

HyREX: Eine Hypermedia-Retrievalengine für XML-Dokumente

Norbert Fuhr
Universität Dortmund
fuhr@cs.uni-dortmund.de

Inhalt

- I. XQuery vs. Information Retrieval
- II. IR-Konzepte für XML
- III. XIRQL
- IV. HyREX-Retrievalengine
- V. Zusammenfassung und Ausblick

I. XQuery vs. Information Retrieval

XQuery: Vorschlag der W3C-Arbeitsgruppe für XML-Anfragesprachen

FOR/LET PathExpression
WHERE AdditionalSelectionCriteria
RETURN ResultConstruction

Daten- vs. Dokument-orientierte Sicht

- Daten-orientierte Sicht
XML als Austauschformat für strukturierte Daten
 - Dokumenten-orientierte Sicht
XML als Format zur Repräsentation der logischen Struktur von Dokumenten
- XQuery fokussiert auf Daten-orientierte Sicht!

IR-Konzepte in XQuery

- Nur boolesches Retrieval
 - keine Gewichtung
 - keine Rangordnungen
- Bislang nur Funktionen zur Suche nach einzelnen Wörtern

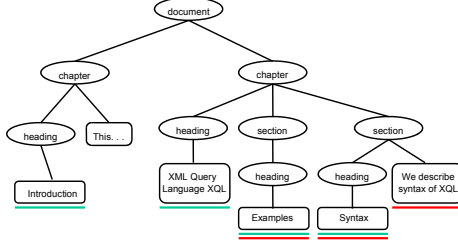
II. IR-Konzepte für XML

1. Gewichtung und Ranking
2. Relevanz-orientierte Suche
3. Datentypen mit vagen Prädikaten
4. Struktureller Relativismus

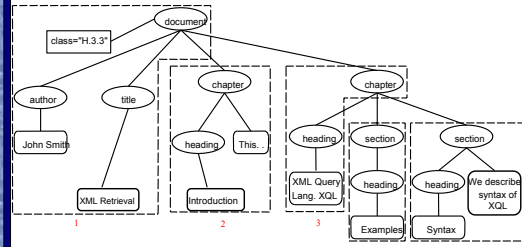
1. Gewichtung und Ranking

Problem: Gewichtung unterschiedlicher Vorkommensformen von Termen

`/document[./heading ∋ "XML" ∨ ./section/* ∋ "XML"]`



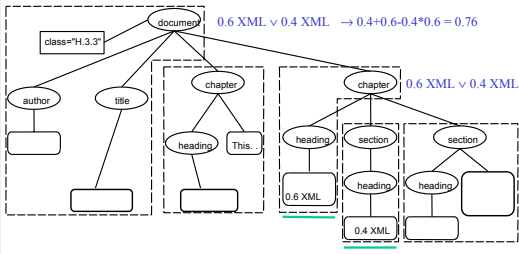
Indexknoten als Einheiten zur Termgewichtung



1. Zerlegung des Dokumentes in disjunkte Teile
2. Anwendung bekannter Indexierungsfunktionen (z.B. $tf \cdot idf$)

Indexknoten als Einheiten zur Termgewichtung

`/document[./heading ∋ "XML" ∨ ./section/* ∋ "XML"]`



2. Relevanz-orientierte Suche

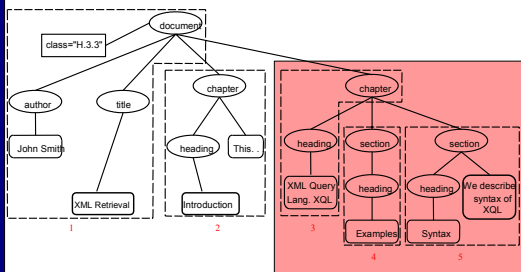
für Anfragen unabhängig von der Dokumentstruktur
(z.B.: "Suche Dokument(teile) über XML-Anfragesprachen")

- Einschränkung der möglichen Antworten (nicht alle Elemente sind geeignet)
- Retrievalstrategie: liefere spezifischsten Teilbaum, der die Anfrage beantwortet
- aber: Verrechnung mit gewichteter Indexierung?

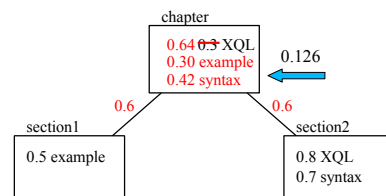
Lösung:

- Indexknoten als Wurzeln von möglichen Antworten
- Augmentierung als Konzept zur Verrechnung des Tradeoff zwischen Indexierungsgewichten und Spezifität von Antworten

Indexknoten für Relevanz-orientierte Suche

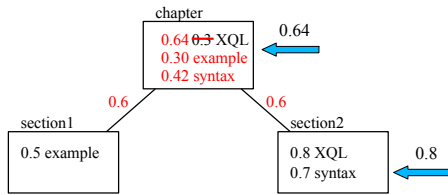


Augmentierung



Beispielanfrage: syntax & example

Augmentierung



Beispielanfrage: XQL

3. Datentypen mit vagen Prädikaten

XML-Markup ermöglicht detaillierte Auszeichnung von Textelementen

- Ausnutzung des Markups für präzisere Suchen
- gleichzeitig Berücksichtigung von Unsicherheit und Vagheit beim Retrieval
- Datentypen mit vagen Prädikaten

„Suche Informationen über das Werk eines Künstlers namens Ulbrich, der um 1900 im Rhein-Main-Gebiet tätig war“

Ernst Olbrich, Darmstadt, 1899

- (Erweiterbare) Datentypen für Dokumenten-zentrierte Sicht

(Personennamen, Datumsangaben, geogr. Bezeichnungen, Klassifikationen / Bilder, Audio, Video,...)

Erweiterbare Typhierarchie

Erweiterbare Typhierarchie mit vagen Prädikaten für jeden Datentyp

1. **text**: substring-Match
2. **westliche Sprache**: Wortsuche, Trunkierung, Wortabstandssuche
3. **deutscher Text**: Grund- und Stammformsuche, Komponenten von Komposita

Datentypen der XML-Elemente werden in XML-Schema definiert

4. Struktureller Relativismus

Unterscheidung Element/Attribut fallenlassen:

~author="Smith"

Suche in allen Elementen eines bestimmten

Datentyps:

#date=2001

III. XIRQL

XML IR Query Language

Erweiterung der Path Expressions von XQuery:

- probabilistisches Retrieval mit gewichteter Dokumentindexierung
- Relevanz-orientierte Suche
- Datentypen mit vagen Prädikaten
- Struktureller Relativismus

XIRQL-Path-Expressions

- Vage Prädikate

```
//text $c-words$ "compute"
```

```
//author $soundslike$ "meier"
```

- Gewichtete Fragebedingungen

```
//*[0.7 . $c-word$ "retrieval" + 0.3 . $c-word$ "XML"]
```

- Relevanz-orientierte Anfragen

```
//inode()[... $c-phrase$ "XML retrieval"]
```

- Struktureller Relativismus:

```
##person $soundslike$ "meier"
```

XIRQL vs. XQuery

XIRQL prozessiert Teilmenge von XQuery-Anfragen:

FOR \$X=PathExpression

RETURN \$X

- Keine Restrukturierung von Antworten
- Keine Werte-basierten Joins zwischen verschiedenen Dokumenten
- Erweiterte Path Expressions für IR

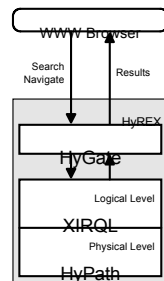
IV. HyREX

Hypermedia Retrieval Engine for XML

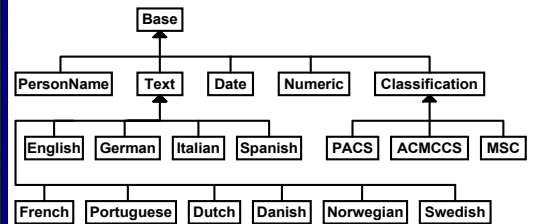
- Open-Source-Software für Information Retrieval in XML-Dokumenten
- Basiert auf der Anfragesprache XIRQL

HyREX-Architektur

- HyGate: Web-Gateway
- XIRQL: Anfragesprache auf der logischen Ebene
- HyPath: Zugriffspfade (physische Ebene)



HyREX-Datentypen



V. Zusammenfassung und Ausblick

IR-Konzepte für XML:

- Gewichtung und Ranking
- Relevanz-orientierte Suche
- Datentypen und vage Prädikate
- Struktureller Relativismus

XIRQL als IR-Erweiterung einer XQuery-Teilmenge

HyREX: Open-Source-Retrievalengine für XML:

ls6-www.cs.uni-dortmund.de/ir/hyrex

Ausblick

DAAD-Projekt FOCUS + EU-NoE DELOS (zusammen mit Mounia Lalmas, Univ. of London, et al.):

- Evaluierung von XML-Retrieval

EU-Projekt CYCLADES (zusammen mit IEI-CNR/Pisa, FhG-Fit/Bonn, FORTH/Heraklion)

- HyREX als Suchmaschine für vernetzte Open Archives

DFG-Projekt CLASSIX, (zusammen mit Gerhard Weikum, Univ. Saarbrücken, ab 1.2.02):

- Entwicklung von Verfahren für effizientes Best-Match-Retrieval für XIRQL
- Kombination von XIRQL und XQuery: probabilistische Variante von XQuery

CARMEN - Next Steps

- **Erstellung von HyREX-Distributionen**
(einfachere Installation, weitere Datentypen und Dokumentformate)
- **Benutzerschnittstelle**
(Anfrageformulierung, Ergebnispräsentation)
- **Erweiterung von HyREX in Richtung XQuery**
(Postprozessierung zur Restrukturierung von Antworten)
- **Integration von HyREX mit Dokument-Management**
(z.B. WebDAV: hierarchische Ordner für Arbeitsgruppen)