



max planck institut
informatik

Semistructured Data Search Evaluation

Ralf Schenkel

Outline

- Introduction
- The INEX Initiative
- INEX Adhoc Evaluation
- Other INEX tracks
- Summary



max planck institut
informatik

OMISE Winter School

February 6, 2

Semistructured Data

- **XML** as most important instance (but also RDF, relational tables/databases, ...)
- Two types of XML documents:
 - „Documents with structure“: **document-centric**
 - „Structured data with text“: **data-centric**

Example for data-centric: DBLP

```
<article key="journals/cacm/Gentry10" mdate="2010-04-26">
  <author>Craig Gentry</author>
  <title>
    Computing arbitrary functions of encrypted data.
  </title>
  <pages>97-105</pages>
  <year>2010</year>
  <volume>53</volume>
  <journal>Commun. ACM</journal>
  <number>3</number>
  <ee>http://doi.acm.org/10.1145/1666420.1666444</ee>
  <url>db/journals/cacm/cacm53.html#Gentry10</url>
</article>
```

**Rather regular structure across documents
not much text per element**

Example for document-centric: Wikipedia

```
<article xmlns:xlink="http://www.w3.org/1999/xlink/">
  <header>
    <title>Wiki markup</title>
    <id>42</id>
    <categories> <category>Markup languages</category>
  </categories>
</header>
<body>
  <section><st>Introduction</st>
    <p><b>Wiki markup</b> is used in <link
      xlink:href=" ../Wi/Wikipedia.xml" xlink:type="simple">
Wikipedia</link>.</p> It allows for a rather rich annotation
  of texts with structure such as tables and lists, links to
  other documents, and much more.
  </section>
  <section>
    <st>Language Components</st>
    <list>
      <entry>tables</entry>
      <entry>lists</entry>
    ...
  </section>
```

**Structure irregular and
different across documents**



Semistructured Data Search/Retrieval

- Why is this different from DB-style queries?
Do not retrieve „all answers“, only „**best answers**“
- Why is this different from document retrieval?
Do not retrieve full documents documents, but document fragments (elements) as results; **focused retrieval**
- Two general querying paradigms:
 - Keywords
 - Structured queries + keywords (XPath FullText)



Example: Document vs. Focused Retrieval

Query: „ticket method“

amazon.com

Shop All Departments

Department
Books

Business & Investing
Job Hunting & Career Guide
Self-Help

+ See All 5 Departments

Shipping Option (What's this?)
Free Super Saver Shipping

Listmania!

Listmania!
create
your list
HERE
learn how

Create a Listmania! list

Search Listmania!

WHAT CAN YOU

Transactional information systems: theory, algorithms, and the practice of ... By Gerhard Weikum, Gottfried Vossen



★★★★★

2 Reviews

Write review

+7 0

About this book

ticket method

Go

Add to My Library ▾



New! Shop for Google eBooks

Go to the Google eBookstore for over 3 million eBooks to read on the Web, Android, iPhone, iPad, Sony and Nook.

Shop for eBooks now »

Get this book

Morgan Kaufmann

Amazon.com - \$118.48

Barnes&Noble.com - \$70.45

Books-A-Million

Find in a library

All sellers »

Related books



Showing 6 results in this book for ticket method - Order by: relevance | pages

Clear search ✕

Page 705 »

... two transactions. Even worse, incompatible orders in which global transactions take their tickets can lead to *global deadlocks* if the participating servers use locking.

For these reasons, an alternative to OTM is the *conservative ticket method* (CTM), which avoids transaction aborts to a large extent. Like OTM, CTM requires subtransactions of global transactions to take tickets at their corresponding sites. However, CTM controls the order in which this happens and ensures that the relative order of *ticket* taking is the same at all participating servers.

Conservative
ticket method
(CTM)

Page 709 »

... step on a single data item as in the page model is not quite accurate from an implementation viewpoint. Rather, tickets need to be tested and manipulated through the servers' interface operations such as SQL commands. On the other hand, this simple observation suffices to "lift" the presented page model *ticket method* to federations with object model servers. All theorems from Section 18.6 still hold, and the *ticket method* is immediately applicable to such richer and realistic systems. In implementation terms, a *ticket* then is actually a tuple in a relational table or an encapsulated object in an object-oriented or object relational database. For data servers other than database systems similar

measures

Page 719 »

Bibliographic Notes 719

18.9 Discuss if and under which conditions the optimistic *ticket method* can be used with servers that provide multiversion serializability for their local schedules (as opposed to conflict serializability).

Hint: The key property that a server has to ensure is that the *ticket order* reflects the local serialization order.

Page 706 »

18.6.2 Implicit Tickets

Both *ticket* methods presented above, the optimistic and the conservative variant, merely require the underlying servers to guarantee local conflict serializability as a "lowest common denominator." Often, servers in a federation may have additional, slightly stronger properties that are, however, not as strong as local rigorosity or the COCSR property. The question then is to what



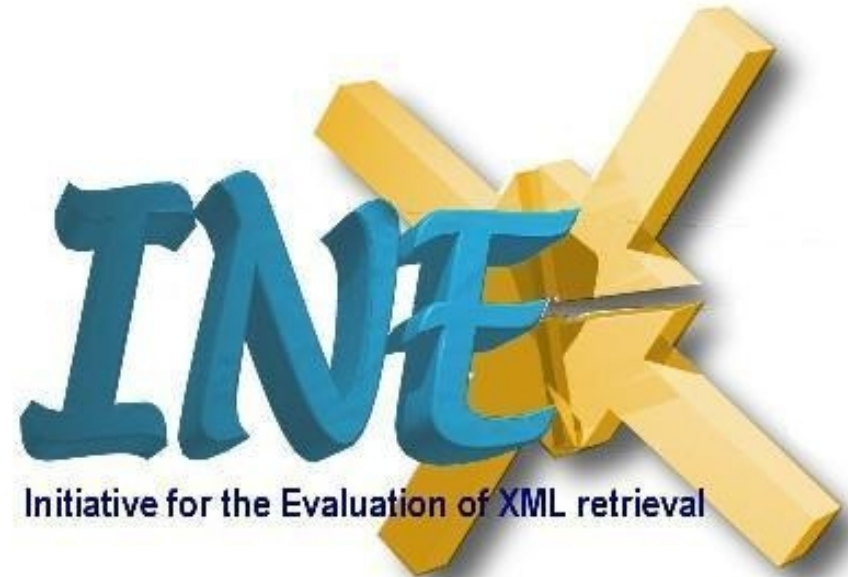
max planck institut
informatik

OMISE Winter School

February 6, 7

The INEX Benchmark Initiative

- started in 2002
- focus on document-centric XML
- large number of participants (>500)
- large number of organizers (100)



(Research) Questions at INEX

- Is focused retrieval better than document retrieval? For which tasks?
- Does document structure help? Are structured queries useful?
- What are good test collections to compare system performance?
- Most important for participants:
Is my system better than the other systems?



Some Tracks at INEX over the Years

- Adhoc Track
- Multimedia Track
keyword-based image search in XML docs
- Heterogeneous Track
search over XML docs with different structure
- Relevance Feedback Track
- Interactive Track
- XML Mining Track
- Efficiency Track
trade off result quality vs. processing time

More on current tracks later

Ingredients of IR Test Collections

- Collection (documents)
- Task
- Topics
- Assessments, relevant results
- Metrics & tools for evaluation

INEX Document collections

- Structured text documents
- 12,227 SGML/XML Articles from IEEE journals
- Wikipedia articles with simple XML markup
- Wikipedia articles with simple XML markup and semantic annotations

IEEE articles

```
<article>
  <fm>
    <ti>IEEE Transactions on ...</ti>
    <atl>Construction of ...</atl>
    <au>
      <fnm>John</fnm>
      <snm>Smith</snm>
      <aff>University of ...</aff>
    </au>
  </fm>
  <bdy>
    <sec>
      <st>...</st>
      <ss1>...</ss1>
      <ss1>...</ss1>
    </sec>
  </bdy>
  <bm>
    <bib>
      <bb>
        <au>...</au><ti>...</ti>
      </bb>
    </bib>
  </bm>
</article>
```

meta data:
title, journal,
author, affiliation

full-text content

cited references



Wikipedia with semantic annotations

Types from WordNet

<article>

```
<group confidence="1.0" wordnetid="26729"
      source="categories">
```

```
<artist confidence="0.75" wordnetid="9187509">
```

```
<header>
```

```
<title>Queen (band)</title>
```

```
<id>42010</id>
```

Information from Infoboxes

```
<Infobox_band>
```

```
<band_name>Queen</band_name>
```

```
<years_active>1971 - Present</years_active>
```

```
<status>Active</status>
```

```
<country confidence="1.0" wordnetid="8023668">
```

```
<link xlink:href=" ../Un/United+K$ingdom.xml"
      xlink:type="simple">
```

```
United Kingdom
```

```
</link>
```

```
</country>
```

```
</Infobox_band>
```

INEX Topics

- Two different types:
 - Content-Only (CO)
 - Content and Structure (CAS)
- Contributed by participants, so diverse in nature

Content-Only (CO) topic

```
<INEX-Topic topic-id="45" query-type="CO" ct-no="056">
```

```
<Title>
```

```
<cw>augmented reality and medicine</cw>
```

```
</Title>
```

```
<Description>
```

```
How virtual (or augmented) reality can contribute to  
improve the medical and surgical practice.
```

```
</Description>
```

```
<Narrative>
```

```
In order to be considered relevant, a document/component  
must include considerations about applications of computer  
graphics and especially augmented (or virtual) reality to  
medicine (including surgery).
```

```
</Narrative>
```

```
<Keywords>
```

```
augmented virtual reality medicine surgery improve  
computer  
assisted aided image
```

```
</Keywords>
```

```
</INEX-Topic>
```



Structured Topics (CAS): INEX 2002

```
<INEX-Topic topic-id="09" query-type="CAS" ct-no="048">
```

```
<Title>
```

```
<te>article</te>
```

```
<cw>non-monotonic reasoning</cw> <ce>bdy/sec</ce>
```

```
<cw>1999 2000</cw> <ce>hdr//yr</ce>
```

```
<cw>-calendar</cw> <ce>tig/at1</ce>
```

```
<cw>belief revision</cw>
```

```
</Title>
```

```
<Description>
```

Retrieve all articles from the years 1999-2000 that deal with works on nonmonotonic reasoning. Do not retrieve articles that are calendar/call for papers.

```
</Description>
```

```
<Narrative>
```

Retrieve all articles from the years 1999-2000 that deal with works on nonmonotonic reasoning. Do not retrieve articles that are calendar/call for papers.

```
</Narrative>
```

```
<Keywords>
```

non-monotonic reasoning belief revision

```
</Keywords>
```

```
</INEX-Topic>
```



max planck institut
informatik

OMISE Winter School

February 6, 17

Structured Topics (CAS): INEX 2003

<inex_topic>

<title>

```
//article[(./fm/yr = '2000' OR ./fm/yr = '1999') AND about(.,  
  'intelligent transportation system')]//sec[about(., 'automation  
+vehicle')]
```

</title>

<description>

Automated vehicle applications in articles from 1999 or 2000 about intelligent transportation systems.

</description>

<narrative>

To be relevant, the target component must be from an article on intelligent transportation systems published in 1999 or 2000 and must include a section which discusses automated vehicle applications, proposed or implemented, in an intelligent transportation system.

</narrative>

<keywords>

intelligent transportation system, automated vehicle, automobile, application, driving assistance, speed, autonomous driving

</keywords>

</inex_topic>

**Extended version of XPath:
about(path,keyword condition)**

**Too complex for IR people:
63% of topics with errors**



Narrowed Extended XPath I (NEXI)

- Restricted axes and composition: only 2 types
 - `//A[B]`
 - `//A[B]//C[D]`
- tag wildcard `*`, tag disjunction (`sec|p`)
- content conditions: `about(path, text)`
- comparison for numeric values only

```
//article[(.//fm//yr = 2000 OR .//fm//yr = 1999)
          AND about(., 'intelligent transportation')]
//sec[about(., 'automation +vehicle')]
```

Strict vs. Loose/vague interpretation

- DB vs. IR interpretation of queries
- **Strict**: exact match of target element (**SCAS**)
- **Vague**: vague match of target element (**VCAS**); requested **article**, but **sec** is valid result
- IR interpretation of query: path specifications considered **hints** as to where to look

XYCAS: different CAS interpretations

- V – Vague
- X – target element
- S – Strict
- Y – support element
- VVCAS: both target and support elements are vague (classic IR view)
- SVCAS: target strict, support elements vague
- VSCAS: target vague, support elements strict
- SSCAS: both target and support elements are strict (classic DB view)

XYCAS example

//**article**[about(., 'XML') and about(//**sec**, 'DB')]]


target


support

SSCAS: only matching article elements with matching sec subelements

VSCAS: any matching elements with matching sec subelements

SVCAS: only matching article elements with any or no matching subelements

VVCAS: any matching elements with any or no matching sec subelements

CO+S topics

```
<inex_topic query_type="CO+S">
  <title>Tolkien languages "lord of the rings"</title>
  <castitle>//article[about(., Tolkien) or about(., "lord of the rings")]
    //sec[about(., Tolkien languages)]</castitle>
  <description>
    Find information about Tolkien languages from the Lord of the Rings.
  </description>
  <narrative>
    The "Lord of the Rings" ... For my own personal interest, I would like to
    learn more background about Tolkien's artificial languages. Later I may
    want to add a section on the influence languages to my fan web page.
    I expect to find relevant information as elements in larger documents that
    deal with Tolkien or Lord of the Rings, e.g., as sections in documents
    about Tolkien or the Lord of the Rings.
    To be relevant an element should discuss Tolkien's artificial languages and
    their influence on the Lord of the Rings books or movies. Information on
    the languages alone without explicit discussion of their impact on the
    books is not relevant; nor is general information on Tolkien or the Lord of
    the Rings.
  </narrative>
  <ontopic_keywords>"High Elvish" ; Quenya ; Sindarin</ontopic_keywords>
  <offtopic_keywords>inspired, film</offtopic_keywords>
</inex_topic>
```

Combines CO and CAS titles



max planck institut
informatik

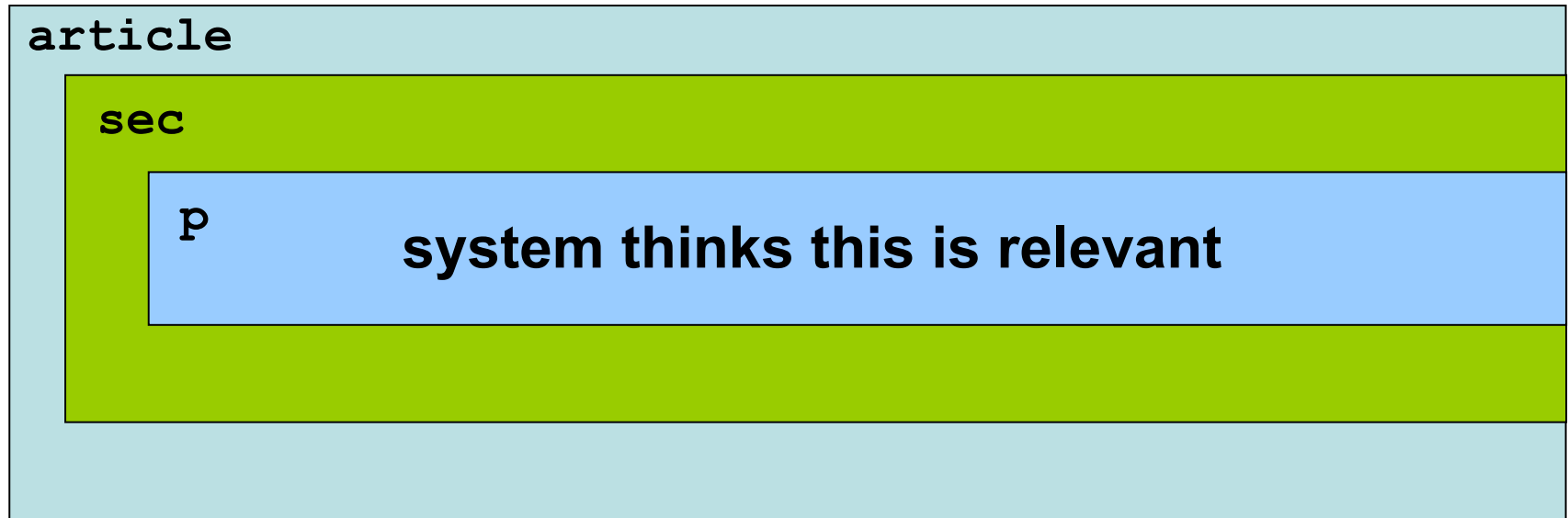
OMISE Winter School

February 6, 23

INEX AdHoc Tasks

- **Thorough**: find all relevant information (elements)
- **Focused**: find all relevant information (elements) without any overlap
- **Relevant-in-context**: document ranking, within each document highlight relevant content
- **Best-in-context**: best entry point into an article

Thorough vs. Focused



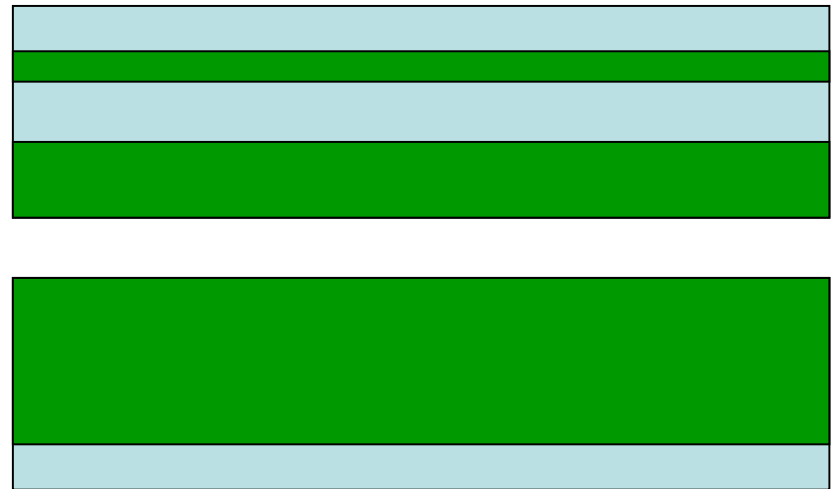
Thorough: should return p, sec, article

**Focused: should return only element
with most relevant content**

Relevant-in-Context Task

For each topic, return ranked list of documents with non-overlapping relevant elements

- rank 1: document 17
`//article[1]/sec[2]/p[1]`
`//article[1]/sec[4]/p[2]`
- rank 2: document 12
`//article[1]/sec[1]`
- ...

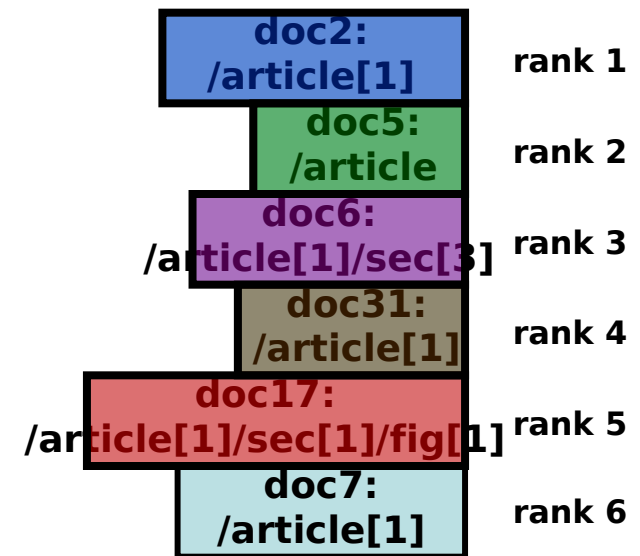


Assessments: Estimate Relevant Results for each Topic

- **General approach**: humans assess all elements for relevance
- But: way too much effort (millions of elements)
- Build **pool of elements/documents to assess** from submitted results
- At INEX: **participants** assess

INEX (document-based) Pooling

- Build pool of **size $S=500$ documents** per topic
- Collect **top-1** result from each run, then **top-2**, ... until S documents found



Build pool of size $S=500$ documents: **doc17, doc2, doc5, doc6, doc56, doc31**

INEX 2004: Two-dimensional relevance

- **Exhaustivity (E)**, which describes the extent to which the document component discusses the topic of request.
- **Specificity (S)**, which describes the extent to which the document component focuses on the topic of request.

4-point scale for exhaustivity

- **Not exhaustive (E0):** the document component does not discuss the topic of request at all.
- **Marginally exhaustive (E1):** the document component discusses only few aspects of the topic of request.
- **Fairly exhaustive (E2):** the document component discusses many aspects of the topic of request.
- **Highly exhaustive (E3):** the document component discusses most or all aspects of the topic of request.

4-point scale for specificity

- **Not specific (S0):** the topic of request is not a theme of the document component.
- **Marginally specific (S1):** the topic of request is a minor theme of the document component (i.e. the component focuses on other, non-relevant topic(s), but contains some relevant information).
- **Fairly specific (S2):** the topic of request is a major theme of the document component (i.e. the component contains mostly relevant content and only some irrelevant content).
- **Highly specific (S3):** the topic of request is the only theme of the document component.

INEX'04 Assessments

The ARPArOme KnowledgeBased Planning and Scheduling Initiative



Processing (and caching) file with stylesheet

```
[ article]
[ fno] x1004
[ doi] 10.1041/x1004s-1995
[ fm]

[ /hdr]
[ /hdr1]
[ /article[1]/fm[1]/hdr[1]]
[ /crl] [ /issn] 0885-9000 /95/$4.00 [ /ccil] [ /onm] © 1995 IEEE
[ /hdr2] [ /obi] [ /volno] Vol. 10 [ /issno] No. 1 [ /pdt] [ /mo] February [ /yr] 1995 [ /pp] pp. 4-9
[ /tig]
```

[/atl] The ARPA-Rome Knowledge-Based Planning and Scheduling Initiative

[/sbt] Guest Editors' Introduction [/pn] pp. 4-9
 [/au] [/fnm] Northrup [/snm] Fowler III ([/aff] [/onm] Rome Laboratory) : [/au] [/fnm] Stephen E.
 [/snm] Cross ([/aff] [/onm] ARPA) : [/au] [/fnm] Chris [/snm] Owens ([/aff] [/onm] BBN Systems and

INEX 2004 assessment tool

[-atl] About this Issue

[-pn] pp. 3-3

[−au] [−fnm] J.A.N. [−snm] Lee, [−role] Editor-in-Chief [?/fm]

Table of Contents

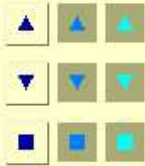
[?bdv]

```

/article[1]/bdy[1]

```

Unknown relevance



Our 17th volume is as diverse in topics as any nontheme issue that we have tried to

t many years. However, it still represents

r than a broader picture of computing in the

[-it] *Annals* are doing their best to bring t

es require authors in other countries to o

n open invitation to authors in other parts

w and help us to follow the lead of our par

Computer Society."

[–p] The five major articles in this issue represent several manu

time, and we are grateful to the authors for having “stuck with us” while we reviewed, re-reviewed, and

reworked their papers. Articles in the field of history do not always present the work of the authors

themselves (though we welcome pioneers to give us their own stories, as in the case of the 1935 article by

John McPherson in this issue); thus, answering the question "is it accurate?" is not always easy. In fact, we

ask our referees to answer the following questions about each manuscript and their responses determine

whether we accept the manuscript “as is” or whether we ask

[-12]

 $[-li]$

[-p] Are the issues addressed in the paper stated clearly

117

●●▲▲? 3■2×0▲0▼0■0▲0▼0■0▲0▼0■0

Assess each element in the pool on 3x3 relevance scale

**High effort
(up to 1 week per topic)**

2006 assessment tool: Highlighting

User demo | Links | Pool | X-Ral > Demo pool 324 > wikien > 1-x-x > 181-x
> File 927344

Ali Baba

Wikipedia link: [Ali Baba](#)

Ali Baba

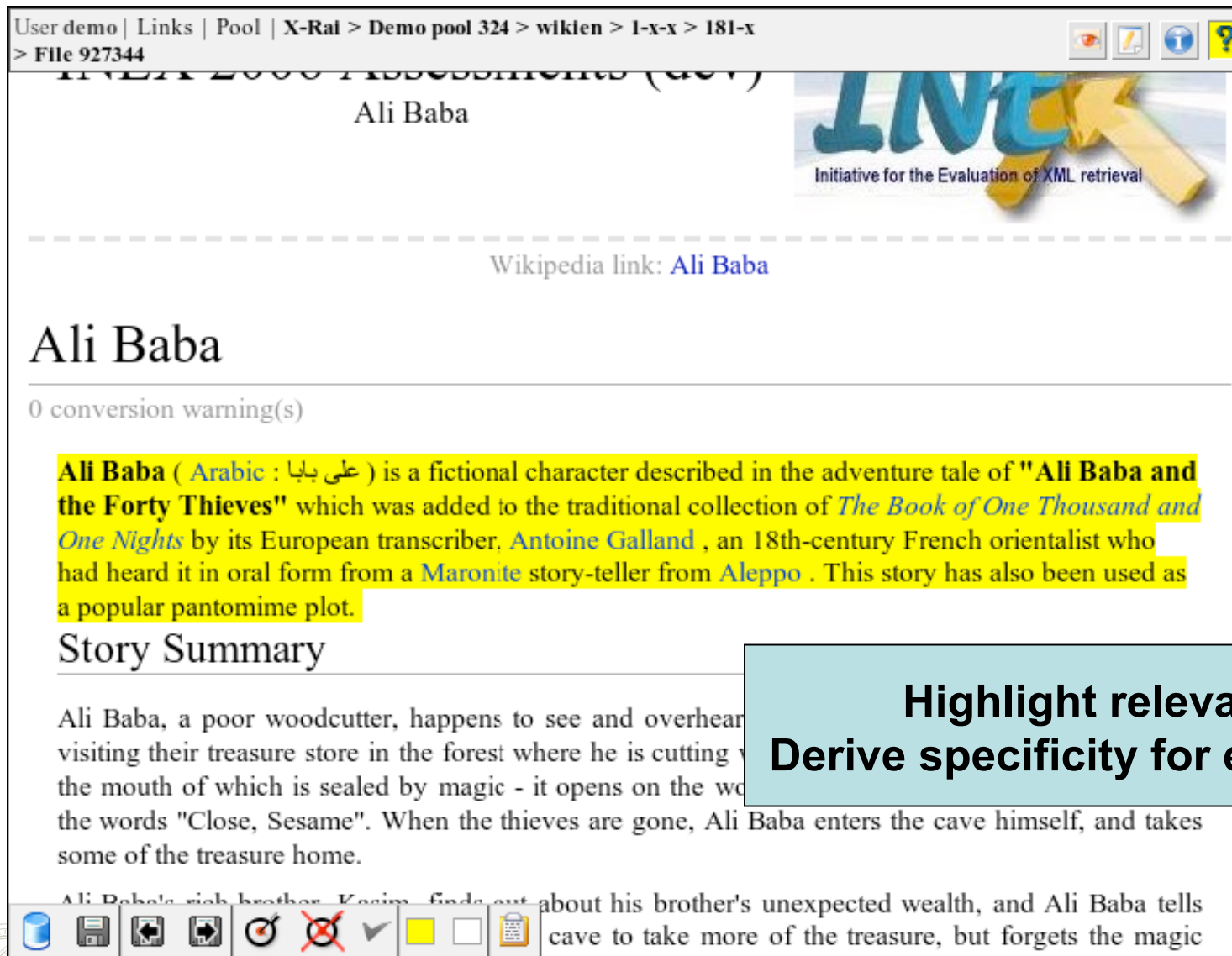
0 conversion warning(s)

Ali Baba (Arabic : علي بابا) is a fictional character described in the adventure tale of "**Ali Baba and the Forty Thieves**" which was added to the traditional collection of *The Book of One Thousand and One Nights* by its European transcriber, **Antoine Galland** , an 18th-century French orientalist who had heard it in oral form from a **Maronite** story-teller from **Aleppo** . This story has also been used as a popular pantomime plot.

Story Summary

Ali Baba, a poor woodcutter, happens to see and overhear visiting their treasure store in the forest where he is cutting the mouth of which is sealed by magic - it opens on the words "Close, Sesame". When the thieves are gone, Ali Baba enters the cave himself, and takes some of the treasure home.

Ali Baba's rich brother, Kasim, finds out about his brother's unexpected wealth, and Ali Baba tells cave to take more of the treasure, but forgets the magic



Highlight relevant text
Derive specificity for each element

INEX 2010++ Assessment Tool

The screenshot shows the INEX 2010++ Assessment Tool interface. The window title is "GPX". The menu bar includes "File: person_130479", "Show Pool", "Hide Keyword...", "Set Keywords", "Keywords:" (with a dropdown), "Help", "Topic", "% assessed", and "0%". The main content area displays XML data for Adrien Brody, including his name, birth date (14 April 1973), height (6' 1"), and a biography. The biography text is highlighted in yellow. A large orange text box is overlaid on the right side of the window, containing the text: "Takes about 1 minute/doc (or 5 turkers on MechTurk à 0,10€". At the bottom of the window, there are four buttons: "Next Document", "Completely Irrelevant", "EXIT", and "DEBUG".

GPX

File: person_130479 Show Pool Hide Keyword... Set Keywords Keywords: Help Topic % assessed 0%

<?xml version="1.0" encoding="UTF-8"?>
<person>
 <name>Adrien Brody
 </name>
 <overview>
 <birth_date>14 April 1973, New York City, New York, USA
 </birth_date>
 <height>6' 1"
 </height>
 <biographies>
 <biography> The son of Hungarian-born photographer Sylvia Plachy and retired history professor Elliot Brody, Adrien Brody grew up an only child in the Woodhaven section of Queens, New York, where he accompanied his mother on assignments for the Village Voice. He credits her with making him feel comfortable in front of the camera. He attended the American Academy of Dramatic Arts and LaGuardia High School for the Performing Arts in New York. Despite a strong performance in _The Thin Red Line (1998)_ (qv), time constraints forced the director to edit out much of Adrien's part. In spite of his later work with 'Spike Lee' (qv) and 'Barry Levinson' (I)' (qv), he never became the star many expected he would become until 'Roman Polanski' (qv) called on him to play a celebrated Jewish pianist in Nazi-occupied Warsaw. He pulled off a brilliant performance in _The Pianist (2002)_ (qv), drawing on the heritage and rare dialect of his Polish grandmother, as well as his father, who lost family members during the Holocaust, and his mother, who fled Communist Hungary as a child during the 1956 uprising against the Soviet Union.
 <by>Red Hawk Management
 </by>
 </biography>
 </biographies>
 <trivias>
 <trivia>Is the only actor to win a Best Actor Oscar when nominated alongside four other nominees.
 </trivia>
 <trivia>In 2003, at the age of 29, he replaced 'Richard Dreyfuss' (qv) as the young John F. Kennedy in _The Assassination of John F. Kennedy (2000)_ (qv).
 </trivia>
 <trivia>Has great appreciation for hip-hop music. Like the Beatles, one of his favorite groups is The Beatles.
 </trivia>
 <trivia>Performed magic shows at children's birthdays as a child as the Amazing A.
 </trivia>
 <trivia>In 2004, Esquire Magazine named him the Best Dressed Man in America.
 </trivia>
 <trivia>Owned and drove a Hummer, which is famous in Hollywood for being 'Arnold Schwarzenegger' (qv)'s favorite vehicle. This ecologically unfriendly vehicle gets under ten miles per gallon and was originally designed for warfare.
 </trivia>
 <trivia>Was considered for a role in _Pearl Harbor (2001)_ (qv).
 </trivia>
 <trivia>Was dressed by Zegna for the Academy Awards.
 </trivia>
 <trivia>Is a big hip-hop fan and plans on becoming a producer. He is being mentored by The 'RZA' (I)' (qv).
 </trivia>
 <trivia>Ranked #21 on VH1's 100 Hottest Hotties.
 </trivia>
 <trivia>His father, Elliot Brody, is a retired history teacher of Polish-Jewish descent who lost family members in the Holocaust. His mother, Sylvia Plachy, the renown photographer, fled Communist Hungary as a child during the 1956 uprising against the Soviet Union.
 </trivias>
</person>

OOPS!

Next Document Completely Irrelevant EXIT DEBUG

Takes about 1 minute/doc (or 5 turkers on MechTurk à 0,10€



Metrics: General principles

- Quantization Q: Map (E,S) value to $[0,1]$
- Recall-based evaluation for each topic
- For a run: Average metrics value for each topic

Quantizations (INEX 2004)

- **Strict**: map to 1 only for (3,3) results (INEX 2002)
- **Generalized** (INEX 2002): graded relevance
- **Specificity-oriented generalized**: more focus on specificity component
- **Specificity-oriented**: map to 1 only for (3,*) results
- **Exhaustivity-oriented**: map to 1 only for (*,3) results

INEX 2002 Thorough metrics: Recall

Consider recall base: set of all elements with $Q > 0$

$$P(rel \mid retr)(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}}$$

x : recall point $0, 0.01, \dots, 1$ (point in the run where fraction x of relevant elements are found)

n : number of relevant elements

esl : expected search length (number of nonrelevant elements at recall x , more difficult when ranking includes ties)

INEX 2006 Thorough metrics: xCG

- Consider ideal run: elements from recall base in descending order of their Q value (here: fraction of relevant characters in element)
- Compute extended cumulated gain of run (xCG) and ideal run (xCI) at rank i

$$xCG[i] = \sum_{j=1}^i xG[j]$$

$$xG[i] = Q(result(i))$$

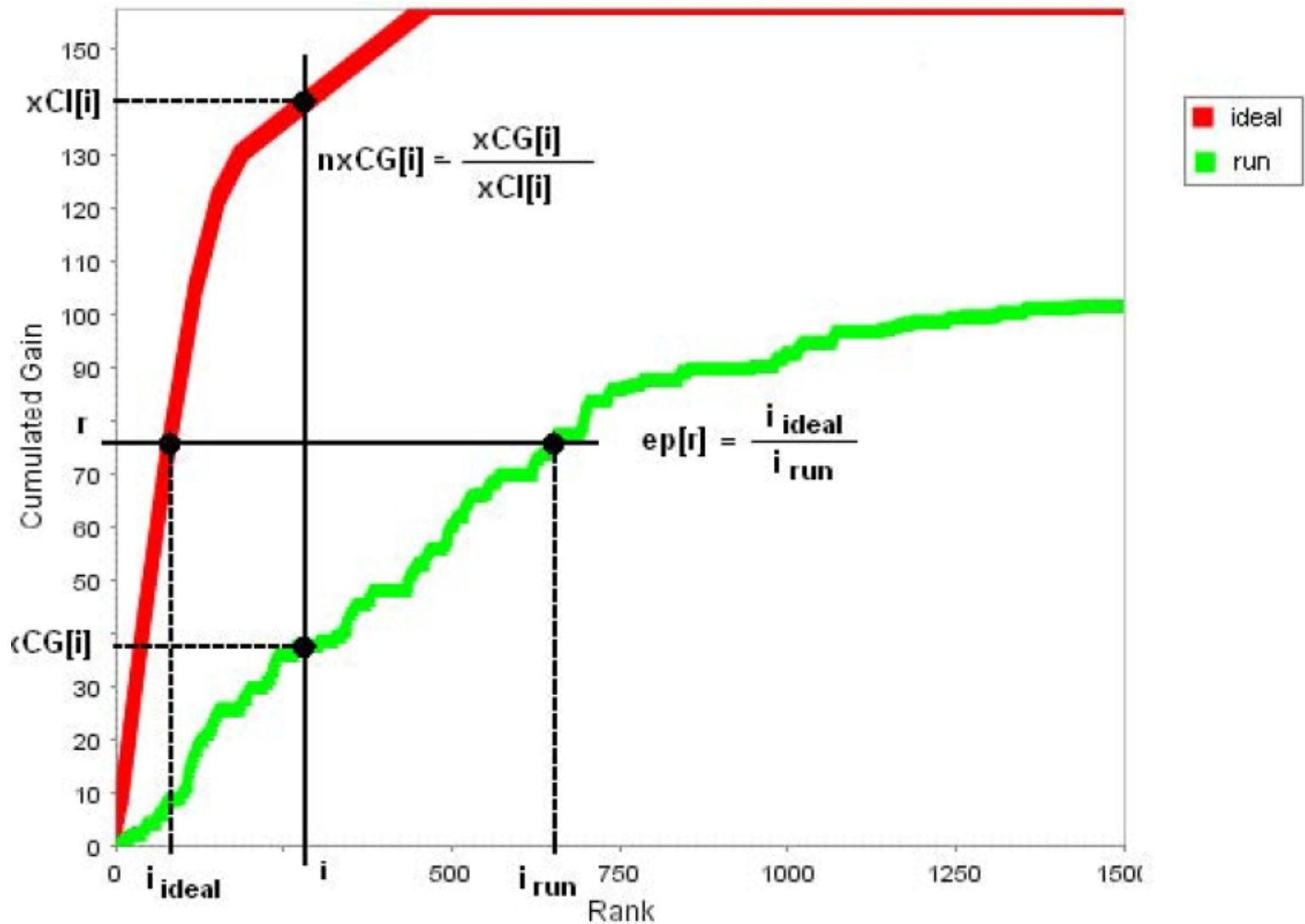
- Compute relative effort (rank) to achieve target gain r :

$$ep[r] = \frac{i_{ideal}}{i_{run}}$$

- Use normalized xCG:

$$nxCG[i] = \frac{xCG[i]}{xCI[i]}$$

INEX 2006 metrics



Not yet considered: Overlap of results

doc12/article[1]

doc12/article[1]/sec[1]

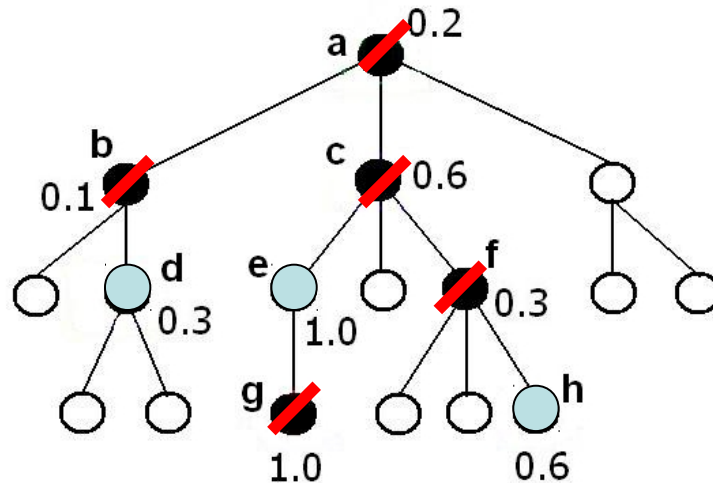
doc12/article[1]/sec[2]

doc12/article[1]/sec[1]/par[1]

- Results beyond 1 do not contribute new content, so are useless for user
- But: included in recall base, must be returned for high **Precall** or **ep[r]** value

INEX 2006 Focused metrics

- Compute ideal **overlap-free** recall base
 - Select elements with highest Q
 - Break ties by choosing element toward the top of the XML tree



From Elements to Passages

Major insight around 2007:

Elements as results are too restrictive since

- boundaries are arbitrary
- Relevant content (aka highlighted text) independent of element boundaries

**Natural consequence:
retrieve text passages instead of elements
(XML structure only hints!)**

INEX 2007 Focused Measures

- Based on retrieved relevant text
- For a result p
 $size(p)$: number of characters in p
 $rsize(p)$: number of unseen relevant characters in p
- Rank-based measures precision & recall

$$P[r] = \frac{\sum_{i=1}^r rsize(p_i)}{\sum_{i=1}^r size(p_i)}$$

$$R[r] = \frac{\sum_{i=1}^r rsize(p_i)}{Trel(q)}$$

where $Trel(q)$ = number of relevant characters for q

INEX 2007: interpolated precision



Precision:

$$p[0.01] = \frac{\# \text{ rel chars at 1\% recall}}{\# \text{ chars at 1\% recall}}$$

Interpolated Precision:

$$iP[r] = \max_{r' \geq r} \{ p[r'] \}$$

Average Interpolated Precision:

$$AiP = \frac{1}{101} \sum_{r=0.0}^{1.0} iP[r]$$

[MAiP: Mean over many topics]

How do we get this



Relevant passages

Doc1:1-510

Doc17:1-90; 200-300

Doc141:81-400

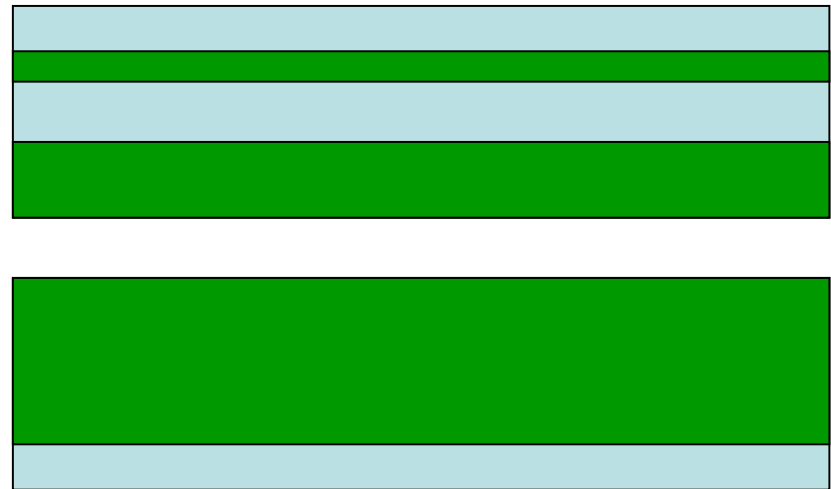
...

10,000 relevant chars

Reminder: Relevant-in-Context Task

For each topic, return ranked list of documents with non-overlapping relevant elements

- rank 1: document 17
`//article[1]/sec[2]/p[1]`
`//article[1]/sec[4]/p[2]`
- Rank 2: document 12
`//article[1]/sec[1]`
- ...



Two-step metrics for relevant in context

- Per-document score $S(d)$: F-measure

$$P(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{\sum_{p \in \mathcal{P}_d} size(p)} \quad R(d) = \frac{\sum_{p \in \mathcal{P}_d} rsize(p)}{Trel(d)} \quad F(d) = \frac{2 \cdot P(d) \cdot R(d)}{P(d) + R(d)}$$

- Per-topic score: generalized precision &

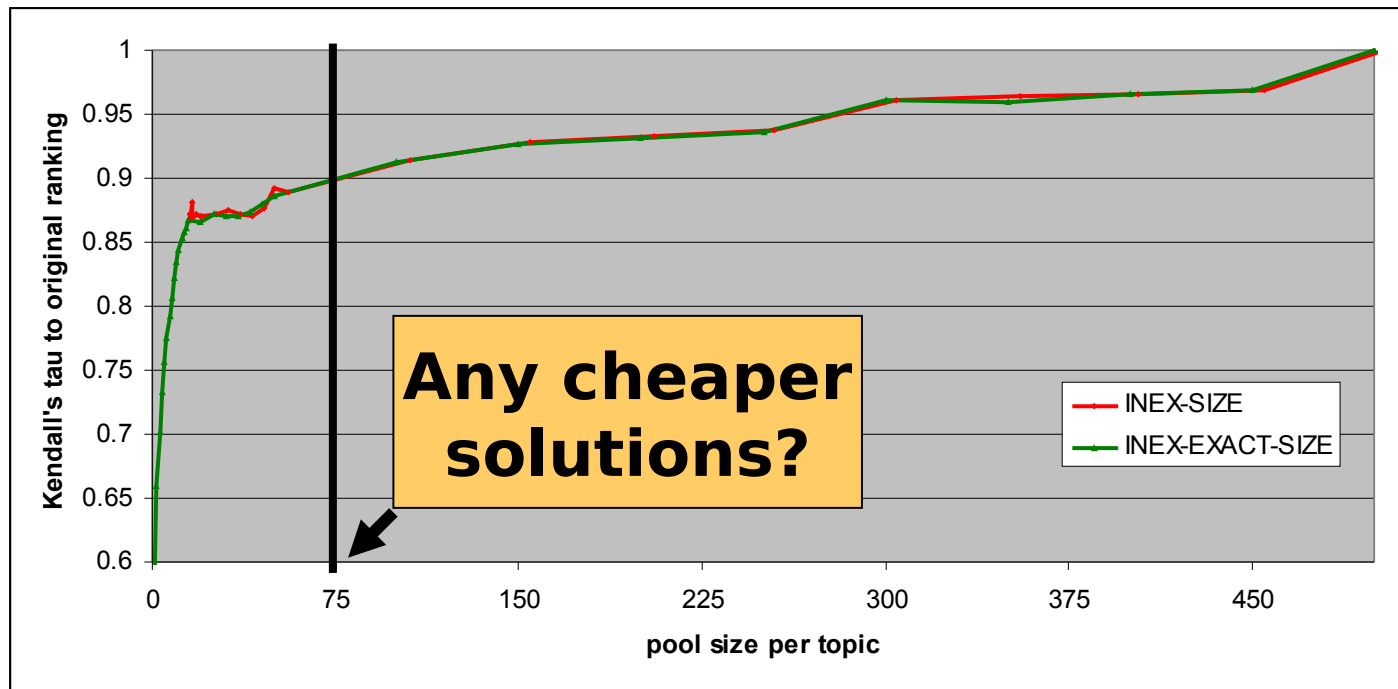
$$gP[r] = \frac{\sum_{i=1}^r S(d_i)}{r} \quad gR[r] = \frac{\sum_{i=1}^r IsRel(d_i)}{Nrel}$$

$$AgP = \frac{\sum_{r=1}^{|\mathcal{L}|} IsRel(d_r) \cdot gP[r]}{Nrel}$$

Effect of Pool Size (INEX 09)

How good is ranking with a smaller pool?

- Consider **Kendall's tau** of run ranking to original ranking (with poolsize=500 for assessments)



General agreement: $\tau=0.9$ is good enough

Solutions for Text Retrieval: MTC

Find **Minimal Test Collection** (to assess)

- For document retrieval
- For rank-based metrics (e.g., precision@10, AvgPrec)

Impact of document d on run's AvgPrec:

- Depends only on rank $r(d)$ of d in the run
- Can be precomputed:
contributes $1/R$ to precision@ R for each rank $R \geq r$
- Select documents to assess based on best impact for run1, best impact for run2, etc.

Stop when

- best run found (threshold)
- best run found with high probability (under assumptions)

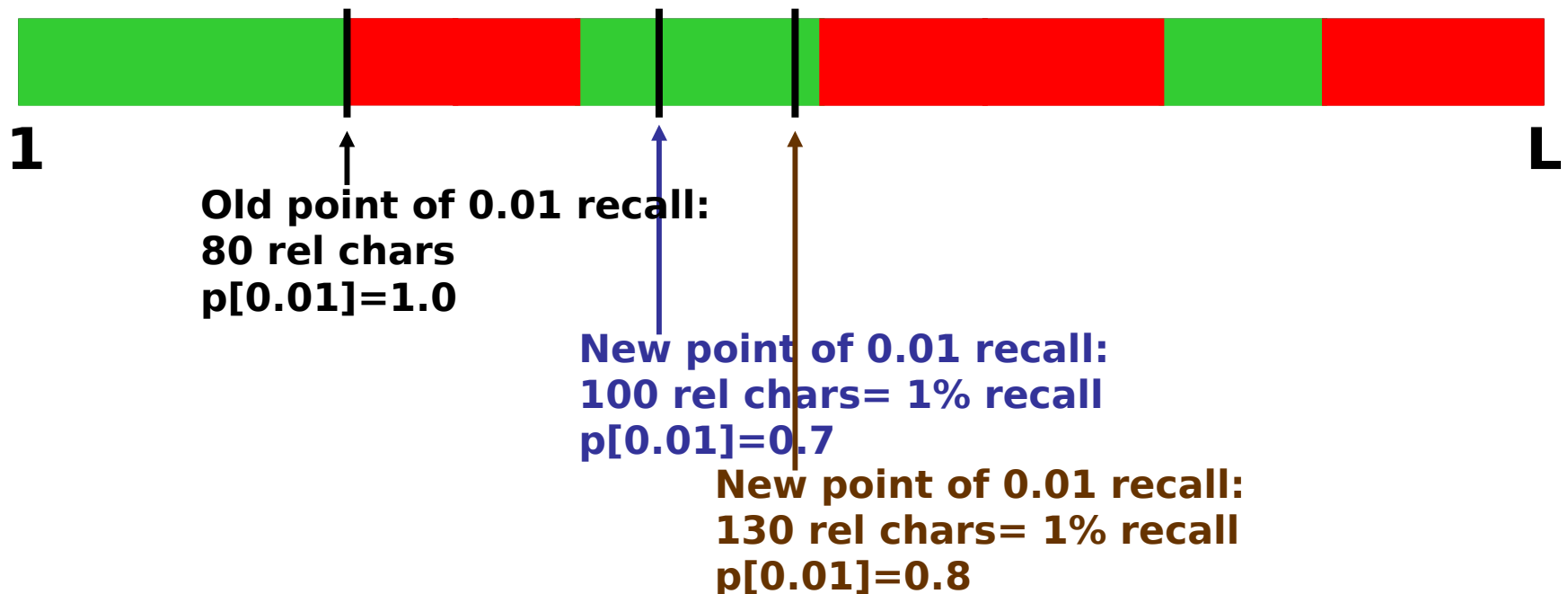
[Carterette et al, SIGIR 2006]



Impact of Assessing a Document

Problem with $iP[0.01]$ (a **recall-based** metric):

- value may **reduce** with more assessments (even docs not included in a run!)



Impact of Assessing a Document

Problem with $iP[0.01]$ (a **recall-based** metric):

- value may **reduce** with more assessments (even docs not included in a run!)

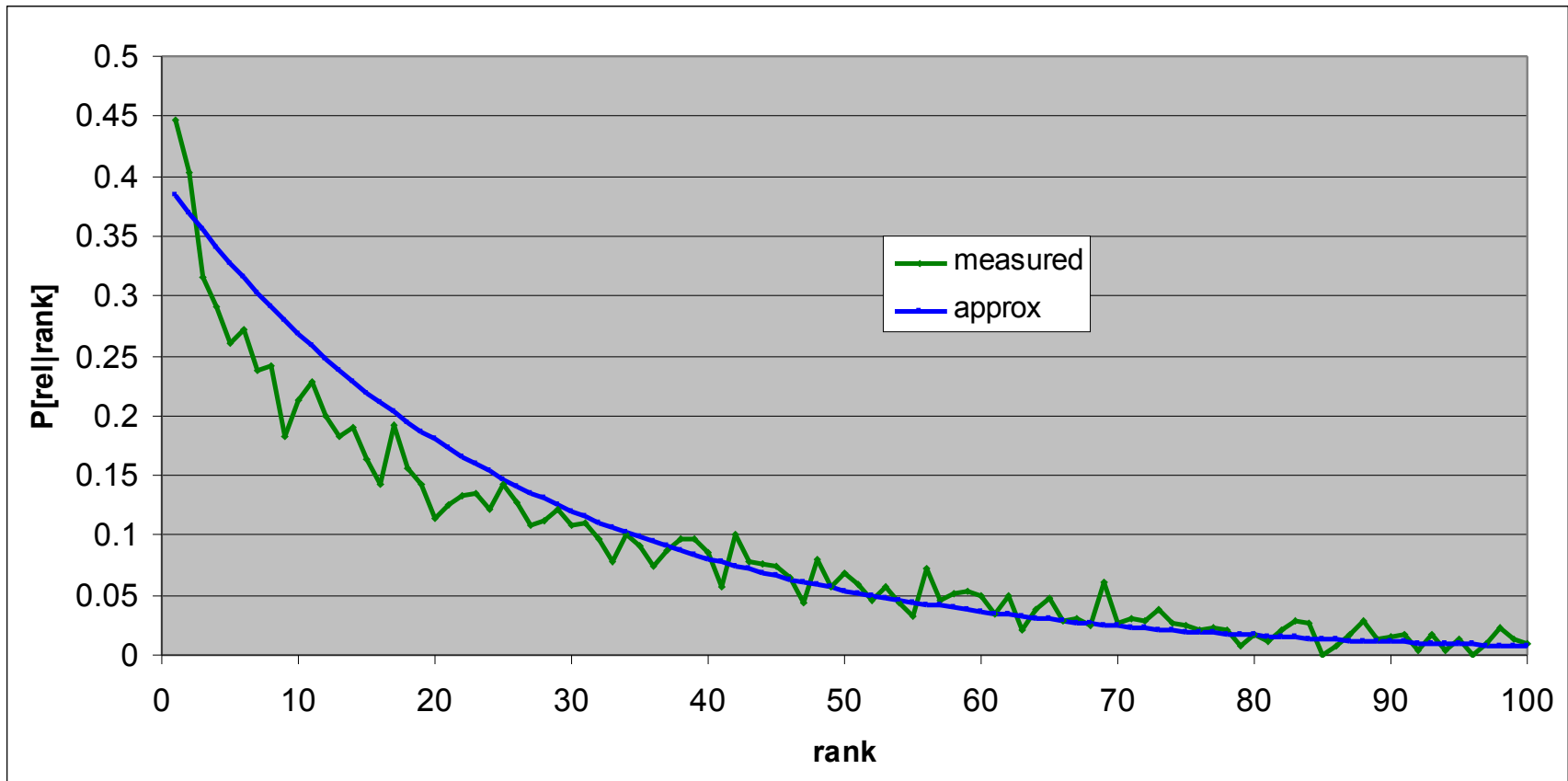
Problem with passage-level assessment:

- **Which part** of the document will be relevant?

Solution:

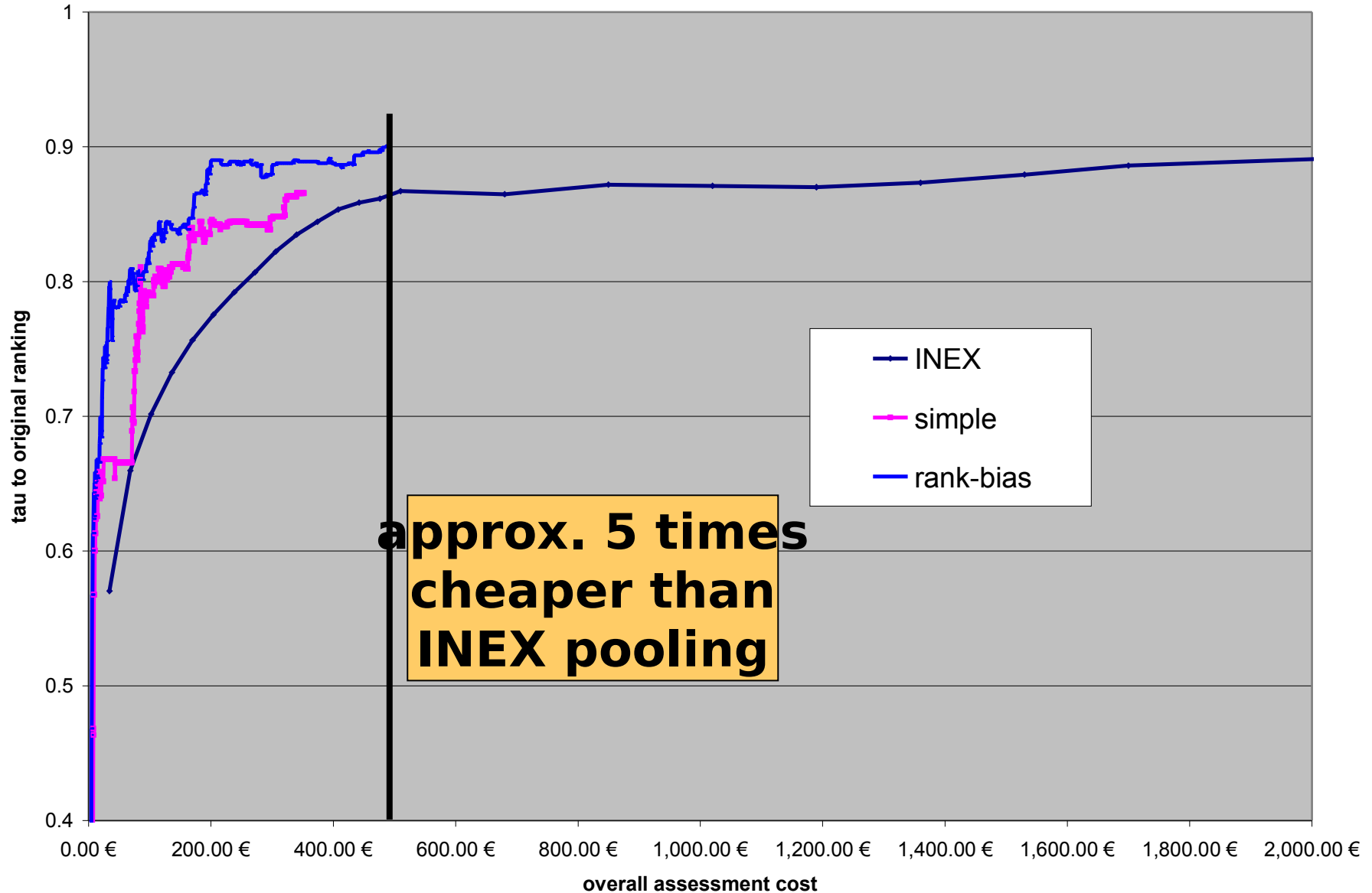
- Consider every retrieved fragment f plus whole doc
- $\text{impact}(f)$ = absolute change in $iP[0.01]$ when f relev.
- $\text{impact}(\text{doc})$ = max impact of any fragment
- Assess document with highest impact next

Rank-Based Relevance Probability



- Approximate $P[\text{rel}|\text{rank}]$ through exponential function
- Weight $\text{impact}(f)$ by minimal rank of f in any run

Experiment (INEX09)



Outline

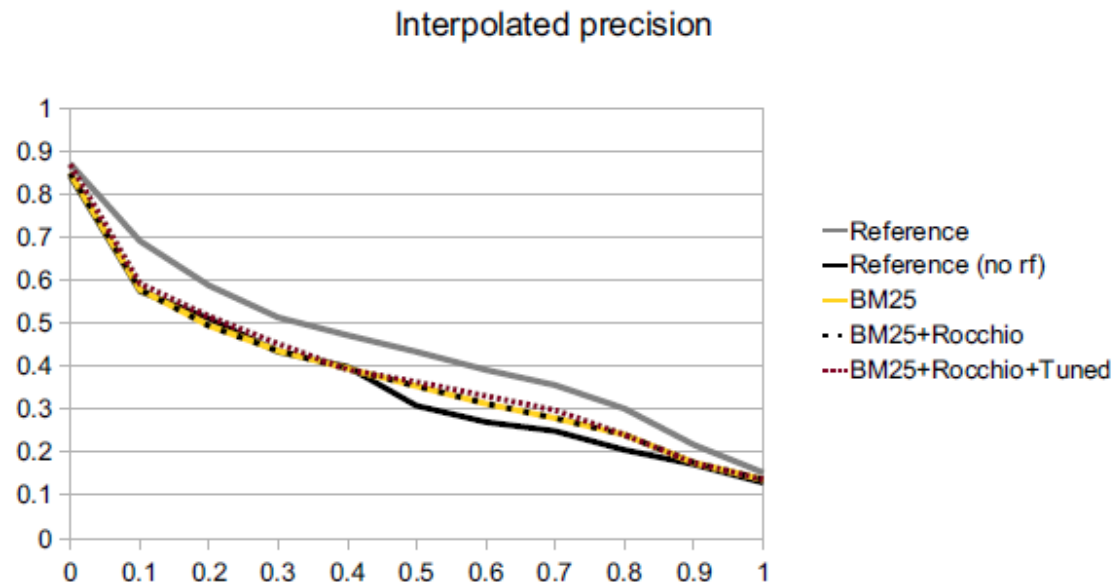
- Introduction
- The INEX Initiative
- INEX Adhoc Evaluation
- **Other INEX tracks**
- Summary

Relevance Feedback Track

- Goal: use **user's feedback** to some results for improving further results of the same query
 - Evaluation non-trivial: relevance of some results is known
 - Traditional approaches:
 - Freeze known results at top of result list
 - Remove known results from result pool
- Both used for the INEX 2006 RF track

RF track 2010-2012

- Interleaved retrieval & feedback
- Variant of freezing with many rounds
- Track provided interface to feedback module
- Submit implementation, not results
- Evaluate resulting list of results with standard tools



Natural Language Query Track

- Goal: Create structured NEXI query from description in natural language

Find sections about compression in articles about information retrieval.

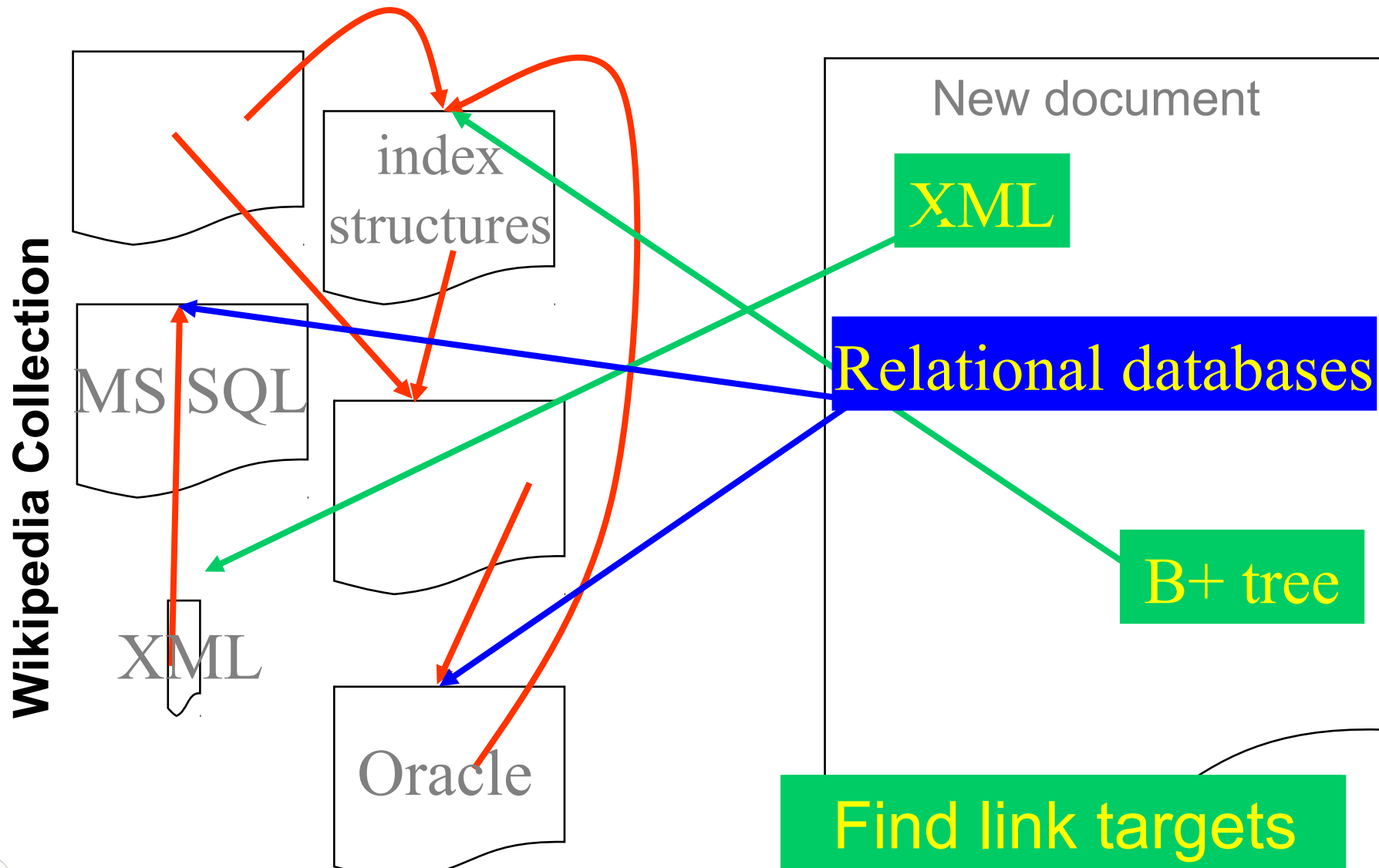


`/article[about(. , IR)]/sec[about(. , compression)]`

- Evaluation: Process resulting queries with search engine and compare result quality to CAS query in the topic

Link the Wiki

Goal: automated discovery of document links



Book Search

- Goals: Investigate
 - book-specific relevance ranking strategies
 - user interface issues

```
<document>
<page pageNumber="1" label="PT CHAPTER" [coords] key="0" id="0">
<region regionType="Text" [coords] key="0" id="0">
<section label="SEC BODY" key="408" id="0">
<line [coords] key="0" id="0">
<word [coords] key="0" id="0" val="Moby"/>
<word [coords] key="1" id="1" val="Dick"/>
</line>
<line [...]><word [...] val="Melville"/>[...]</line>[...]
</section> [...]
</region> [...]
</page> [...]
</document>
```

Social Book Search Task

- Goal: study relative value of authoritative metadata and user-generated content
- Collection:
 - meta data for 2.8 million books from Amazon
 - tags, ratings, reviews from LibraryThing (LT)
- Task: **Recommend books to read based on request in LT forum**
- Additional input: LT profile of requesting user
- Assessments:
 - books recommended by others
 - pooling + Mechanical Turk



Multimedia Track (now at CLEF)

- 2 Collections:
 - Wikipedia XML including image files (60GB)
 - Image Metadata (from Wikipedia)
- Additional precomputed information provided:
 - Classification scores for 101 categories (Aircraft, Racing, Walking, ...)
 - 120-dimensional feature vector (based on natural images statistics)

Multimedia Track

- Task 1: Retrieve document fragments for an info need with a multimedia character
`//section[about(.//figure//image,concept:maps)]`

(may include example images)
- Task 2: Pure image retrieval
(from the metadata collection)

Tweet Contextualization Track

Example: @alfred **#AlfredNobelPrize** ceremony this evening in **Scottsdale, AZ**

1 The Alfred Noble Prize is an award presented by the combined engineering societies of the United States, given each year to a person not over thirty-five for a paper published in one of the journals of the participating societies.

2 The prize was established in 1929 in honor of Alfred Noble, Past President of the American Society of Civil Engineers.

3 It has no connection to the Nobel Prize, although the two are often confused due to their similar spellings.

Snippet Retrieval Track

- **Goal:** generate informative snippets for search results
- Two-stage assessment:
 - Use snippet to predict relevance of result
 - Use result document to determine relevance

Linked Data Track

```
<jeopardy_clue>Niagara Falls has its source of origin  
from this lake. </jeopardy_clue>  
<keyword_title>Niagara Falls source lake</keyword_title>  
<sparql_ft>  
  select ?q Where {  
    <http://dbpedia.org/resource/Niagara_Falls>  
    <http://dbpedia.org/property/watercourse> ?o .  
    ?o <http://dbpedia.org/ontology/origin> ?q .  
    filter FTContains(?o, "river water course niagara") .  
    filter FTContains(?q, "lake origin of")}  
</sparql_ft>
```

Summary – Lessons Learned

Two main insights of the AdHoc track:

- Advantage (if any) of structured queries over content-only queries depends on collection & information need
- Focused retrieval is often not better than document (aka article-level) retrieval

Good understanding how to evaluate
adhoc search tasks on document-centric XML

Retrieval of data-centric documents
(or relational tables) largely unexplored

References

- INEX proceedings 2002-2012, <https://inex.mmci.uni-saarland.de/>
- Gabriella Kazai, Mounia Lalmas: eXtended cumulated gain measures for the evaluation of content-oriented XML retrieval. ACM Trans. Inf. Syst. 24(4): 503-542 (2006)
- Norbert Gövert, Norbert Fuhr, Mounia Lalmas, Gabriella Kazai: Evaluating the effectiveness of content-oriented XML retrieval methods. Inf. Retr. 9(6): 699-722 (2006)
- Mounia Lalmas, Gabriella Kazai, Jaap Kamps, Jovan Pehcevski, Benjamin Piwowarski, Stephen Robertson: INEX 2006 Evaluation Measures. INEX 2006: 20-34
- Gabriella Kazai, Mounia Lalmas, Arjen P. de Vries: The overlap problem in content-oriented XML retrieval evaluation. SIGIR 2004: 72-79
- Benjamin Piwowarski, Andrew Trotman, Mounia Lalmas: Sound and complete relevance assessment for XML retrieval. ACM Trans. Inf. Syst. 27(1) (2008)
- Mounia Lalmas, Anastasios Tombros: INEX 2002 - 2006: Understanding XML Retrieval Evaluation. DELOS Conference 2007: 187-196
- Sihem Amer-Yahia, Mounia Lalmas: XML search: languages, INEX and scoring. SIGMOD Record 35(4): 16-23 (2006)
- Andrew Trotman, Mounia Lalmas: Why structural hints in queries do not help XML-retrieval. SIGIR 2006: 711-712
- Andrew Trotman, Mounia Lalmas: Strict and vague interpretation of XML-retrieval queries. SIGIR 2006: 709-710
- Besnik Fetahu, Ralf Schenkel: Retrieval evaluation on focused tasks. SIGIR 2012: 1135-1136

