



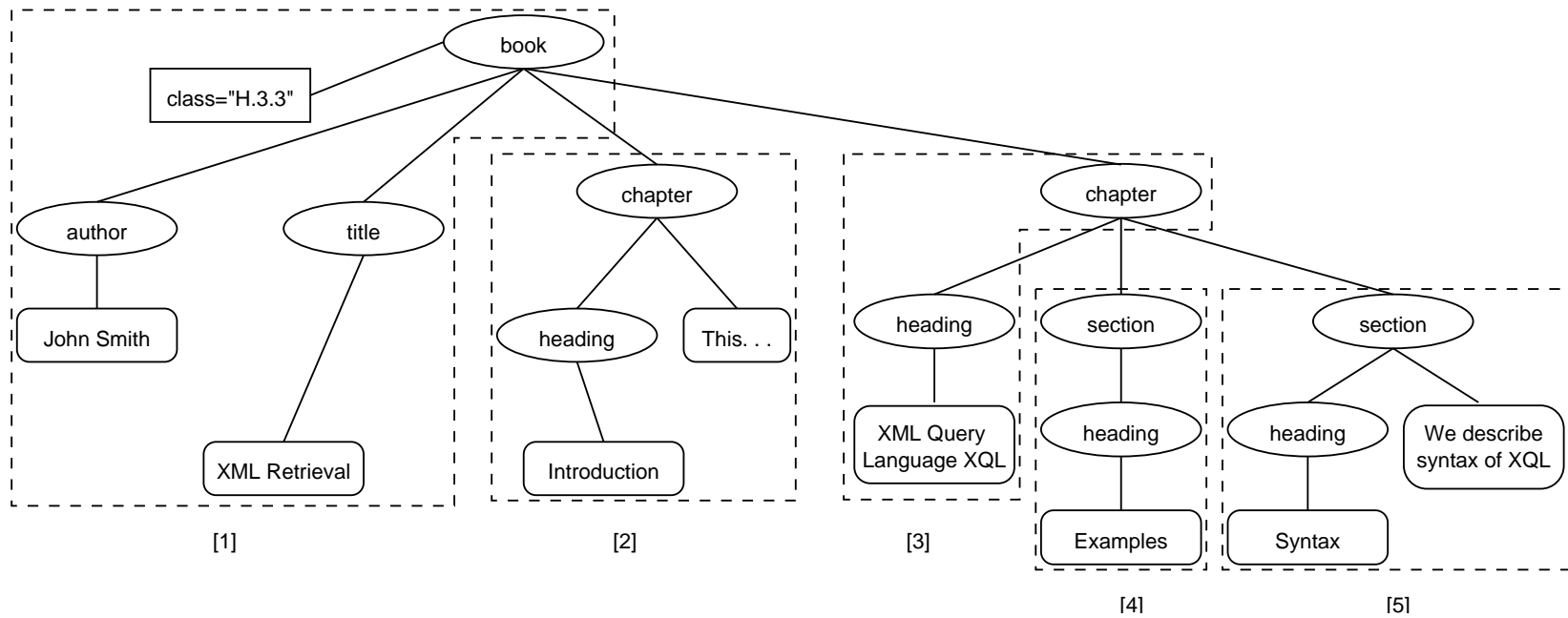
HyREX @ INEX 2003 (Content-Only Queries)

Mohammad Abolhassani, Norbert Fuhr, Saadia Malik
University of Duisburg-Essen, Germany

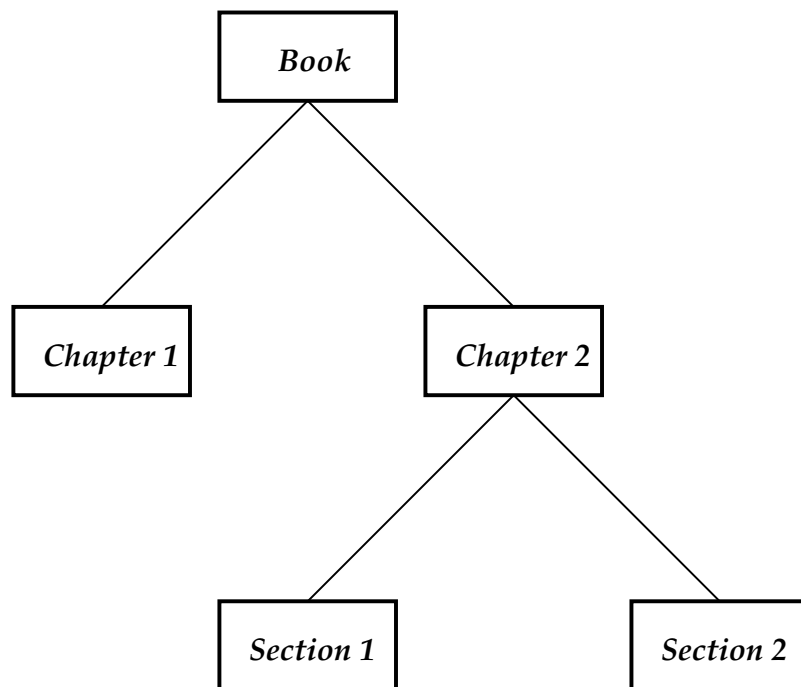
Content-Only Queries

- **Index Nodes**
 - a subtree of the document tree
 - meaningful as retrieval answer
 - defined based on the DTD
- **Retrieval approaches**
 - **Augmentation** (HyREX @ INEX 2002)
 - **DFR: Divergence From Randomness** [Amati/Rijsbergen 2002]

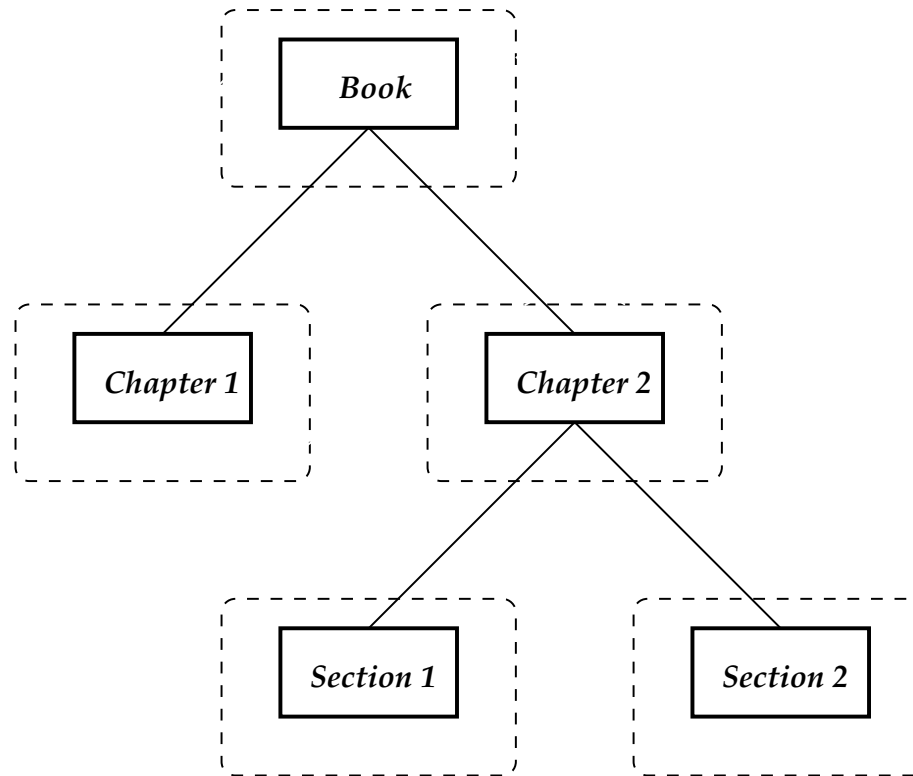
Index Nodes



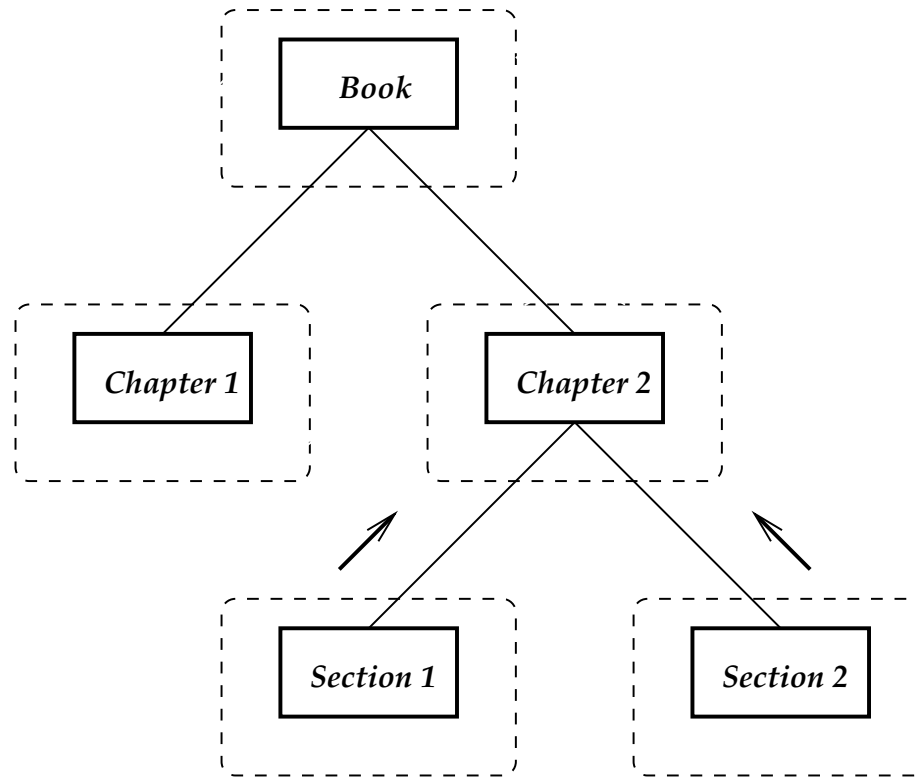
Index Nodes and Augmentation



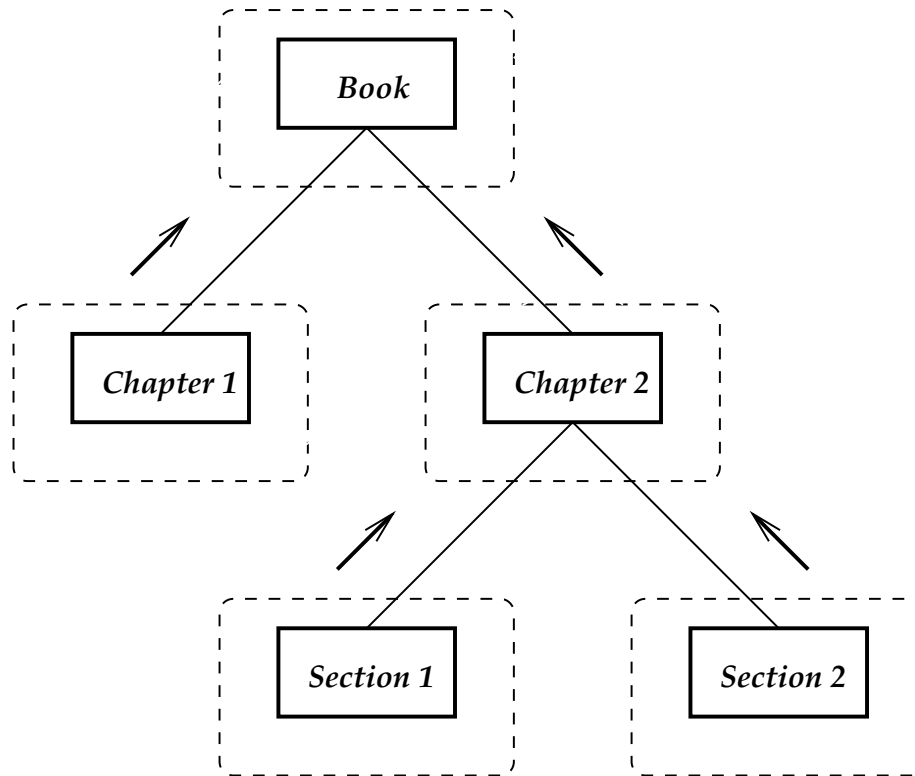
Index Nodes and Augmentation



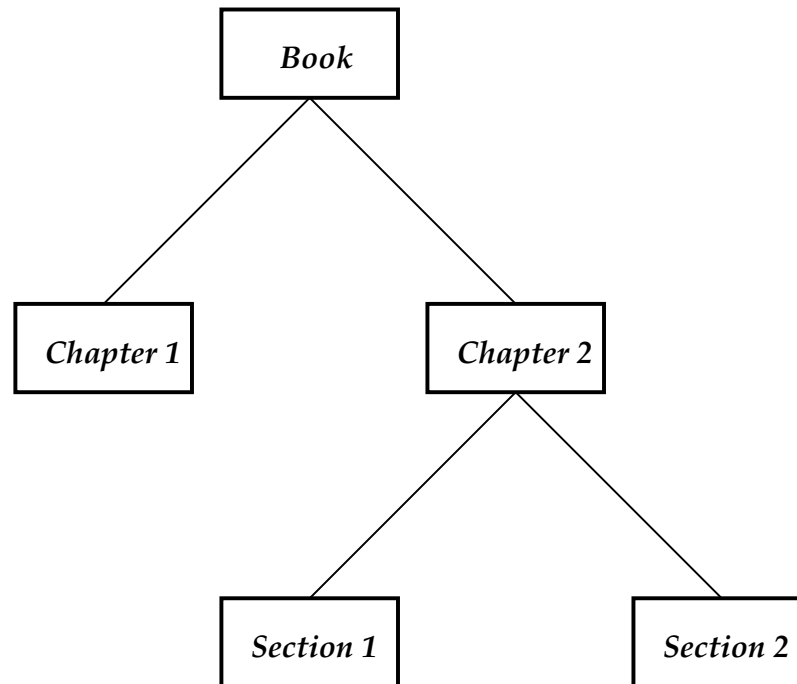
Index Nodes and Augmentation



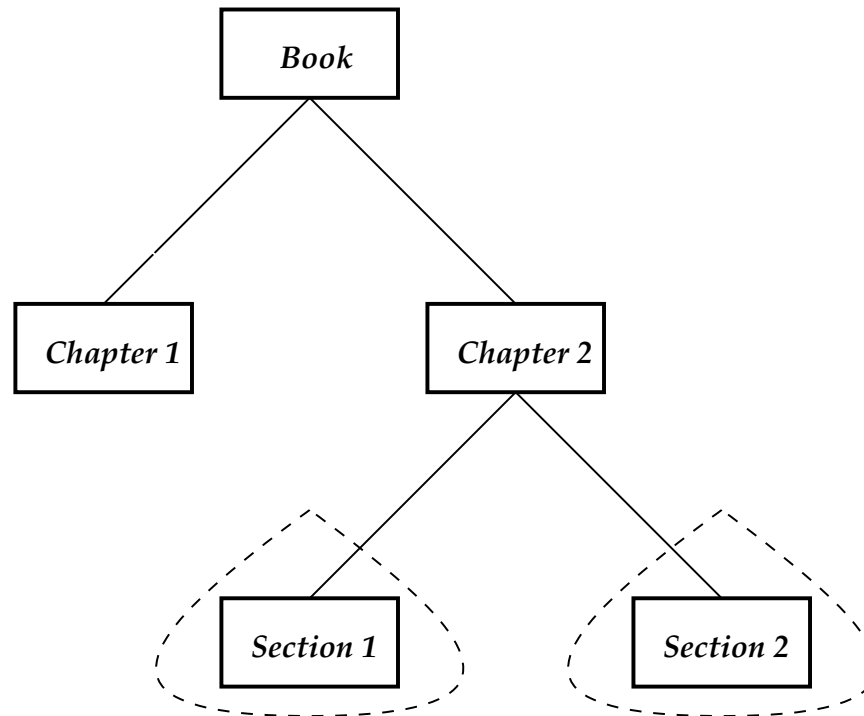
Index Nodes and Augmentation



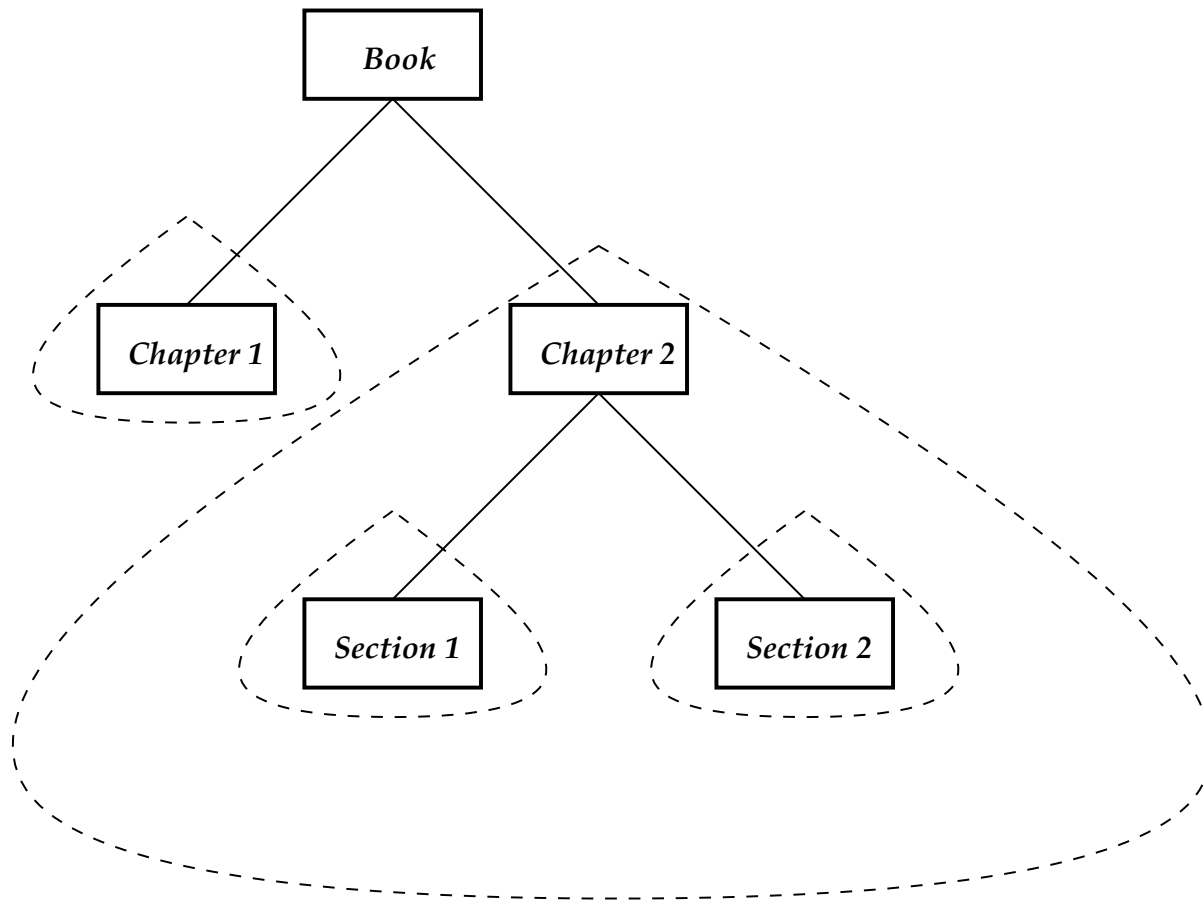
Index Nodes and DFR



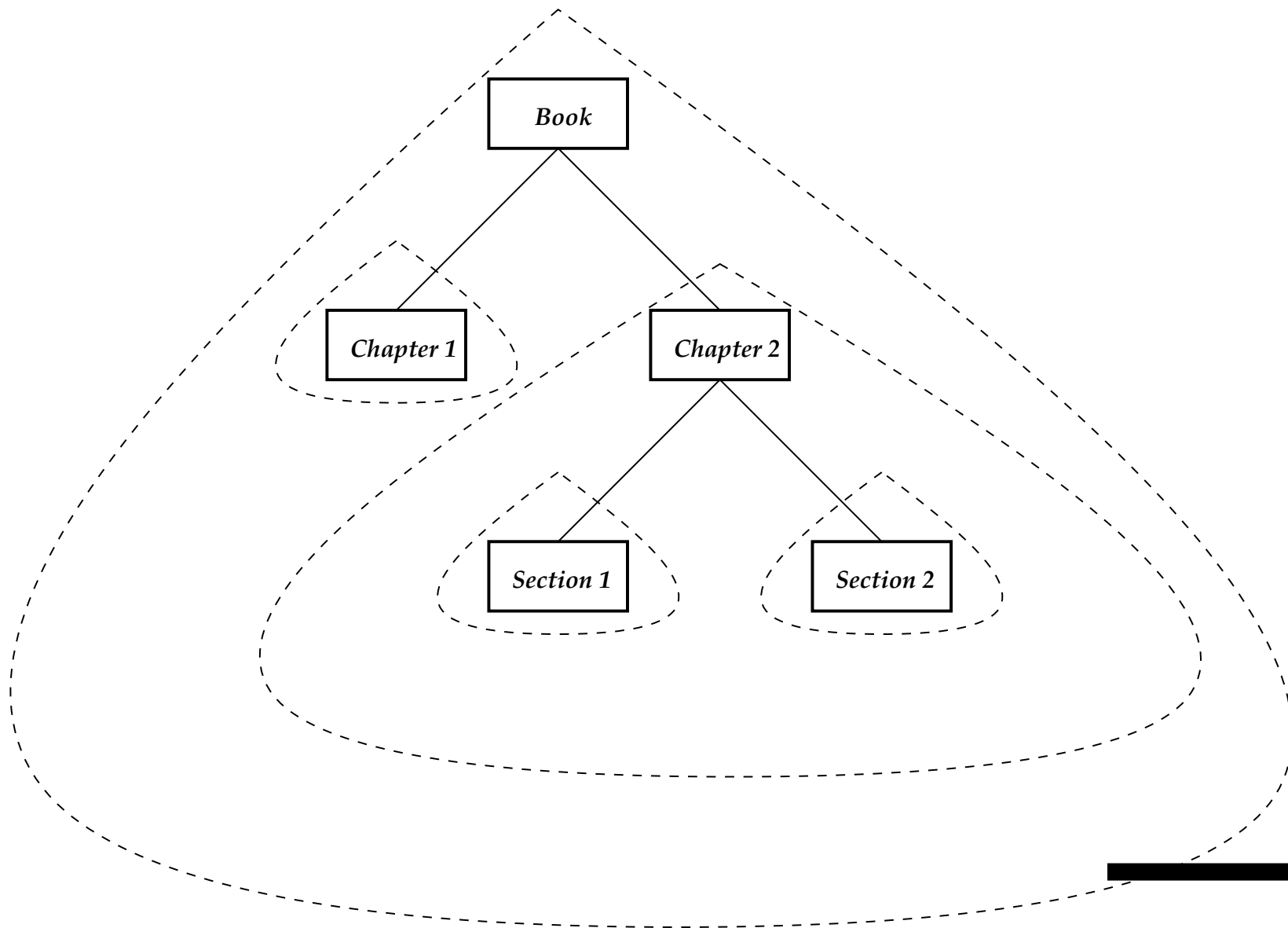
Index Nodes and DFR



Index Nodes and DFR



Index Nodes and DFR



A language model for XML IR

Divergence From Randomness (DFR) Basic Model

[Amati & Rijsbergen 2002]:

framework for deriving probabilistic models of IR, based on the *language model* approach.

Divergence from randomness

Term weighting:

measuring the divergence of the actual term distribution from a random process

$$\rightarrow w = \ln f_1 \cdot \ln f_2$$

- $\ln f_1$: models for distribution of terms over a collection of N documents of equal size (*Bose-Einstein, Bernoulli*)
- $\ln f_2$: models for multiple occurrences of a term within a document belonging to the *elite set*, (set of documents containing the term) (*Bernoulli, Laplace*)

Applying the DFR model

- applying document length normalisation (*second normalisation*) to "term frequency":

$$\rho(l) = c \cdot l^\beta \quad (\text{term density in document})$$

$$tf_n = \int_{l(d)}^{l(d)+avl} \rho(l) dl$$

- mapping tf to *normalised* term frequency (tf_n)
- use tf_n for computing Inf_1 and Inf_2
- applying a linear retrieval function:

$$R(q, d) = \sum_{t \in q} qt f \cdot Inf_2(tf_2) \cdot Inf_1(tf_1)$$

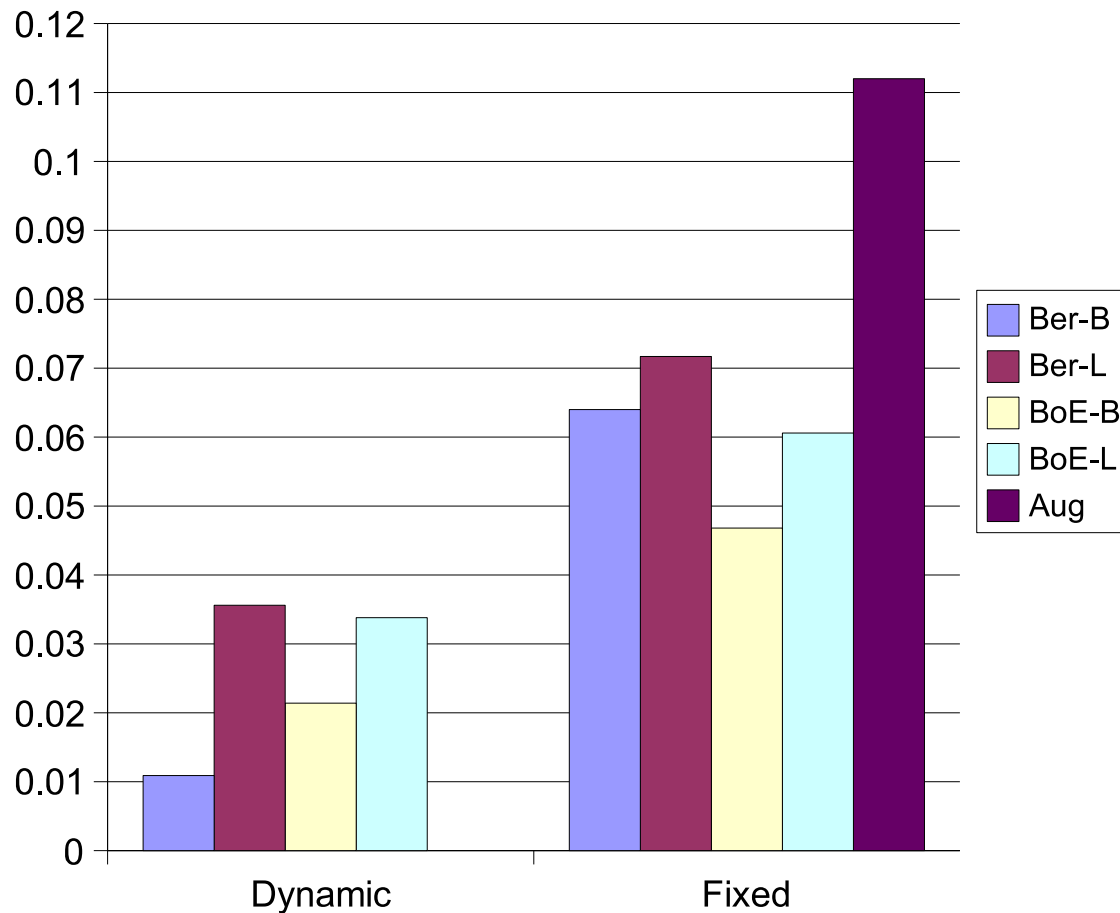
Experiments

Dynamic vs. fixed document length

- Dynamic document length: assume that collection consists of documents having the same size as current index node: $N = L/l(d)$
- Fixed document length: average document length = average length of index node

Experiments

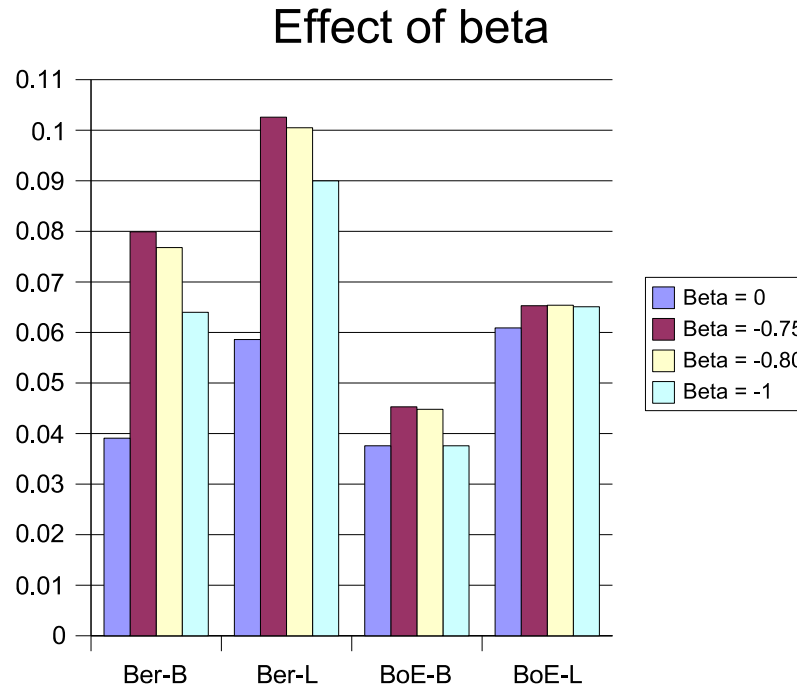
First DFR experiments vs. Augmentation



Experiments

Document length normalization

(term density: $\rho = c \cdot l^\beta$)



→ retrieval quality still below augmentation approach!

Considering document structure

third normalisation:

effect of "different levels" in the index node hierarchy
(root has level 1)

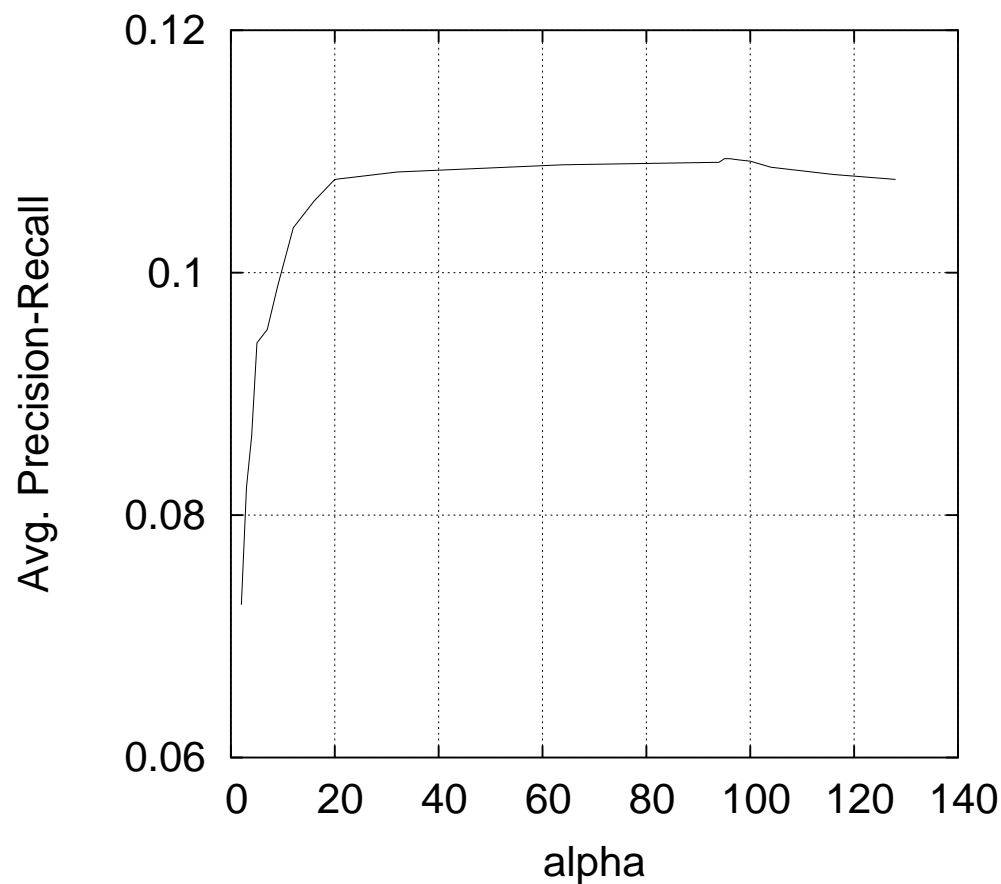
$$tf'_n = tf_n \cdot \frac{lev}{\alpha}$$

replacing tf_n with tf'_n for computing Inf_2 :

$$Inf_2 = \frac{1}{\frac{1}{\alpha} \cdot h(d) \cdot tf_n + 1}$$

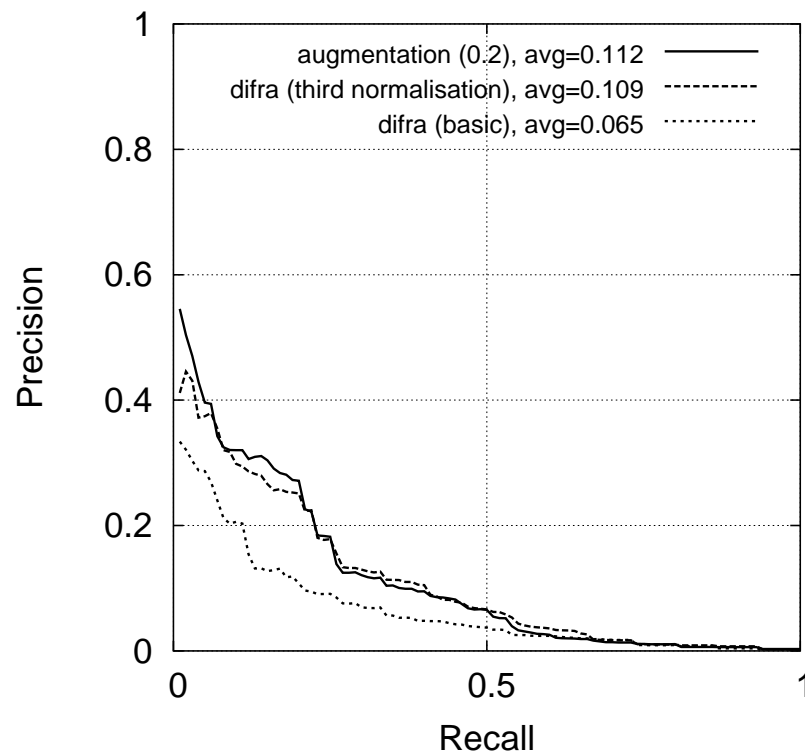
Experiments

Results for the Bose-Einstein L Norm combination with the third function using various values of α :



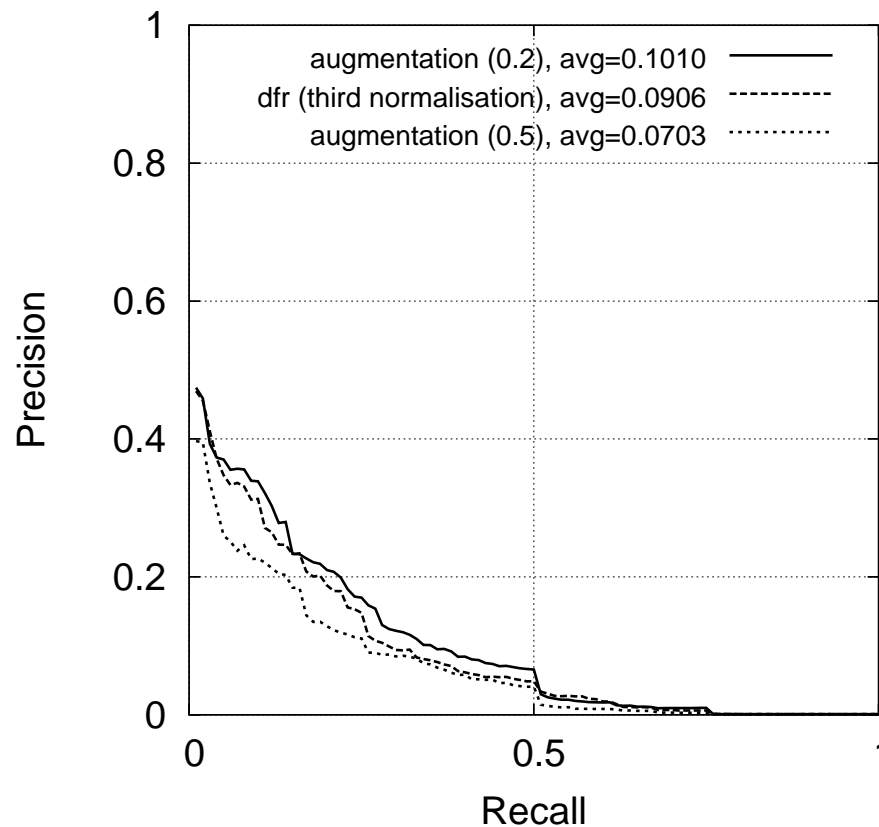
Experiments

Effect of third normalization
in comparison with augmentation approach
(Retrieval results for INEX 2002 collection)



INEX 2003

DFR with “best parameters” (from INEX 2002), in comparison to augmentation (factors 0.5 and 0.2:)



Conclusion

- new XML retrieval model, based on the "divergence from randomness" model
- importance of considering hierarchic structure of XML documents
- further research needed for theoretical justification of "third normalization"