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]
This is a diagram of a book structure with various sections and headings. The book has an author named John Smith and a title "XML Retrieval".

1. Author: John Smith
2. Title: XML Retrieval
3. Chapter 1: Introduction
   - Heading: This...
4. Chapter 2: XML Query Language XQL
   - Section: Examples
   - Section: Syntax
5. Chapter 3: We describe syntax of XQL

The diagram shows the hierarchical structure of the book with nodes labeled as "chapter", "heading", and "section".
Index Nodes and Augmentation

- Book
  - Chapter 1
  - Chapter 2
    - Section 1
    - Section 2
Index Nodes and Augmentation

- **Book**
  - **Chapter 1**
  - **Chapter 2**
    - **Section 1**
    - **Section 2**
Index Nodes and Augmentation

- Book
  - Chapter 1
  - Chapter 2
  - Section 1
  - Section 2
Index Nodes and Augmentation

- Book
  - Chapter 1
  - Chapter 2
    - Section 1
    - Section 2
Index Nodes and DFR

- Book
  - Chapter 1
  - Chapter 2
    - Section 1
    - Section 2
Index Nodes and DFR

- Book
  - Chapter 1
  - Chapter 2
    - Section 1
    - Section 2
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

\[ w = \text{Inf}_1 \cdot \text{Inf}_2 \]

- \( \text{Inf}_1 \): models for distribution of terms over a collection of \( N \) documents of equal size (Bose-Einstein, Bernoulli)

- \( \text{Inf}_2 \): 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":
  \[ \rho(l) = c \cdot l^\beta \] (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:
  \[ R(q, d) = \sum_{t \in q} q tf \cdot Inf_2(tf_2) \cdot Inf_1(tf_1) \]
Dynamic vs. fixed document length

- Dynamic document length: assume that collection consists of documents having the same size as current index node: \( N = \frac{L}{l(d)} \)

- 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!
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 $\alpha$: 

![Graph showing Avg. Precision-Recall against alpha values. The graph starts at 0.06 and goes up to 0.12, with alpha values ranging from 0 to 140.]
Experiments

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