > \text{start}_1, \text{end}_2 < \text{end}_1} R_1) \]

\[ R = \pi_{\text{start}_2, \text{end}_2, \text{name}_2} (R_2 \bowtie_{\text{start}_2 < \text{start}_1, \text{end}_2 > \text{end}_1} R_1) \]

- "Containment join" (.intersection and intersection)
Example query

```
//article//bdy[about(//sec, structured)
and about(//sec, documents)]
```

\[
\sigma_{bdy}(N) \quad \sigma_{article}(N) \quad \sigma_{sec}(N) \quad \sigma_{structured}(W) \quad \sigma_{sec}(N) \quad \sigma_{documents}(W)
\]
Intermezzo: Logical Algebra

- **Relational algebra:**
  - New operators for IR-like queries
  - Relational query plan highly dependant on relational storage
  - Not XML (structure) aware

- **Logical Algebra**
  - Right level of abstraction for IR operators
  - Data independence
  - Query rewriting and optimization on logical level
  - IR understanding and IR operator optimization
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
Region Algebra

Domain $R = \{r | r = (s, e, n, t)\}$

Operators

- select $\sigma_{\text{name}}(R)$
- containing $
- contained by $
- intersection $
- union $

WIRD’04
Example logical expression

\[ \sigma_{n=\text{bdy}}(C) \sigma_{n=\text{article}}(C) \sigma_{n=\text{sec}}(C) \sigma_{n=\text{structured}}(C) \sigma_{n=\text{sec}}(C) \sigma_{n=\text{documents}}(C) \]

\[ \text{ARTICLE} = \sigma_{n=\text{article}}(C) \]
Logical query plan

//article//bdy[about(., region) and about(., algebra)]
[about(./sec, XML) //p[about(., information) and about(., retrieval)]

R1 = BDY □ ARTICLE
R2 = ((R1 □ REGION) ▭ (R1 □ ALGEBRA)) ▭ (SEC ▭ XML)
R3 = P □ R2
R4 = (R3 ▭ INFORMATION) ▭ (R3 ▭ RETRIEVAL)

ARTICLE = σ_{n=article}(C)
Algebra Operator Properties (1)

### Regular
- Identity \( \{(\cap, C), (\cup, \emptyset)\} \)
- Commutativity \( \{\cap, \cup\} \)
- Associativity \( \{\cap, \cup\} \)
- Distributivity \( \{(\subseteq, \cup), (\subseteq, \cup), (\cap, \cup), (\cup, \cap)\} \)
Algebra Operator Properties (2)

Special cases

1. \( \text{op1} = \{ \sqcap, \sqcup \}; \text{op2} = \{ \sqcup, \sqcap \} \)
   
   \((R_1 \text{op1} R_2) \text{op2} R_3 = (R_1 \text{op2} R_3) \text{op1} R_2 \)
   
   \((R_1 \text{op1} R_2) \text{op2} R_3 = (R_1 \text{op1} R_2) \sqcap (R_1 \text{op2} R_3) \)

2. \( \text{op1} = \{ \sqcap, \sqcup \}; \text{op2} = \{ \sqcup, \sqcap \} \)
   
   \((R_1 \text{op1} R_2) \text{op2} R_3 = (R_1 \text{op2} R_3) \text{op1} (R_2 \text{op2} R_3) \)
Properties in action (1)

1. \((\text{BDY} \sqsubset \text{ARTICLE}) \sqsubset ((\text{SEC} \sqsubset \text{STRUCTURED}) \sqcap \text{SEC} \sqsubset \text{DOCUMENTS}))\)

2. \((\text{BDY} \sqsubset ((\text{SEC} \sqsubset \text{STRUCTURED}) \sqcap \text{SEC} \sqsubset \text{DOCUMENTS})) \sqsubset \text{ARTICLE}\)

1. \((\text{BDY} \sqsubset ((\text{SEC} \sqsubset \text{DOCUMENTS}) \sqsubset \text{STRUCTURED})) \sqsubset \text{ARTICLE}\)

1. \((\text{BDY} \sqsubset ((\text{SEC} \sqsubset \text{DOCUMENTS}) \sqsubset \text{STRUCTURED})) \sqsubset \text{ARTICLE}\)
Properties in action (2)

((ARTICLE ⊑ REGION) ⊑ (ARTICLE ⊑ ALGEBRA)) ⊑ (SEC ⊑ XML)

1

((ARTICLE ⊑ REGION) ⊑ (SEC ⊑ XML)) ⊑
((ARTICLE ⊑ ALGEBRA) ⊑ (SEC ⊑ XML))

2

((ARTICLE ⊑ (SEC ⊑ XML)) ⊑ REGION) ⊑
((ARTICLE ⊑ (SEC ⊑ XML)) ⊑ ALGEBRA)

3

((ARTICLE ⊑ (SEC ⊑ XML)) ⊑ REGION) ⊑ ALGEBRA

WIRD’04
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
Scoring in region algebra

Domain $R = \{r | r = (s, e, n, t, p)\}$

New operators
- ranked containing $\sqsupseteq_p$
- ranked contained by $\sqsubseteq_p$
- ranked intersection $\sqcap_p$
- ranked union $\sqcup_p$
Scoring operators

\[
\begin{align*}
\# R_1 \boxdot_p R_2 & \quad p = p_1 \bullet f_\boxdot(r_1, R_2) \\
\# R_1 \odot_p R_2 & \quad p = p_1 \bullet f_\odot(r_1, R_2) \\
\# R_1 \bigoplus_p R_2 & \quad p = p_1 \bigoplus p_2 \\
\# R_1 \bigodot_p R_2 & \quad p = p_1 \bigotimes p_2 \\
\end{align*}
\]
Scoring functions and operators

- Scoring functions:
  - structural relation
  - score values

- Abstract operators:
  - “and” expression
  - “or” expression

\[
\begin{align*}
  f_{\triangledown}(r, R) \\
  f_{\subseteq}(r, R) \\
  \otimes = \{\cdot, \text{min}, \ldots\} \\
  \oplus = \{+, \text{max}, \ldots\}
\end{align*}
\]
Complex scoring functions

\[ f_\subset (r, R) = \sum_{\bar{r} \in R \subset R'} (g_\subset (\bar{r}, r) \cdot \overline{p}) \]

\[ f_\supset (r, R) = \sum_{\bar{r} \in R \subset R'} (g_\supset (\bar{r}, r) \cdot \overline{p}) \]

\[ R' = \{ r \} \]

Possible implementation of \( g \):

\[ g_\supset (\bar{r}, r) = \frac{\text{size}(\bar{r})}{\text{size}(r)} \]

\[ g_\subset (\bar{r}, r) = \frac{\text{size}(\bar{r})}{\sum_{\bar{r}} \text{size}(r)} \]
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
Properties of scoring operators

\[(R \sqcap_p C) = (C \sqcap_p R)\]
\[(R \cup_p \emptyset) = (\emptyset \cup_p R)\]
\[p \cdot 1 = 1 \cdot p = p\]
\[p + 0 = 0 + p = p\]
\[\forall r \in R\]

\[R_1 \cap_p (R_2 \cup_p R_3) = (R_1 \cap_p R_2) \cup_p (R_1 \cap_p R_3)\]

\[(R_1 \text{ op1 } R_2) \text{ op2 } R_3 = (R_1 \text{ op2 } R_3) \text{ op1 } R_2\]

\[\text{op1} = \{\sqcap_p, \sqcup_p\} \quad \text{op2} = \{\sqcap_p, \sqcup_p\}\]

\[p = (p_1 \cdot f(r_1, R_2)) \cdot f(r_1, R_3) = (p_1 \cdot f(r_1, R_3)) \cdot f(r_1, R_2)\]

only if \[f(r, R) = f(s, n, t, R)\]
... conditional properties(1) ...

(R1 \text{ op1} R2) \text{ op2} R3 = (R1 \text{ op1} R2) \cap_p (R1 \text{ op2} R3)

\text{ op1} = \{\Box_p, \bigcirc_p\} \quad \text{ op2} = \{\boxdot_p, \boxtimes_p\}

p = (1 \bullet \sum r \ (g(\bar{r}, r_1)) \bullet p_2) \bullet \sum r \ (g(\bar{r}, r_1)) \bullet p_3

= (1 \bullet \sum r \ (g(\bar{r}, r_1)) \bullet p_2)) \bullet (1 \bullet \sum r \ (g(\bar{r}, r_1)) \bullet p_3)
Conditional properties - example

$$((P \square_p \text{SEC}) \square_p \text{INFORMATION}) \square_p$$

$$((P \square_p \text{SEC}) \square_p \text{RETRIEVAL})$$

$$((P \square_p \text{INFORMATION}) \square_p (P \square_p \text{RETRIEVAL}))$$

$$\square_p \text{SEC}$$

$$((P \square_p \text{INFORMATION}) \square_p \text{RETRIEVAL})$$

$$\square_p \text{SEC}$$
Conditional properties (2)

\[ \text{op1} = \{ \square_p, \blacklozenge_p \} \]

\[ (R_1 \square_p R_2) \text{ op1 } R_3 = (R_1 \text{ op1 } R_2) \square_p (R_2 \text{ op1 } R_3) \]

\[ (p_1 \cdot p_2) \cdot f(r_{1,2}, R_3) = (p_1 \cdot f(r_{1,2}, R_3)) \cdot (p_2 \cdot f(r_{1,2}, R_3)) \]

\[ (R_1 \blacklozenge_p R_2) \text{ op1 } R_3 = (R_1 \text{ op1 } R_2) \blacklozenge_p (R_2 \text{ op1 } R_3) \]

\[ (p_1 + p_2) \cdot f(r_{\parallel 2}, R_3) = (p_1 \cdot f(r_{\parallel 2}, R_3)) + (p_2 \cdot f(r_{\parallel 2}, R_3)) \]
Outline

- Motivation
- XML and Relational Databases
- Region Algebra & Operator Properties
- Region Algebra & Relevance Ranking
- Properties of Ranking Operators
- Conclusions and Future Work
Conclusions

- Problem: Execution of IR-like queries over XML documents stored in relational database
- Usefulness of intermediate logical level based on region algebra (with score computation)
  - Data independence between levels
  - Right level of abstraction (understanding IR)
  - Opportunities for query optimization on logical level (including ranking operators)
... still to come

- Experimental evaluation: benefits of intermediate logical level
- The definition of score operators => operator properties
- Usage of different retrieval models
- Theoretical foundation for the definition of score operators