SIGIR 2004

Workshop on the Integration of Information Retrieval and Databases (WIRD'04)

An XML-IR-DB Sandwich

Is it better with an Algebra in Between?

Vojkan Mihajlovic

Djoerd Hiemstra

Henk Ernst Blok

Peter M. G. Apers

University of Twente
department of computer science

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>
                <p>Structured documents ...</p>
                <p> Text search ...</p>
        </sec>
        ...
   </bdy>
   ...
</article>
```

XML example - Index (step 1)

<article^{0} lang^{1}=“en”^{2} date^{3}=“10/02/04”^{4}>

<title>^{5}Region^{6} algebra^{7}</title>^{8}

<bdy>^{9}

<sec>^{10}

<p>^{11}Structured^{12} documents^{13} …</p>^{54}

<p>^{55}Text^{56} search^{57} …</p>^{575}

</sec>^{576}

…

</bdy>^{9876}

…

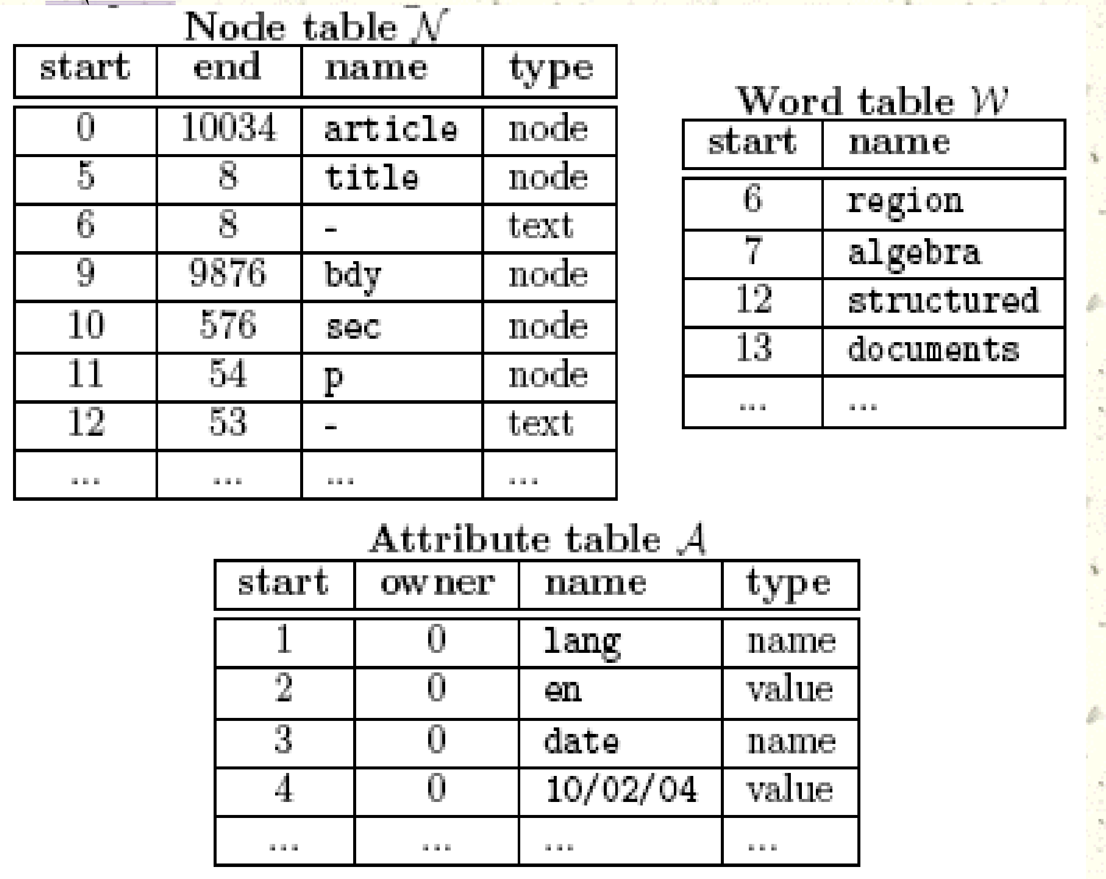
</article>^{10034}

- ■ Start tag
- ■ End tag
- ■ Attribute name
- ■ Attribute value
- ■ Term
- …

Indexed XML example (step 2)

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

The Storage of XML



Node table $\mathcal{N}$

start	end	name	type
0	10034	article	node
5	8	title	node
6	8	-	text
9	9876	bdy	node
10	576	sec	node
11	54	p	node
12	53	-	text
...	...	...	...

Word table $\mathcal{W}$

start	name
6	region
7	algebra
12	structured
13	documents
...	...

Attribute table $\mathcal{A}$

start	owner	name	type
1	0	lang	name
2	0	en	value
3	0	date	name
4	0	10/02/04	value
...	...	...	...

- 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

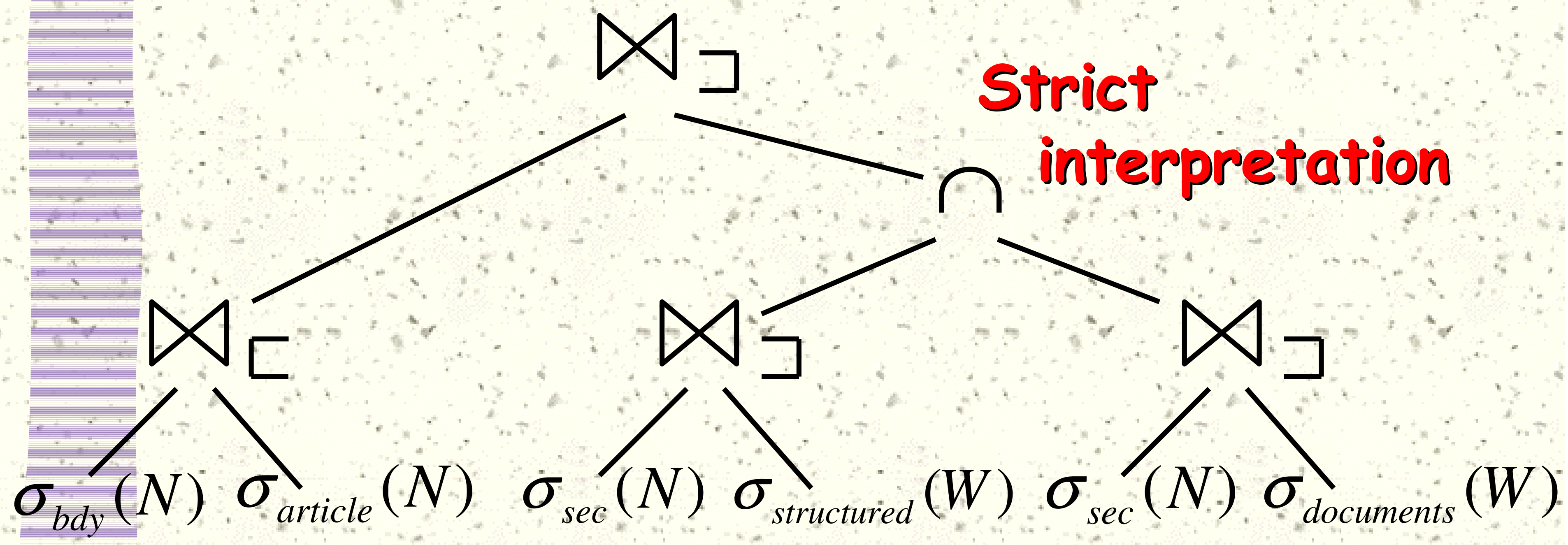
$$R = \pi_{start_2,end_2,name_2}(R_2 \rhd\lhd_{start_2>start_1,end_2<end_1} R_1)$$

$$R = \pi_{start_2,end_2,name_2}(R_2 \rhd\lhd_{start_2<start_1,end_2>end_1} R_1)$$

  - “Containment join” ($\bowtie_{\sqsupset}$ and $\bowtie_{\sqsubset}$)

Example query

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



Strict interpretation

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

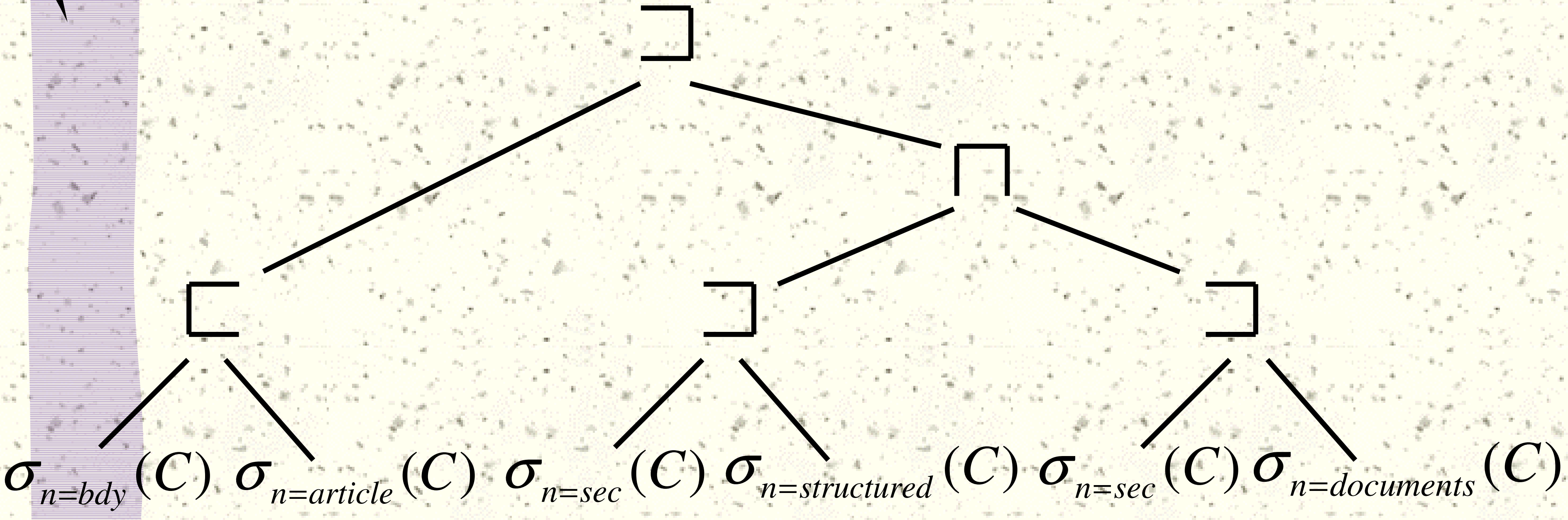
Outline

- 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 $\sigma_{n=name}(R)$
  - containing $\sqsupset$
  - contained by $\sqsubset$
  - intersection $\sqcap$
  - union $\sqcup$

Example logical expression



$\sqsupset$

$\sqcap$

$\sqsubset$ $\sqsupset$ $\sqsupset$

$\sigma_{n=bdy}(C)$ $\sigma_{n=article}(C)$ $\sigma_{n=sec}(C)$ $\sigma_{n=structured}(C)$ $\sigma_{n=sec}(C)$ $\sigma_{n=documents}(C)$

ARTICLE = $\sigma_{n=article}(C)$

Logical query plan

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

R1 = BDY $\sqsubset$ ARTICLE

R2 = ((R1 $\sqsupset$ REGION) $\sqcap$ (R1 $\sqsupset$ ALGEBRA)) $\sqsupset$ (SEC $\sqsupset$ XML)

R3 = P $\sqsubset$ R2

R4 = (R3 $\sqsupset$ INFORMATION) $\sqcap$ (R3 $\sqsupset$ RETRIEVAL)

ARTICLE = $\sigma_{n=article}(C)$

Algebra Operator Properties (1)

- Regular
  - Identity $\{(\sqcap, C), (\sqcup, \emptyset)\}$
  - Commutativity $\{\sqcap, \sqcup\}$
  - Associativity $\{\sqcap, \sqcup\}$
  - Distributivity $\{(\sqsupset, \sqcup), (\sqsubset, \sqcup), (\sqcap, \sqcup), (\sqcup, \sqcap)\}$

Algebra Operator Properties (2)

- Special cases

  - $op1 = \{\sqsupset, \sqsubset\};\ op2 = \{\sqsupset, \sqsubset\}$

1

$(R1\ op1\ R2)\ op2\ R3 = (R1\ op2\ R3)\ op1\ R2$

$(R1\ op1\ R2)\ op2\ R3 = (R1\ op1\ R2) \sqcap (R1\ op2\ R3)$

2

  - $op1 = \{\sqcap, \sqcup\};\ op2 = \{\sqsupset, \sqsubset\}$

$(R1\ op1\ R2)\ op2\ R3 = (R1\ op2\ R3)\ op1\ (R2\ op2\ R3)$

3

Properties in action (1)

(BDY ⊏ ARTICLE) ⊐((SEC ⊐STRUCTURED) ⊓
(SEC ⊐DOCUMENTS))

1

(BDY ⊐((SEC ⊐STRUCTURED) ⊓
(SEC ⊐DOCUMENTS))) ⊏ARTICLE

2

(BDY ⊐((SEC ⊐STRUCTURED) ⊐DOCUMENTS))
⊏ ARTICLE

1

(BDY ⊐((SEC ⊐DOCUMENTS) ⊐STRUCTURED))
⊏ ARTICLE

Properties in action (2)

((ARTICLE ⊐ REGION) ⊓ (ARTICLE ⊐ ALGEBRA))
⊐ (SEC ⊐ XML)

3

((ARTICLE ⊐ REGION) ⊐ (SEC ⊐ XML)) ⊓
((ARTICLE ⊐ ALGEBRA) ⊐ (SEC ⊐ XML))

1

((ARTICLE ⊐ (SEC ⊐ XML)) ⊐ REGION) ⊓
((ARTICLE ⊐ (SEC ⊐ XML)) ⊐ ALGEBRA)

2

((ARTICLE ⊐ (SEC ⊐ XML)) ⊐ REGION) ⊐ ALGEBRA

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 $\sqsupset_p$
  - ranked contained by $\sqsubset_p$
  - ranked intersection $\sqcap_p$
  - ranked union $\sqcup_p$

Scoring operators

- R1 $\sqsupset_p$ R2 $\quad p = p_1 \bullet f_{\supset}(r_1, R_2)$
- R1 $\sqsubset_p$ R2 $\quad p = p_1 \bullet f_{\subset}(r_1, R_2)$
- R1 $\sqcap_p$ R2 $\quad p = p_1 \otimes p_2$
- R1 $\sqcup_p$ R2 $\quad p = p_1 \oplus p_2$

Scoring functions and operators

- Scoring functions:
  - structural relation $f_{\supset}(r,R)$
  - score values $f_{\subset}(r,R)$
- Abstract operators:
  - "and" expression $\otimes = \enclose{circle}{\{\bullet,}\min,\ldots\}$
  - "or" expression $\oplus = \enclose{circle}{\{+,}\max,\ldots\}$

Complex scoring functions

$$f_{\subset}(r,R) = \sum_{\bar{r} \in R \subset R'} (g_{\subset}(\bar{r}, r) \bullet \bar{p})$$

$$f_{\supset}(r,R) = \sum_{\bar{r} \in R \subset R'} (g_{\supset}(\bar{r}, r) \bullet \bar{p})$$

$$R' = \{r\}$$

$$g_{\supset}(\bar{r}, r) = \frac{size(\bar{r})}{size(r)}$$

Possible implementation of g

$$g_{\subset}(\bar{r}, r) = \frac{size(\bar{r})}{\sum_{\bar{r}} size(r)}$$

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

$(R \sqcap_p \mathcal{C}) = (\mathcal{C} \sqcap_p R)$ $p * 1 = 1 * p = p$

$(R \sqcup_p \emptyset) = (\emptyset \sqcup_p R)$ $p + 0 = 0 + p = p$

$\forall r \in R$

$R1 \sqcap_p (R2 \sqcup_p R3) = (R1 \sqcap_p R2) \sqcup_p (R1 \sqcap_p R3)$

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

$\text{op1} = \{\sqsupset_p, \sqsubset_p\}$ $\text{op2} = \{\sqsupset_p, \sqsubset_p\}$

$$p = (p_1 \bullet f(r_1, R_2)) \bullet f(r_1, R_3) = (p_1 \bullet f(r_1, R_3)) \bullet f(r_1, R_2)$$

only if $f(r, R) = f(s, n, t, R)$

… conditional properties(1) …

(R1 op1 R2) op2 R3 = (R1 op1 R2) $\sqcap_p$ (R1 op2 R3)

op1 = {$\sqsupset_p$, $\sqsubset_p$}   op2 = {$\sqsupset_p$, $\sqsubset_p$ }

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

$$= (1 \bullet \sum_{\bar{r}} (g(\bar{r}, r_1)) \bullet p_2)) \bullet (1 \bullet \sum_{\bar{r}} (g(\bar{r}, r_1)) \bullet p_3)$$

4

Conditional properties - example

$((P \sqsubseteq_p SEC) \sqsupseteq_p INFORMATION) \sqcap_p$
$((P \sqsubseteq_p SEC) \sqsupseteq_p RETRIEVAL)$ $\cancel{\boxed{4}}$

$((P \sqsupseteq_p INFORMATION) \sqcap_p (P \sqsupseteq_p RETRIEVAL))$
$\sqsubseteq_p SEC$

$\boxed{4}$

$((P \sqsupseteq_p INFORMATION) \sqsupseteq_p RETRIEVAL)$
$\sqsubseteq_p SEC$

Conditional properties (2)

op1 = {⊐$_p$, ⊏$_p$}

~~**(R1⊓$_p$R2) op1 R3 = (R1 op1 R2)⊓$_p$(R2 op1 R3)**~~

$$\cancel{(p_1 \bullet p_2) \bullet f(r_{1,2}, R_3) = (p_1 \bullet f(r_{1,2}, R_3)) \bullet (p_2 \bullet f(r_{1,2}, R_3))}$$

(R1⊔$_p$R2) op1 R3 = (R1 op1 R2)⊔$_p$(R2 op1 R3)

$$(p_1 + p_2) \bullet f(r_{1\|2}, R_3) = (p_1 \bullet f(r_{1\|2}, R_3)) + (p_2 \bullet f(r_{1\|2}, R_3))$$

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