Applying the Divergence From Randomness Approach for Content-Only Search in XML Documents

Mohammad Abolhassani, Norbert Fuhr

University of Duisburg-Essen, Germany

XML Retrieval

INEX tasks:
(Initiative for the evaluation of XML retrieval)

XML Retrieval

INEX tasks:
(Initiative for the evaluation of XML retrieval)

- Content-and-structure queries: Query conditions referring to content and structure of XML elements to be retrieved.

XML Retrieval

INEX tasks:
(Initiative for the evaluation of XML retrieval)

- Content-and-structure queries: Query conditions referring to content and structure of XML elements to be retrieved.
- Content-only queries:
  - query refers only to element content
  - IR system should retrieve most specific elements satisfying the query

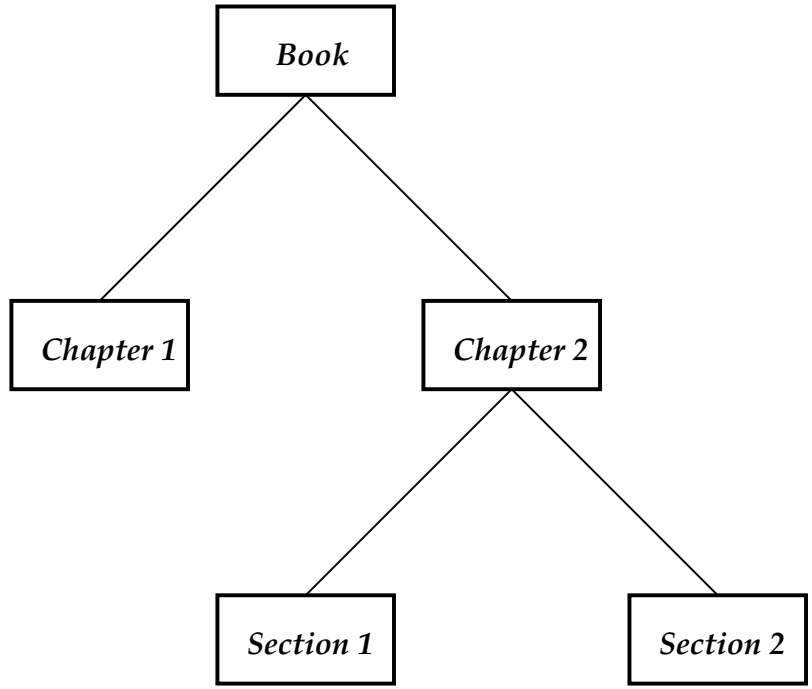
Content-Only Queries

- **Index Nodes**
  - a subtree of the document tree
  - meaningful as retrieval answer
  - defined based on the DTD

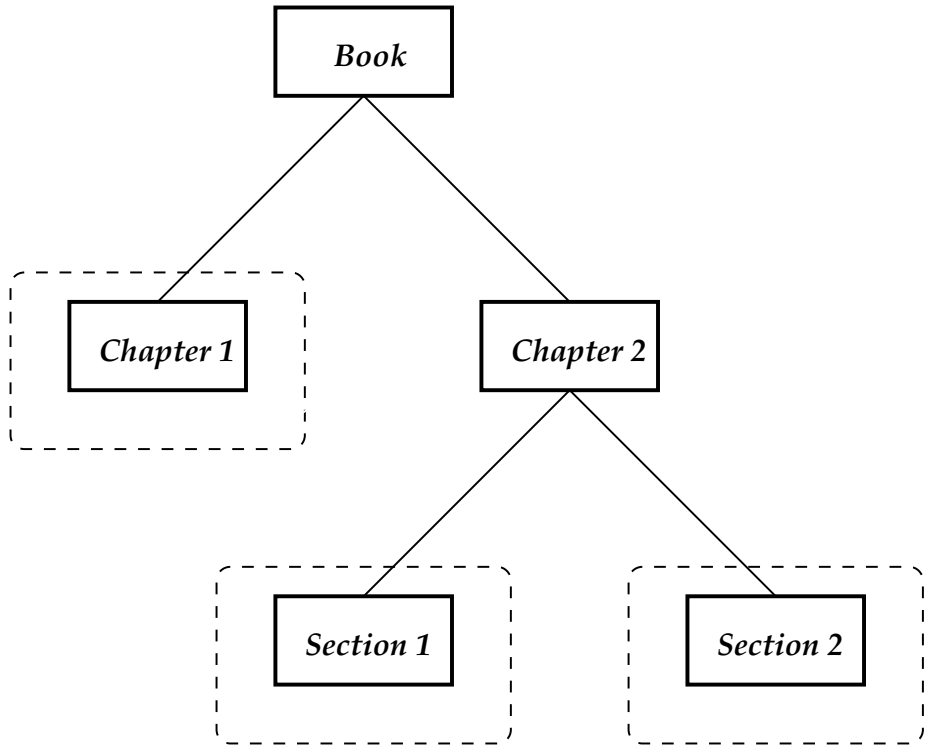
Content-Only Queries

- **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]

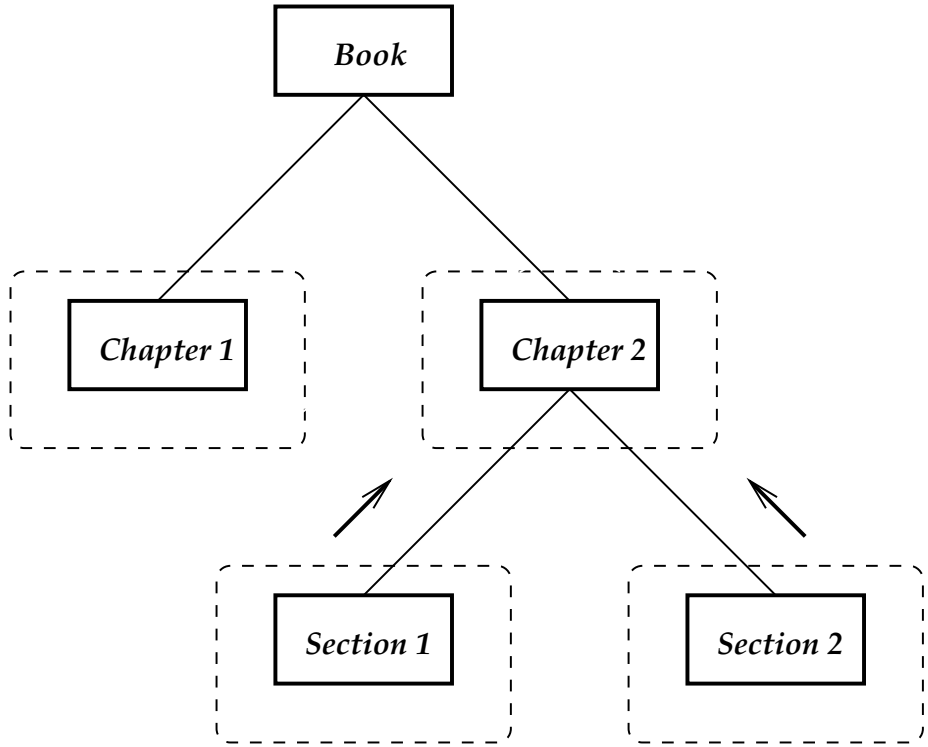
Index Nodes and Augmentation



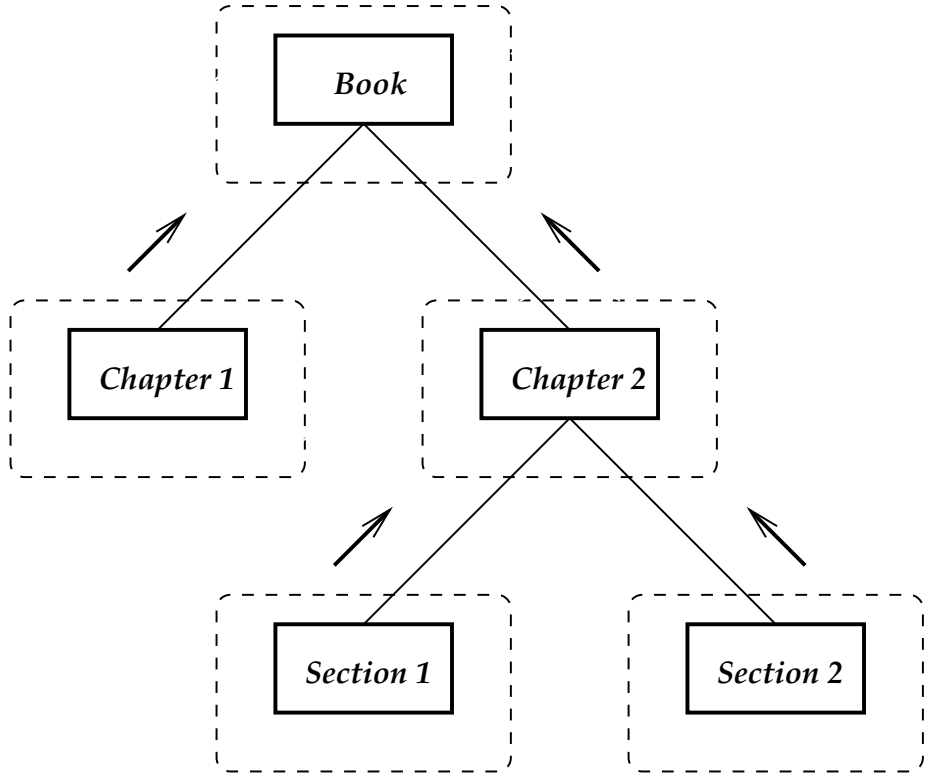
Index Nodes and Augmentation



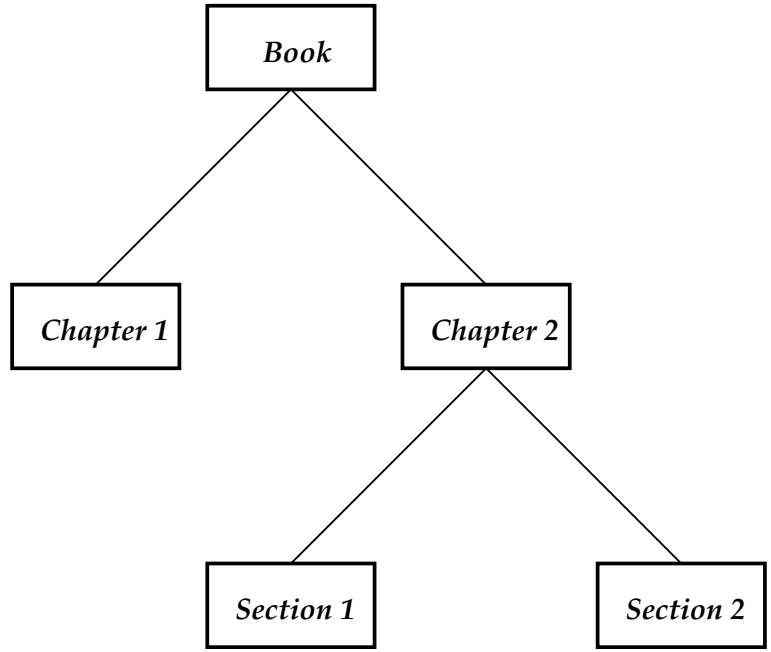
Index Nodes and Augmentation



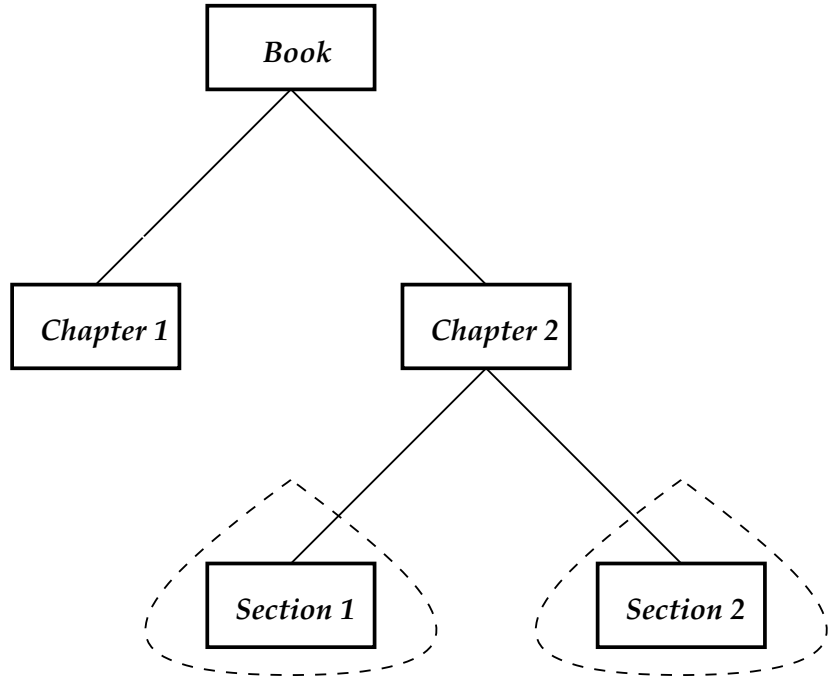
Index Nodes and Augmentation



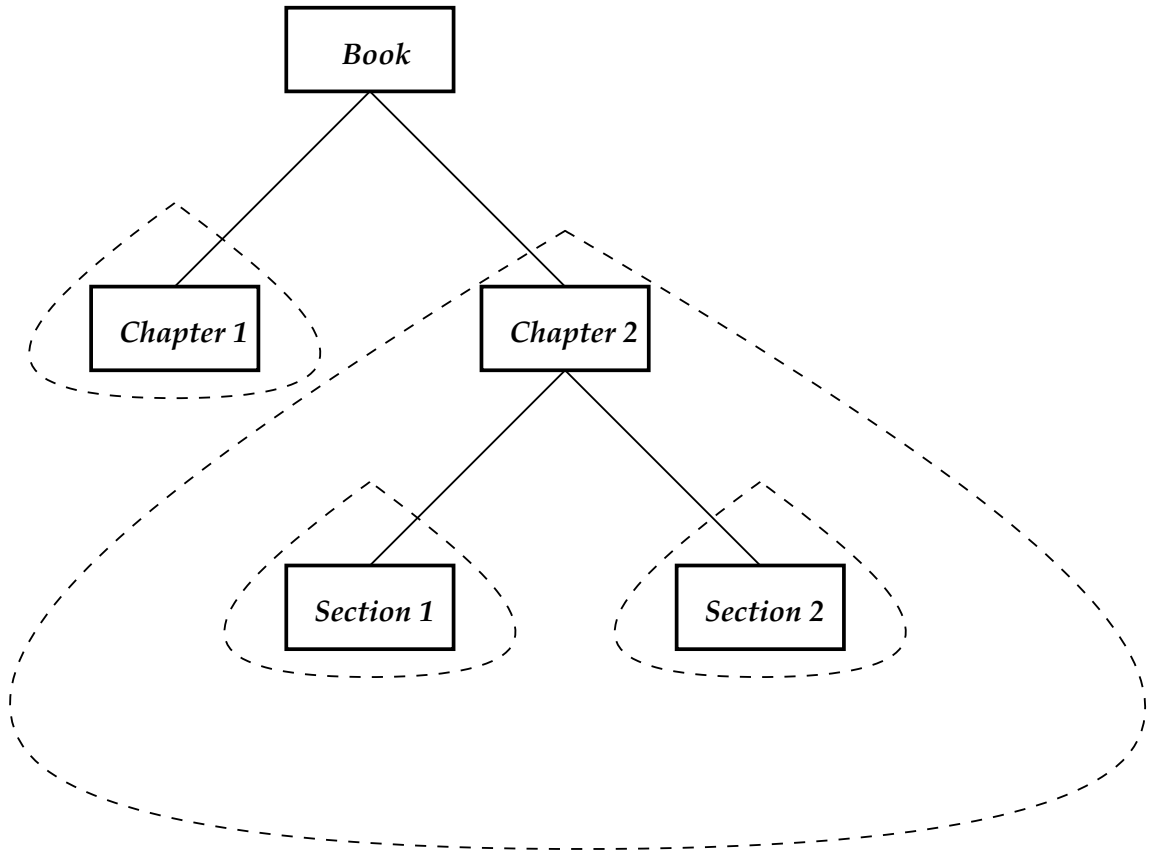
Index Nodes and DFR



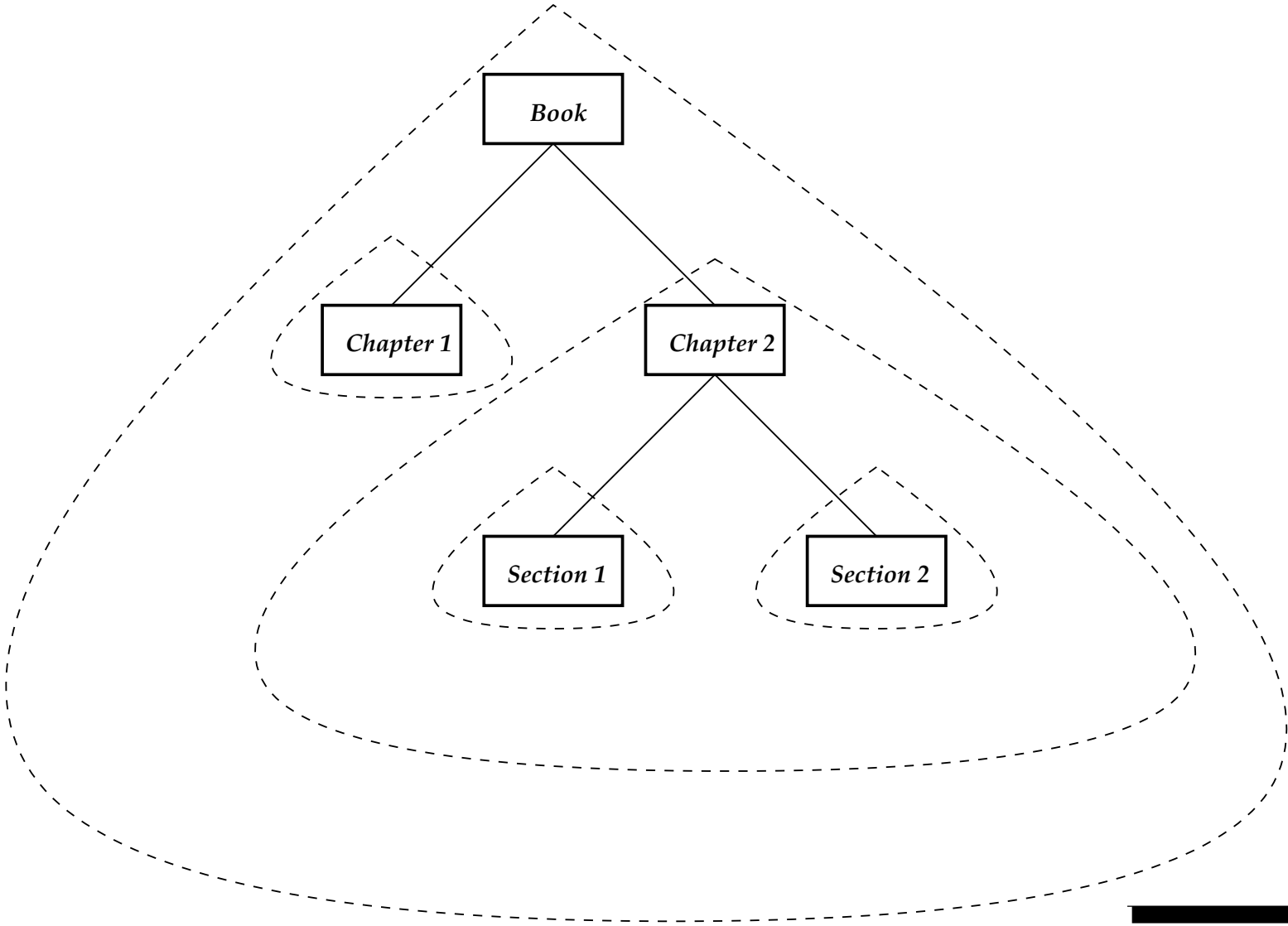
Index Nodes and DFR



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

$\rightarrow w = Inf_1 \cdot Inf_2$

Divergence from randomness

Term weighting:
measuring the divergence of the actual term distribution from a random process
$\rightarrow w = Inf_1 \cdot Inf_2$

- Inf_1: models for distribution of terms over a collection of N documents of equal size
  (Bose-Einstein, Bernoulli)

Divergence from randomness

Term weighting:
measuring the divergence of the actual term distribution from a random process
$\rightarrow w = Inf_1 \cdot Inf_2$

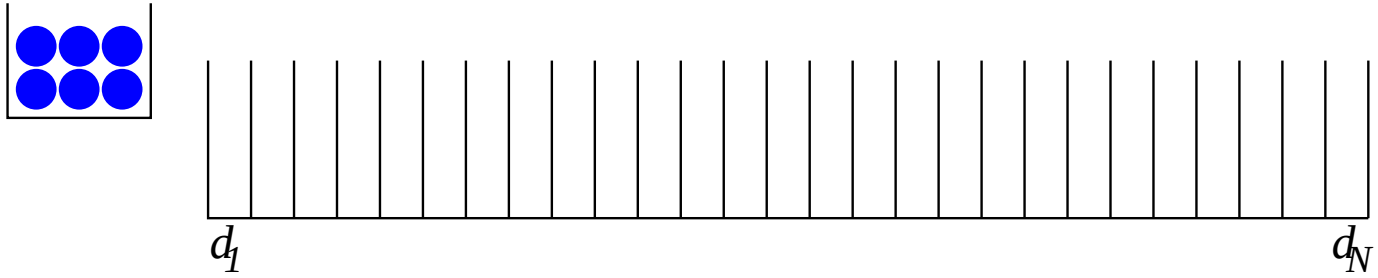
- Inf_1: models for distribution of terms over a collection of N documents of equal size
  (Bose-Einstein, Bernoulli)
- Inf_2: models for multiple occurences of a term within a document belonging to the *elite set*, (set of documents containing the term)
  (Bernoulli, Laplace)

Models of randomness

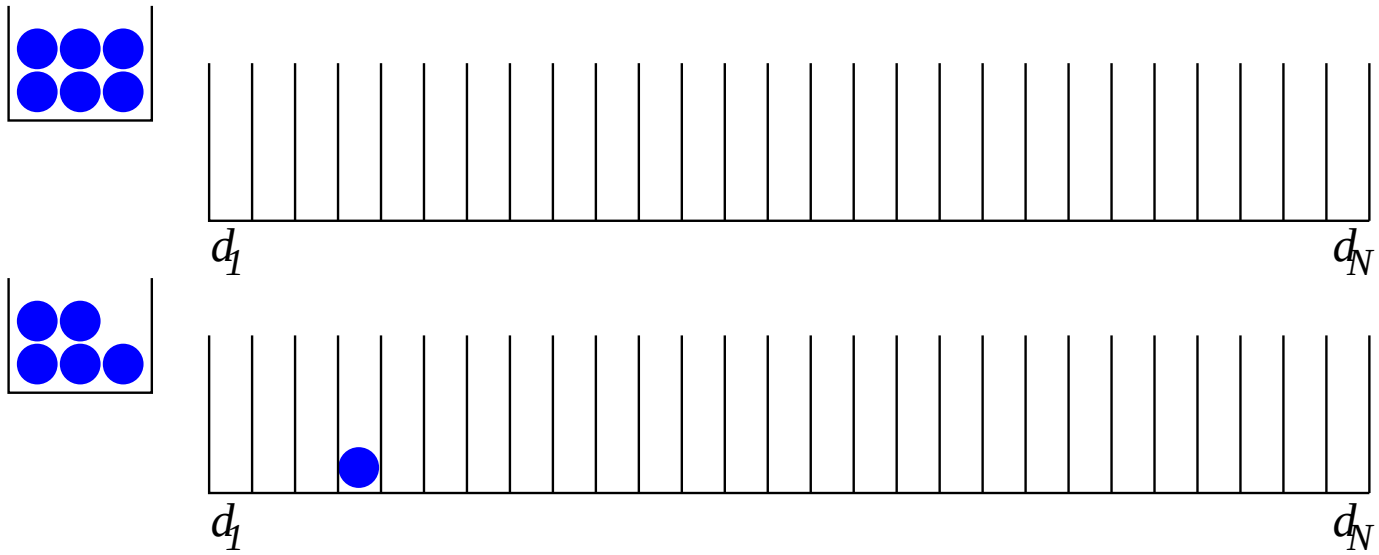
depending on assumptions about event space (definition of equiprobable events)

- Binomial model
- Bose-Einstein model

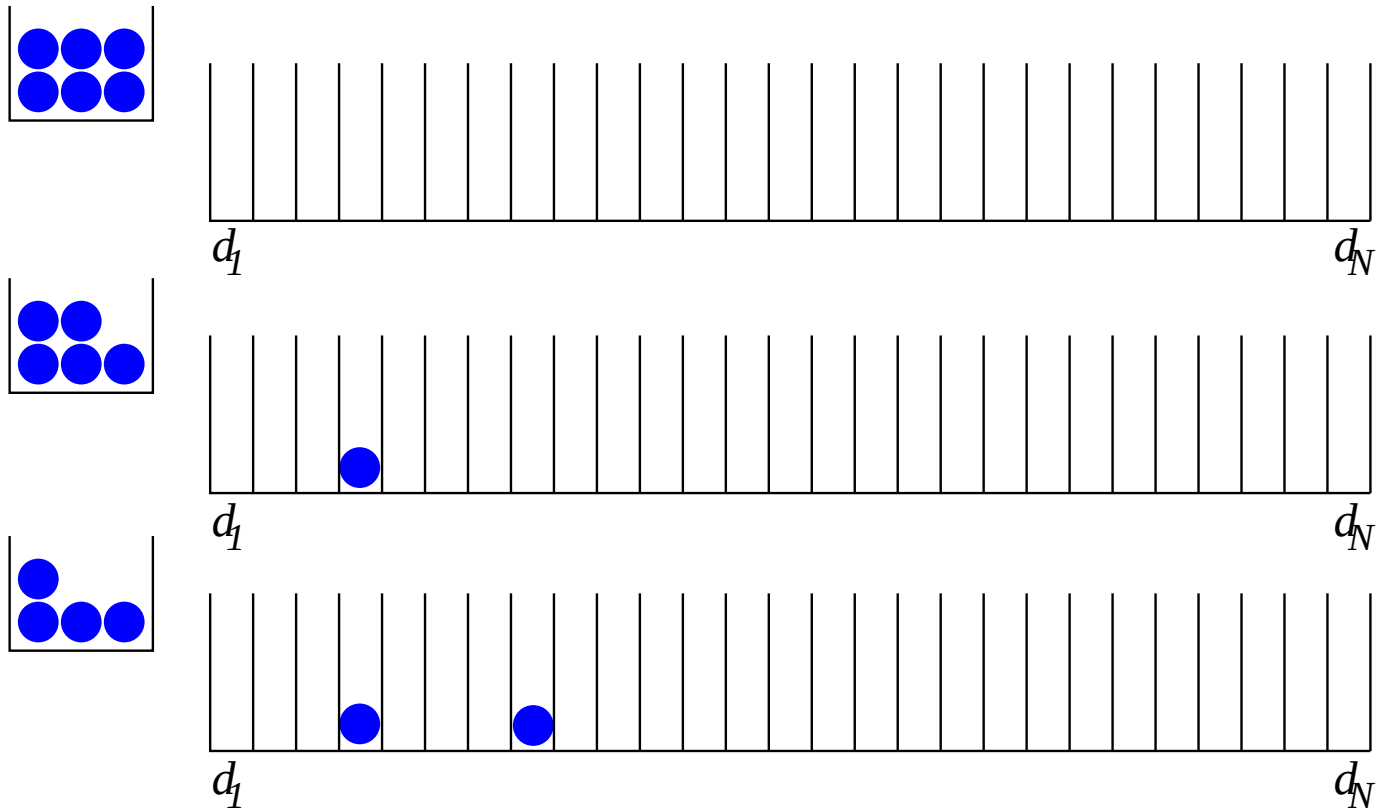
Binomial model



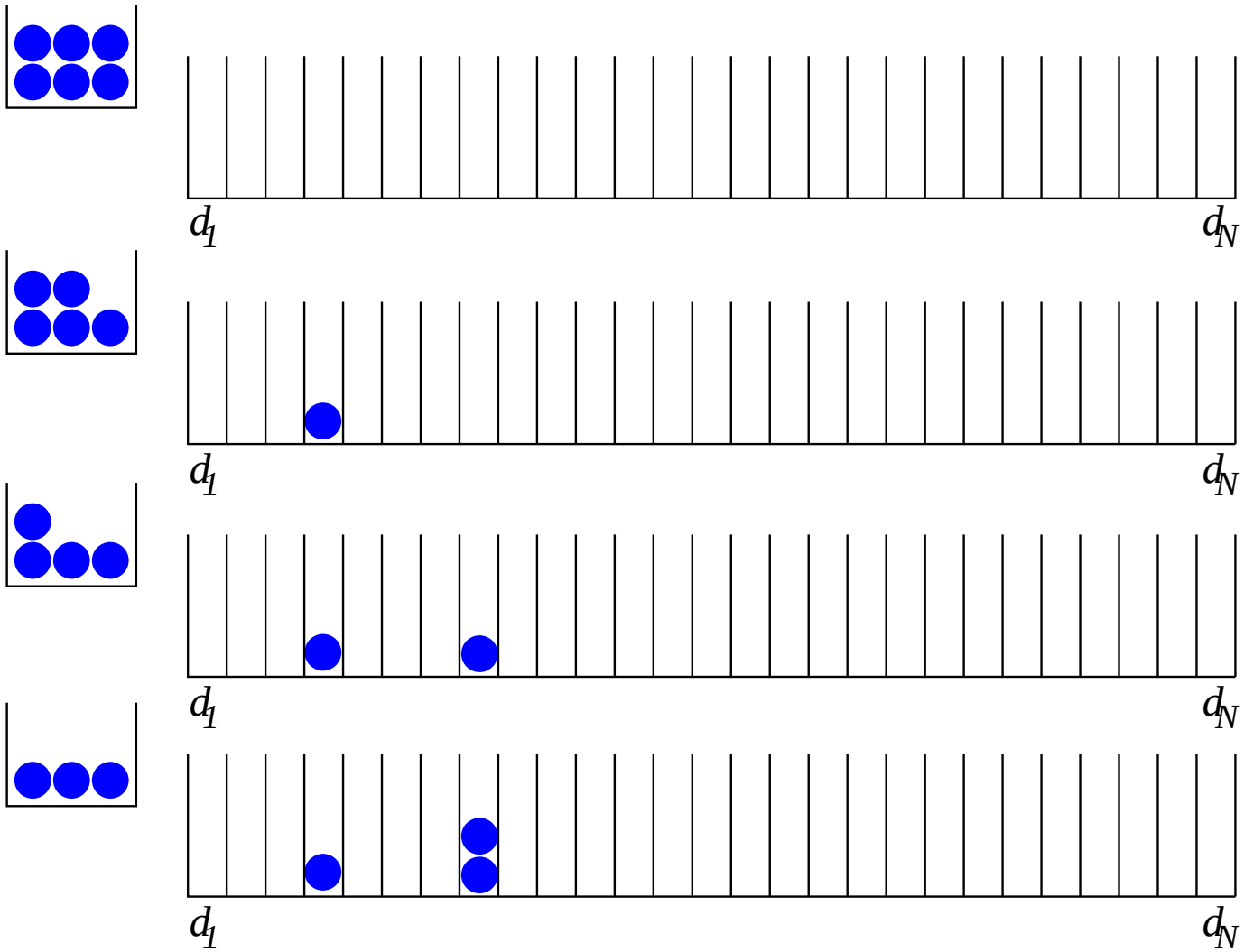
Binomial model



Binomial model



Binomial model



Binomial model

Basic event:

occurrence of a single term in a document

Bernoulli process with $p = \frac{1}{N}$,

N = # documents in the collection

F = # total occurrences of a term

tf = # occurrences of the term in a single doc.

Binomial model

Basic event:

occurrence of a single term in a document

Bernoulli process with $p = \frac{1}{N}$,

N = # documents in the collection

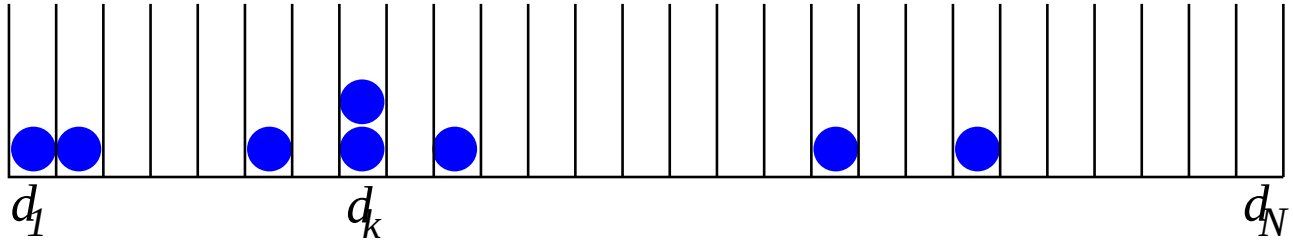
F = # total occurrences of a term

tf = # occurrences of the term in a single doc.

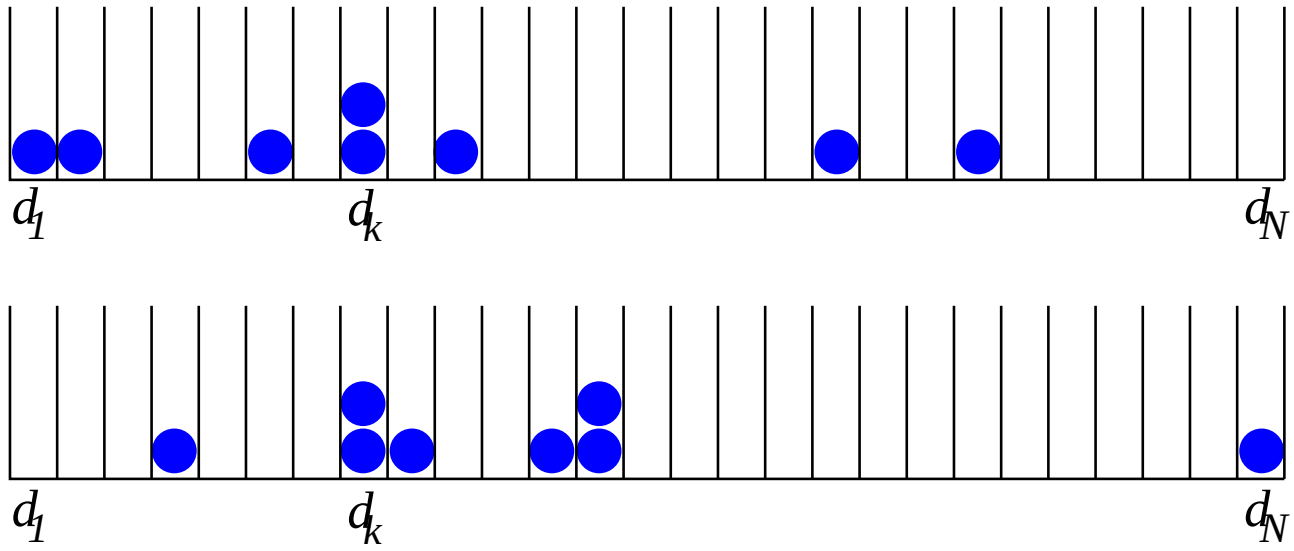
$$Prob_1(tf) = Prob_1 = B(N, F, tf) = \binom{F}{tf} p^{tf} q^{F-tf}$$

with $p = \frac{1}{N}$ and $q = \frac{N-1}{N}$.

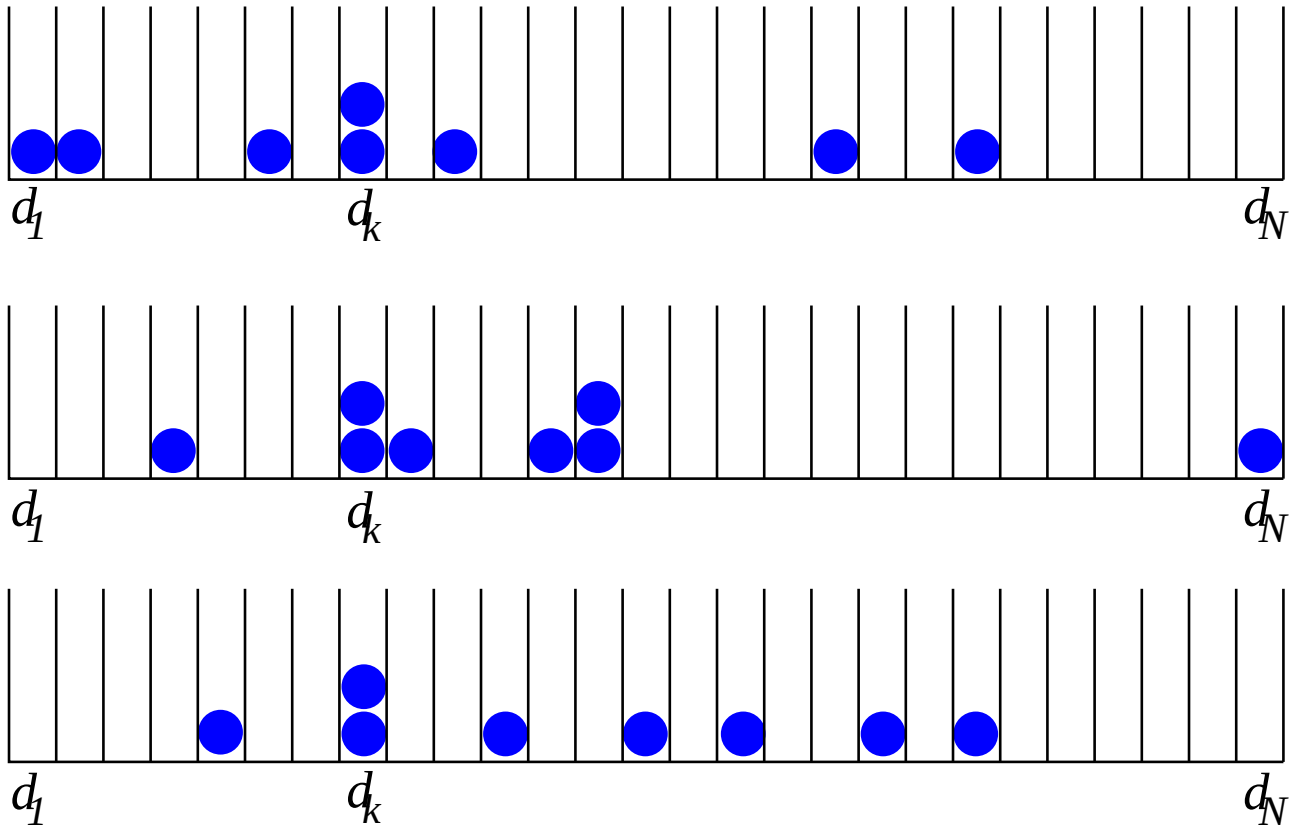
Bose-Einstein model



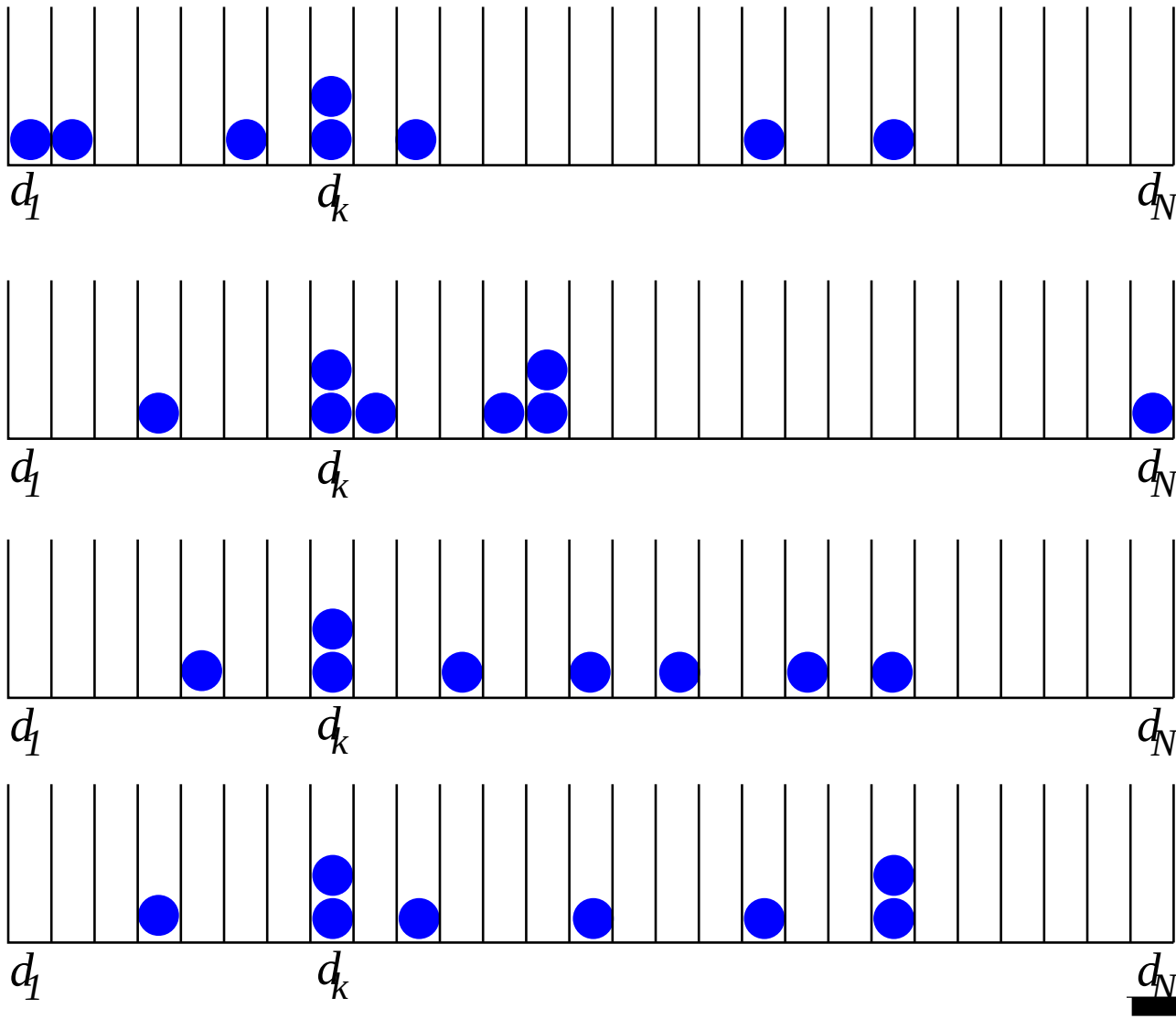
Bose-Einstein model



Bose-Einstein model



Bose-Einstein model



Bose-Einstein model

- Place randomly F tokens of a word in N docs.
  tf_k: term frequency in the k-th document
  event is described by occupancy numbers
  $tf_1, \ldots, tf_N$
  (each occupancy equiprobable)

Bose-Einstein model

- Place randomly F tokens of a word in N docs.
  tf_k: term frequency in the k-th document
  event is described by occupancy numbers
  $tf_1, \ldots, tf_N$
  (each occupancy equiprobable)
- Regard occupancy problem:
  all N-tuples satisfying the equation

$$tf_1 + \ldots + tf_k + \ldots + tf_N = F$$

  vs. all allocations where k-th document has term frequency tf

First normalization

Resizing the information content by the *aftereffect* of sampling:

- If a document contains at least one occurrence of the term:
  the probability of further occurrences is higher than probability of first occurrence

First normalization

Resizing the information content by the *aftereffect* of sampling:

- If a document contains at least one occurrence of the term:
  the probability of further occurrences is higher than probability of first occurrence
- Regard probability of additional occurrence:

$$Prob_2(tf) = p(tf + 1|tf, d)$$

First Normalization (2)

- Laplace:

$$Prob_2 = p(tf|tf-1,d) \sim \frac{tf}{tf+1}$$

First Normalization (2)

- Laplace:

$$Prob_2 = p(tf|tf-1,d) \sim \frac{tf}{tf+1}$$

- Bernoulli:
  regard probability that an additional token falls into the observed document

$$Prob_2 = \frac{B(n,F+1,tf+1)}{B(n,F,tf)}$$

Second normaliz.: document length

$l(d)$ - document length
$\rho(l)$ – term density function

Hypotheses:

Second normaliz.: document length

$l(d)$ - document length
$\rho(l)$ – term density function

Hypotheses:

H1 The distribution of a term is uniform in the document: $\rho(l) = c$

Second normaliz.: document length

$l(d)$ - document length
$\rho(l)$ – term density function

Hypotheses:

H1 The distribution of a term is uniform in the document: $\rho(l) = c$

H2 The term frequency density $\rho(l)$ is a decreasing function of the length: $\rho(l) = c/l$.

where c is determined by $tf = \int_0^{l(d)} \rho(l)dl$

Applying the DFR model

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

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

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

Applying the DFR model

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

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

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

  1. map tf to *normalized* term frequency (tfn)

Applying the DFR model

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

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

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

  1. map tf to *normalized* term frequency (tfn)
  2. use tfn for computing Inf_1 and Inf_2

Applying the DFR model

- apply document length normalisation (*second normalisation*) to "term frequency":
  $\rho(l) = c \cdot l^{\beta}$ (term density in document)
  $tfn = \int_{l(d)}^{l(d)+avl} \rho(l)dl$
  1. map tf to *normalized* term frequency (tfn)
  2. use tfn for computing Inf_1 and Inf_2
- apply linear retrieval function:
  $R(q,d) = \sum_{t \in q} qtf \cdot Inf_2(tfn) \cdot Inf_1(tfn)$

Applying DFR to XML documents

DFR: developed for atomic documents
XML: nested index nodes

Applying DFR to XML documents

DFR: developed for atomic documents
XML: nested index nodes

Dynamic vs. fixed document length

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

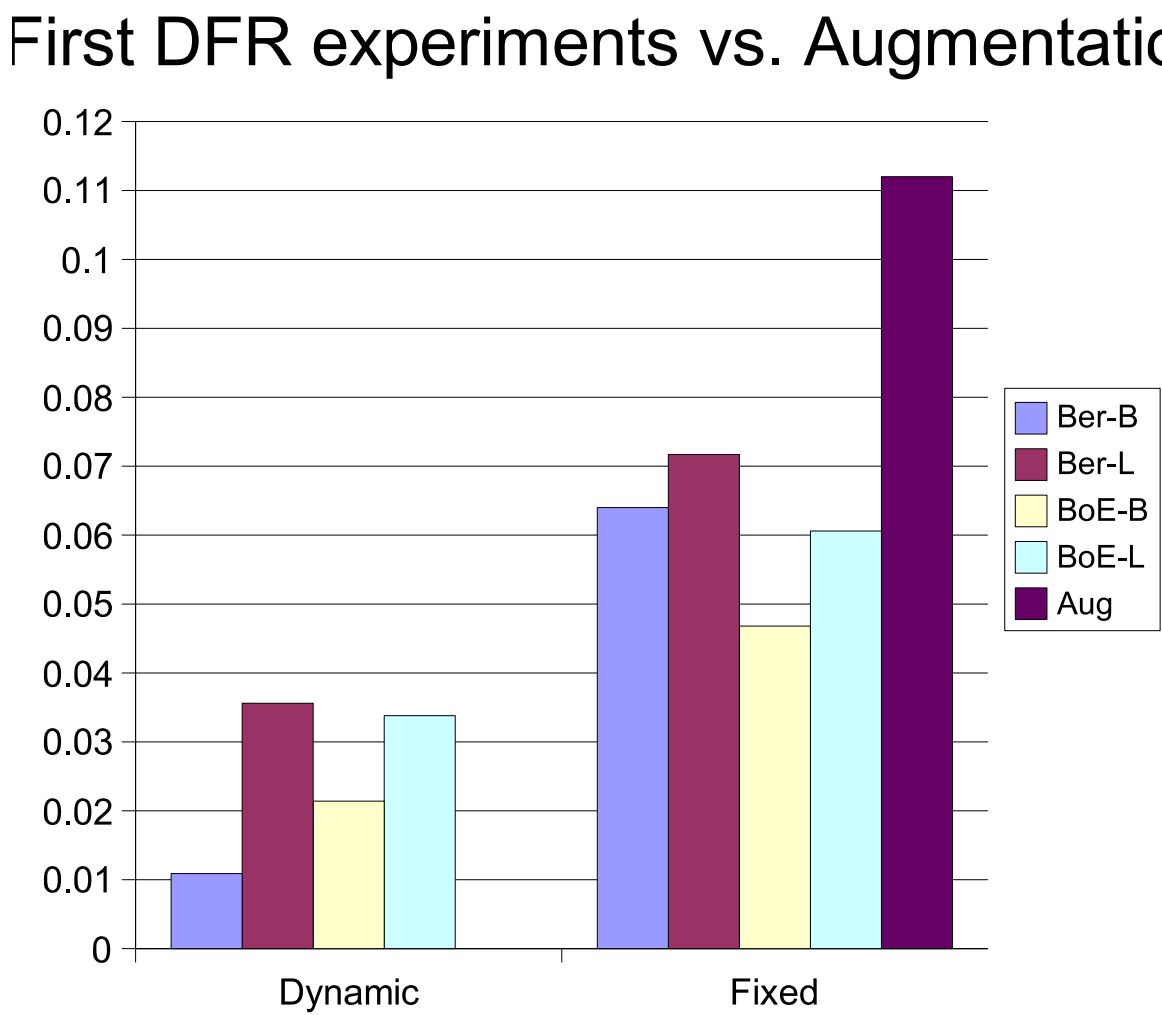
Applying DFR to XML documents

DFR: developed for atomic documents
XML: nested index nodes

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 nodes

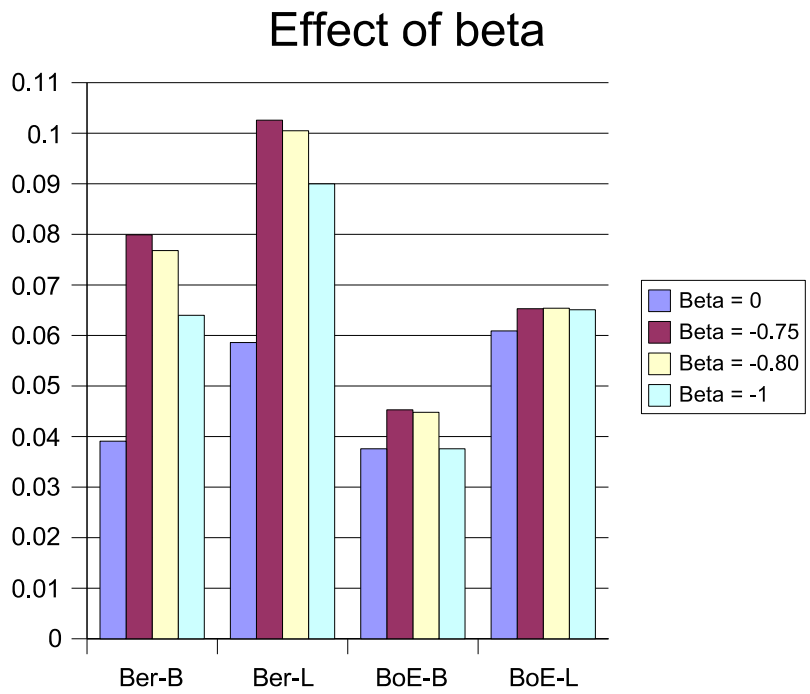
Experiments



Experiments

Document length normalization

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



$\longrightarrow$ retrieval quality still below augmentation approach!

Considering document structure

third normalisation:
effect of "different levels" $h(d)$ in the index node hierarchy
(root has level 1)

$$tfn' = tfn \cdot \frac{h(d)}{\alpha}$$

Considering document structure

third normalisation:
effect of "different levels" $h(d)$ in the index node hierarchy
(root has level 1)

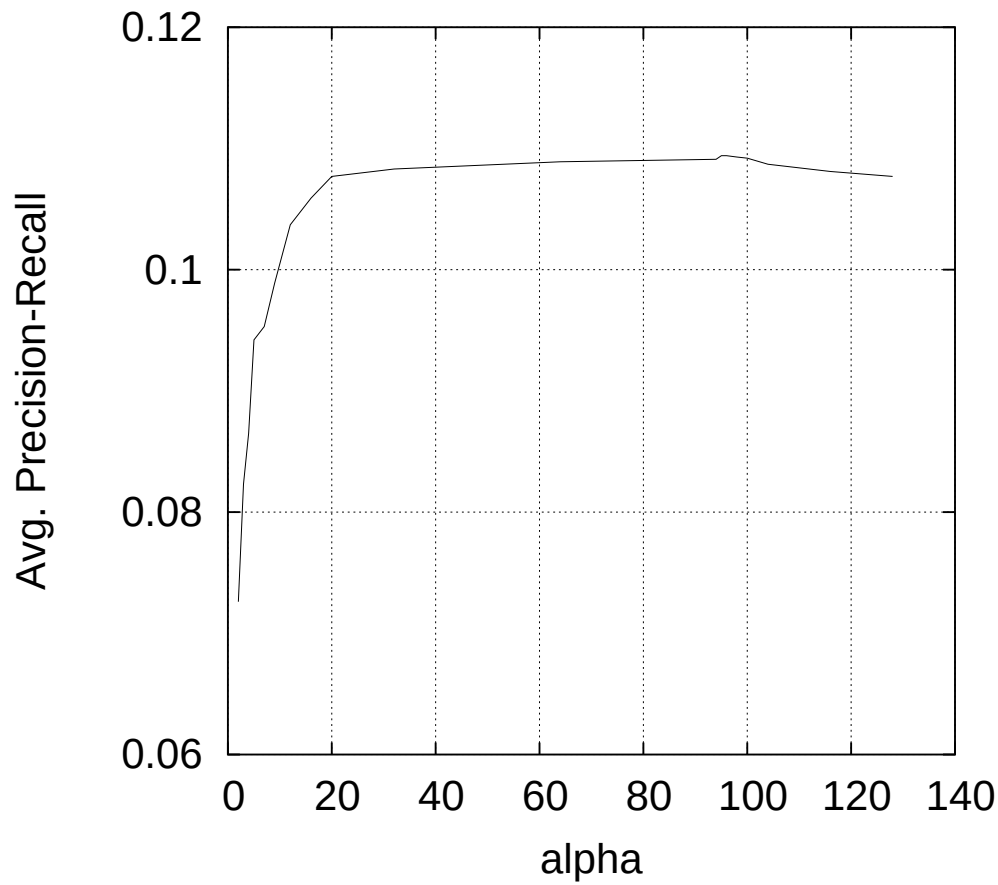
$$tfn' = tfn \cdot \frac{h(d)}{\alpha}$$

replacing tfn with tfn' for computing Inf_2:

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

Experiments

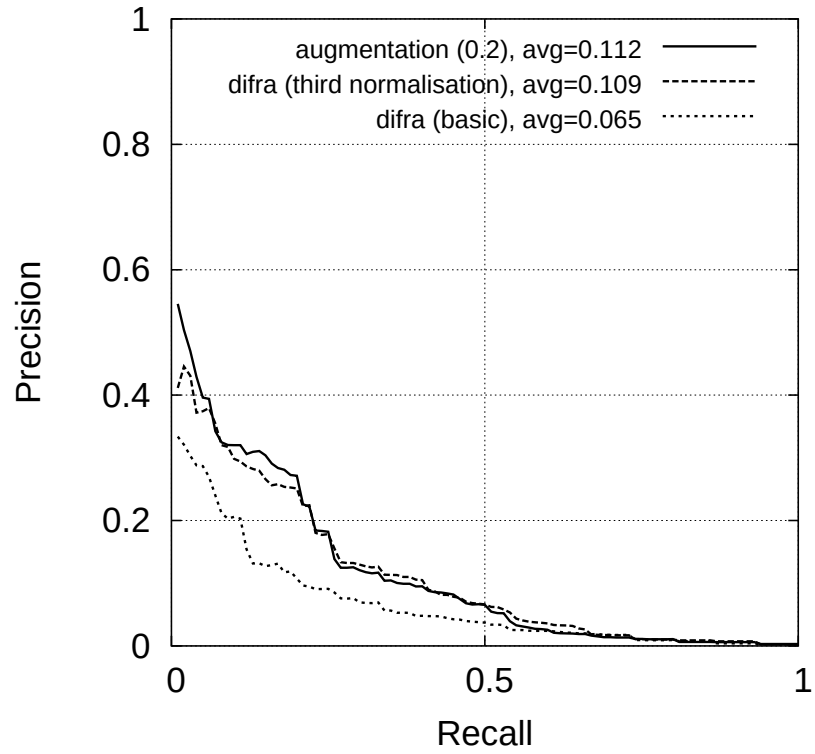
Results for the Bose-Einstein L Norm combination with the third function using various values of α:



Experiments

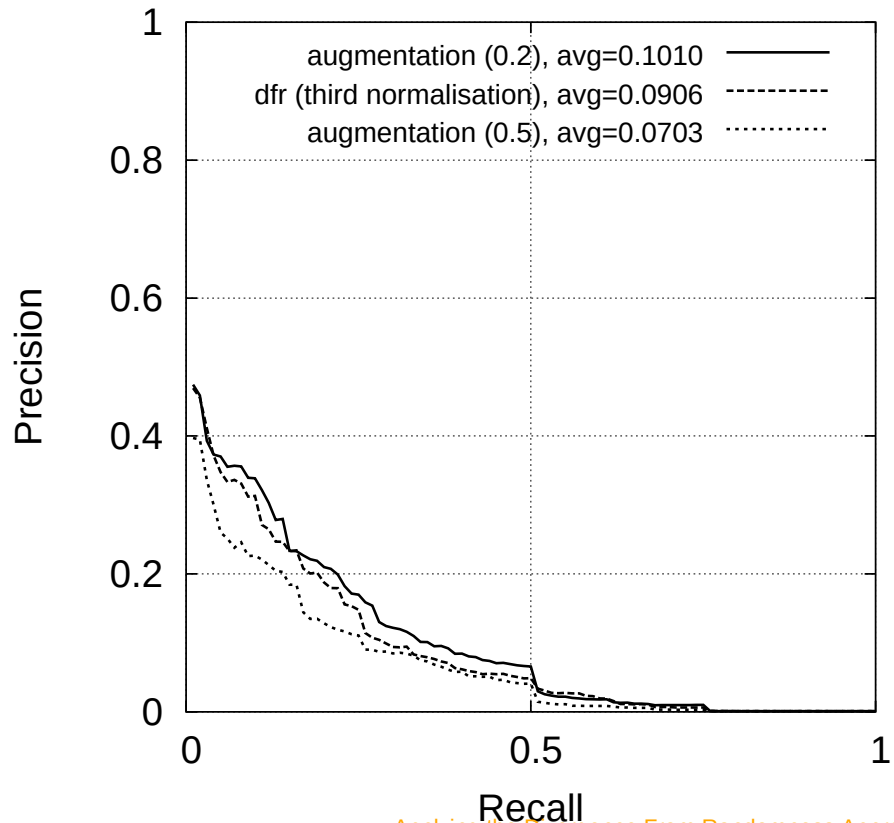
Effect of third normalization

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”

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”

INEX Website:

http://www.is.informatik.uni-duisburg.de/projects/inex/