XML Retrieval

Agenda

- Introduction
- Basic XML concepts
- Challenges in XML IR
- Vector space model for XML IR
- Evaluation of XML IR
- Text-centric vs. data-centric XML retrieval

Introduction – IR and relational databases

IR systems are often contrasted with relational databases (RDB).

- Traditionally, IR systems retrieve information from *unstructured text* ("raw" text without markup).
- RDB systems are used for querying *relational data*: sets of records that have values for predefined attributes such as employee number, title and salary.

	RDB search	unstructured IR
objects	records	unstructured docs
main data structure	table	inverted index
model	relational model	vector space & others
queries	SQL	free text queries

Some structured data sources containing text are best modeled as structured documents rather than relational data (Structured retrieval).

Structured Retrieval

Basic setting: queries are structured or unstructured; documents are structured.

Applications of structured retrieval

Digital libraries, patent databases, blogs, tagged text with entities like persons and locations (named entity tagging)

Example

- Digital libraries: *give me a full-length article on fast fourier transforms*
- Patents: *give me patens whose claims mention RSA public key encryption and that cite US patent 4,405,829*
- Entity-tagged text: *give me articles about sightseeing tours of the Vatican and the Coliseum*

Why RDB is not suitable in this case

Three main problems

1. An unranked system (DB) would return a potentially large number of articles that mention the Vatican, the Coliseum and sightseeing tours without ranking them by relevance to query.
2. Difficult for users to precisely state structural constraints – may not know which structured elements are supported by the system.

   tours AND (COUNTRY: *Vatican* OR

   LANDMARK: *Coliseum*)?

   tours AND (STATE: *Vatican* OR BUILDING: *Coliseum*)?
3. Users may be completely unfamiliar with structured search and advanced search interfaces or unwilling to use them.

Solution: adapt ranked retrieval to structured documents to address these problems.

Structured Retrieval

RDB search, Unstructured IR, Structured IR

	RDB search	unstructured retrieval	structured retrieval
objects	records	unstructured docs	trees with text at leaves
main data structure	table	inverted index	?
model	relational model	vector space & others	?
queries	SQL	free text queries	?

Standard for encoding structured documents: Extensible Markup Language (XML)

- structured IR → XML IR
- also applicable to other types of markup (HTML, SGML, ...)

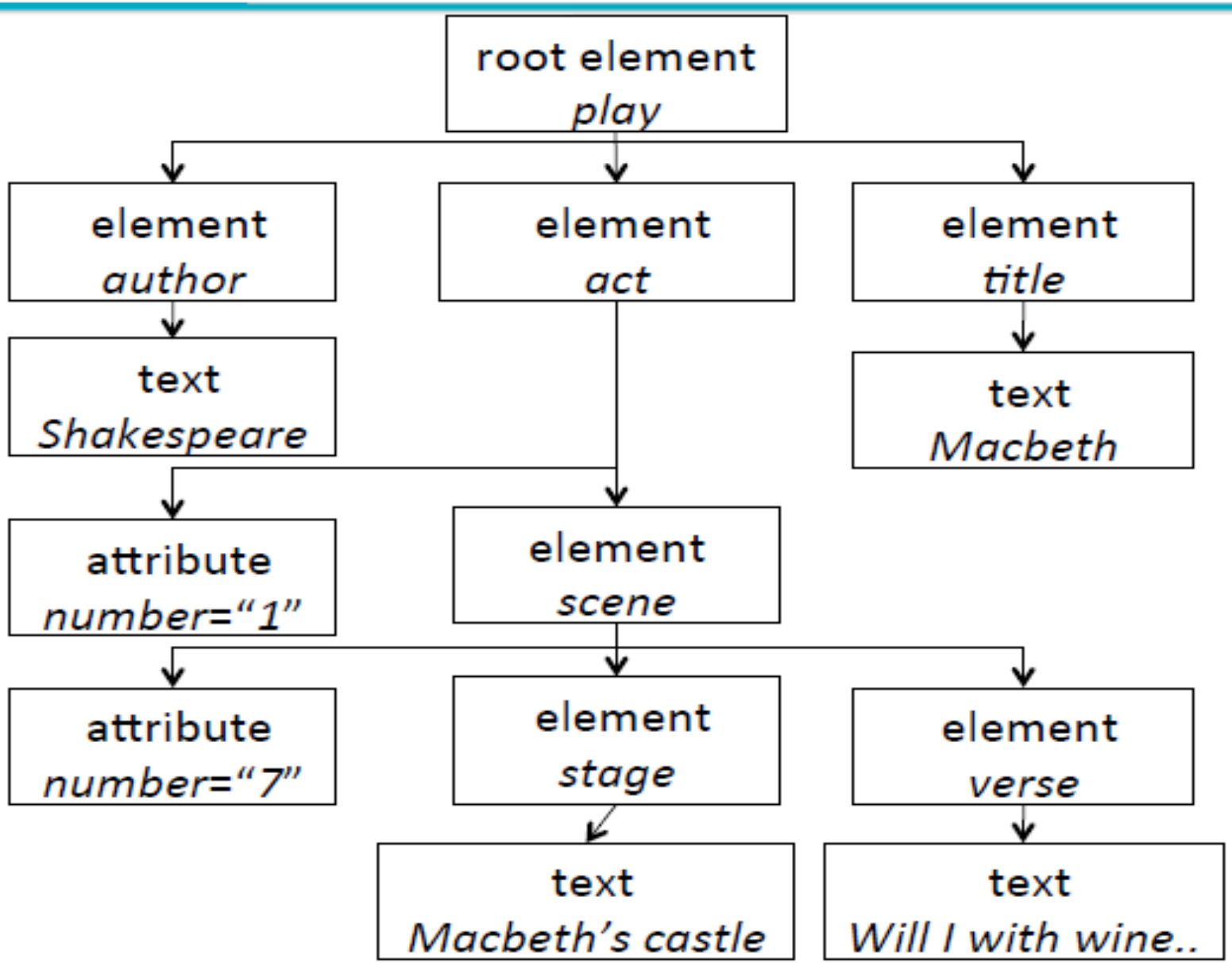
We call XML retrieval *structured retrieval* in this chapter. Some researchers prefer the term *semistructured retrieval* to distinguish XML retrieval from database querying.

XML document

- Ordered, labeled tree
- Each node of the tree is an XML element, written with an opening and closing XML tag (e.g. <title…>, </title…>)
- An element can have one or more XML attributes (e.g. number)
- Attributes can have values (e.g. 7)
- Attributes can have child elements (e.g. title, verse)

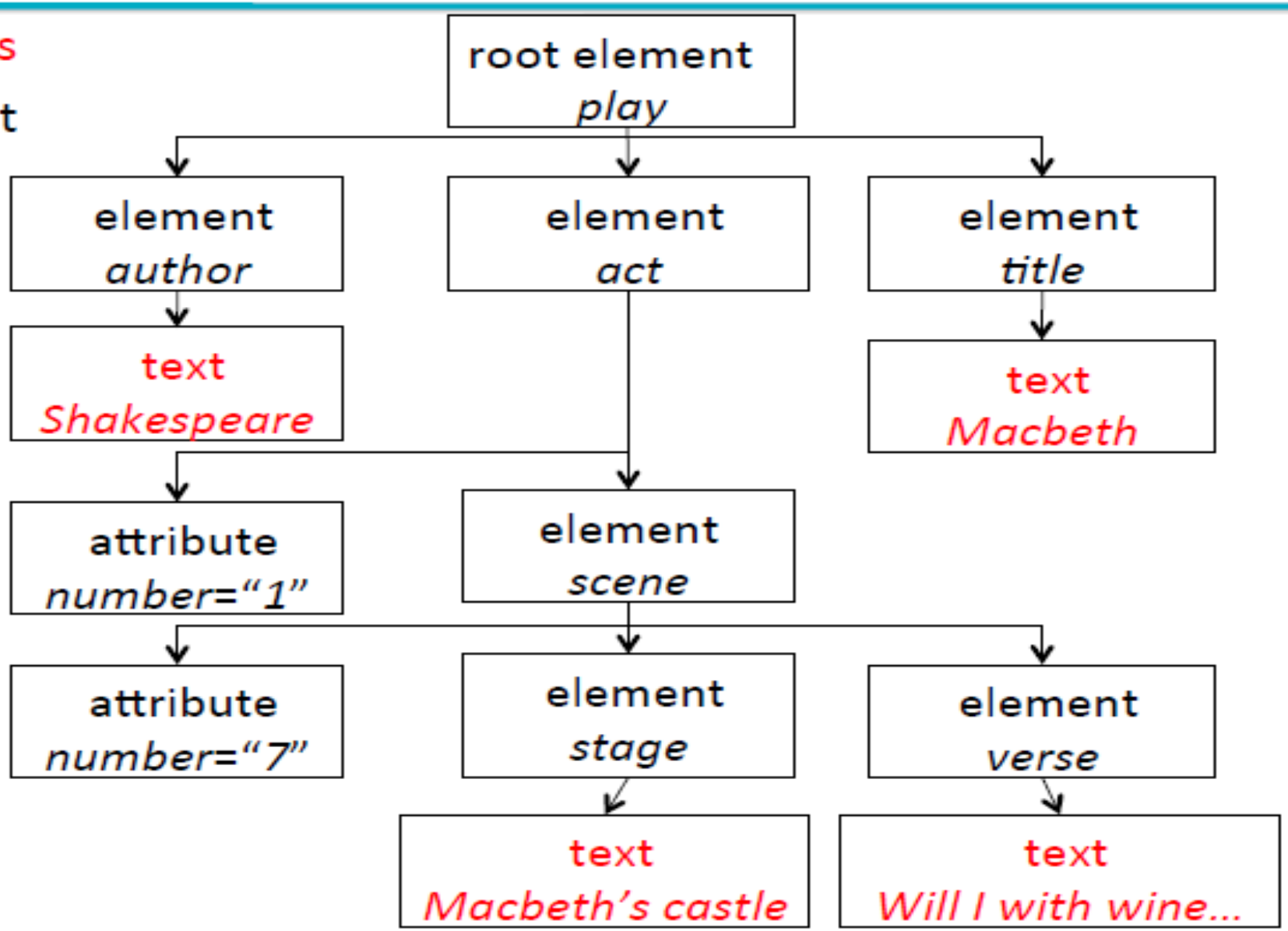
```
<play>
<author>Shakespeare</author>
<title>Macbeth</title>
<act number="1">
<scene  number=""7">
<stage>Macbeth's castle</stage>
<verse>Will I with wine
…</verse>
</scene>
</act>
</play>
```

XML document



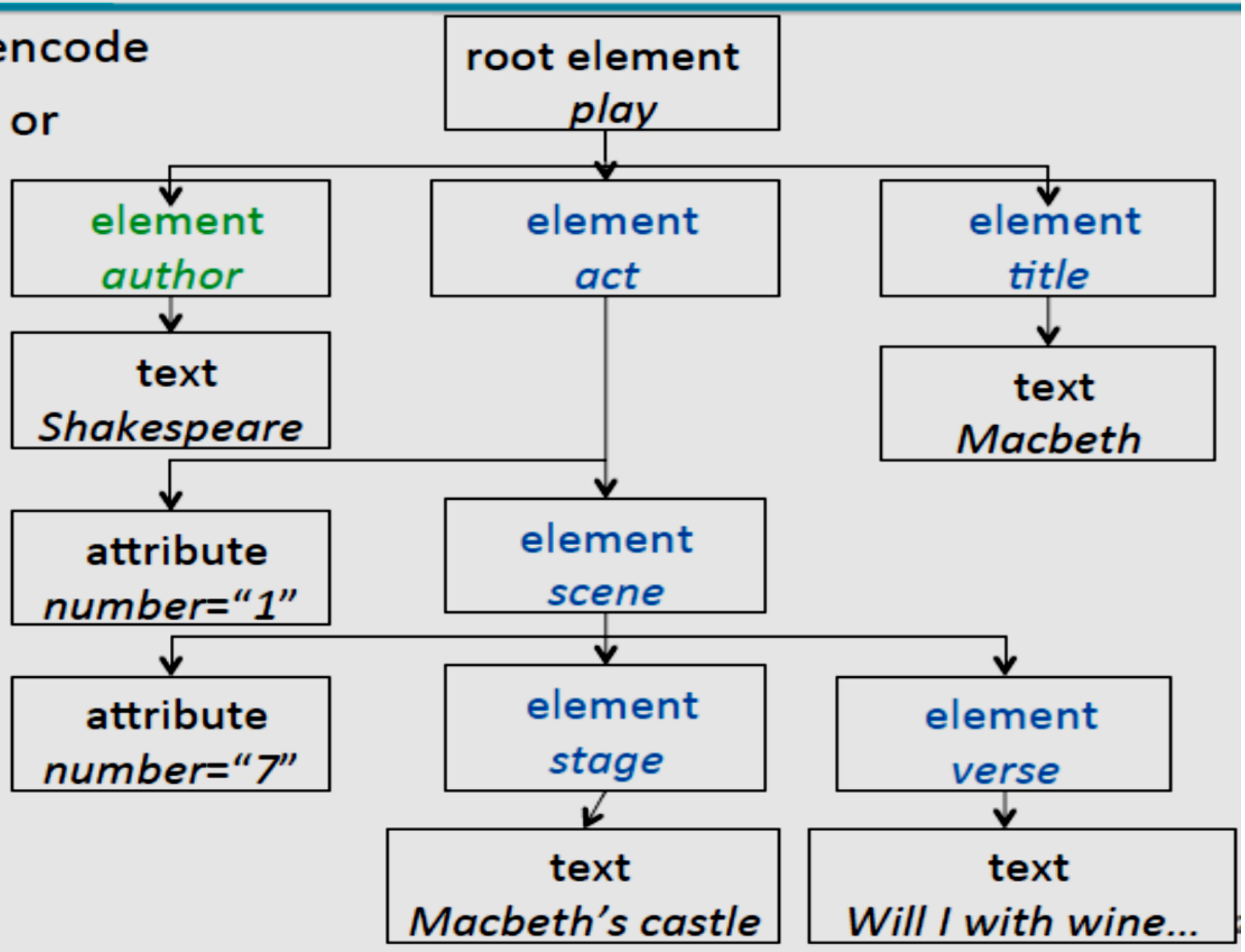
XML document

The leaf nodes consist of text



XML document

The internal nodes encode document structure or metadata functions



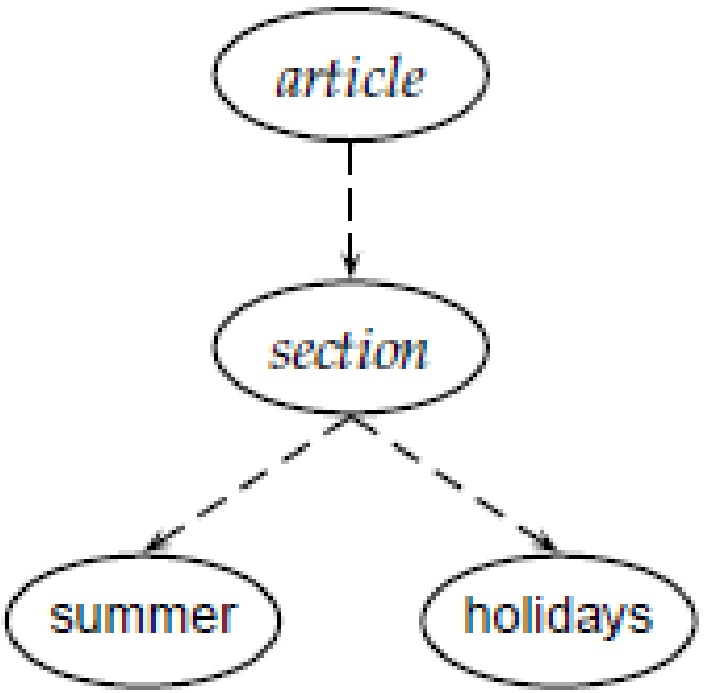
XML basics

- **Goal**: separate *layout* from *presentation* (syntax vs. semantics)
- **XML Documents Object Model (XML DOM)**: standard for accessing and processing XML documents
  - The DOM represents elements, attributes and text within elements as nodes in a tree.
  - With a DOM API, we can process an XML documents by starting at the root element and then descending down the tree from parents to children.
- **XPath**: standard for enumerating path in an XML document collection.
  - We will also refer to paths as XML contexts or simply contexts
- **Schema**: puts constraints on the structure of allowable XML documents. E.g. a schema for Shakespeare's plays: scenes can occur as children of acts.
  - Two standards for schemas for XML documents are: XML DTD (document type definition) and XML Schema

NEXI Format

A common format for XML queries is *NEXI* (Narrowed NEXI Extended Xpath I).

```
//article
[.//yr = 2001 or .//yr = 2002]
//section
[about(.,summer holidays)]
```



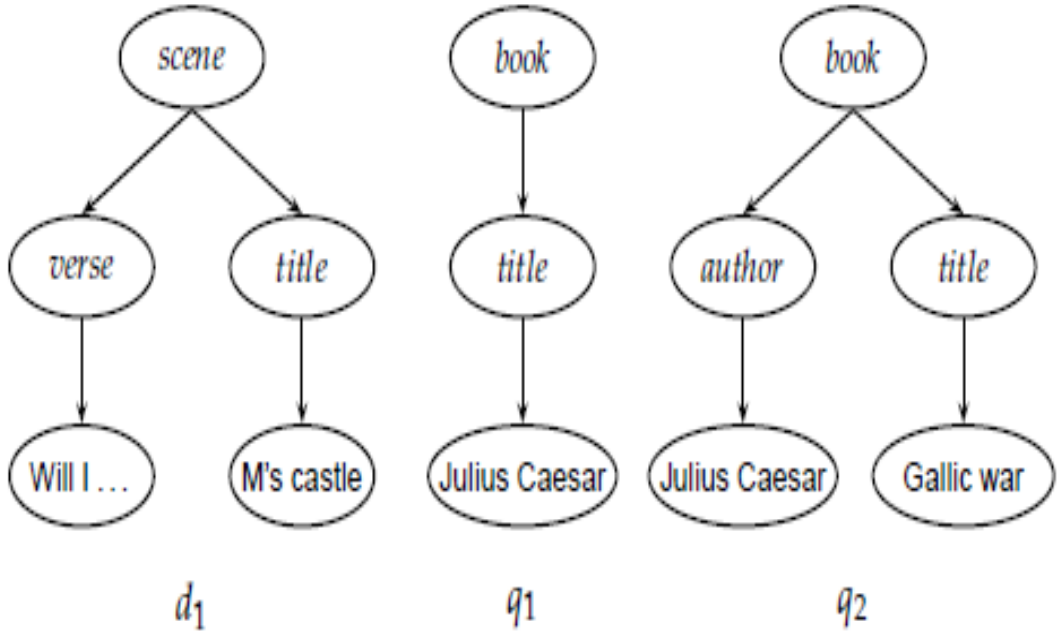
► **Figure 10.3** An XML query in NEXI format and its partial representation as a tree.

Tree representation of XML documents and queries

```
<play>
<author>Shakespeare</author>
<title>Macbeth</title>
<act number="I">
<scene number="vii">
<title>Macbeth's castle</title>
<verse>Will I with wine and wassail ...</verse>
</scene>
</act>
</play>
```

▶ **Figure 10.1** An XML document.

If we discard relational attributes, we can represent documents as trees with only one type of node: element nodes. In other words, we remove all attribute nodes from the XML document, such as the *number* attribute in Figure 10.1. Figure 10.4 shows a subtree of the document in Figure 10.1 as an element-node tree (labeled d_1).



▶ **Figure 10.4** Tree representation of XML documents and queries.

We can represent queries as trees in the same way. This is a query-by-example approach to query language design because users pose queries by creating objects that satisfy the same formal description as documents. In Figure 10.4, q_1 is a search for books whose titles score highly for the keywords Julius Caesar. q_2 is a search for books whose author elements score highly for Julius Caesar and whose title elements score highly for Gallic war.[3]

Challenges in XML IR

First challenge: document parts to retrieve

Structured or XML retrieval: users want us to return parts of documents (i.e., XML elements), not entire documents as IR systems usually do in unstructured retrieval.

Example

If we query Shakespeare's plays for *Macbeth's castle*, should we return the scene, the act or the entire play?

- In this case, the user is probably looking for the scene.
- However, an otherwise unspecified search for *Macbeth* should return the play of this name, not a subunit.

Solution: structured document retrieval principle

Structured document retrieval principle

Structured document retrieval principle

One criterion for selecting the most appropriate part of a document:

A system should always retrieve the most specific part of a document answering the query.

- Motivates a retrieval strategy that returns the smallest unit that contains the information sought, but does not go below this level.
- Hard to implement this principle algorithmically. E.g. query: title:*Macbeth* can match both the title of the tragedy, *Macbeth*, and the title of Act I, Scene vii, *Macbeth's castle*.
  - But in this case, the title of the tragedy (higher node) is preferred.
  - Difficult to decide which level of the tree satisfies the query.

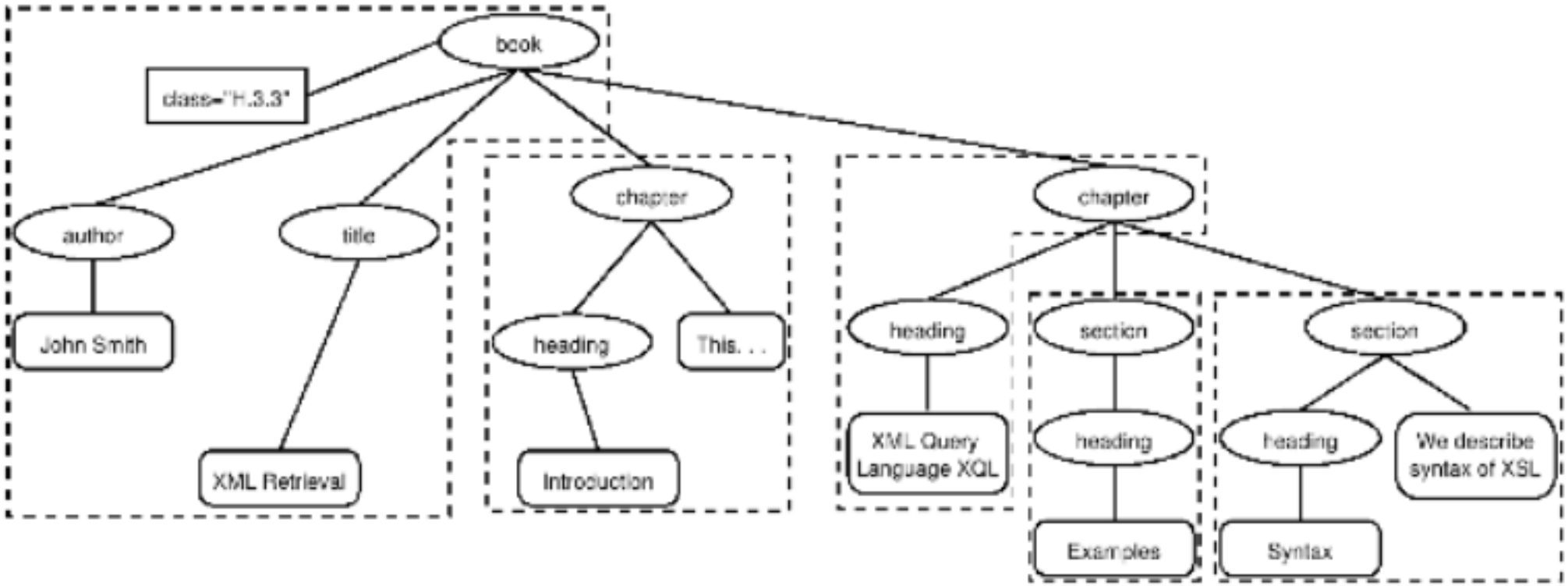
Second challenge: document parts to index

Central notion for indexing and ranking in IR: documents unit or **indexing unit**.

- In unstructured retrieval, usually straightforward: files on your desktop, email massages, web pages on the web etc.
- In structured retrieval, there are four main different approaches to defining the indexing unit
  1. non-overlapping pseudodocuments
  2. top down
  3. bottom up
  4. all

XML indexing unit: approach 1

Group nodes into non-overlapping pseudodocuments.



Indexing units: books, chapters, section, but without overlap.

Disadvantage: pseudodocuments may not make sense to the user because they are not coherent units.

XML indexing unit: approach 2

Top down (2-stage process):

1. Start with one of the latest elements as the indexing unit, e.g. the book element in a collection of books
2. Then, postprocess search results to find for each book the subelement that is the best hit.

This two-stage retrieval process often fails to return the best subelement because the relevance of a whole book is often not a good predictor of the relevance of small subelements within it.

XML indexing unit: approach 3

Bottom up:

Instead of retrieving large units and identifying subelements (top down), we can search all leaves, select the most relevant ones and then extend them to larger units in postprocessing.

Similar problem as top down: the relevance of a leaf element is often not a good predictor of the relevance of elements it is contained in.

XML indexing unit: approach 4

Index all elements: the least restrictive approach. Also problematic:

- Many XML elements are not meaningful search results, e.g., an ISBN number.
- Indexing all elements means that search results will be highly redundant.

Example

For the query *Macbeth's castle* we would return all of the *play*, *act*, *scene* and *stage* elements on the path between the root node and *Macbeth's castle*. The leaf node would then occur 4 times in the result set: 1 directly and 3 as part of other elements.

We call elements that are contained within each other **nested elements**. Returning redundant nested elements in a list of returned hits is not very user-friendly.

Third challenge: nested elements

Because of the redundancy caused by the nested elements it is common to restrict the set of elements eligible for retrieval.

Restriction strategies include:

- discard all small elements
- discard all element types that users do not look at (working XML retrieval system logs)
- discard all element types that assessors generally do not judge to be relevant (if relevance assessments are available)
- only keep element types that a system designer or librarian has deemed to be useful search results

In most of these approaches, result sets will still contain nested elements.

Third challenge: nested elements

Further techniques:

- remove nested elements in a postprocessing step to reduce redundancy.
- collapse several nested elements in the results list and use highlighting of query terms to draw the user's attention to the relevant passages.

Highlighting

- Gain 1: enables users to scan medium-sized elements (e.g., a section); thus, if the section and the paragraph both occur in the results list, it is sufficient to show the section.
- Gain 2: paragraphs are presented in-context (i.e., their embedding section). This context may be helpful in interpreting the paragraph.

Nested elements and term statistics

Further challenge related to nesting: we may need to distinguish different contexts of a term when we compute term statistics for ranking, in particular inverse document frequency (idf).

Example

The term *Gates* under the node *author* is unrelated to an occurrence under a content node like *section* if used to refer to the plural of *gate*. It makes little sense to compute a single document frequency for *Gates* in this example.

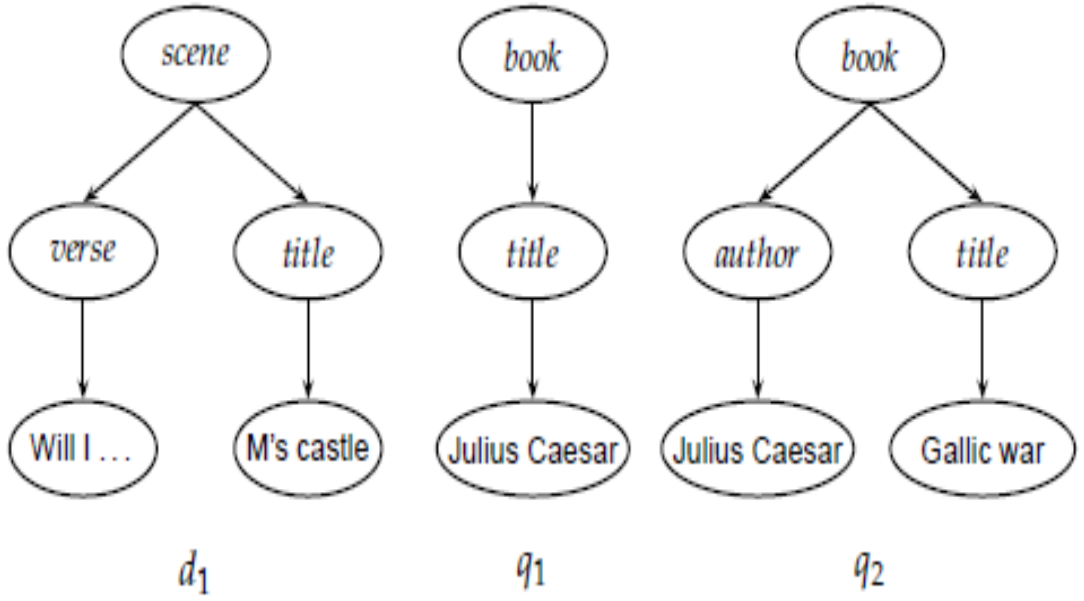
Solution: compute idf for XML-context term pairs.

- sparse data problems (many XML-context pairs occur too rarely to reliably estimate df)
- compromise: consider the parent node x of the term and not the rest of the path from the root to x to distinguish contexts.

- In many cases, several different XML schemas occur in a collection since the XML documents in an IR application often come from more than one source. This phenomenon is called *schema heterogeneity* or *schema diversity* and presents yet another challenge.

Vector space model for XML IR

- Lexicalized subtrees

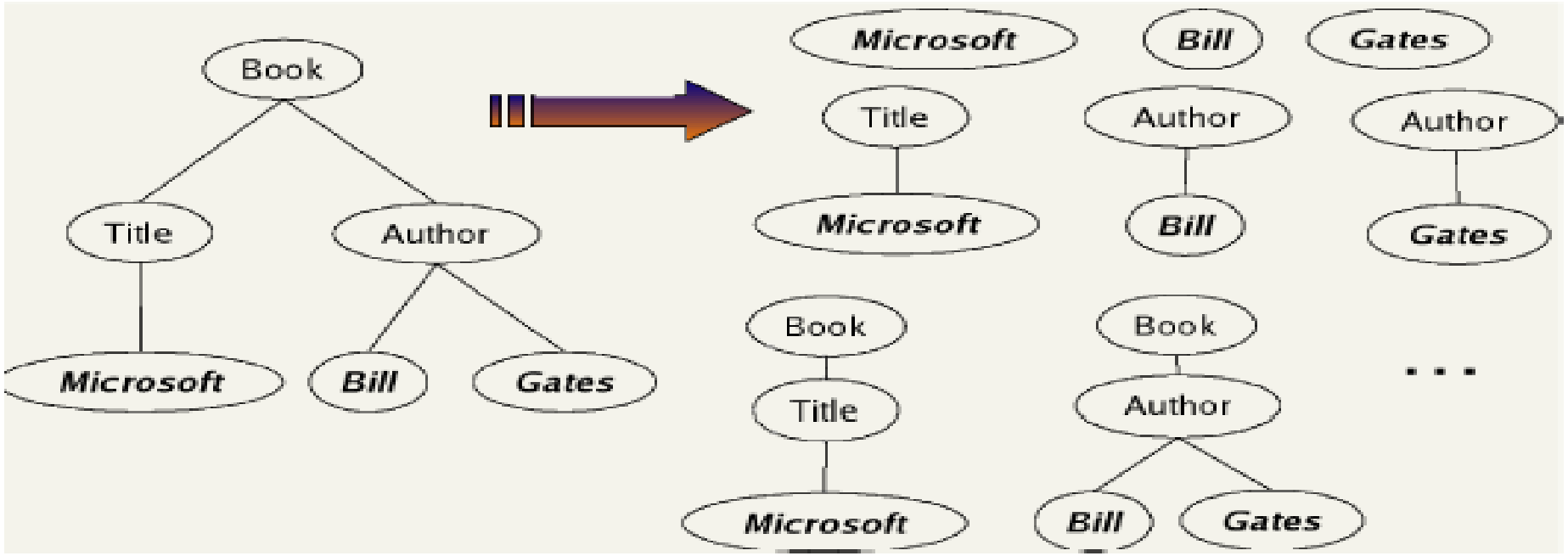


We want a book entitled *Julius Caesar* to be a match for $q1$ and no match (or a lower weighted match) for $q2$. In unstructured retrieval, there would be a single dimension of the vector space for Caesar. In XML retrieval, we must separate the title word Caesar from the author name Caesar.

Aim: to have each dimension of the vector space encode a word together with its position within the XML tree.

How: Map XML documents to lexicalized subtrees.

1. Take each text node (leaf) and break it into multiple nodes, one for each word. E.g. split *Bill Gates* into *Bill* and *Gates*
2. Define the dimensions of the vector space to be lexicalized subtrees of documents – subtrees that contain at least one vocabulary term.



► **Figure 10.8** A mapping of an XML document (left) to a set of lexicalized subtrees (right).

Lexicalized subtrees

We can now represent queries and documents as vectors in this space of lexicalized subtrees and compute matches between them, e.g. using the vector space formalism.

Vector space formalism in unstructured VS. structured IR

The main difference is that the dimensions of vector space in unstructured retrieval are vocabulary terms whereas they are lexicalized subtrees in XML retrieval.

Structural term

There is a tradeoff between the dimensionality of the space and the accuracy of query results.

- If we restrict dimensions to vocabulary terms, then we have a standard vector space retrieval system that will retrieve many documents that do not match the structure of the query (e.g., *Gates* in the title as opposed to the author element).
- If we create a separate dimension for each lexicalized subtree occurring in the collection, the dimensionality of the space becomes too large.

Compromise: index all paths that end in a single vocabulary term, in other words all XML-context term pairs. We call such an XML-context term pair a structural term and denote it by <*c*, *t*>: a pair of XML-context *c* and vocabulary term *t*.

Context resemblance

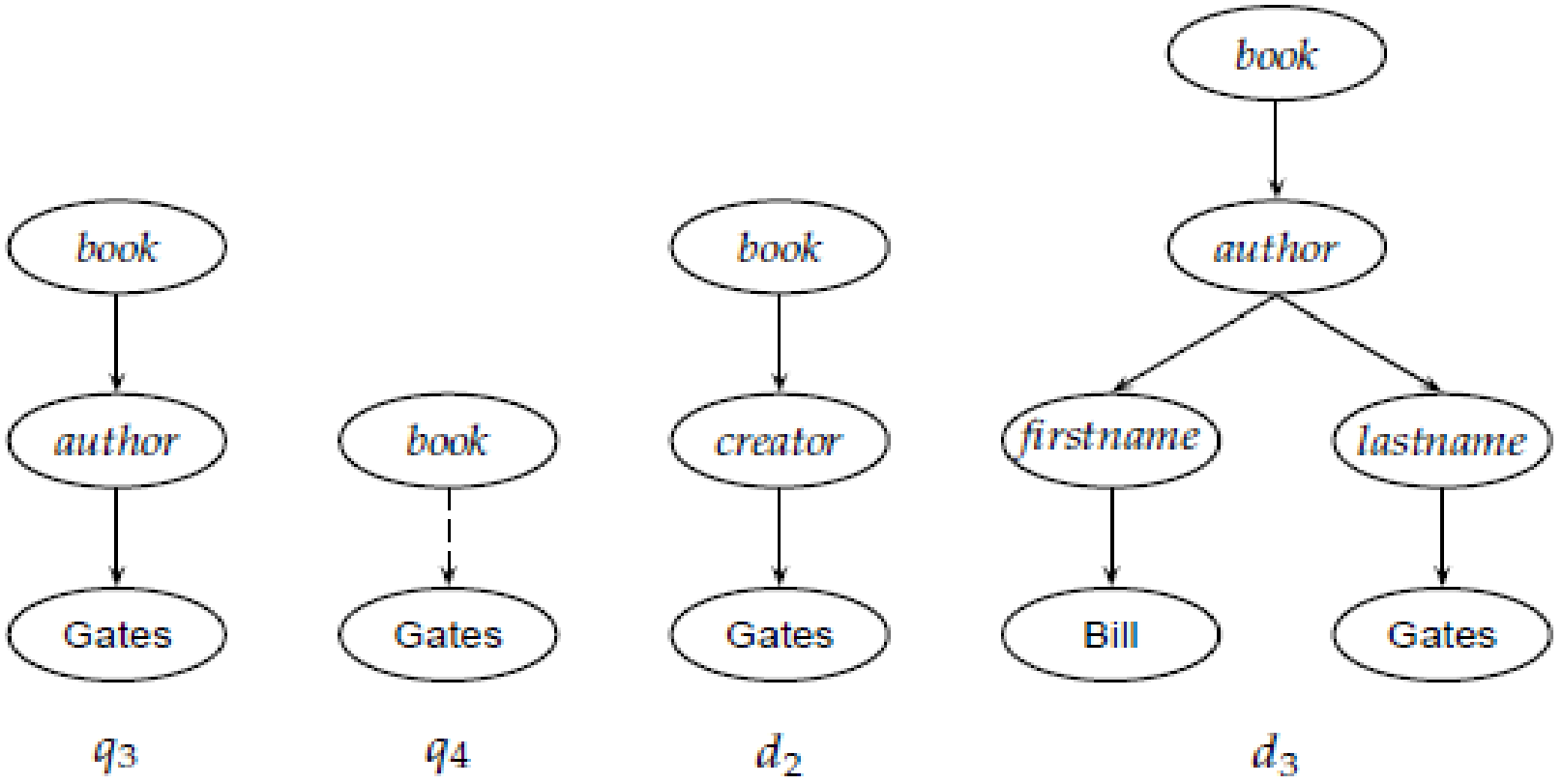
A simple measure of the similarity of a path c_q in a query and a path c_d in a document is the following *context resemblance* function C_R:

$$C_R(c_q, c_d) = \begin{cases} \frac{1+|c_q|}{1+|c_d|} & \text{if } c_q \text{ matches } c_d \\ 0 & \text{if } c_q \text{ does not match } c_d \end{cases}$$

$|c_q|$ and $|c_d|$ are the number of nodes in the query path and document path, resp.

c_q matches c_d iff we can transform c_q into c_d by inserting additional nodes.

Context resemblance example



► **Figure 10.6** Schema heterogeneity: intervening nodes and mismatched names.

- CR(*cq*4 , *cd*2) = 3/4 = 0.75 and CR(*cq*4 , *cd*3) = 3/5 = 0.6

  where *cq*4 , *cd*2 and *cd*3 are the relevant paths from top to leaf node in *q*4, *d*2 and *d*3, respectively.

- The value of CR(*cq*, *cd*) is 1.0 if *q* and *d* are identical.

Document similarity measure

The final score for a document is computed as a variant of the cosine measure, which we call SIMNOMERGE.

SIMNOMERGE(q, d) =

$$\sum_{c_k \in B} \sum_{c_l \in B} \text{CR}(c_k, c_l) \sum_{t \in V} \text{weight}(q, t, c_k) \frac{\text{weight}(d, t, c_l)}{\sqrt{\sum_{c \in B, t \in V} \text{weight}^2(d, t, c)}}$$

- V is the vocabulary of non-structural terms
- B is the set of all XML contexts
- weight (q, t, c), weight(d, t, c) are the weights of term t in XML context c in query q and document d, resp. (standard weighting e.g. idf_t x $\text{wf}_{t,d}$, where idf_t depends on which elements we use to compute df_t.)

SIMNOMERGE(q, d) is not a true cosine measure since its value can be larger than 1.0.

SIMNOMERGE algorithm

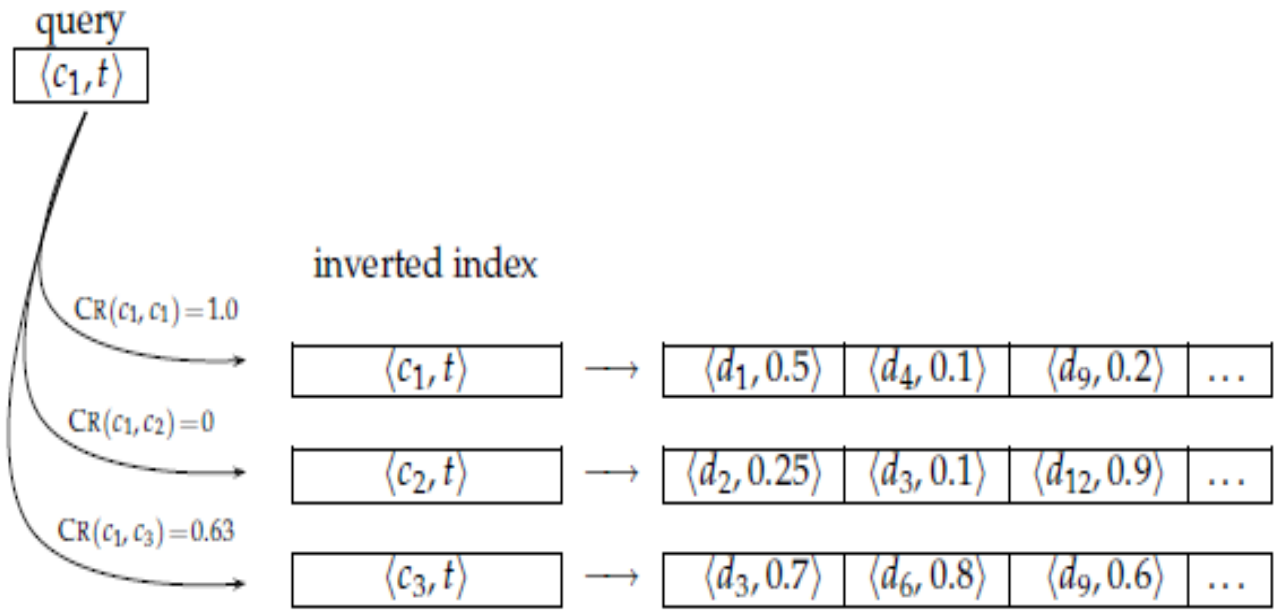
SCOREDOCUMENTSWITHSIMNOMERGE(q, B, V, N, *normalizer*)

```
1  for n ← 1 to N
2  do score[n] ← 0
3  for each ⟨c_q, t⟩ ∈ q
4  do w_q ← WEIGHT(q, t, c_q)
5      for each c ∈ B
6      do if CR(c_q, c) > 0
7             then postings ← GETPOSTINGS(⟨c, t⟩)
8                  for each posting ∈ postings
9                  do x ← CR(c_q, c) * w_q * weight(posting)
10                    score[docID(posting)] += x
11 for n ← 1 to N
12 do score[n] ← score[n]/normalizer[n]
13 return score
```

- The query-document similarity function in is called SIMNOMERGE because different XML contexts are kept separate for the purpose of weighting.

SIMNOMERGE computation of query-document similarities

- <$c1$, t> is one of the structural terms in the query.
- We successively retrieve all postings lists for structural terms <c', t> with the same vocabulary term t.
- Three example postings lists are shown.
- For the first one, we have CR($c1$, $c1$) = 1.0 since the two contexts are identical.
- The next context has no context resemblance with $c1$: CR($c1$, $c2$) = 0 and the corresponding postings list is ignored. The context match of $c1$ with $c3$ is 0.63>0 and it will be processed.



▶ **Figure 10.10** Scoring of a query with one structural term in SIMNOMERGE.

In this example, the highest ranking document is $d9$ with a similarity of 1.0 ×0.2 +0.63× 0.6 = 0.578.
To simplify the figure, the query weight of <c_1, t> is assumed to be 1.0.

Evaluation of XML IR

Initiative for the Evaluation of XML retrieval (INEX)

INEX: standard benchmark evaluation (yearly) that has produced test collections (documents, sets of queries, and relevance judgments).
Based on IEEE journal collection (since 2006 INEX uses the much larger English Wikipedia test collection).
The relevance of documents is judged by human assessors.

INEX 2002 collection statistics	
12,107	number of documents
494 MB	size
1995—2002	time of publication of articles
1,532	average number of XML nodes per document
6.9	average depth of a node
30	number of CAS topics
30	number of CO topics

INEX topics

Two types:

1. content-only or **CO topics**: regular keyword queries as in unstructured information retrieval
2. content-and-structure or **CAS topics**: have structural constraints in addition to keywords

Since CAS queries have both structural and content criteria, relevance assessments are more complicated than in unstructured retrieval

INEX topics

- Queries are specified using a simplified version of XPath called NEXI
- NEXI constructs include *paths* and *path filters*
  - A path is a specification of an element (or node) in the XML tree structure
  - A path filter restricts the results to those that satisfy textual or numerical constraints

NEXI Examples

- //A//B : any B element that is a descendant of an A element in the XML tree. A descendant will be contained in the ancestor element.
- //A/* : any descendant of an A element
- //A[about(.//B,"topic")] : elements that contain a B element that is "about" "topic". The *about* predicate is not defined but is implemented using some retrieval model. .//B is a relative path.
- //A[.//B = 777] : A elements that contain a B element with value equal to 777.

INEX Example Queries

- //article[.//fm/yr < 2000]//sec[about(.,"search engines")]
  - Find articles published before 2000 (fm is the "front matter" of the article) that contain sections discussing "search engines"
- //article[about(.//st,+comparison) AND about(.//bib,"machine learning")]
  - Find articles with a section title containing the word "comparison" and with a bibliography that discusses "machine learning"
- //*[about(.//fgc, corba architecture) AND about(.//p, figure corba architecture)]
  - Find any elements that contain a figure caption about "corba architecture" and a paragraph mentioning "figure corba architecture".

INEX relevance assessments

INEX 2002 defined component coverage and topical relevance as orthogonal dimensions of relevance.

Component coverage

Evaluates whether the element retrieved is "structurally" correct, i.e., neither too low nor too high in the tree.

We distinguish four cases. Could these apply to unstructured text?

1. Exact coverage (E): The information sought is the main topic of the component and the component is a meaningful unit of information.
2. Too small (S): The information sought is the main topic of the component, but the component is not a meaningful (self-contained) unit of information.
3. Too large (L): The information sought is present in the component, but is not the main topic.
4. No coverage (N): The information sought is not a topic of the component.

INEX relevance assessments

The **topical relevance** dimension also has four levels: highly relevant (3), fairly relevant (2), marginally relevant (1) and nonrelevant (0).

Combining the relevance dimensions

Components are judged on both dimensions and the judgments are then combined into a digit-letter code, e.g. 2S is a fairly relevant component that is too small. In theory, there are 16 combinations of coverage and relevance, but many cannot occur. For example, a nonrelevant component cannot have exact coverage, so the combination 3N is not possible.

INEX relevance assessments

The relevance-coverage combinations are quantized as follows:

$$\mathbf{Q}(rel, cov) = \begin{cases} 1.00 & \text{if} \quad (rel, cov) = 3\text{E} \\ 0.75 & \text{if} \quad (rel, cov) \in \{2\text{E}, 3\text{L}\} \\ 0.50 & \text{if} \quad (rel, cov) \in \{1\text{E}, 2\text{L}, 2\text{S}\} \\ 0.25 & \text{if} \quad (rel, cov) \in \{1\text{S}, 1\text{L}\} \\ 0.00 & \text{if} \quad (rel, cov) = 0\text{N} \end{cases}$$

This evaluation scheme takes account of the fact that binary relevance judgments, which are standard in unstructured IR, are not appropriate for XML retrieval. The quantization function **Q** does not impose a binary choice relevant/ nonrelevant and instead allows us to grade the component as partially relevant. The number of relevant components in a retrieved set *A* of components can then be computed as:

$$\#(\text{relevant items retrieved}) = \sum_{c \in A} \mathbf{Q}(rel(c), cov(c))$$

INEX evaluation measures

As an approximation, the standard definitions of precision and recall can be applied to this modified definition of relevant items retrieved, with some subtleties because we sum graded as opposed to binary relevance assessments.

Drawback

Overlap is not accounted for. Accentuated by the problem of multiple nested elements occurring in a search result.

Recent INEX focus: develop algorithms and evaluation measures that return non-redundant results lists and evaluate them properly.

Text-centric vs. data-centric XML retrieval

- TEXT-CENTRIC XML : XML structure serves as a framework within which we match the text of the query with the text of the XML documents. This is a system that is optimized for *text-centric XML*. While both text and structure are important, we give higher priority to text.
- Text-centric approaches are appropriate for data that are essentially text documents, marked up as XML to capture document structure. This is becoming a de facto standard for publishing text databases since most text documents have some form of interesting structure – paragraphs, sections, footnotes etc.
- A text-centric retrieval engine will have a hard time with proteomic data in bioinformatics or with the representation of a city map
- Two other types of queries that are difficult to handle in a text-centric structured retrieval model are joins and ordering constraints.
  - The query for employees with unchanged salary requires a join.
  - The query to retrieve the chapter of the book *Introduction to algorithms* that follows the chapter *Binomial heaps* requires ordering constraint
  - But, Xquery can handle it.

- In contrast, *data-centric XML* mainly encodes numerical and non-text attribute value data. When querying data-centric XML, exact match conditions is required in most cases.
- Eg. Find employees whose salary is the same this month as it was 12months ago.
  - This query requires no ranking. It is purely structural and an exact matching of the salaries in the two time periods is probably sufficient to meet the user's information need.
- Data-centric approaches are commonly used for data collections with complex structures that mainly contain non-text data.

- Relational databases are better equipped to handle many structural constraints, particularly joins (but ordering is also difficult in a database framework) For this reason, most data-centric XML retrieval systems are extensions of relational databases