Structured (XML) Document Retrieval

XML DB course - CC68K

Part IV

General Outline

Outline

Motivations for SDR

Fact

- *Traditional IR is about finding relevant documents to a user's information need, e.g. entire book.*
- *SDR allows users to retrieve document components that are more focussed to their information needs (ex. a chapter of a book instead of an entire book).*
- *The structure of documents is exploited to identify which document components to retrieve.*

Aims of SDR

Aim of SDR is to return

- document components of varying granularity (e.g. a book, a chapter, a section, a paragraph, a table, a figure, etc)
- relevant to the user's information need both with regards to content and structure

Fact

- *SDR involves the same tasks as in the conceptual model for IR*
- *but with different inner functionality (e.g. indexing, query formulation, retrieval, result presentation, feedback, ...)*

SDR Concepts

Like in IR

- Indexation of queries and documents into an adequate representation
- A score (RSV) between the query and the document representations
- Feedback can be used both to update document or query representations

But

- Document and possibly queries are structured
  Vector Space Models are not anymore adequate
- Feedback is (almost) *not* used

Many open questions

- What to search?
- How to express the search?
- How to search?
- How to know if what we found is relevant?

Evaluation of XML Retrieval: INEX

- Since 2002
- System-centred evaluation of effectiveness of XML retrieval approaches from ~40 institutions.
- This is a collaborative effort, where participants contribute to the development of the collection, i.e. the queries and the relevance assessments
- Similar methodology as for TREC is followed, but adapted to XML retrieval.
- More information: http://inex.is.informatik.uni-duisburg.de
  `http://inex.is.informatik.uni-duisburg.de`

Outline

Queries for SDR I

Content-only (CO) queries

- Standard IR queries but here we are retrieving document components
- “Santiago metro”

Structure-only queries

- Usually not that useful from an IR perspective
- “Paragraph containing a diagram next to a table”

Queries for SDR II

Content-and-structure (CAS) queries

- Put on constraints on which types of components are to be retrieved
  E.g. “Sections of an article in the Mercurio about congestion charges”
- E.g. “Articles that contain sections about congestion charges in Santiago, and that contain a picture of a hole in the road”

Documents

In general, any document can be considered structured according to one or more structure-type

- Linear order of words, sentences, paragraphs
- Hierarchy or logical structure of a book's chapters, sections
- Links (hyperlink), cross-references, citations
- Temporal and spatial relationships in multimedia documents

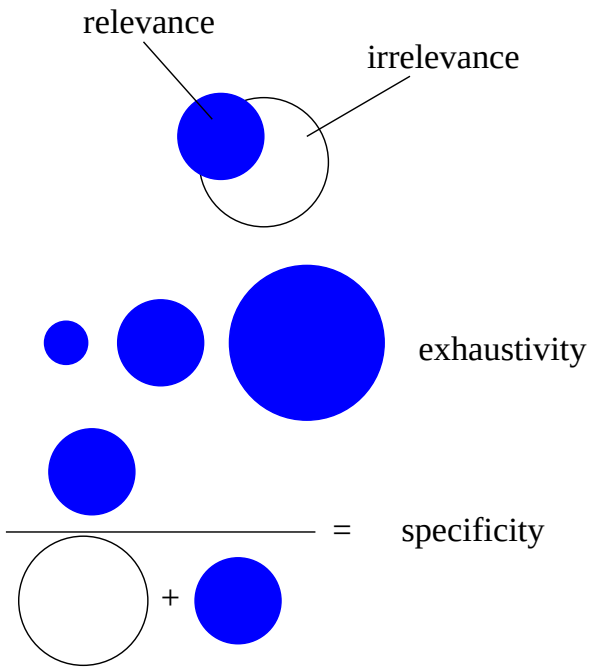
Fact

- *We only consider the logical structure*
- *Documents are in XML (e**X**tended **M**arkup **L**anguage)*
- *Query languages:*
  - *Keywords*
  - *XPath-like (XPath, XQL, XQuery)*
  - *Proximal nodes*

Relevance

Definition

- **Exhaustivity**: describes the extent to which the document component *discusses* the query.
- **Specificity**: describes the extent to which the document component *focuses* on the query.



Outline

What to search?

Other IR paradigms

- Document/Web retrieval
- Passage retrieval

in XML IR

The object is more complex, and so is the answer. In INEX, there are

Outline

XQuery Full-Text I

Facts

- Extension of XQuery
- A “score” keyword
- Full text operators: ftcontains, ftand, ftor, etc.

Example

XQuery FT

```
for $book in /book[.//author ftcontains "Marigold"]
let score $score := $book/title ftcontains "Web Site
Usability"
where $score > 0.8 order by $score descending return
$book/@number
```

XQuery Full-Text II

Example

XQuery FT Options

```
/book/title ftcontains "usability" case insensitive
diacritics insensitive without stemming without
thesaurus without stop words language "none" without
wildcards
```

NEXI I

Facts

- Narrowed Extented XPath for INEX
- Since 2003 in INEX
- Based on XPath but much simpler
- Only addition: an “about” clause

NEXI: examples I

CO query

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_topic topic_id="162" query_type="CO" ct_no="1">
<title> Text and Index Compression Algorithms </title>
<description>Any type of coding algorithm for text and index
compression</description>
<narrative>We have developed an information retrieval system
implementing compression techniques for indexing documents.  We are
interested in improving the compression rate of the system preserving a
fast access and decoding of the data.  A relevant document/component
should introduce new algorithms or compares the performance of existing
text-coding techniques for text and index compression.  A
document/component discussing the cost of text compression for text
coding and decoding is highly relevant.  Strategies for dictionary
compression are not relevant.</narrative>
<keywords>text compression, text coding, index compression
algorithm</keywords>
</inex_topic>
```

NEXI: examples II

CAS query

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE inex_topic SYSTEM "topic.dtd">
<inex_topic topic_id="128" query_type="CAS" ct_no="22">
<title>//article[about(., intelligent transport systems)]//sec[about(.,
on-board route planning navigation system for automobiles)]</title>
<description>Find discussions about on-board route planning or
navigation systems which are in publications about intelligent
transport systems for automobiles.</description>
<narrative>I'm interested in information about on board route planning
or navigation systems for automobiles.  Relevant elements discuss
either a requirement analysis or a concrete implementation of such a
system.  Elements about navigation or route planning systems that
cannot be accessed within the automobile will not be considered
relevant.  Systems of other phenomena than automobiles will also not be
judged relevant.</narrative>
<keywords>in-vehicle systems, vehicle intelligence, vehicle information
systems, traffic information services, vehicle-mounted
equipment</keywords>
</inex_topic>
```

Outline

AdHoc (since 2002) I

Ad hoc retrieval is the task of searching a *static collection* for relevant units given an *information need*.

Content only (CO)

- Focussed
- Thorough
- Fetch & Browse

Content with structural hints (CO+S)

- Same tasks but using the NEXI query

Content And Structure (CAS)

AdHoc (since 2002) II

- NEXI queries are composed of support + target
  `//article[about(.,santiago)]//p[about(.,cinema)]`
- 4 subtasks: Vague/Strict (support) Vague/Strict (target) CAS (VVCAS, SVCAS, VSCAS, SSCAS)

Relevance Feedback (since 2004) I

Motivations

- It is difficult to ask directly “the perfect query”
- User interaction refines the query
- A lot of existing work in IR: typically, the bag of word representation of the query is updated
- Nothing (almost) exists in XML IR: in 2004, one adaptation of the Rocchio algorithm (update of query weigths):

$$Q' = \alpha Q + \frac{\beta}{n_1} \sum_{i=1}^{n_1} R_i - \frac{\gamma}{n_2} \sum_{i=1}^{n_2} NR_i$$

Natural Language (since 2004)

"The ultimate goal is to design and build software that will analyse, understand, and generate results in response to queries that humans express naturally"

Two subtasks

NLQ2NEXI Translation to NEXI

NLQ Answering directly (might be NLQ2NEXI + use of a SDR system)

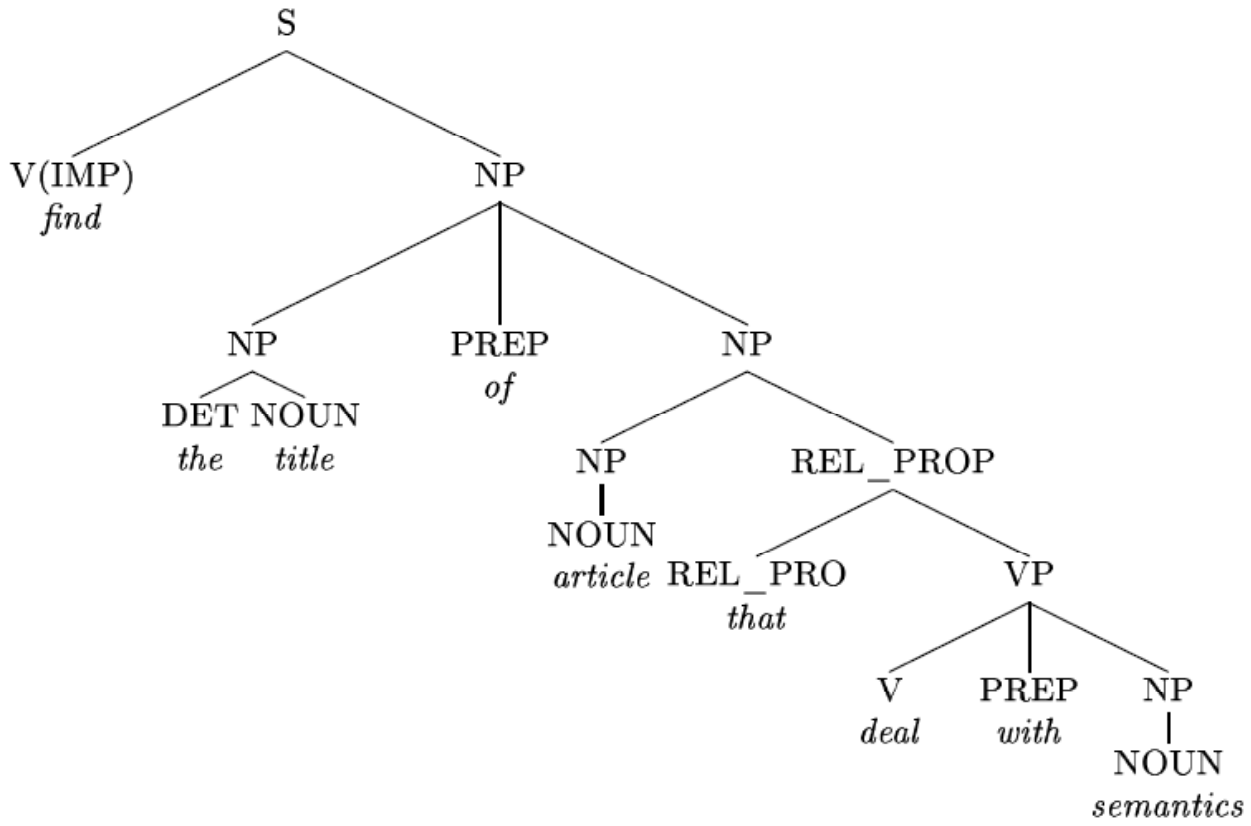
NLQ2NEXI: example I

Methodology

- Part of speech (POS) tagging (assign a category to each word)
- Grammar over the tags (NP -> DET? NOUN) gives us a tree
- The tree is then converted into a set using the Discourse Representation Theory (DRT)
- And a lot of (specific) rewriting, rule reducing (axioms)
- The logical formula is converted into NEXI

```
Find the title of articles that deal with semantics
V    DET NOUN PREP NOUN     REL  V    PREP  NOUN
```

NLQ2NEXI: example II



Using the DRT:
event(e1,find)
object(e1,x)
title(x)
event(e2,deal)
agent(e2,y)
article(y)
of(x,y)
with(e2,z)
semantics(z)

that gives, with some inferences like “article(y) semantics(z) event(e2,y) event(e2,deal) with(e2,z)” implies “about(y,z)”:
//article[about(.,semantics)]//title

Heterogeneous (since 2004)

Problem

- In uncontrolled environments, the structure of documents is variable and unknown
  - It is impossible to predefine the retrieval units using tag names.
  - For CAS, how to translate the queries queries?

Methodology

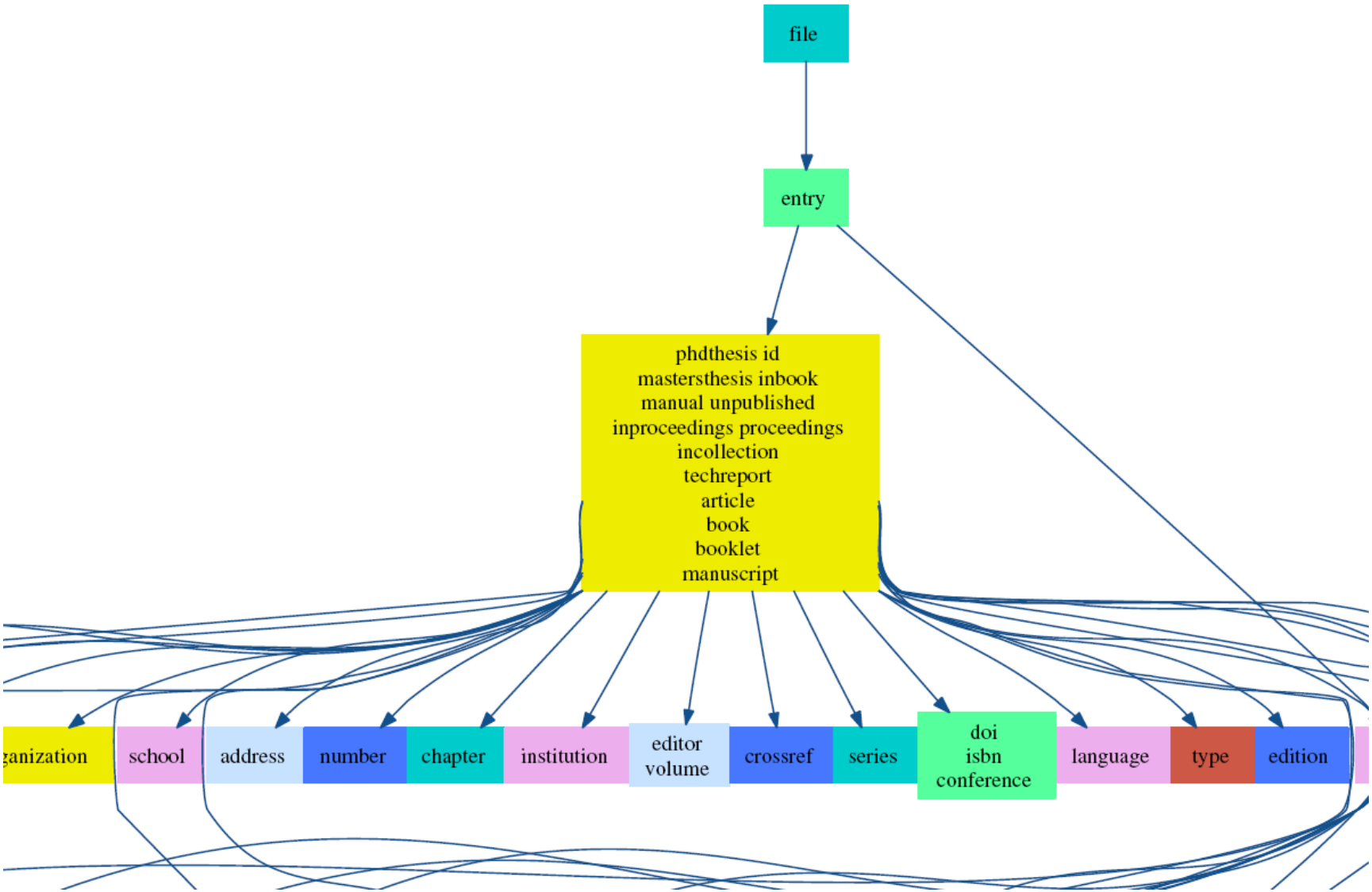
- Same as for the adhoc but...
- with an heterogeneous test collection

Heterogeneous (example)

Example: Extirp

- Minimum and maximum size of an element
- Ratio number of text/element nodes to distinguish between text-oriented and data-oriented elements.

Hetereogenous (Visualisation: XSum)

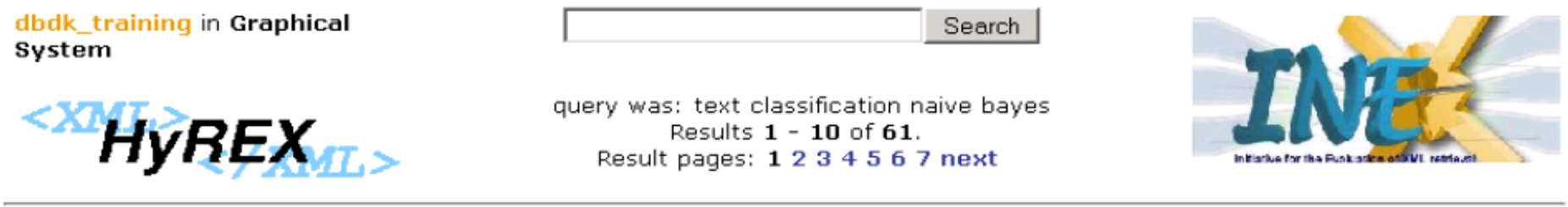


Interactive (since 2004) I

Motivations

- Investigate the behaviour of users when interacting with components of XML documents
- To investigate and develop approaches for XML retrieval which are effective in user-based environments.
- Very important for the evaluation of SDR systems

Interactive (since 2004) II



Search Results

1: **Scalable Feature Mining for Sequential Data**
Neal Lesh Mitsubishi Electric Research Lab Mohammed J. Zaki Rensselaer Polytechnic Institute Mitsunori Ogihara University of Rochester

2: **Probability and Agents**
Marco G. Valtorta University of South Carolina mgv@cse.sc.edu Michael N. Huhns University of South Carolina huhns@sc.edu

3: **Combining Image Compression and Classification Using Vector Quantization**
Karen L. Oehler Member IEEE Robert M. Gray Fellow IEEE

4: 7 hit(s) **Text-Learning and Related Intelligent Agents: A Survey**
Dunja Mladenic J. Stefan Institute

5: **Detecting Faces in Images: A Survey**
Ming-Hsuan Yang Member IEEE David J. Kriegman Senior Member IEEE Narendra Ahuja Fellow IEEE

Interactive (since 2004) III

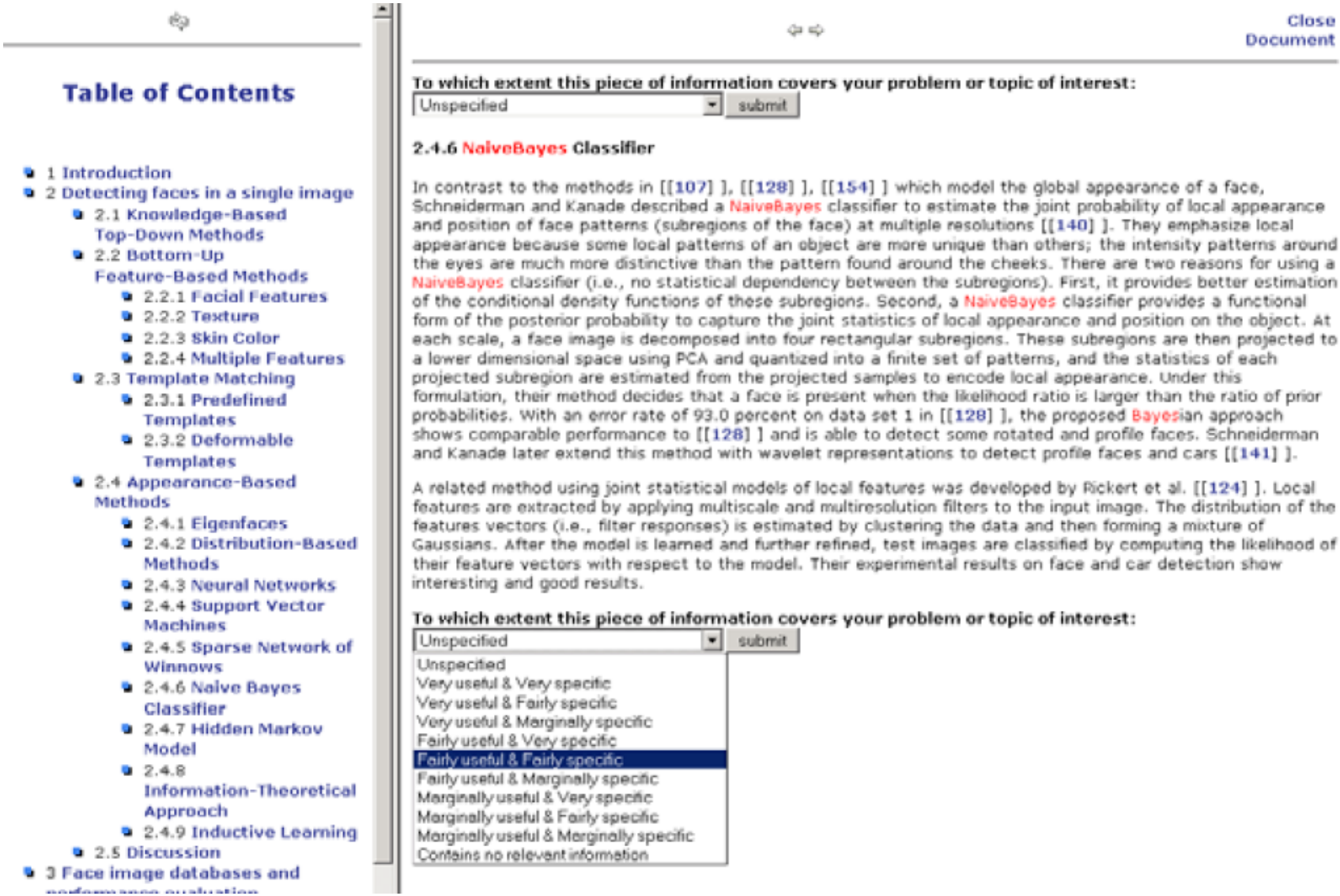


Table of Contents

- 1 Introduction
- 2 Detecting faces in a single image
  - 2.1 Knowledge-Based Top-Down Methods
  - 2.2 Bottom-Up Feature-Based Methods
    - 2.2.1 Facial Features
    - 2.2.2 Texture
    - 2.2.3 Skin Color
    - 2.2.4 Multiple Features
  - 2.3 Template Matching
    - 2.3.1 Predefined Templates
    - 2.3.2 Deformable Templates
  - 2.4 Appearance-Based Methods
    - 2.4.1 Eigenfaces
    - 2.4.2 Distribution-Based Methods
    - 2.4.3 Neural Networks
    - 2.4.4 Support Vector Machines
    - 2.4.5 Sparse Network of Winnows
    - 2.4.6 Naive Bayes Classifier
    - 2.4.7 Hidden Markov Model
    - 2.4.8 Information-Theoretical Approach
    - 2.4.9 Inductive Learning
  - 2.5 Discussion
- 3 Face image databases and

Close Document

To which extent this piece of information covers your problem or topic of interest:
Unspecified submit

2.4.6 NaiveBayes Classifier

In contrast to the methods in [[107]], [[128]], [[154]] which model the global appearance of a face, Schneiderman and Kanade described a NaiveBayes classifier to estimate the joint probability of local appearance and position of face patterns (subregions of the face) at multiple resolutions [[140]]. They emphasize local appearance because some local patterns of an object are more unique than others; the intensity patterns around the eyes are much more distinctive than the pattern found around the cheeks. There are two reasons for using a NaiveBayes classifier (i.e., no statistical dependency between the subregions). First, it provides better estimation of the conditional density functions of these subregions. Second, a NaiveBayes classifier provides a functional form of the posterior probability to capture the joint statistics of local appearance and position on the object. At each scale, a face image is decomposed into four rectangular subregions. These subregions are then projected to a lower dimensional space using PCA and quantized into a finite set of patterns, and the statistics of each projected subregion are estimated from the projected samples to encode local appearance. Under this formulation, their method decides that a face is present when the likelihood ratio is larger than the ratio of prior probabilities. With an error rate of 93.0 percent on data set 1 in [[128]], the proposed Bayesian approach shows comparable performance to [[128]] and is able to detect some rotated and profile faces. Schneiderman and Kanade later extend this method with wavelet representations to detect profile faces and cars [[141]].

A related method using joint statistical models of local features was developed by Rickert et al. [[124]]. Local features are extracted by applying multiscale and multiresolution filters to the input image. The distribution of the features vectors (i.e., filter responses) is estimated by clustering the data and then forming a mixture of Gaussians. After the model is learned and further refined, test images are classified by computing the likelihood of their feature vectors with respect to the model. Their experimental results on face and car detection show interesting and good results.

To which extent this piece of information covers your problem or topic of interest:
Unspecified submit

Unspecified
Very useful & Very specific
Very useful & Fairly specific
Very useful & Marginally specific
Fairly useful & Very specific
Fairly useful & Fairly specific
Fairly useful & Marginally specific
Marginally useful & Very specific
Marginally useful & Fairly specific
Marginally useful & Marginally specific
Contains no relevant information

Document mining (since 2005)

1. Clustering and classification have been used in many IR applications. Clustering is particularly important when dealing with heterogeneous structural collections (different document DTDs for example). One important issue is how to handle the heterogeneity which characterises XML collections gathered from different sources.
2. Is the structural information contained in the document trees relevant for organizing collections into homogeneous clusters and how to use structure information for organizing document collections ?
3. What are the complementarities between the information brought by the logical structure and the textual content for classification and clustering ?
4. Problems of multiple labels: labels from different DTDs might be different whereas the documents organisation is similar. How to handle this heterogeneity problem?

Outline

Models

Term Weight Propagation

- Term Selection
- Aggregation
  $\rightarrow$maximum, augmentation, LM, ...

Score Propagation

- Extension of boolean models (p-norm)
- Extension of VSM
- Bayesian networks

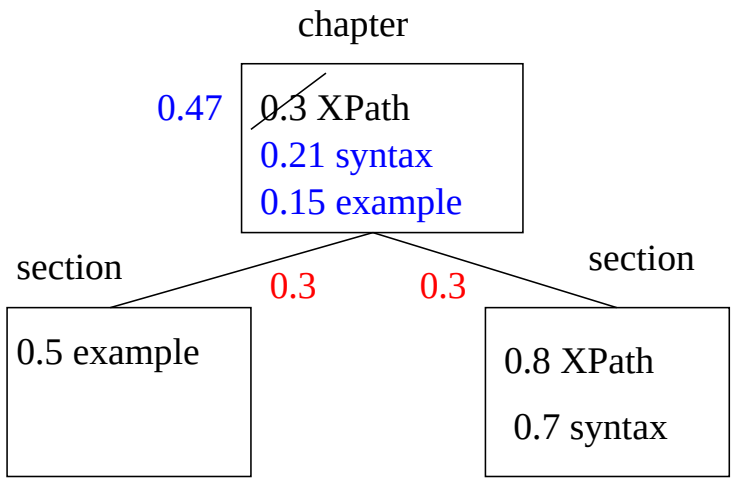
”Moving” Corpus

- The elements are grouped in e-collections
- Statitistics are computed on these e-collections

Augmentation (term weight propagation)

Principle

- Some nodes are **elementary** elements (answers)
- Aggregate weights of children (begining with **elementary** elements)

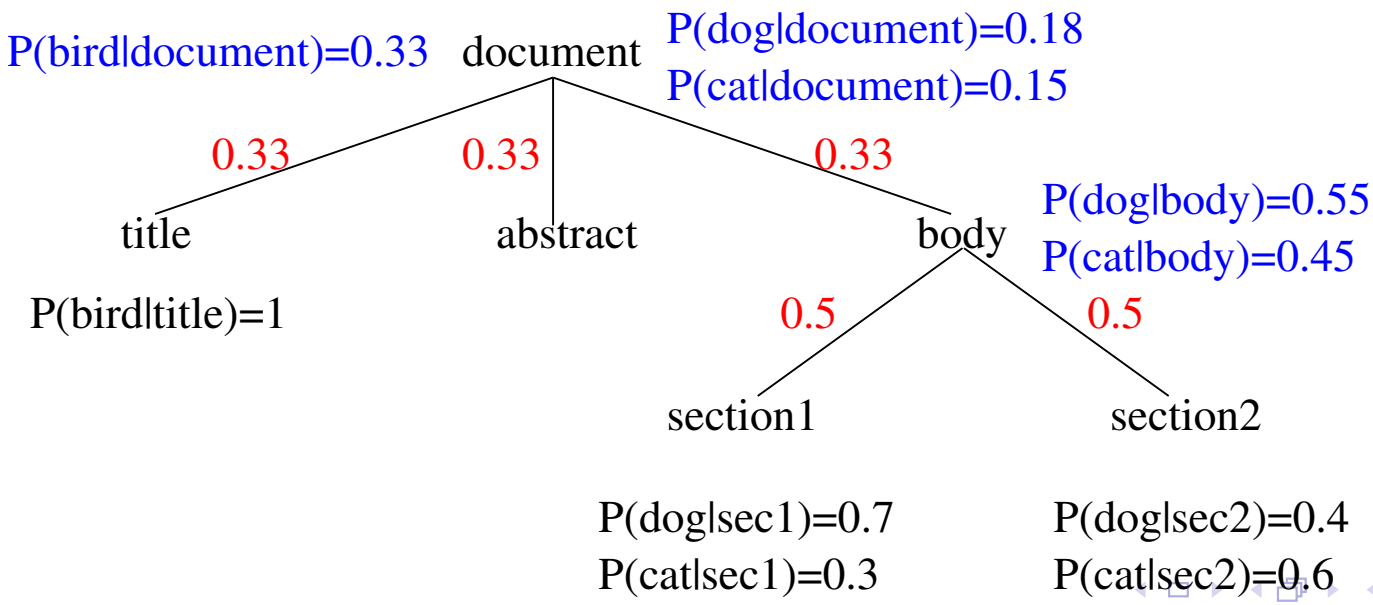


Language Models (term weight propagation)

$$P(Q|\Theta_E) = \prod_{\omega \in \{q_1,\ldots,q_n\}} P(\omega|\theta_E)$$

Estimating $P(\omega|\theta_E)$

- Mixture of element- and collection-specific estimates
- Then, mixture of language models



Entropy and term selection (term weight propagation)

- The idea is that a term appears only once in any path
- $$\text{weigth of t} = \log(1 + tf) \times \begin{cases} \log \frac{\text{\# of documents}}{\text{\# of documents with t}} & \text{if leaf} \\ \frac{-\sum_{child} tf(child) \times \log \frac{tf(child)}{tf}}{-tf \times \log \frac{1}{\text{\# of children}}} & \text{otherwise} \end{cases}$$
- Then we try to “pull up” the terms. Condition

$$\text{weight} \geq \text{child average weight} + \text{child weight variance}$$

Score propagation

Many different methods!

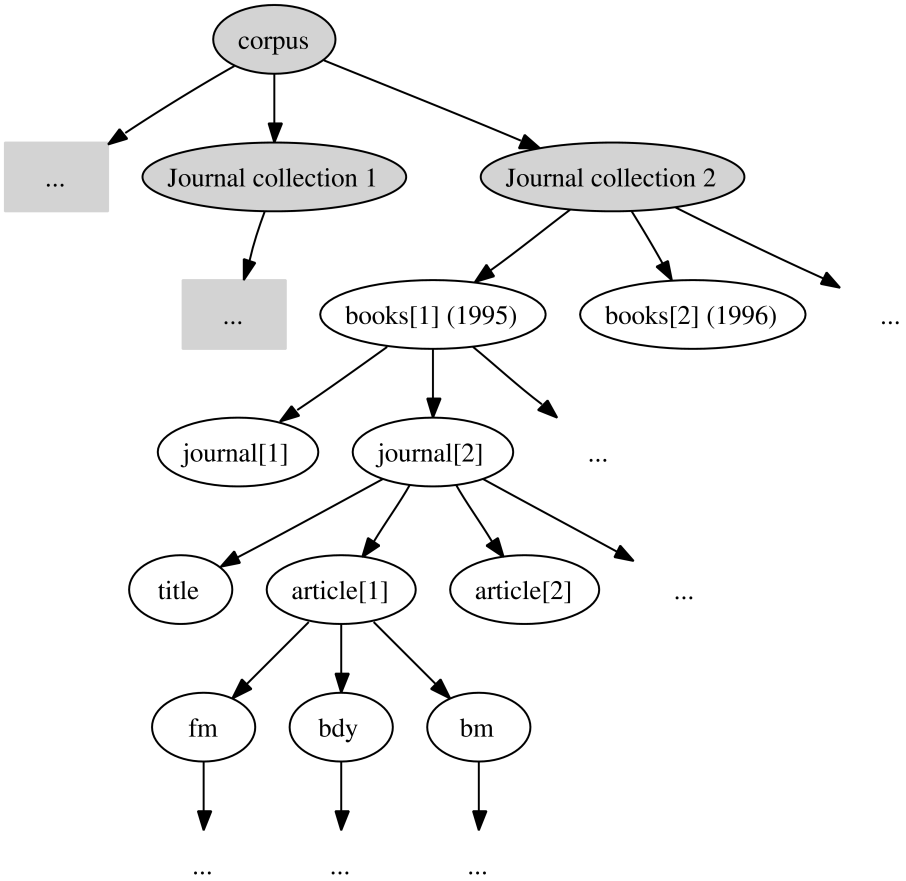
- Aggregation of children score

$$RSV(q, e) = \sum_{child} k_{child} \mathrm{RSV}(q, child)$$

- Aggregation of descendants

$$RSV(q, e) = \sum_{desc} RSV(q, desc) \prod_{x \text{ in path}(e,desc)} weight(x)$$

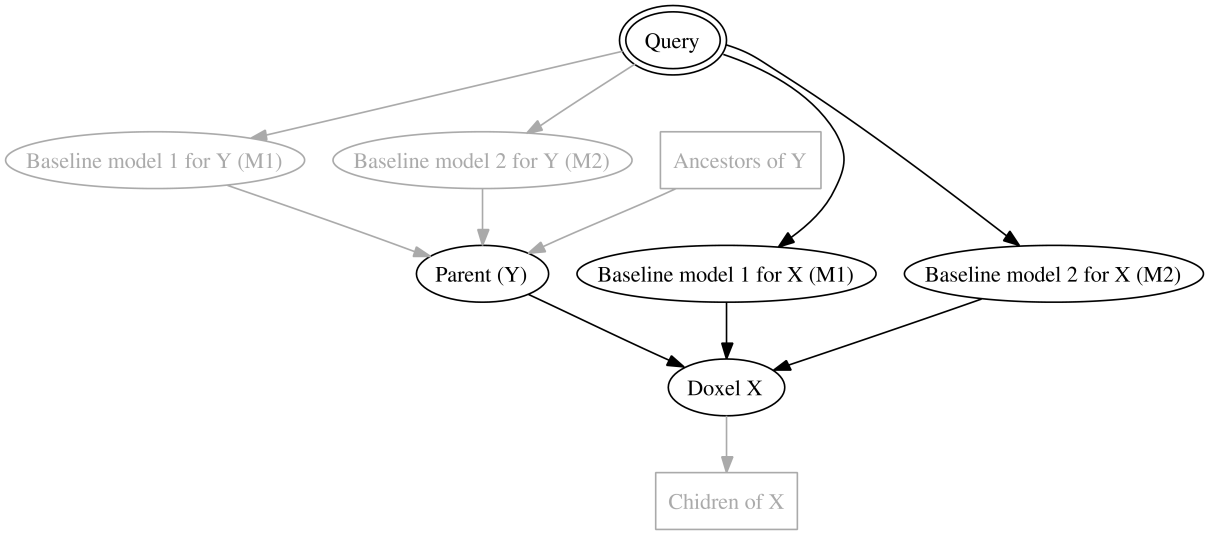
Bayesian Networks: Structure (score propagation)



Components

- Fixed structure = corpus structure
- Parameters
- Baseline models

Bayesian Networks: Local Inference (score propagation)



Variables

- Query: vector of frequencies
- Baseline models: binary {relevant, not relevant}
- Element: {not relevant, too big, SDR-relevant}

Bayesian Networks: learning (score propagation)

What?

- Parameters ($\Longrightarrow$ CPT)
- Adaptation to specific corpora/query types

How?

- Set of queries + associated assessments
- Algorithms
  - Expectation/Maximisation (EM)
  - Cross-Entropy with gradient ascent
  - Order-based criterions

Outline

Models

Fragment Queries

Query Fragment of an XML document

Search Match of the two representations

XPath / Algebra based

Query An XPath-like expression

Search

1. Transformation into an algebraic expression
2. An event is associated to each element
3. Score = probability of the event

Fragments: JuruXML

A modified VSM

$$\text{RSV}(q,d) = \frac{1}{|q||d|} \sum_{(t_i,c_i^q)\in q} \sum_{(t_j,c_{i_j}^d)\in q} \omega^q(t_i,c_i^q)\cdot\omega^d(t_i,c_i^d)\cdot cr(c_i^q,c_i^d)$$

noting $c_\bullet^\bullet$ a path and $t_\bullet$ a term.

Example

$$cr(c_1,c_2) = \begin{cases} \frac{1+\text{length}(c_1)}{1+\text{length}(c_2)} & \text{if } c_1 \text{ is a subsequence of } c_2 \\ 0 & \text{otherwise} \end{cases}$$

Fragments: Language Models / Dynamic TF-IDF

Idea

- Take into account the structural conditions
- The term weight depends on the element types

TF-IDF The collection is defined by elements sharing the same “path”

LM Element-specific LM

Algebra: XIRQL / S-BN

Extension of XPath

- Weighting and ranking
- Data types with vague predicates

Principle

- A query is transformed into an event for each retrievable element
- The probability of the event is the score of the element

XIRQL An event $\sim$ a term occurence

S-BN Using a BN network (event = relevance to a query composed of keywords)

Algebra: example

```
//image[../p[about(.,"cat pictures")]]
```

$$\downarrow$$

$$\text{child}(\mathit{rel}(\text{cat picture}) \cap \mathit{label}(\text{p})) \cap \mathit{label}(\text{image}) \cap \text{desc}(d)$$

$$\downarrow$$

$$a \text{ is relevant} \equiv a \in \mathit{label}(\text{image}) \wedge \bigvee_{b \in pa(a)} (b \in \mathit{rel}(q_1) \wedge b \in \mathit{label}(\text{p}) \wedge b \in \text{desc}(d))$$

Problematic

GOAL
Develop collections to evaluate systems

But... contrary to IR: elements are nested

- Binary relevance scale is not enough
  $\Longrightarrow$ A new scale
- Elements relevance are interdependents
  $\Longrightarrow$ Constraints on assessments
- Standard metrics are not adapted
  $\Longrightarrow$ new metrics

INEX collection

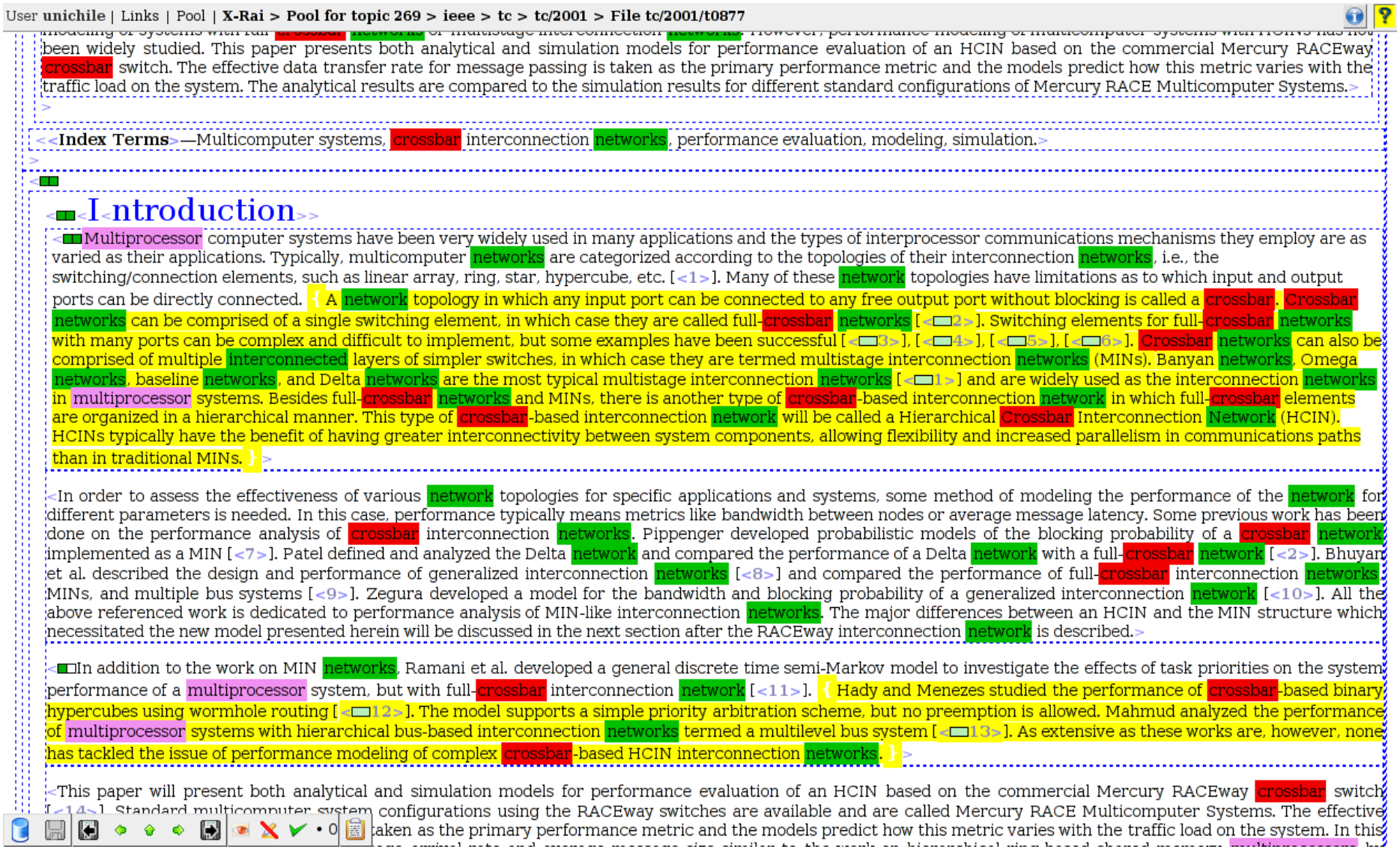
Fact

- *Documents (~500MB), 12,107 articles in XML format from the IEEE Computer Society*
- *Topics:*

  2002 *30 CO and 30 CAS*
  2003 *32 CO and 32 CAS*
  2004 *33 CO and 24 CAS (for now)*

Outline

The Interface



Active Rules

Ensure the *consistency*

- Elements within a document are *not* independant
- Help the user to assess
- Consistency of assessments

Some rules

- $\sum_i E_{y_i} \geq E_x \geq \max_i (E_{y_i})$

Outline

XML-IR Metrics

The proposed metrics

Recall-precision like "Quantised" precision/recall, "Norbert Gövert" (NG) precision/recall

Cumulated Gain eXtended Cumulated Gain (xCG, nxCG, EP-GR)

User Model Tolerance To Irrelevance (T2I), Precision Recall with User Modeling (GR, PRUM, EPRUM)

How to know if a metric is good?

What

- Fidelity: do they measure what we want?
- Stability: how sensitive are they?

How

- **USER EXPERIMENTS**
- Stereotypical runs
- Double asssessed topics
- Varying the number of topics

Recall-precision

User Model

- The user consults every element in list order
- (S)he is “happy” with every kind of relevant information, even
  - if (s)he has already seen *the same content*
  - **if (s)he has already seen it entirely or partly (nesting)**

$$P(\text{Relevant}|\text{Retrieved}, \text{Wanted} = r)$$

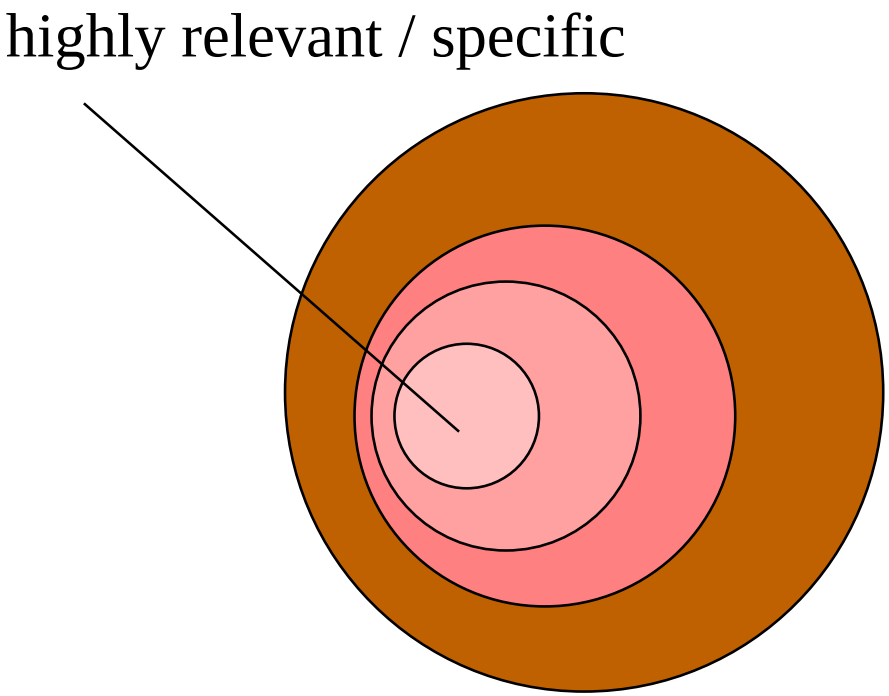
Quantisation

$$f_{strict}(e,s) = \begin{cases} 1 & \text{if } (e,s) = (3,3) \\ 0 & \text{otherwise} \end{cases}$$

$$f_{gen}(e,s) = \begin{cases} 1 & \text{if } (e,s) = (3,3) \\ 0.75 & \text{if } (e,s) \in \{(2,3),(3,2),(3,1)\} \\ 0.5 & \text{if } (e,s) \in \{(1,3),(2,2),(2,1)\} \\ 0.25 & \text{if } (e,s) \in \{(1,2),(1,1)\} \\ 0 & \text{otherwise} \end{cases}$$

(...) and 5 others one in INEX 2004!

The “Recall Base”



Recall-Precision NG

User Model: classical model + ...

- No more relevance for already retrieved elements

$$\text{recall} = \frac{\sum_e \text{rel}(e)(1 - \frac{\text{size(seen part of } e)}{\text{size}(e)})}{\sum_e \text{rel}(e)}$$

$$\text{precision} = \frac{\sum_e \text{spe}(e)(1 - \frac{\text{size(seen part of } e)}{\text{size}(e)})}{\sum_e (1 - \frac{\text{size(seen part of } e)}{\text{size}(e)})}$$

Problems

- The measure is very instable
- Theoritical fundations?

xCG (official metric for 2005)

$$\mathrm{xCG}(n) = \sum_{k=1}^{n} rv(x_k)$$

where x_k is the k^{th} element of the list consulted by the user. The relevance value, rv, is then defined as:

$$rv(x) = \begin{cases} (1-\alpha)\, q(x) & \text{ancestor seen} \\ \alpha \sum\limits_{y \text{ child of } x} \frac{rv(y) \times \mathrm{size}(y)}{\mathrm{size}(x)} + (1-\alpha)\, q(x) & \text{descendants seen} \\ q(x) & \text{otherwise (no overlap)} \end{cases}$$

where $q(x)$ is a quantisation of the assessment of element x. In order to prevent to reward a set of descendant elements more than their ideal ancestor, the following constraint is used for any ideal element y:

$$\sum_{x \in S} rv(x) \leq rv(y)$$

PRUM metrics

User Model

- R/P model
- Stochastic user behaviour
  $\Longrightarrow$ the user can navigate in the document
  $\Longrightarrow$ the user may find an element relevant or not
- Relevant Information = Highly Specific elements only

Limitations

- Some parameters have to be validated
- Complexity

Tolerance to Irrelevance (T2I)

User Model

- R/P model
- The user reads sequentially and stops after a certain amount of irrelevant information

Limitations

- (No implementation)
- Some theorical and pratical problems have to be solved
- Some parameters have to be validated

Outline

General references I

- http://inex.is.informatik.uni-duisburg.de:2004/
- SIGIR 2002 and 2004 workshops on XML retrieval
- Special issue of JASIST on XML and Information Retrieval, Volume 56(2), 2002.
- Proceedings of the INEX Worshop (2002, 2003 and 2004)
- Robert Luk, H.V. Leong, Tharam Dillon, Alvin Chan, W. Bruce Croft, and James Allan. A survey in indexing and searching XML documents. *JASIS*, 6(53) 415–437, March 2002.

Models I

- N. Fuhr, Großjohann. *XIRQL: An XML query language based on information retrieval concepts.* ACM Transactions on Information Systems (TOIS), 22(2), 313-356, 2004.
- Chinenyanga, T. & Kushmerick, N. (2001) An expressive and efficient language for XML information retrieval. J. American Society for Information Science & Technology (special issue on XML and Information Retrieval).
- Y. Mass, M. Mandelbrod, E. Amitay, Maarek Y., and A. So er. JuruXML - an XML retrieval system at INEX 02, INEX 2003 proceedings, pages 73-90.
- P. Ogilvie and J. Callan. Language models and structure document retrieval (In INEX 2003 Proceedings)

Models II


T. Grabs and H.-J. Schek. Flexible information retrieval from XML with PowerDBXML (In INEX 2003 proceedings)


Benjamin Piwowarski and Patrick Gallinari. A bayesian network for XML information retrieval: Searching and learning with the INEX collection. *Information Retrieval*, December 2004.

Evaluation I

- Report of the INEX'03 metrics working group", pp. 184-190 of the INEX'03 proceedings.
- Benjamin Piwowarski and Mounia Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the Thirteenth Conference on Information and Knowledge Management (CIKM 2004)*, Washington D.C., U.S.A., November 2004.
- G. Kazai, M. Lalmas, and Arjen P. Vries. The Overlap Problem in Content-Oriented XML Retrieval. In *SIGIR 2004*.
- N. Gövert, G. Kazai, N. Fuhr, and M. Lalmas. Evaluating the effectiveness of content-oriented XML retrieval University of Dortmund, Computer Science, 2003.

Evaluation II

A. P. de Vries, G. Kazai, M. Lalmas,Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit

Piwowarski, B., and Gallinari, P. Expected ratio of relevant units: A measure for structured information retrieval. In INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the Second INEX Workshop (Dagstuhl, France, Dec. 2003), N. Fuhr, M. Lalmas, and S. Malik, Eds.