The Role of Document Structure in Querying, Scoring and Evaluating XML Full-Text Search

Sihem Amer-Yahia
AT&T Labs Research - USA
Database Department

Talk at the Universities of Toronto and Waterloo
Nov. 9th and 10th, 2005

Outline

- Introduction
- Querying
- Scoring
- Evaluation
- Open Issues

Outline

- Introduction
  - *IR vs. Structured Document Retrieval (SDR)*
  - *XML vs. IR Search*
- Querying
- Scoring
- Evaluation
- Open Issues

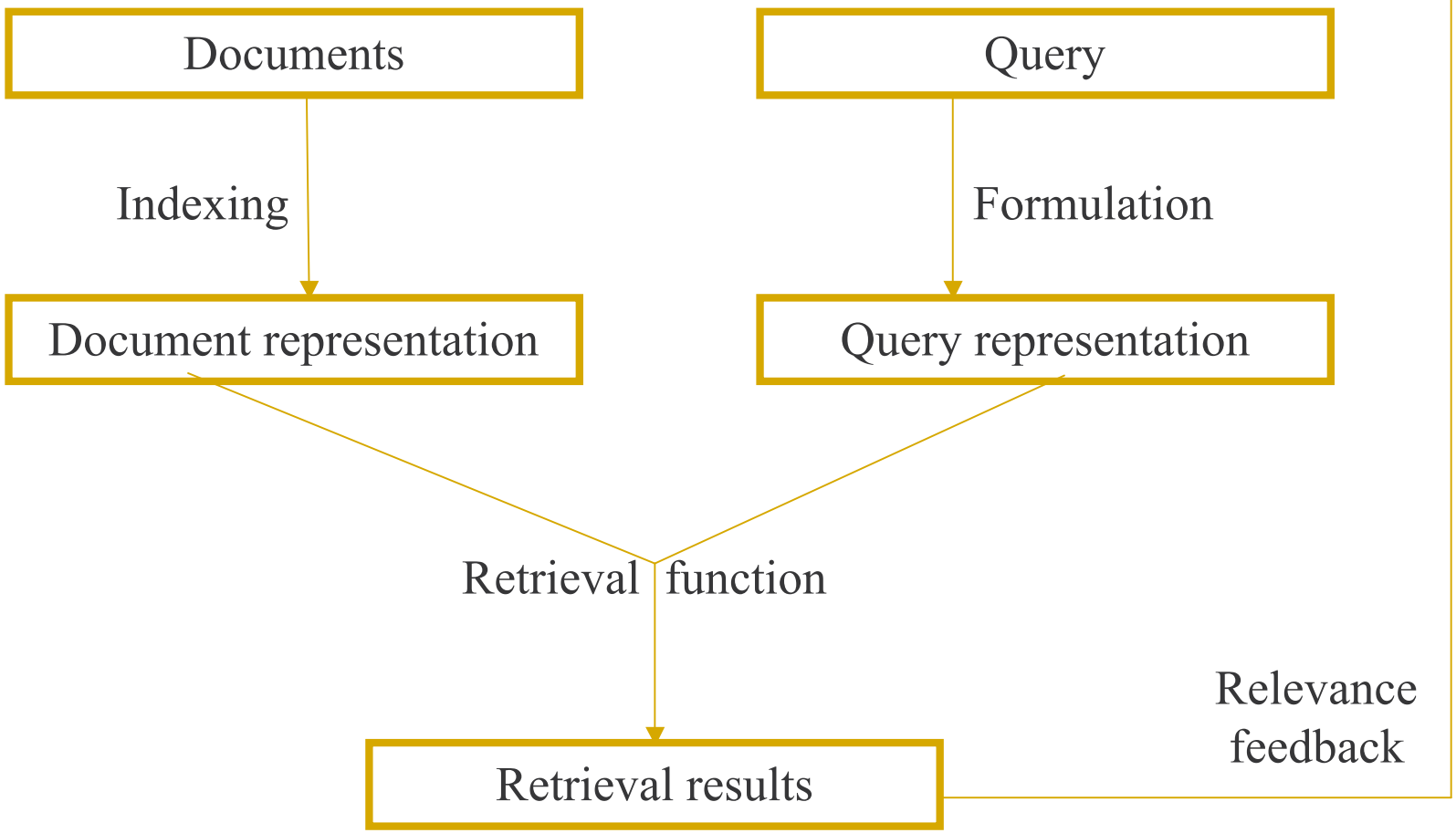
IR vs SDR

- Traditional IR is about finding relevant documents to a user's information need, e.g., entire book.

- SDR allows users to retrieve **document components** that are more focussed on their information needs, e.g., a chapter, a page.

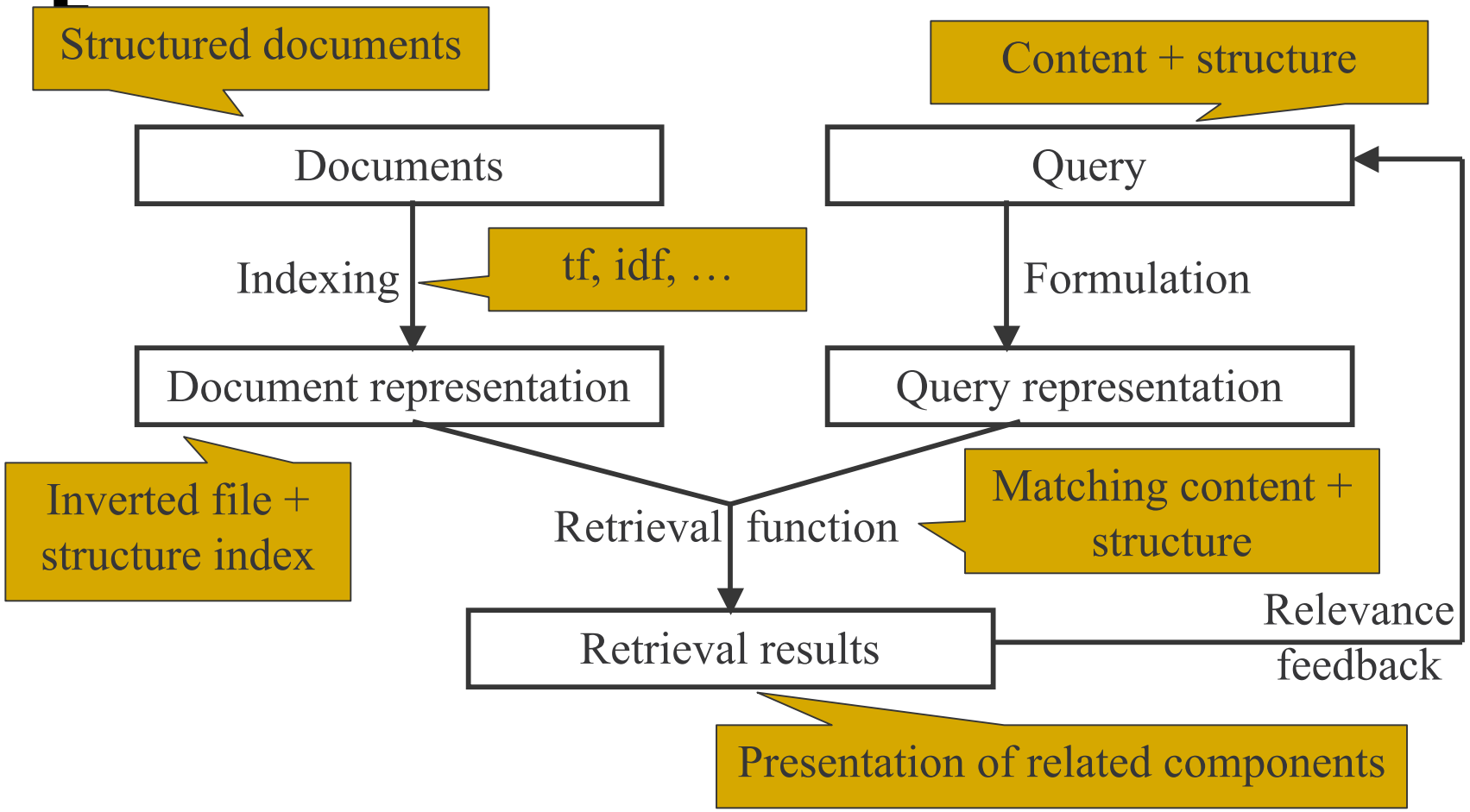
- Improve precision
- Exploit visual memory

Conceptual Model for IR



(Van Rijsbergen 1979)

Conceptual Model for SDR



Conceptual Model for SDR (XML)

Structured documents

XML adopted to represent a mix of structure and text
(e.g., Library of Congress bills, IEEE INEX data collection)

tf, idf, …

Scoring may capture document structure

structure index captures in which document component the term occurs (e.g. title, section), as well as the type of document components (e.g. XML tags)

Inverted file + structure index

Content + structure

query languages referring to both content and structure are being developed for accessing XML documents, e.g.
XIRQL, NEXI, XQUERY FT

additional constraints are imposed on structure

Matching content + structure

e.g. a chapter and its sections may be retrieved

Presentation of related components

109TH CONGRESS
1ST SESSION

H. R. 2739

To address rising college tuition by strengthening the compact between the States, the Federal Government, and institutions of higher education to make college more affordable.

IN THE HOUSE OF REPRESENTATIVES

MAY 26, 2005

Mr. TIERNEY (for himself, Ms. MCCOLLUM of Minnesota, Mr. GEORGE MILLER of California, Mr. KILDEE, Mr. EMANUEL, Mr. BISHOP of New York, Mr. PAYNE, Ms. WOOLSEY, Mrs. MCCARTHY, Mr. WU, Mr. DAVIS of Illinois, Mr. GRIJALVA, Mr. MEEHAN, Mr. BECERRA, Mr. REYES, Mr. GONZALEZ, Ms. LINDA T. SÁNCHEZ of California, Mr. MCGOVERN, Ms. DELAURO, Mr. OWENS, Mr. HINOJOSA, Mr. KUCINICH, Mr. HOLT, Mr. CASE, Mr. VAN HOLLEN, Mr. ORTIZ, Mr. GUTIERREZ, Mr. CARDOZA, Mrs. JONES of Ohio, Ms. BALDWIN, Mr. WEXLER, Mr. BARROW, Mr. JEFFERSON, Mr. RYAN of Ohio, Ms. SOLIS, Ms. VELÁZQUEZ, and Ms. SCHAKOWSKY) introduced the following bill; which was referred to the Committee on Education and the Workforce

---

A BILL

To address rising college tuition by strengthening the compact between the States, the Federal Government, and institutions of higher education to make college more affordable.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

XML Document Example

http://thomas.loc.gov/home/gpoxmlc109/h2739_ih.xml

```
<bill bill-stage="Introduced-in-House">
  <congress>109th CONGRESS</congress> <session>1st
     Session</session>
    <legis-num>H. R. 2739</legis-num>
    <current-chamber>IN THE HOUSE OF
     REPRESENTATIVES</current-chamber>
    <action>
      <action-date date="20050526">May 26, 2005</action-date>
      <action-desc><sponsor name-id="T000266">Mr. Tierney</sponsor>
     (for himself, <cosponsor name-id="M001143">Ms. McCollum of
     Minnesota</cosponsor>, <cosponsor name-id="M000725">Mr.
     George Miller of California</cosponsor>) introduced the following bill;
     which was referred to the <committee-name committee-
     id="HED00">Committee on Education and the
     Workforce</committee-name>
      </action-desc>
    </action>
```

…

THOMAS: Library of Congress

Outline

- Introduction
- Querying
  - search context: XML nodes vs entire document.
  - search result: XML nodes or newly constructed answers vs entire document.
  - search expression: keyword search, Boolean operators, proximity distance, scoping, thesaurus, stop words, stemming.
  - document structure: explicitly specified in query or used in query semantics.
- Scoring
- Evaluation
- Open Issues

Languages for XML Search

- Keyword search (CO Queries)
  - “xml”
- Tag + Keyword search
  - book: xml
- Path Expression + Keyword search (CAS Queries)
  - /book[./title about “xml db”]
- XQuery + Complex full-text search
  - for $b in /book
    let score $s := $b ftcontains “xml” && “db” distance 5

XRank

```
<workshop date="28 July 2000">
    <title> XML and Information Retrieval: A SIGIR 2000 Workshop </title>
    <editors> David Carmel, Yoelle Maarek, Aya Soffer </editors>
    <proceedings>
        <paper id="1">
            <title> XQL and Proximal Nodes </title>
            <author> Ricardo Baeza-Yates </author>
            <author> Gonzalo Navarro </author>
            <abstract> We consider the recently proposed language … </abstract>
            <section name="Introduction">
                Searching on structured text is becoming more important with XML …
                <subsection name="Related Work">
                    The XQL language …
                </subsection>
            </section>
            <cite xmlns:xlink="http://www.acm.org/www8/paper/xmlql> … </cite>
        </paper>
```

(Guo et al, SIGMOD 2003)

XRank

```
<workshop date="28 July 2000">
    <title> XML and Information Retrieval: A SIGIR 2000 Workshop </title>
    <editors> David Carmel, Yoelle Maarek, Aya Soffer </editors>
    <proceedings>
        <paper id="1">
            <title> XQL and Proximal Nodes </title>
            <author> Ricardo Baeza-Yates </author>
            <author> Gonzalo Navarro </author>
            <abstract> We consider the recently proposed language … </abstract>
            <section name="Introduction">
                Searching on structured text is becoming more important with XML …
                <subsection name="Related Work">
                    The XQL language …
                </subsection>
            </section>
            <cite xmlns:xlink="http://www.acm.org/www8/paper/xmlql> … </cite>
        </paper>
```

XIRQL

```
<workshop date="28 July 2000">
    <title> XML and Information Retrieval: A SIGIR 2000 Workshop </title>
    <editors> David Carmel, Yoelle Maarek, Aya Soffer </editors>
    <proceedings>
        <paper id="1">
            <title> XQL and Proximal Nodes </title>
            <author> Ricardo Baeza-Yates </author>
            <author> Gonzalo Navarro </author>
            <abstract> We consider the recently proposed language … </abstract>
            <section name="Introduction">
                Searching on structured text is becoming more important with XML …
                <em> The XQL language </em>
            </section>
            …
            <cite xmlns:xlink="http://www.acm.org/www8/paper/xmlql> … </cite>
        </paper>
        …
```

index nodes

(Fuhr & Großjohann, SIGIR 2001)

Similar Notion of Results

- Nearest Concept Queries
  - (Schmidt et al, ICDE 2002)
- XKSearch
  - (Xu & Papakonstantinou, SIGMOD 2005)

Languages for XML Search

- Keyword search (CO Queries)
  - “xml”
- Tag + Keyword search
  - book: xml
- Path Expression + Keyword search (CAS Queries)
  - /book[./title about “xml db”]
- XQuery + Complex full-text search
  - for $b in /book
    let score $s := $b ftcontains “xml” && “db” distance 5

XSearch

<workshop date="28 July 2000">
<title> XML and Information Retrieval: A SIGIR 2000 Workshop </title>
<editors> David Carmel, Yoelle Maarek, Aya Soffer </editors>
<proceedings>
<paper id="1">
<title> XQL and Proximal Nodes </title>
<author> Ricardo Baeza-Yates </author>
<author> Gonzalo Navarro </author>
<abstract> We consider the recently proposed language … </abstract>
<section name="Introduction">
Searching on structured text is becoming more important with XML …
…
</paper>
<paper id="2">
<title> XML Indexing </title>
…

Not a "meaningful" result

(Cohen et al, VLDB 2003)

Languages for XML Search

- Keyword search (CO Queries)
  - “xml”
- Tag + Keyword search
  - book: xml
- Path Expression + Keyword search (CAS Queries)
  - /book[./title about “xml db”]
- XQuery + Complex full-text search
  - for $b in /book
    let score $s := $b ftcontains “xml” && “db” distance 5

XPath 2.0

- fn:contains($e, string)
  returns true iff $e contains string

```
//section[fn:contains(./title, “XML Indexing”)]
```

(W3C 2005)

XIRQL

- Weighted extension to XQL (precursor to XPath)

```
//section[0.6 · .//* $cw$ "XQL" +
          0.4 · .//section $cw$ "syntax"]
```

(Fuhr & Großjohann, SIGIR 2001)

XXL

- Introduces a similarity operator ~

```
Select  Z
From   http://www.myzoos.edu/zoos.html
Where  zoos.#.zoo As Z and
          Z.animals.(animal)?.specimen as A and
          A.species ~ “lion” and
          A.birthplace.#.country as B and
          A.region ~ B.content
```

(Theobald & Weikum, EDBT 2002)

NEXI

- Narrowed Extended XPath I
- INEX Content-and-Structure (CAS) Queries
- Specifically targeted for content-oriented XML search (i.e. “aboutness”)

```
//article[about(.//title, apple) and
          about(.//sec, computer)]
```

(Trotman & Sigurbjornsson, INEX 2004)

Languages for XML Search

- Keyword search (CO Queries)
  - “xml”
- Tag + Keyword search
  - book: xml
- Path Expression + Keyword search (CAS Queries)
  - /book[./title about “xml db”]
- XQuery + Complex full-text search
  - for $b in /book
    let score $s := $b ftcontains “xml” && “db” distance 5

Schema-Free XQuery

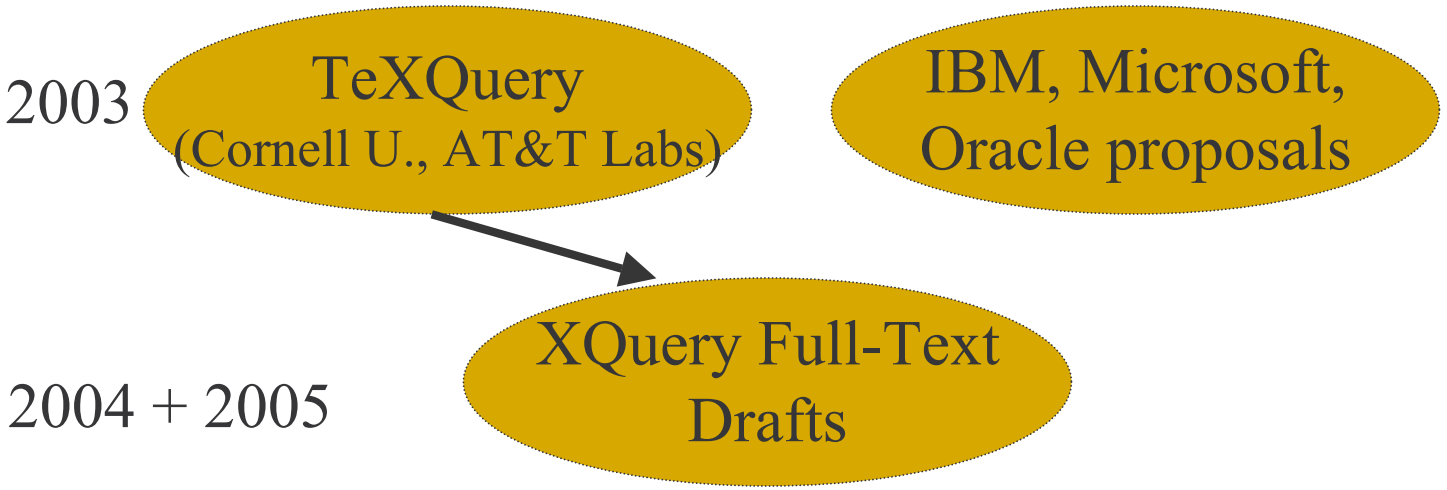
- Meaningful least common ancestor (mlcas)

```
for $a in doc("bib.xml")//author
      $b in doc("bib.xml")//title
      $c in doc("bib.xml")//year
where $a/text() = "Mary" and
          exists mlcas($a,$b,$c)
return <result> {$b,$c} </result>
```

(Li, Yu, Jagadish, VLDB 2003)

TeXQuery and XQuery FT

- Fully composable FT primitives.
- Composable with XPath/XQuery.
- Based on a formal model.
- Scoring and ranking on all predicates.



(Amer-Yahia, Botev, Shanmugasundaram, WWW 2004)
(http://www.w3.org/TR/xquery-full-text/, W3C 2005)

FTSelections and FTMatchoptions

- FTWord | FTAnd | FTOr | FTNot | FTMildNot | FTOrder | FTWindow | FTDistance | FTScope | FTTimes | FTSelection (FTMatchOptions)*
  - *books//title [. ftcontains "usability" case sensitive with thesaurus "synonyms"]*
  - *books//abstract [. ftcontains ("usability" || "web-testing")]*
  - *books//content ftcontains ("usability" && "software") window at most 3 ordered with stopwords*
  - *books//abstract [. ftcontains (("Utilisation" language "French" with stemming && ".?site" with wildcards) same sentence]*
  - *books//title ftcontains "usability" occurs 4 times && "web-testing" with special characters*
  - *books//book/section [. ftcontains books/book/title]/title*

FTScore Clause

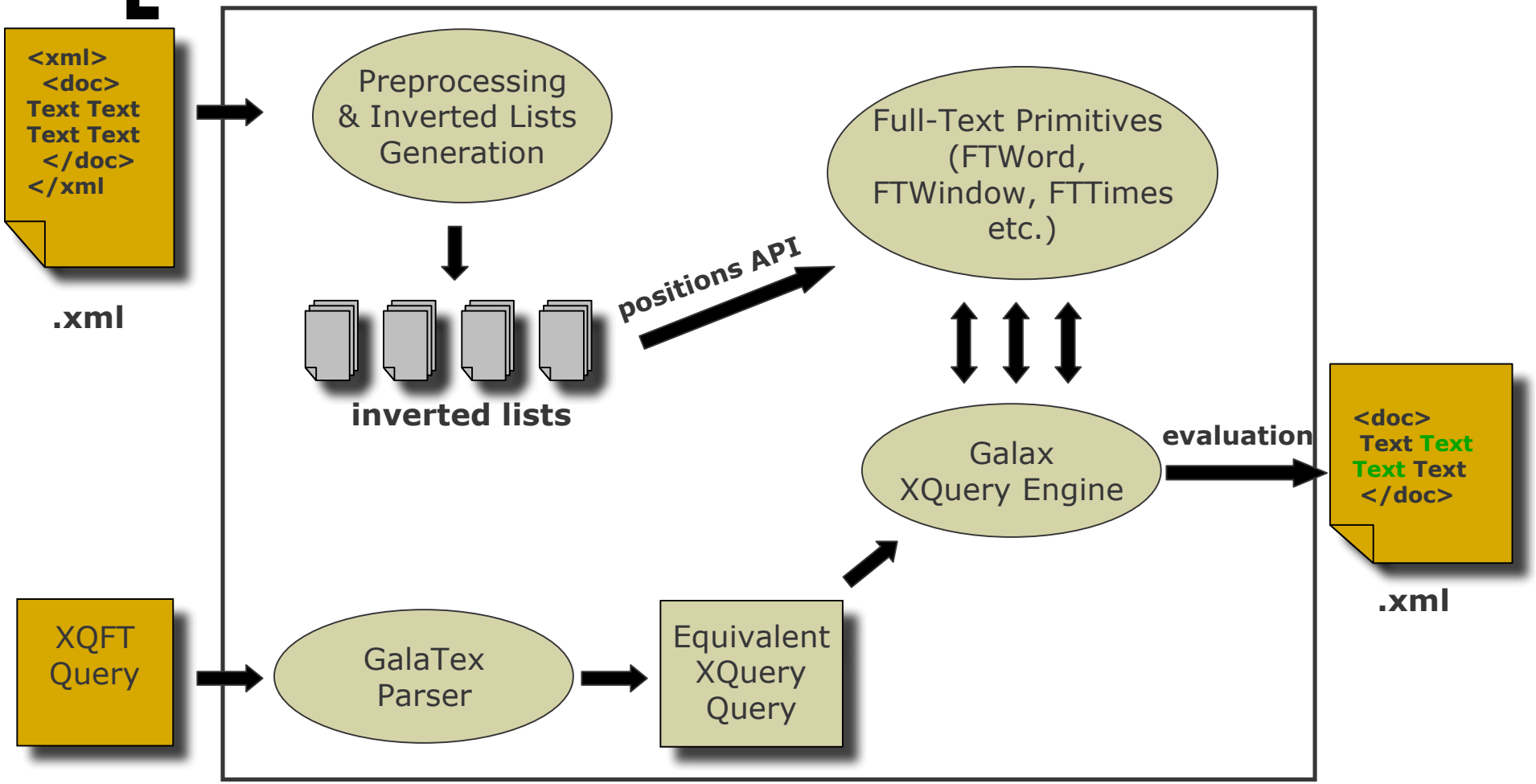
In any order

FOR $v [SCORE $s]? IN [FUZZY] Expr
LET …
WHERE …
ORDER BY …
RETURN

Example

```
FOR $b SCORE $s in FUZZY
            /pub/book[. ftcontains “Usability” && “testing”
                        and ./price < 10.00]
ORDER BY $s
RETURN $b
```

GalaTex Architecture



(http://www.galaxquery.org/galatex)

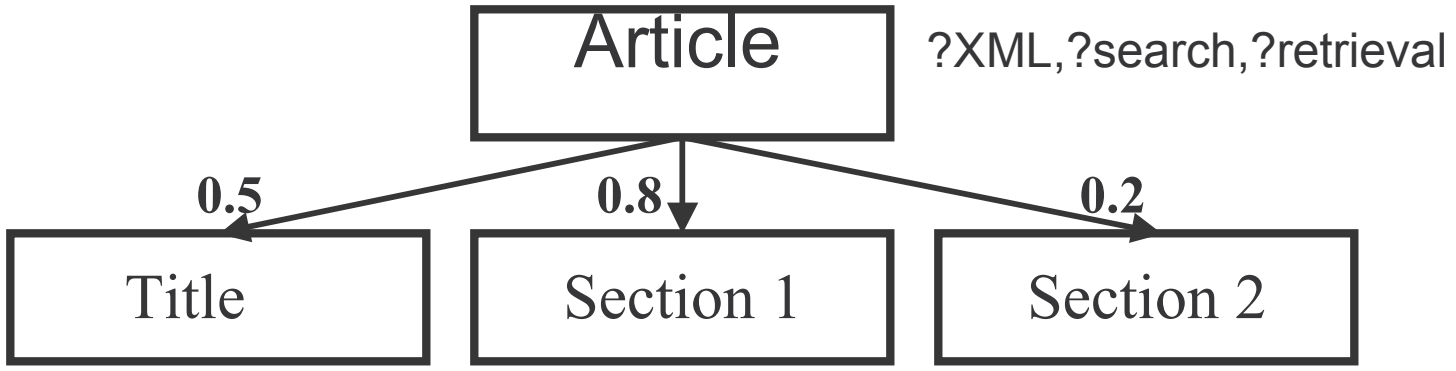
Outline

- Introduction
- Querying
- Scoring
- Evaluation
- Open Issues

Scoring

- Keyword queries and Tag + Keyword queries
  - initial term weights per element.
  - elements with same tag may have same score.
  - score propagation along document structure.
  - overlapping elements.
- Path Expression + Keyword queries
  - initial term weights based on paths.
- XQuery + Complex full-text queries
  - compute scores for (newly constructed) XML fragments satisfying XQuery (structural, full-text and scalar conditions).

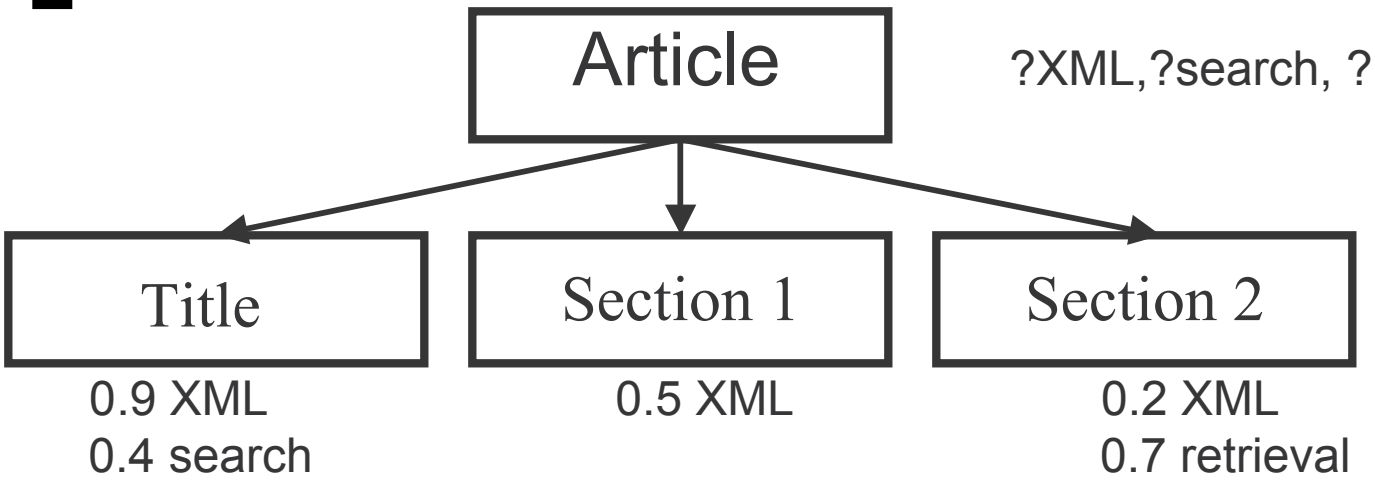
Term Weights



- how to obtain document and collection statistics (e.g., tf, idf)
- how to estimate element scores (frequency, user studies, size)?

- which components contribute best to content of Article?
- do we need edge weights (e.g., size, number of children)?
- is element size an issue?

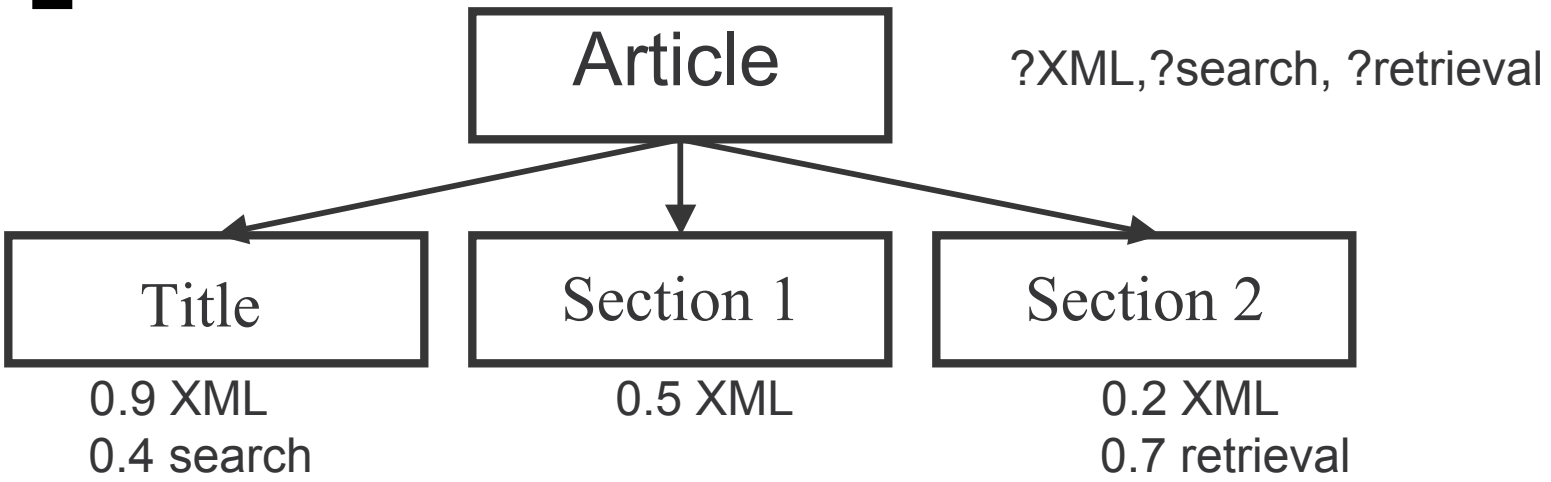
Score Propagation (XXL)



- Compute similar terms with relevance score *r1* using an ontology (weighted distance in the ontology graph).
- Compute *TFIDF* of each term for a given element content with relevance score *r2*.
- Relevance of an element content for a term is *r1*r2*.
- Probabilities of conjunctions multiplied (independence assumption) along elements of same path to compute path score.

(Theobald & Weikum, EDBT 2002)

Overlapping elements



- Section 1 and article are both relevant to "XML retrieval"
- which one to return so that to reduce overlap?
- Should the decision be based on user studies, size, types, etc?

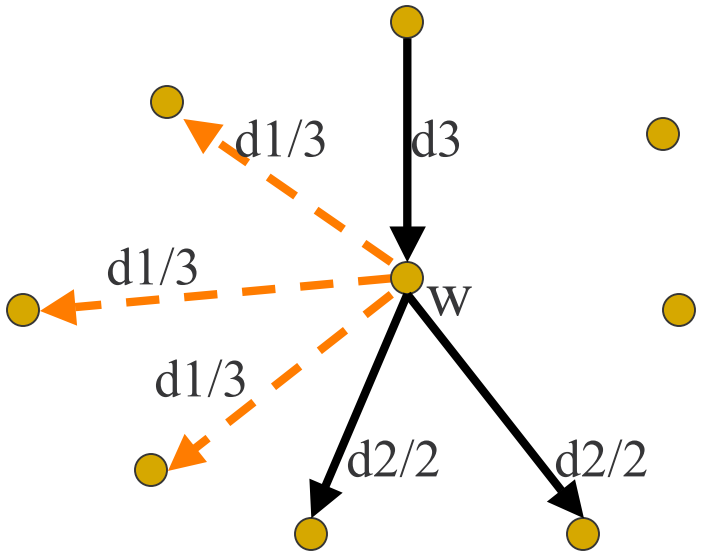
Controlling Overlap

- Start with a component ranking, elements are re-ranked to control overlap.
- Retrieval status values (RSV) of those components containing or contained within higher ranking components are iteratively adjusted.

1. Select the highest ranking component.
2. Adjust the RSV of the other components.
3. Repeat steps 1 and 2 until the top *m* components have been selected.

(Clarke, SIGIR 2005)

ElemRank



: Hyperlink edge

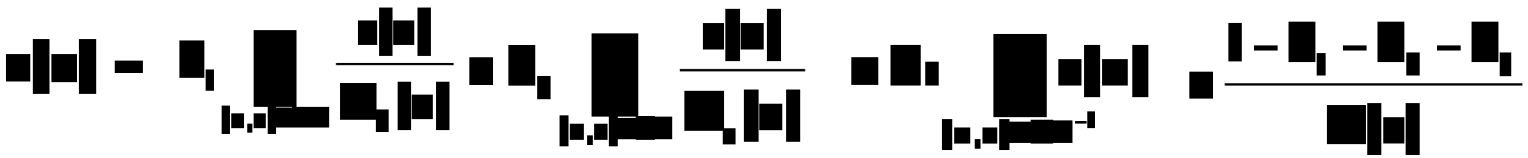
: Containment edge

d1: Probability of following hyperlink

d2: Probability of visiting a subelement

d3: Probability of visiting parent

1-d1-d2-d3: Probability of random jump



(Guo et al, SIGMOD 2003)

Scoring

- Keyword queries
  - compute possibly different scores.
- Tag + Keyword queries
  - compute scores based on tags and keywords.
- Path Expression + Keyword queries
  - compute scores based on paths and keywords.
- XQuery + Complex full-text queries
  - compute scores for (newly constructed) XML fragments satisfying XQuery (structural, full-text and scalar conditions).

Vector–based Scoring (JuruXML)

- Transform query into (term,path) conditions:

  article/bm/bib/bibl/bb[about(., *hypercube mesh torus nonnumerical database*)]

- (term,path)-pairs:

  hypercube, article/bm/bib/bibl/bb

  mesh, article/bm/bib/bibl/bb

  torus, article/bm/bib/bibl/bb

  nonnumerical, article/bm/bib/bibl/bb

  database, article/bm/bib/bibl/bb

- Modified cosine similarity as retrieval function for vague matching of path conditions.

(Mass et al, INEX 2002)

JuruXML Vague Path Matching

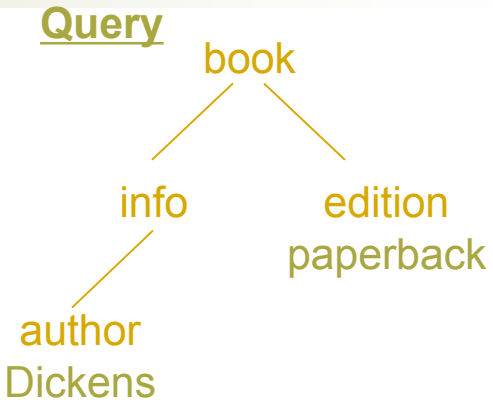
Modified vector-based cosine similarity

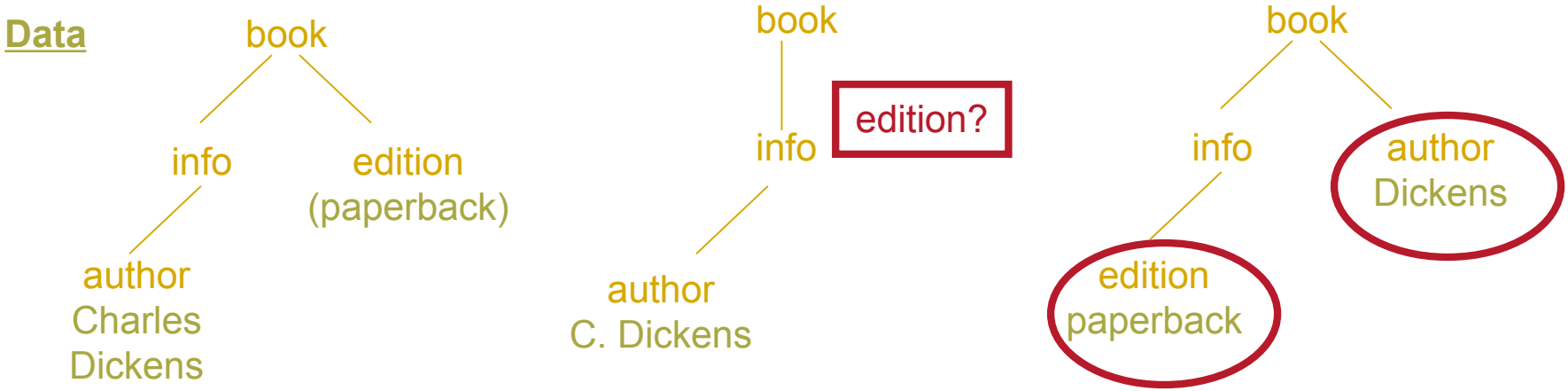
Example of length normalization:

cr (article/bibl, article/bm/bib/bibl/bb) = 3/6 = 0.5

XML Query Relaxation

- Tree pattern relaxations:
  - Leaf node deletion
  - Edge generalization
  - Subtree promotion

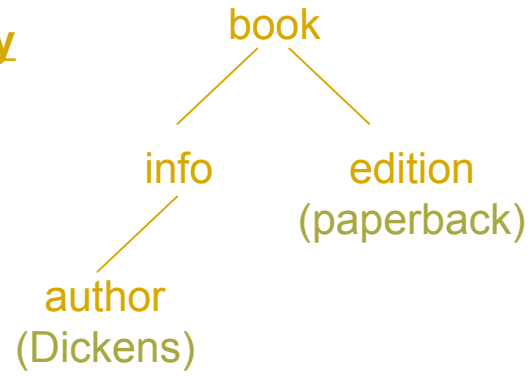




(Schlieder, EDBT 2002)(Delobel & Rousset, 2002)
(Amer-Yahia, Lakshmanan, Pandit, SIGMOD 2004)

A Family of Scoring Methods

- Twig scoring
  - High quality
  - Expensive computation
- Path scoring
- Binary scoring
  - Low quality
  - Fast computation



(Amer-Yahia, Koudas, Marian, Srivastava, Toman, VLDB 2005)

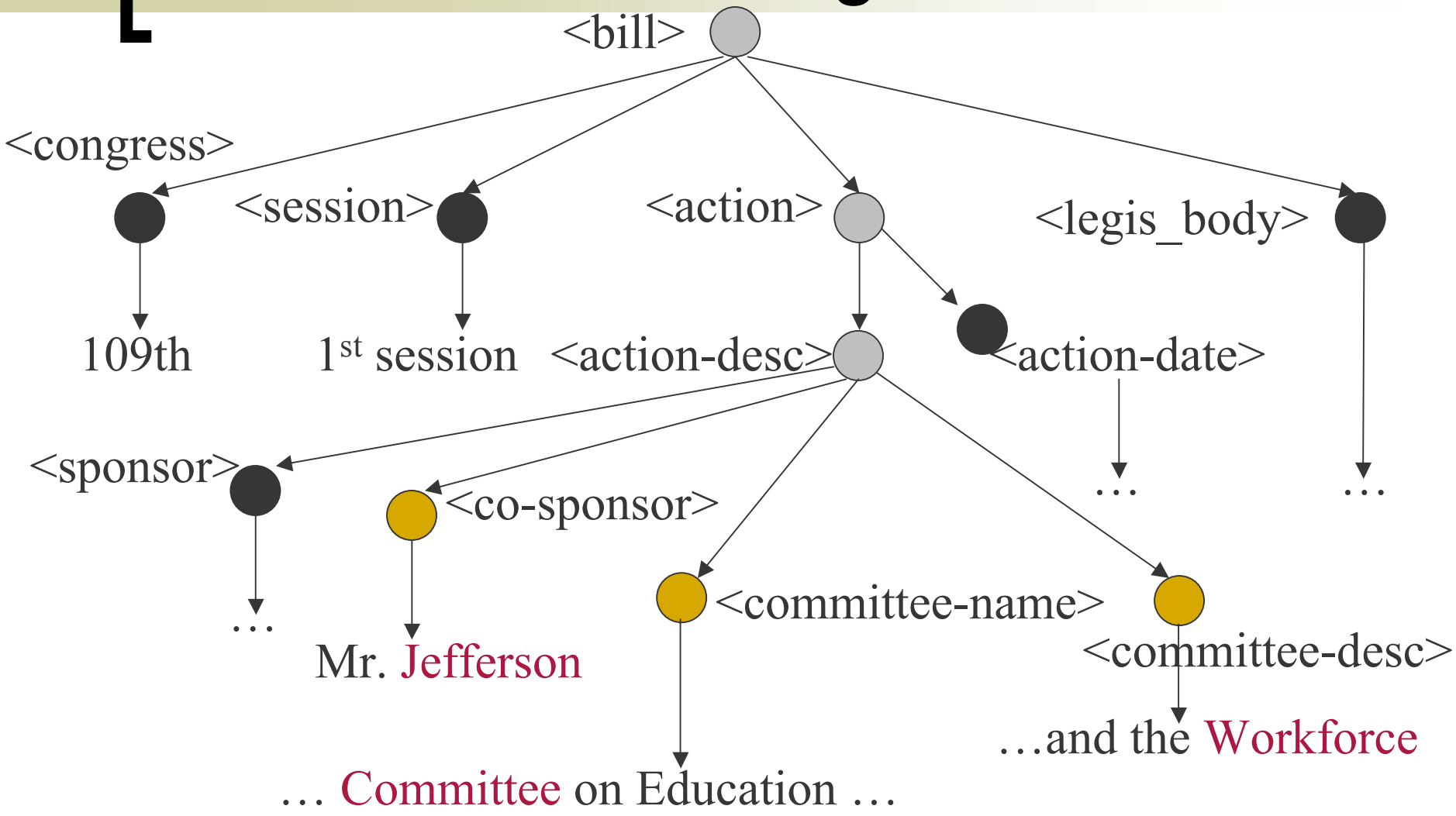
Scoring

- Keyword queries
  - compute possibly different scores.
- Tag + Keyword queries
  - compute scores based on tags and keywords.
- Path Expression + Keyword queries
  - compute scores based on paths and keywords.
  - Evaluate effectiveness of scoring methods.
- XQuery + Complex full-text queries
  - compute scores for (newly constructed) XML fragments satisfying XQuery (structural, full-text and scalar conditions).
  - compose approximation on structure and on text.

Outline

- Introduction
- Querying
- Scoring
- Evaluation
  - Formalization of existing XML search languages
  - Structure-aware evaluation algorithms
  - Implementation in GalaTex
- Open Issues

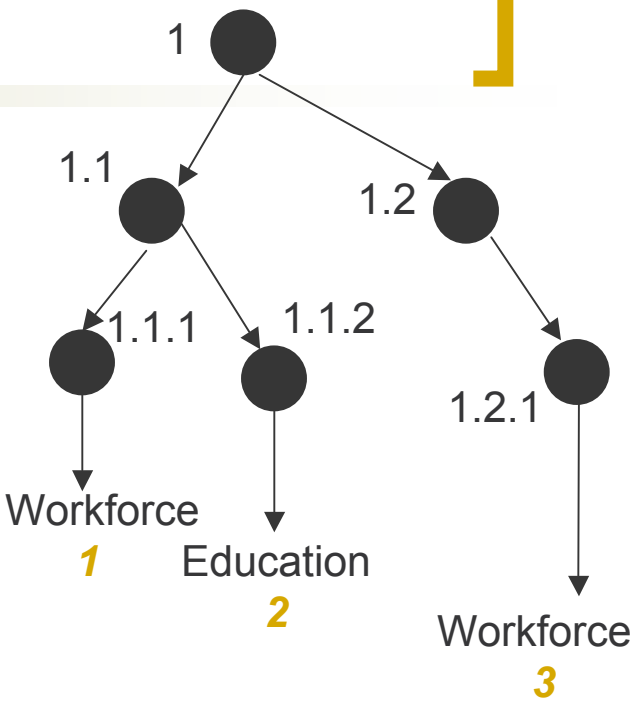
LOC document fragment



Sample Query on LOC

Find *action descriptions of bills introduced by "Jefferson" with a committee name containing the words "education" and "workforce" at a distance of no more than 5 words in the text*

Data model



- R

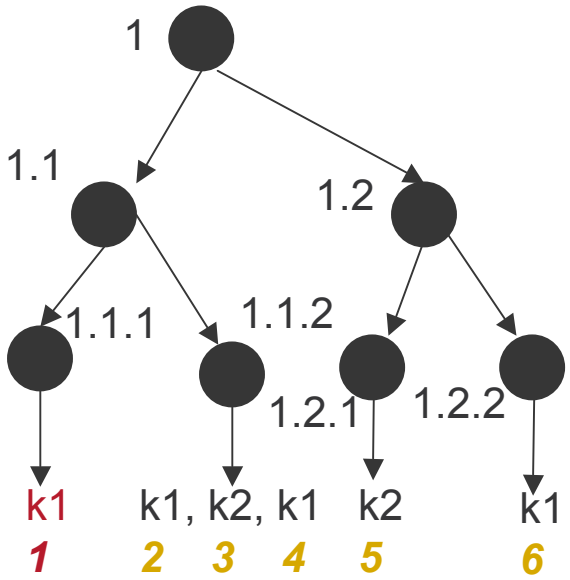
Node	tokPos
1	...
1.1	

- tokPos

word	position list
workforce	{1, 3}
education	{2}

Data model instantiation

- One relation per keyword in the document



Node	tokPos
1.2.2	k1 ; {6}
1.2	k1 ; {6}
1.1.2	k1 ; {2, 4}
1.1.1	k1 ; {1}
1.1	k1 ; {1, 2, 4}
1	k1 ; {1, 2, 4, 6}

Instance 1: R_{k1}

-redundant storage

-each tuple is self-contained

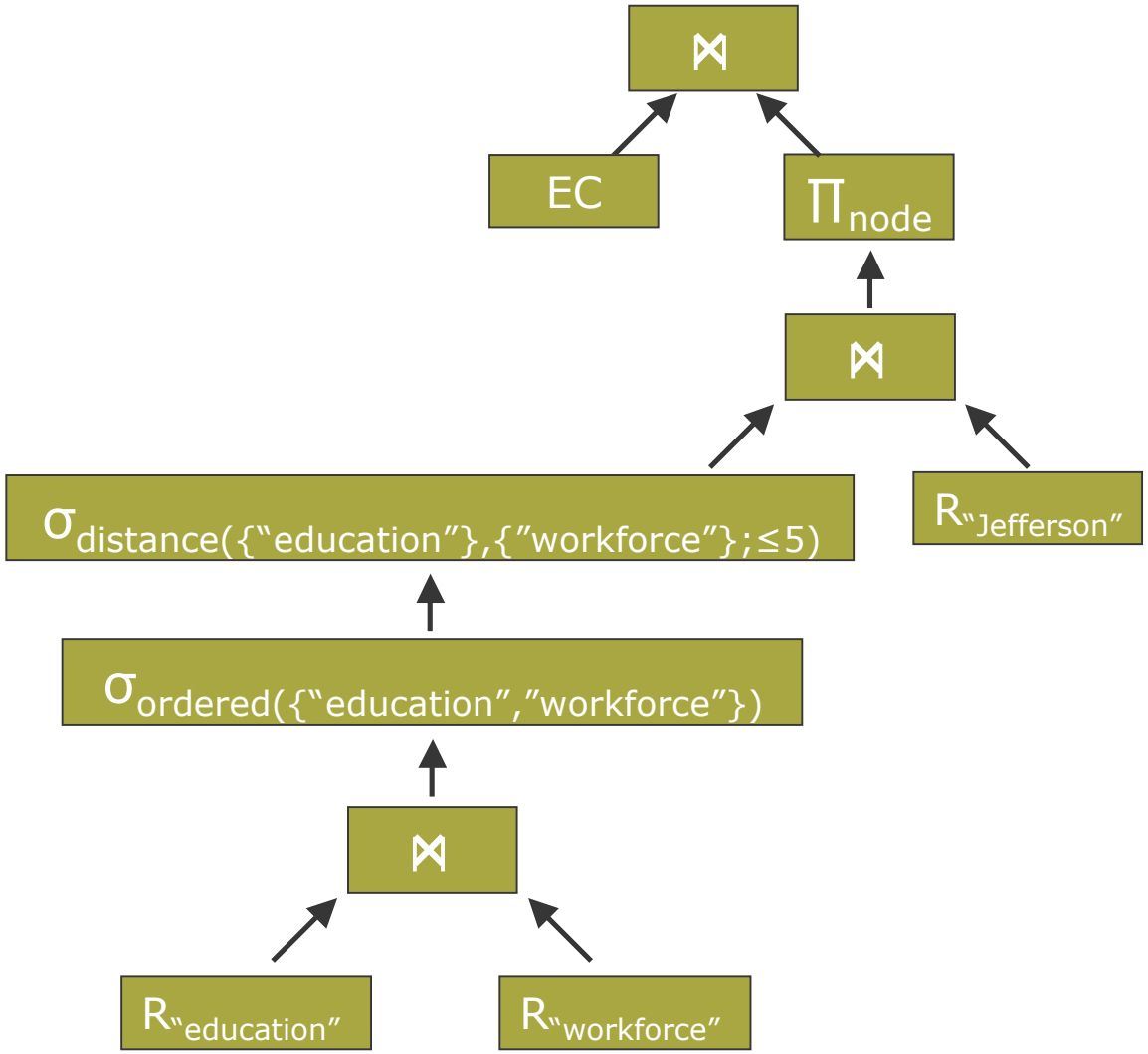
Node	tokPos
1.2.2	k1 ; {6}
1.1.2	k1 ; {2, 4}
1.1.1	k1 ; {1}

Instance 2: $scuR_{k1}$

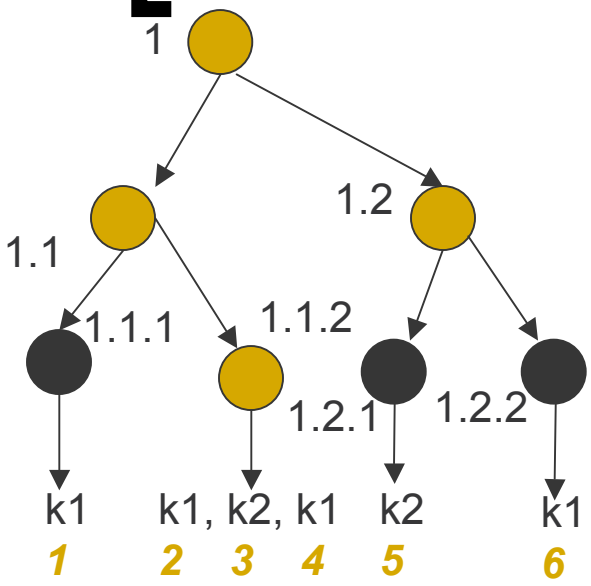
-no redundant positions

-smallest nbr of nodes

FT-Algebra and Query Plan



Join Evaluation



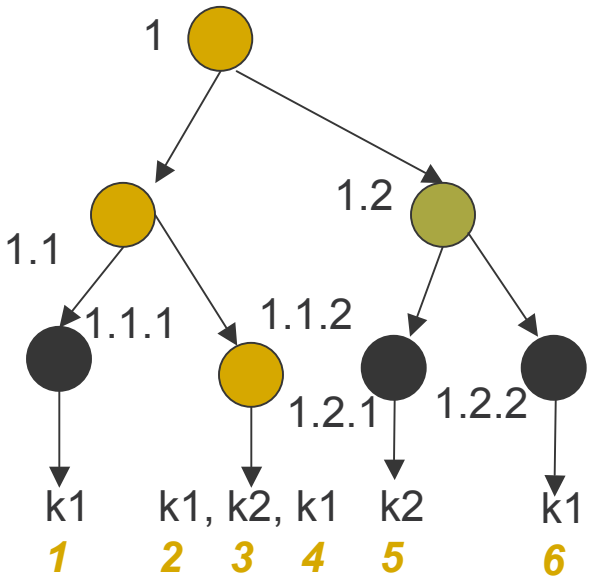
Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1, 2, 4}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

⋈

Node	tokPos
1.2.2	k1 ; {6}
1.2	k1 ; {6}
1.1.2	k1 ; {2, 4}
1.1.1	k1 ; {1}
1.1	k1 ; {1, 2, 4}
1	k1 ; {1, 2, 4, 6}

Node	tokPos
1.2.1	k2 ; {5}
1.2	k2 ; {5}
1.1.2	k2 ; {3}
1.1	k2 ; {3}
1	k2 ; {3, 5}

Join Evaluation on SCU



Node	tokPos
1.1.2	k2 ; {3} k1 ; {2, 4}

⋈

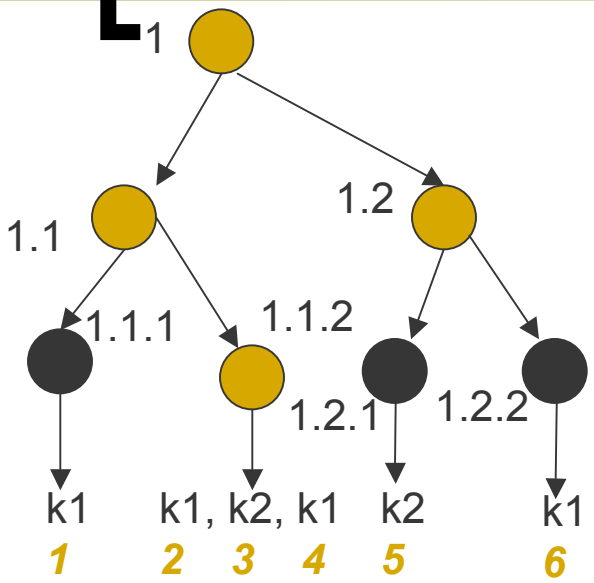
$scuR_{k1}$

Node	tokPos
1.2.2	k1 ; {6}
1.1.2	k1 ; {2, 4}
1.1.1	k1 ; {1}

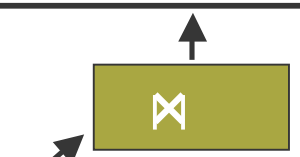
$scuR_{k2}$

Node	tokPos
1.2.1	k2 ; {5}
1.1.2	k2 ; {3}

Need for LCAs



Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}



$scuR_{k1}$

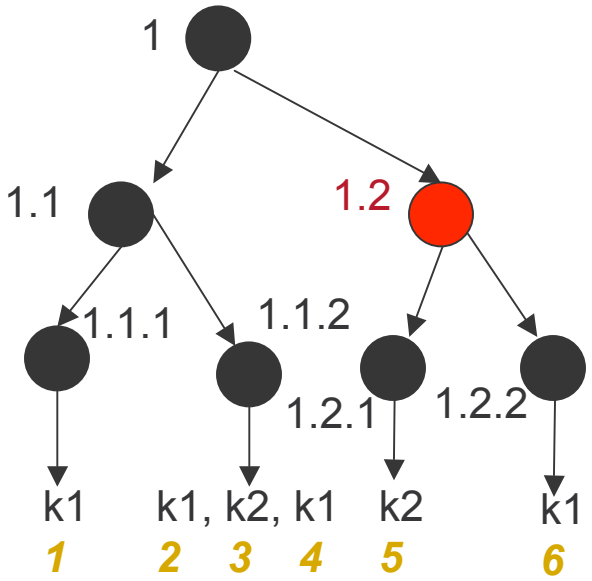
Node	tokPos
1.2.2	k1 ; {6}
1.1.2	k1 ; {2, 4}
1.1.1	k1 ; {1}

$scuR_{k2}$

Node	tokPos
1.2.1	k2 ; {5}
1.1.2	k2 ; {3}

(Schmidt et al, ICDE 2002)(Li, Yu, Jagadish, VLDB 2003)
(Guo et al, SIGMOD 2003)(Xu & Papakonstantinou, SIGMOD 2005)

SCU: is LCA enough?



$\sigma_{distance(\{"k1"\},\{"k2"\};=2)}$ ← fail

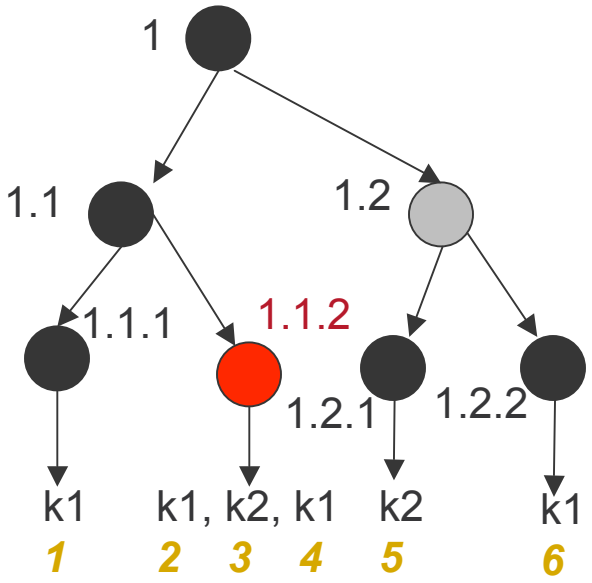
$\sigma_{ordered(\{"k2","k1"\})}$ ← pass

Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

$\bowtie$

$scuR_{k1}$ $scuR_{k2}$

SCU: is LCA enough?



$\sigma_{distance(\{"k1"\},\{"k2"\};=2)}$ ← fail

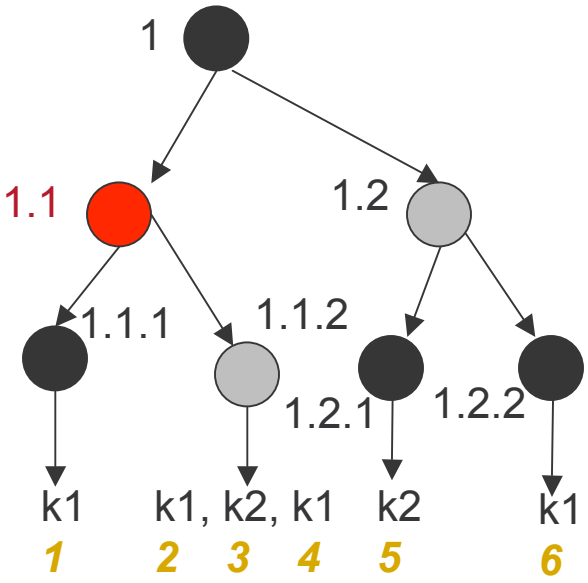
$\sigma_{ordered(\{"k2","k1"\})}$ ← pass

Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

⋈

$scuR_{k1}$ $scuR_{k2}$

SCU: is LCA enough?

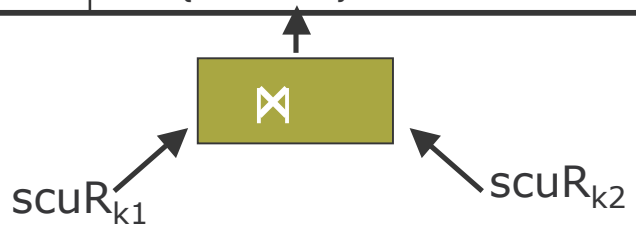


$\sigma_{distance(\{"k1"\},\{"k2"\};=2)}$

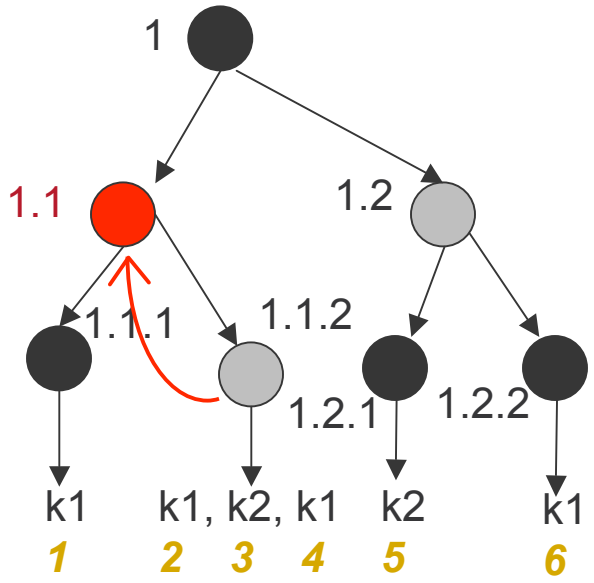
$\sigma_{ordered(\{"k2","k1"\})}$ ← fail

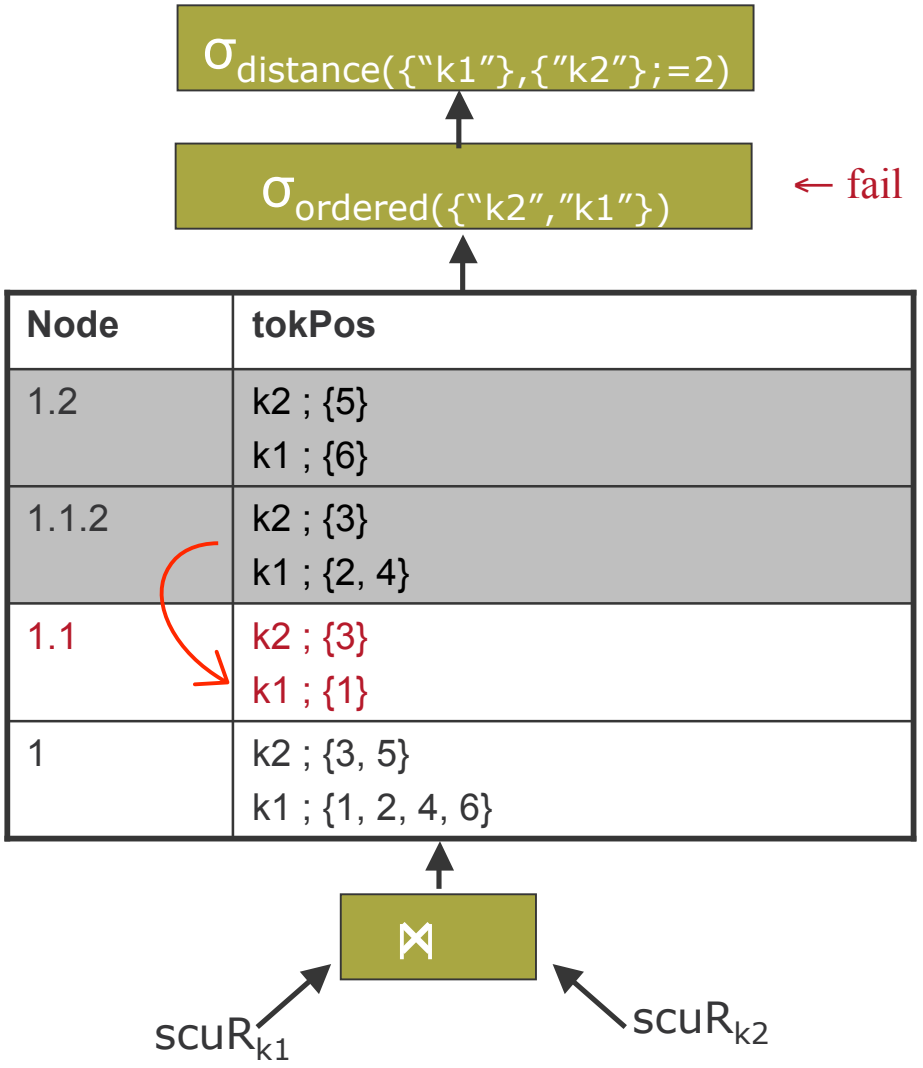
Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

Does not satisfy 'ordered' alone, but it should be an answer!

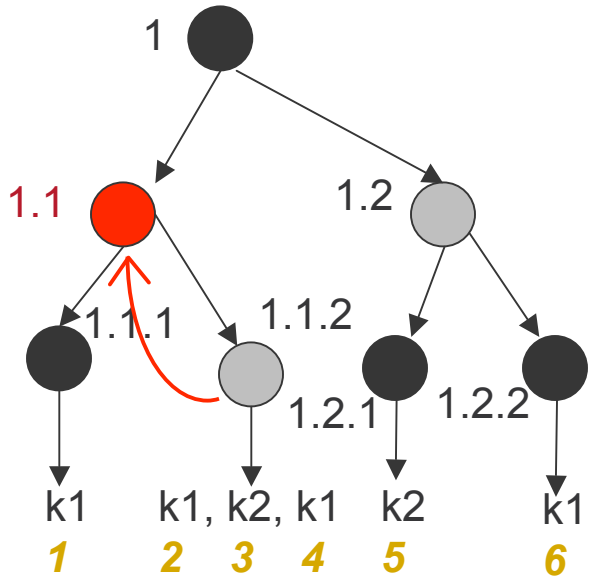


SCU: is LCA enough?





SCU: position propagation



Node	tokPos
1.1	k2 ; {3} k1 ; {1, 2, 4}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

$\sigma_{distance(\{"k1"\},\{"k2"\};=2)}$ ← pass

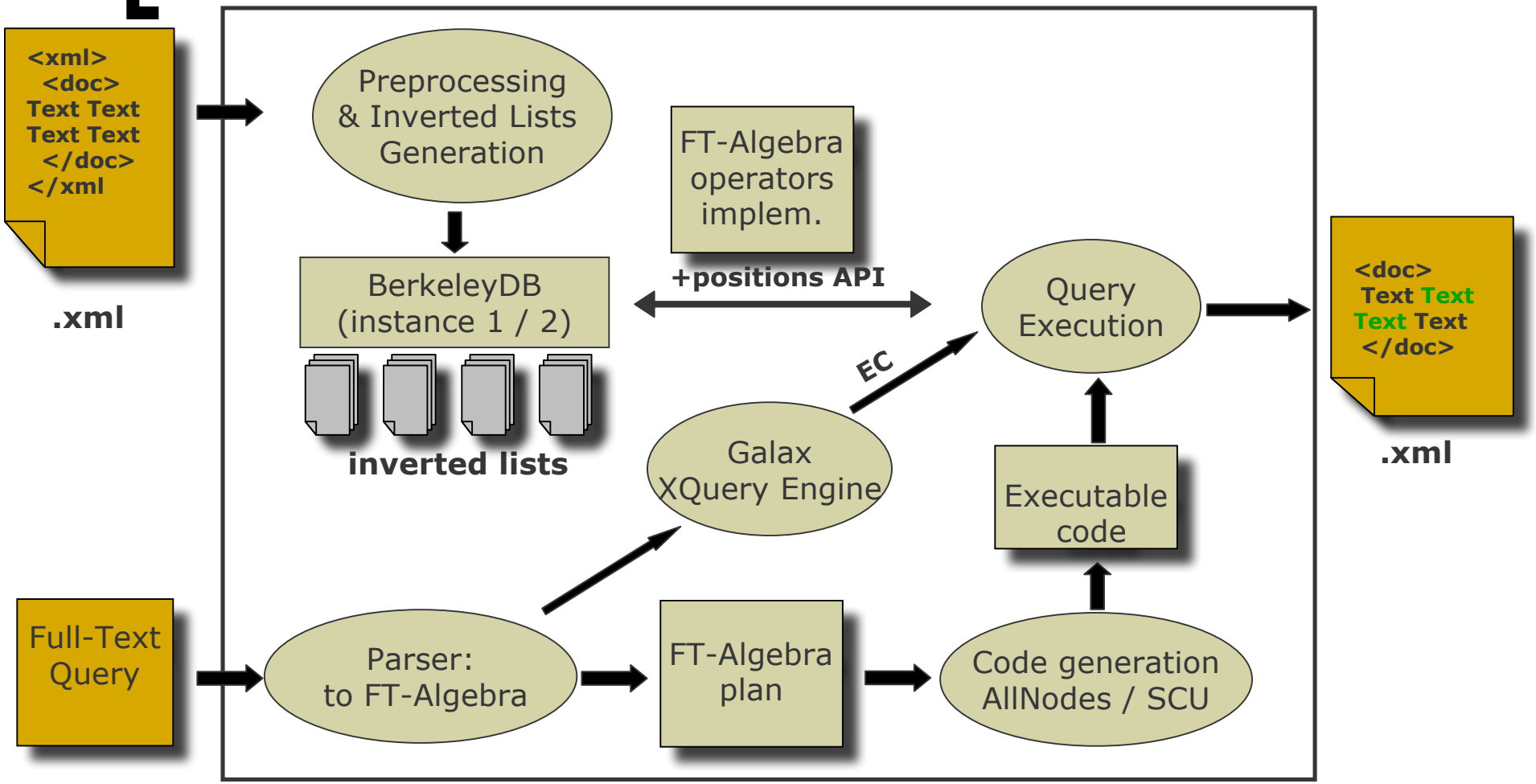
$\sigma_{ordered(\{"k2","k1"\})}$ ← pass

Node	tokPos
1.2	k2 ; {5} k1 ; {6}
1.1.2	k2 ; {3} k1 ; {2, 4}
1.1	k2 ; {3} k1 ; {1}
1	k2 ; {3, 5} k1 ; {1, 2, 4, 6}

SCU Summary

- Key ideas
  - $R_1 \ltimes_{SCU} R_2 \rightarrow$ find LCA
  - $\sigma_{SCU}(R) \rightarrow$ propagation along doc. structure
    - if node satisfies σ predicate, output node
    - o/w propagate its tokPos to its first ancestor in R
- Benefit: reduces size of intermediate results
- Challenge: minimize computation overhead
  - selections
    - additional column in R for direct access to ancestors
    - TRIE structures
  - joins
    - record highest ancestor in EC of each node in scuR and use sort-merge

GalaTex Architecture: in progress



Open Issues (in no particular order)

- Difficult research issues in XML retrieval are not 'just' about the effective retrieval of XML documents, but also about what and how to evaluate!
- System architecture: DB on top of IR, IR on top of DB, true merging?
- Experimental evaluation of scoring methods (INEX).
- Score-aware algebra for XML for the joint optimization of queries on both structure and text.
- More details: http://www.research.att.com/~sihem