Structured Document Retrieval

Benjamin Piwowarski

DCC

October 28, 2004

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 (for now) *not* used

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 Joaquin Jose Lavin Infante”

Queries: 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>
```

Queries: 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>
```

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

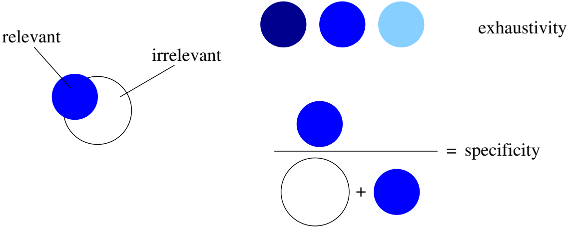
Fact

- *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

Models

Score Propagation

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

Term Weight Propagation

- Term Selection
- Aggregation
  →maximum, augmentation, LM, ...

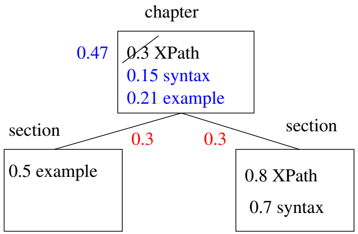
”Moving” Corpus

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

Augmentation

Principle

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

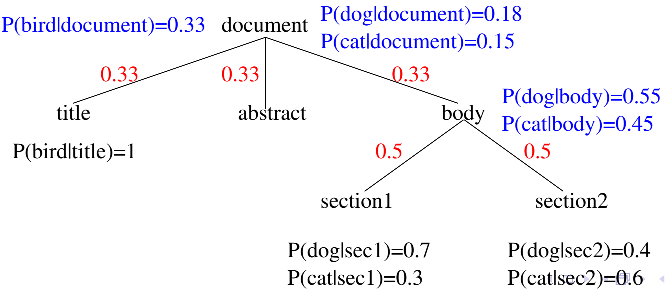


Language Models

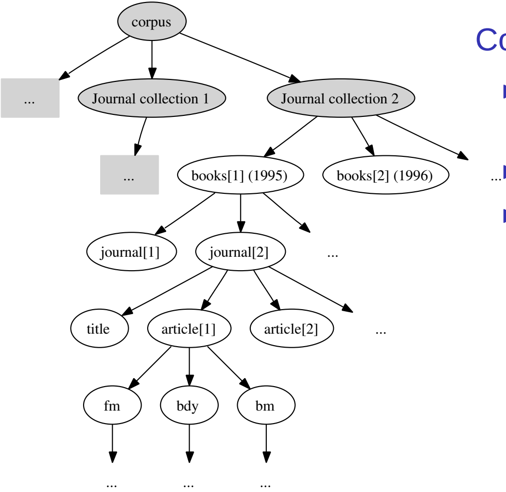
$$P(Q|\Theta_E) = \prod_{\omega \in \{q_1,\ldots,q_n\}} P(\omega|\theta_E)$$

Estimating $P(\omega|\theta_E)$

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



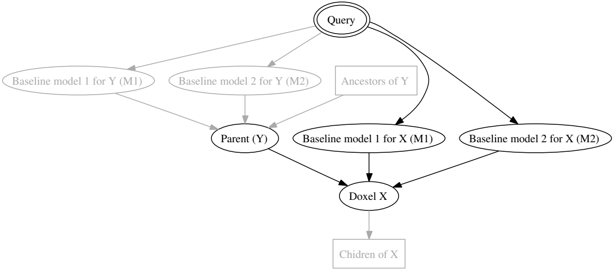
Bayesian Networks: Structure



Components

- Fixed structure = corpus structure
- Parameters
- Baseline models

Bayesian Networks: Local Inference



Variables

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

Bayesian Networks: learning

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

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

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

Example

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

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: ELIXIR

ELIXIR

- An extension of WHIRL
- Path-based language similar to XQuery
- Vague predicate for text (~)

$$\mathrm{RSV}(q,d) = \prod \cos(v_j, c) \prod \cos(v_j, v_k)$$

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 initiative

INitiative for the Evaluation of XML Retrieval

- Since 2002
- System-centred evaluation of effectiveness of XML retrieval approaches
- 30 to 40 institutions each year
- Collaborative effort (participants contribute to the development of the collection)
- Similar methodology as for TREC is followed, but adapted to XML retrieval.

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

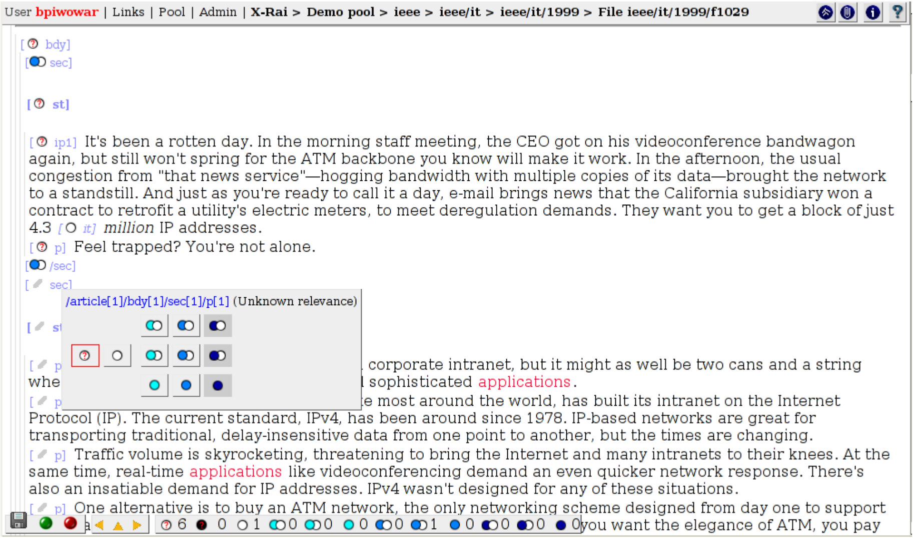
Structured Document Retrieval
- Motivations
- Concepts

Retrieval Systems
- “Content Only” queries
- “Content And Structure” queries

Evaluation
- Assessments
- Metrics

Conclusion
- Summary
- Bibliography

The Interface



Passive Rules

Ensure the *exhaustivity*

- Assess the relevance of all the elements (>8M in IEEE)?
  THIS IS NOT POSSIBLE
- Hypothesis:
  Highly specific elements *might* be near to a submitted element

Some rules

- When an element has been assessed as not relevant (E0S0), no element is added to the pool.
- When an element has been assessed as highly specific (E*S3), only its ancestors are added to the pool.

Active Rules

Ensure the *consistency*

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

Some rules

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

Outline

XML-IR Metrics

The proposed metrics

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

Recall generalisation Expected Ratio of Relevant Elements

other Tolerance To Irrelevance (T2I), Cumulated Gain

Stereotypical runs

Idea: emphasis on metrics differences and caveats

1. Perfect
2. Parent
3. Ancestors
4. First Child
5. Biggest Child

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)**

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

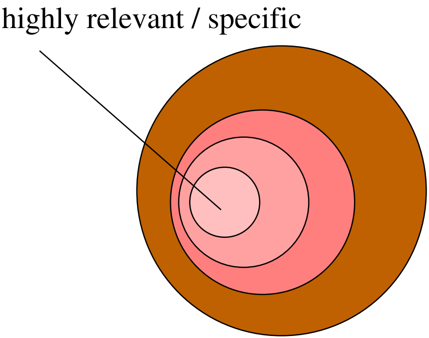
Quantisation

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

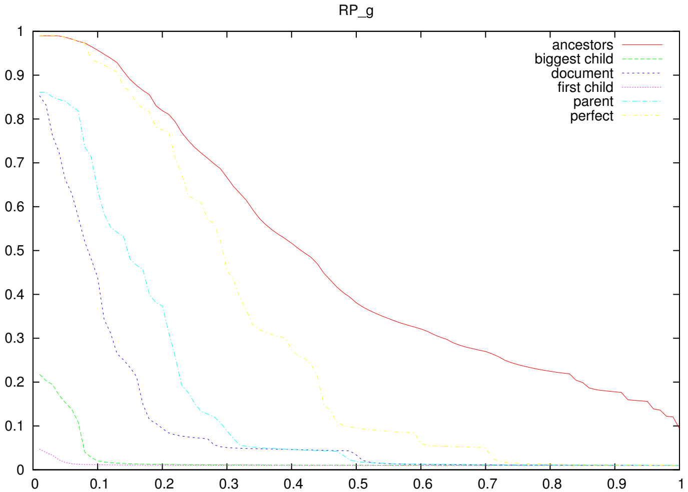
$$f_{gen}(e,s) = \begin{cases} 1 & \text{if } (e,s) = (3,3) \\ 0.75 & \text{if } (e,s) \in \{(2,3),(3,2),(3,1)\} \\ 0.5 & \text{if } (e,s) \in \{(1,3),(2,2),(2,1)\} \\ 0.25 & \text{if } (e,s) \in \{(1,2),(1,1)\} \\ 0 & \text{otherwise} \end{cases}$$

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

The “Recall Base”



Recall-Precision limits



Recall-Precision NG

User Model: classical model + ...

- No more relevance for already retrieved elements

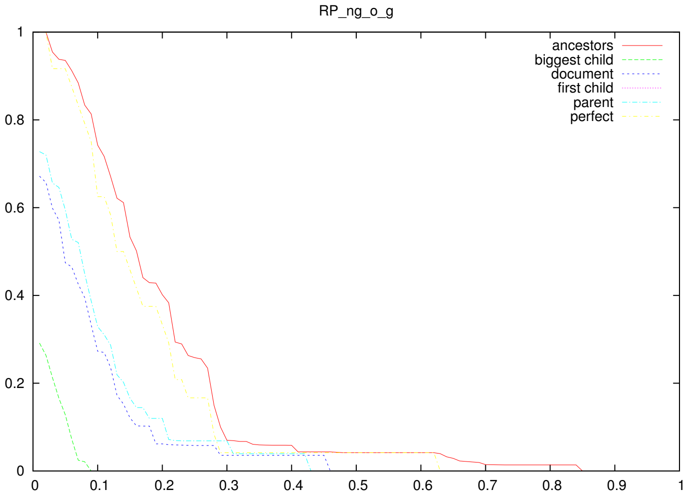
$$\text{recall} = \frac{\sum_e \text{rel}(e)(1 - \frac{\text{size}(\text{seen part of } e)}{\text{size}(e)})}{\sum_e \text{rel}(e)}$$

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

Problems

- The measure is very instable
- Theoritical fundations?

Recall-Precision NG



Generalised Recall

User Model

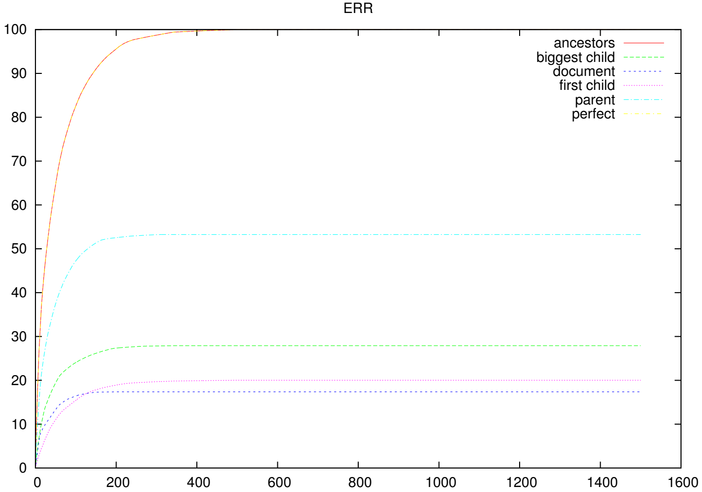
- R/P model
- Stochastic user behaviour
  $\Longrightarrow$ the user can navigate in the document
  $\Longrightarrow$ the user may find an element relevant or not
- Relevant Information = Highly Specific elements only

$$GR(n) = \frac{E(\text{Number of seen relevant elements})}{E(\text{Number of relevant elements})}$$

Limitations

- An equivalent of precision is missing
- Some parameters have to be validated

Generalised Recall Runs



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

Models and methods

CO Search

- Well-defined task
- Various approaches using extensions of classical models

CAS Search

- The query language is still under development
- Vague interpretation of every type of condition

New tasks

- Natural Language (NLP) Queries
- Relevance Feedback
- Heterogeneous collections
- Interactive Retrieval

Evaluation

INEX

- XML documents: IEEE (+ others)
- 3 years of assessments
  - 95 CO topics
  - 86 CAS topics

Metrics

- Precision/recall
- Precision/recall - NG
- Generalised Recall
- *Tolerance To Irrelevance*
- *Cumulated Gain*

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.