Intelligente und effiziente Suche auf semistrukturierten Daten

Gerhard Weikum

weikum@mpi-sb.mpg.de

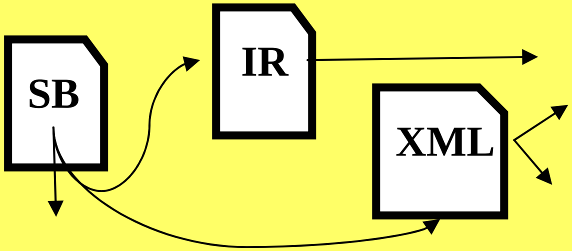
http://www.mpi-sb.mpg.de/~weikum/

Outline

- Motivation and Challenges
- XXL & XXL-light: IR on XML Data
- Role of Ontologies
- Efficient Evaluation of Top-k Queries
- Ongoing and Future Work

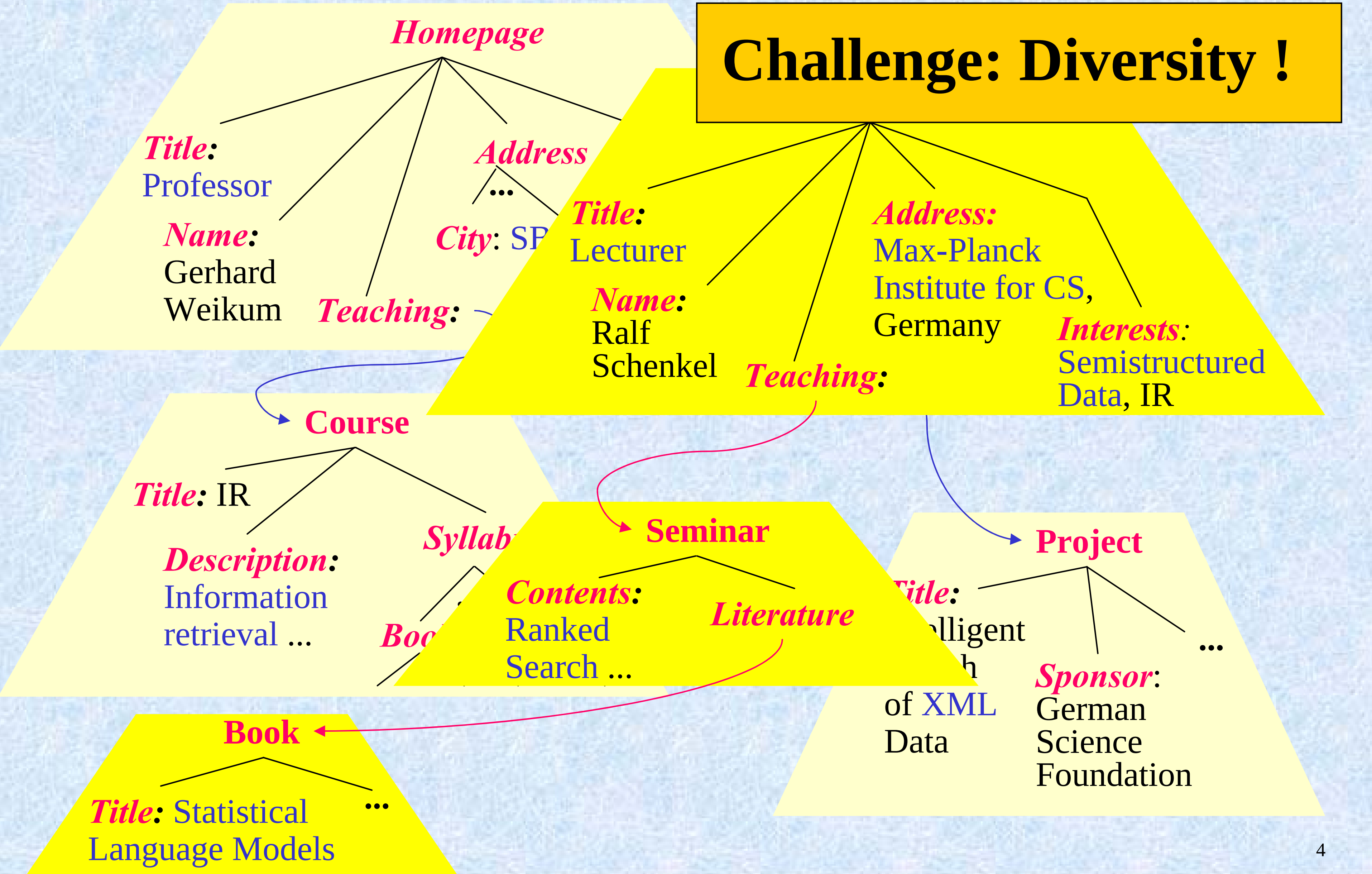
A Few Challenging Queries
(on Web / Deep Web / Intranet / Personal Info)

- **Which professors from Saarbruecken (SB) are teaching IR and have research projects on XML?**



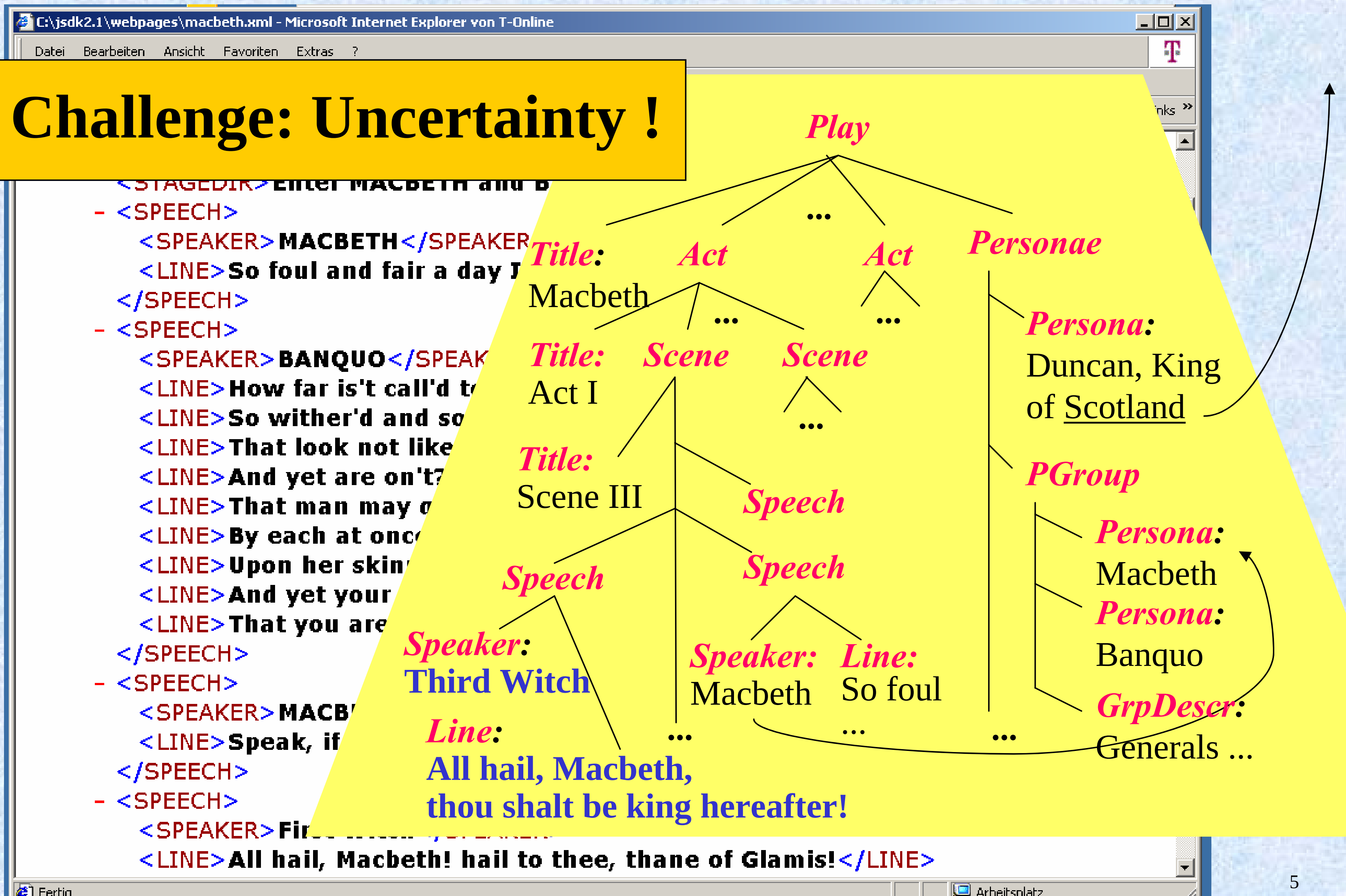
- **Which gene expression data from Barrett tissue in the esophagus exhibit high levels of gene A01g?**
- **What are the most important results on large deviation theory?**
- **Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king?**
- **Who was the French woman that I met at the PC meeting where Paolo Atzeni was PC Chair?**
- **Are there any published theorems that are equivalent to or subsume my latest mathematical conjecture?**

What if the Semantic Web Existed and All Information Were in XML?

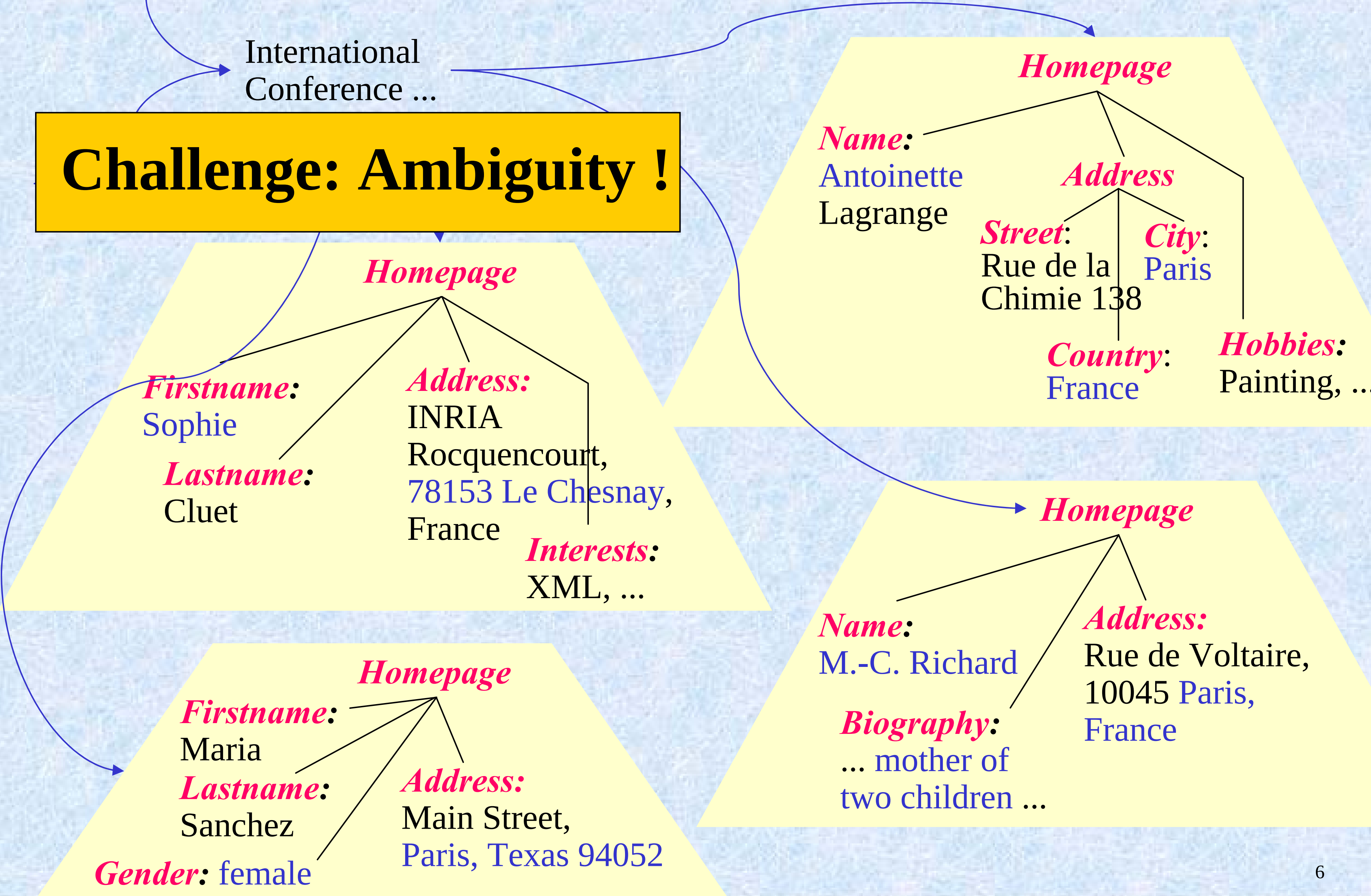


What if the Semantic Web Existed and All Information Were in XML?

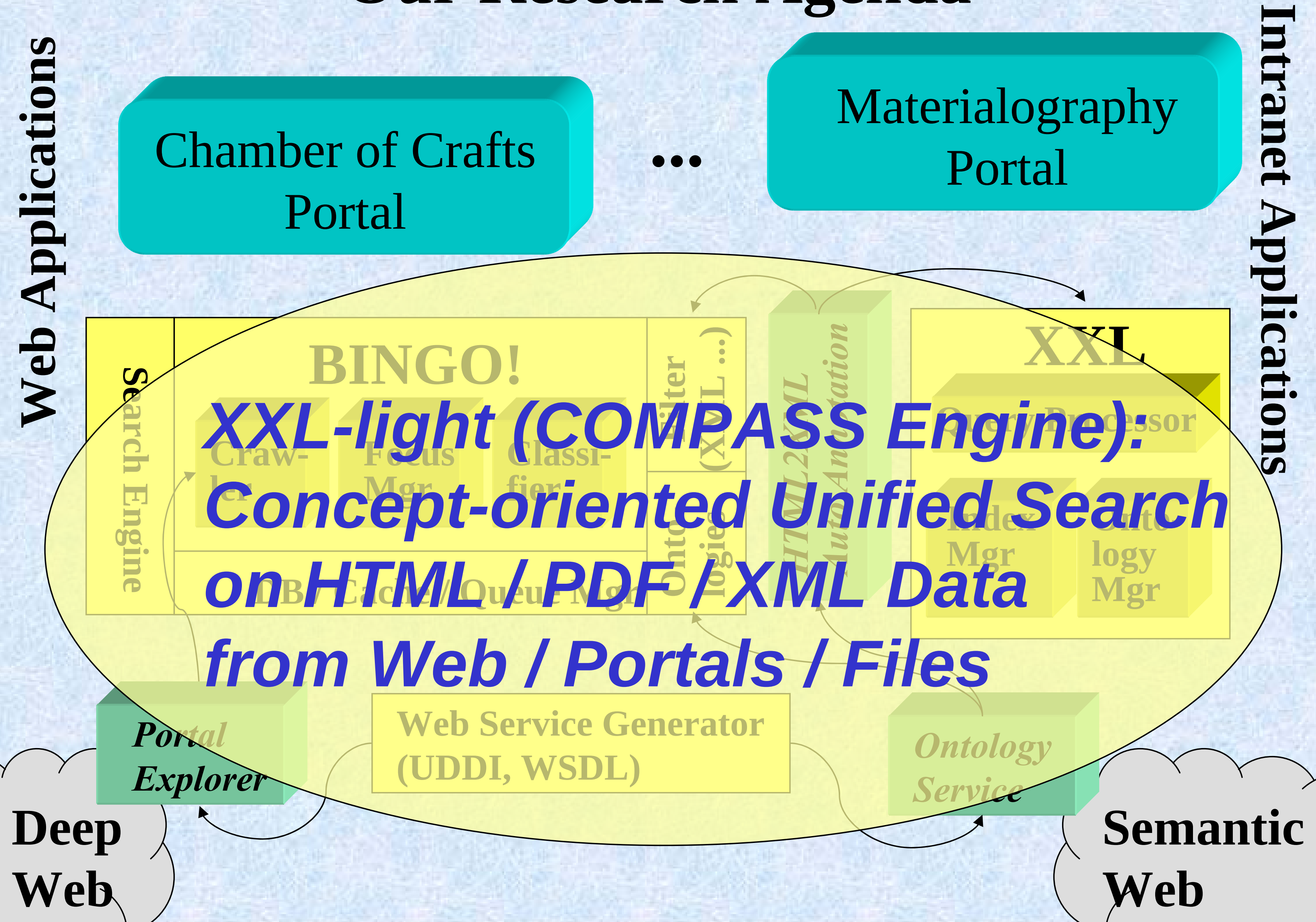
Challenge: Uncertainty !



What if the Semantic Web Existed and All Information Were in XML?



Our Research Agenda

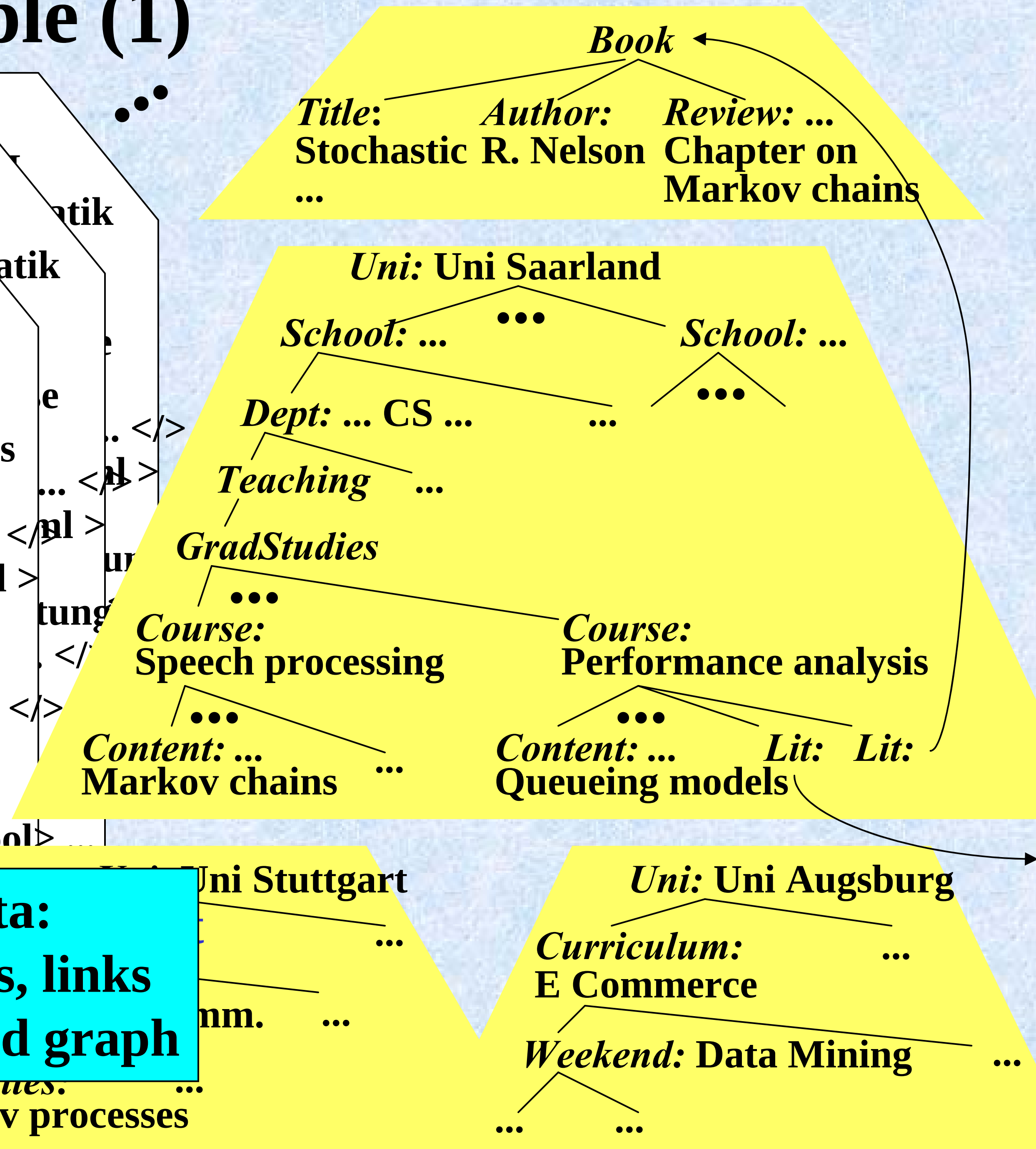


Outline

- ✓ Motivation and Challenges
- XXL & XXL-light: IR on XML Data
- Role of Ontologies
- Efficient Evaluation of Top-k Queries
- Ongoing and Future Work

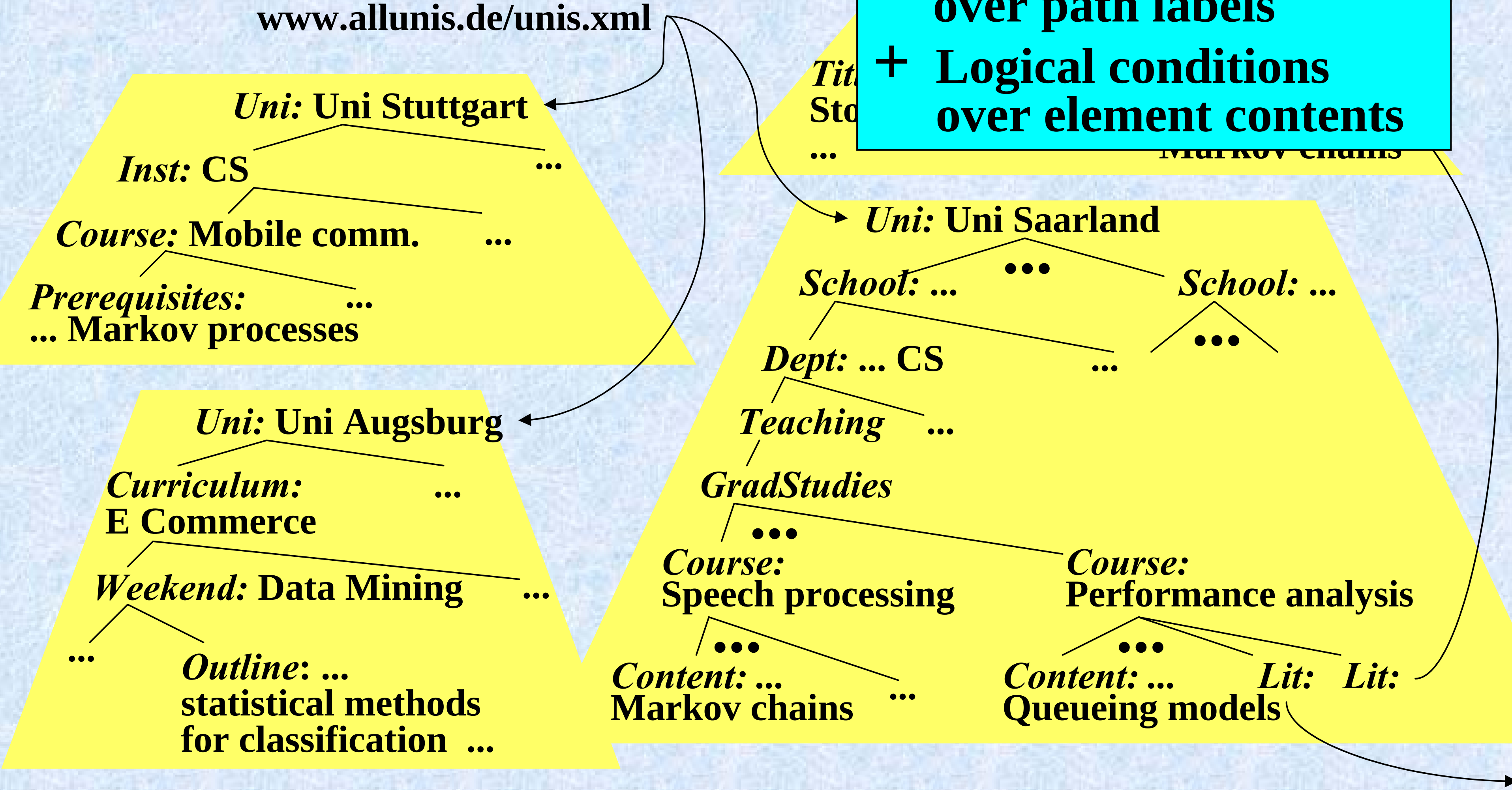
XML-IR Example (1)

<*Uni*> Uni Saarland
<*School*> Math & Engineering
<*Dept*> CS
<*Teaching*> ...
<*GradStudies*>
<*Course*> Performance analysis
<*Lecturer*> ... </>
<*Content*> Queueing models .. </>
<*Lit* href=springer/nelson.xml >
<*Lit* href=... > </*Course*>
<*Course*> Speech processing
<*Content*> ... Markov chains... </>
</*Course*>
...
</*Teaching*> .. </Dept> .. </School> ...
</Uni...



Semistructured data: elements, attributes, links organized as labeled graph

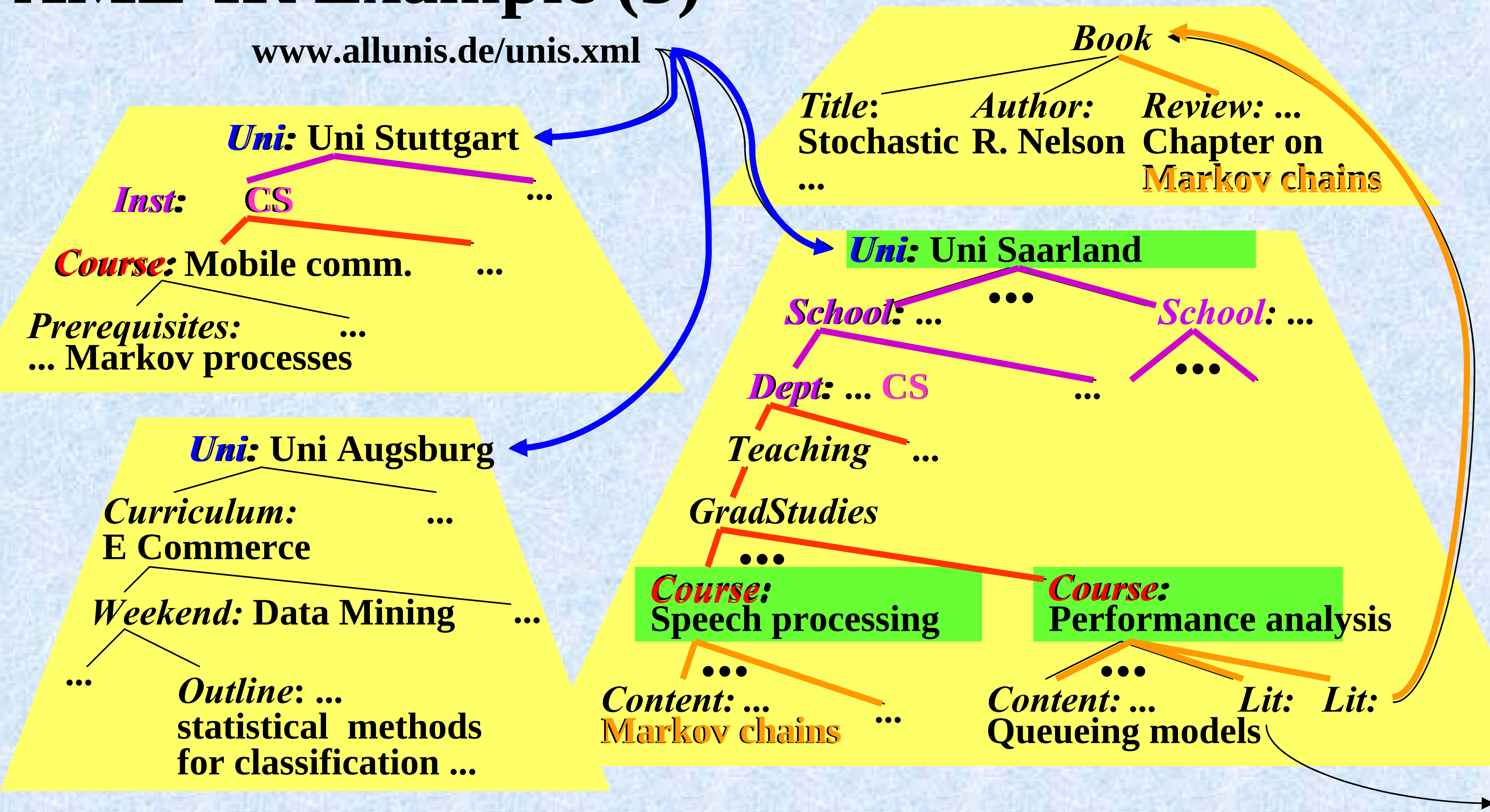
XML-IR Example (2)



Regular expressions over path labels
+ Logical conditions over element contents

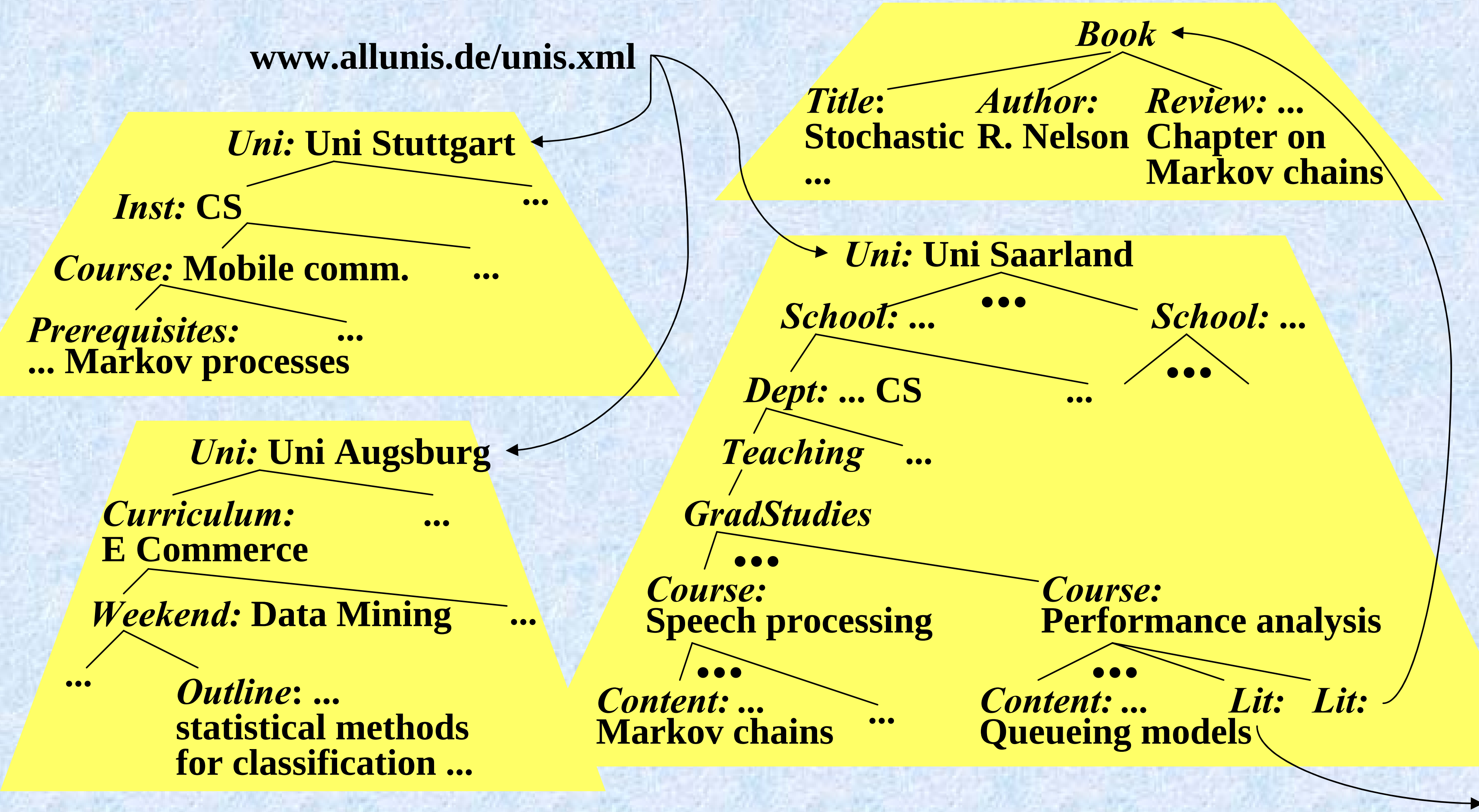
Select U, C From www.allunis.de/unis.xml Where Uni As U And U.#.School?.#.(Inst | Dept)+ As D And D Like „%CS%“ And D.#.Course As C And C.# Like „%Markov chain%“

XML-IR Example (3)



Select U, C From www.allunis.de/unis.xml Where Uni As U
And U.#.School?.#.(Inst | Dept)+ As D And D Like „%CS%“
And D.#.Course As C And C.# Like „%Markov chain%“

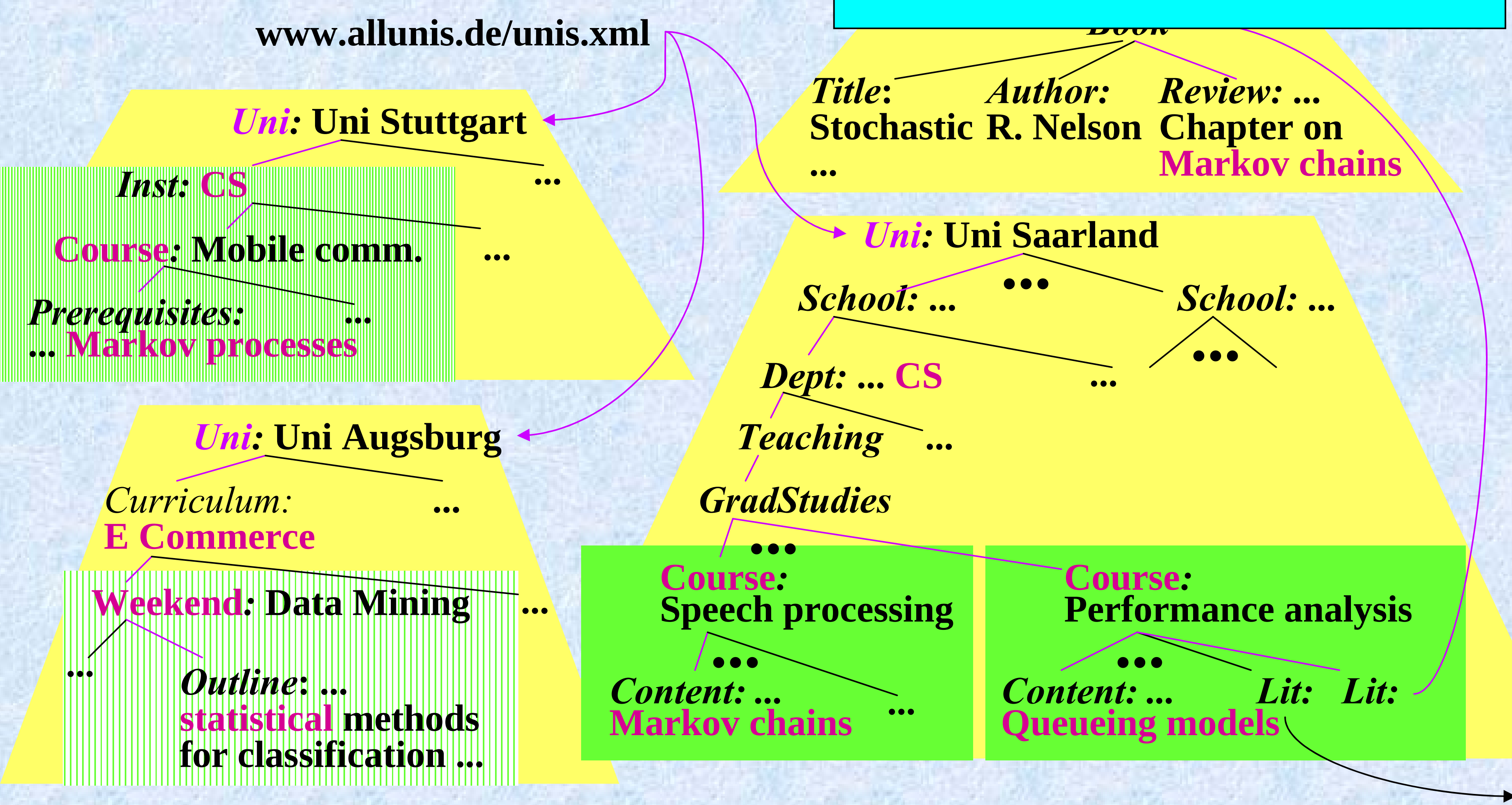
XML-IR Example (4)



Select U, C From www.allunis.de/unis.xml Where Uni As U
And U.# As D And D ~ „CS“
And D.#.~Course As C And C.# ~ „Markov chain“

XML-IR Example (5)

Result ranking of XML data based on semantic similarity



Select U, C From www.allunis.de/unis.xml Where Uni As U
And U.# As D And D ~ „CS“
And D.#.~Course As C And C.# ~ „Markov chain“

XML-IR Concepts

Example COMPASS
(Concept-Oriented Multi-Format Por
simple, extensible core language – app

Query Semantics:
- **query is a pattern with relaxable conditions**
- **results are approximate matches to query with similarity scores**

Where clause: conjunction of restricte
with binding of variabl

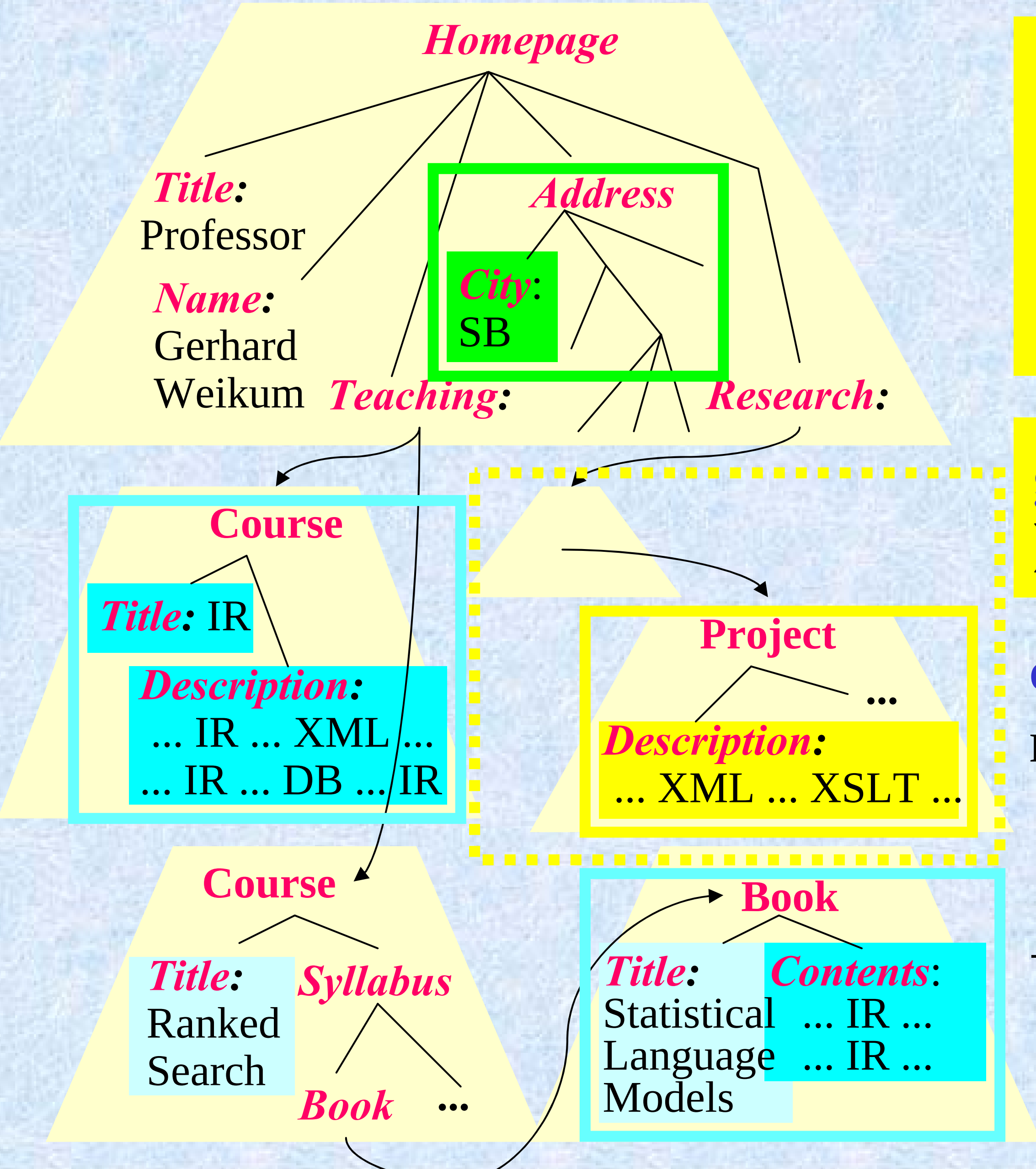
Elementary conditions on names and contents

Select *P, C, R* From *index*
Where *~professor* As *P*
And *P = „Saarbruecken“*
And *P//~course = „Information Retrieval“* As *C*
And *P//~research = „~XML“* As *R*

Semantic similarity conditions on names and contents
~research = „~XML“

Relevance scoring based on
tf*idf similarity of contents,
ontological similarity of names,
probabilistic combination of conditions

XML-IR Scoring Model

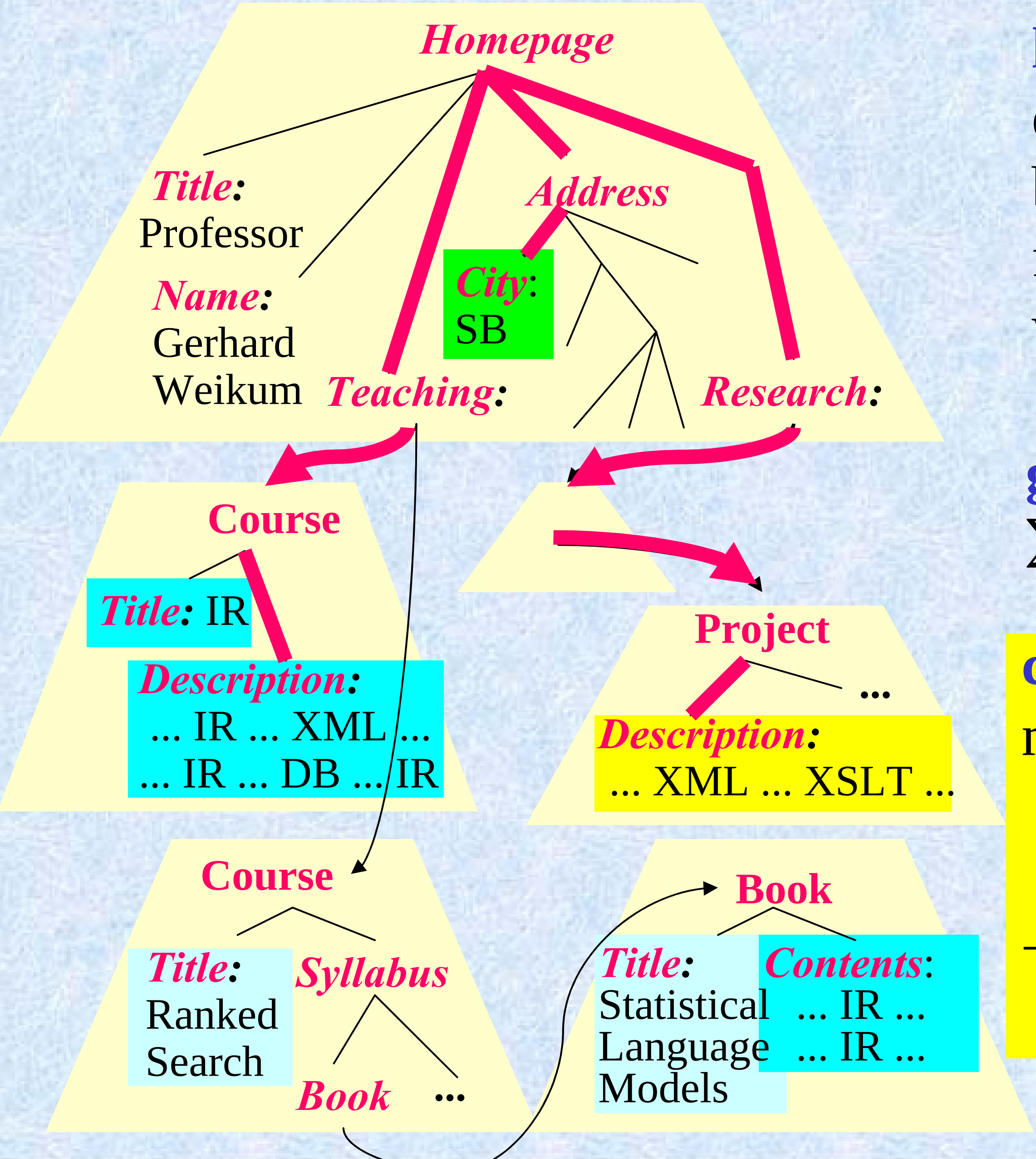


local score for elementary condition: based on tf*idf-style statistics for node or node context with score propagation

global score for query: $\sum$ local scores * compactness

compactness of result:
max{$\sum$ node & edge weights | graph connecting matching nodes}
→ generalized MST (related to Steiner trees)

XML-IR Scoring Model

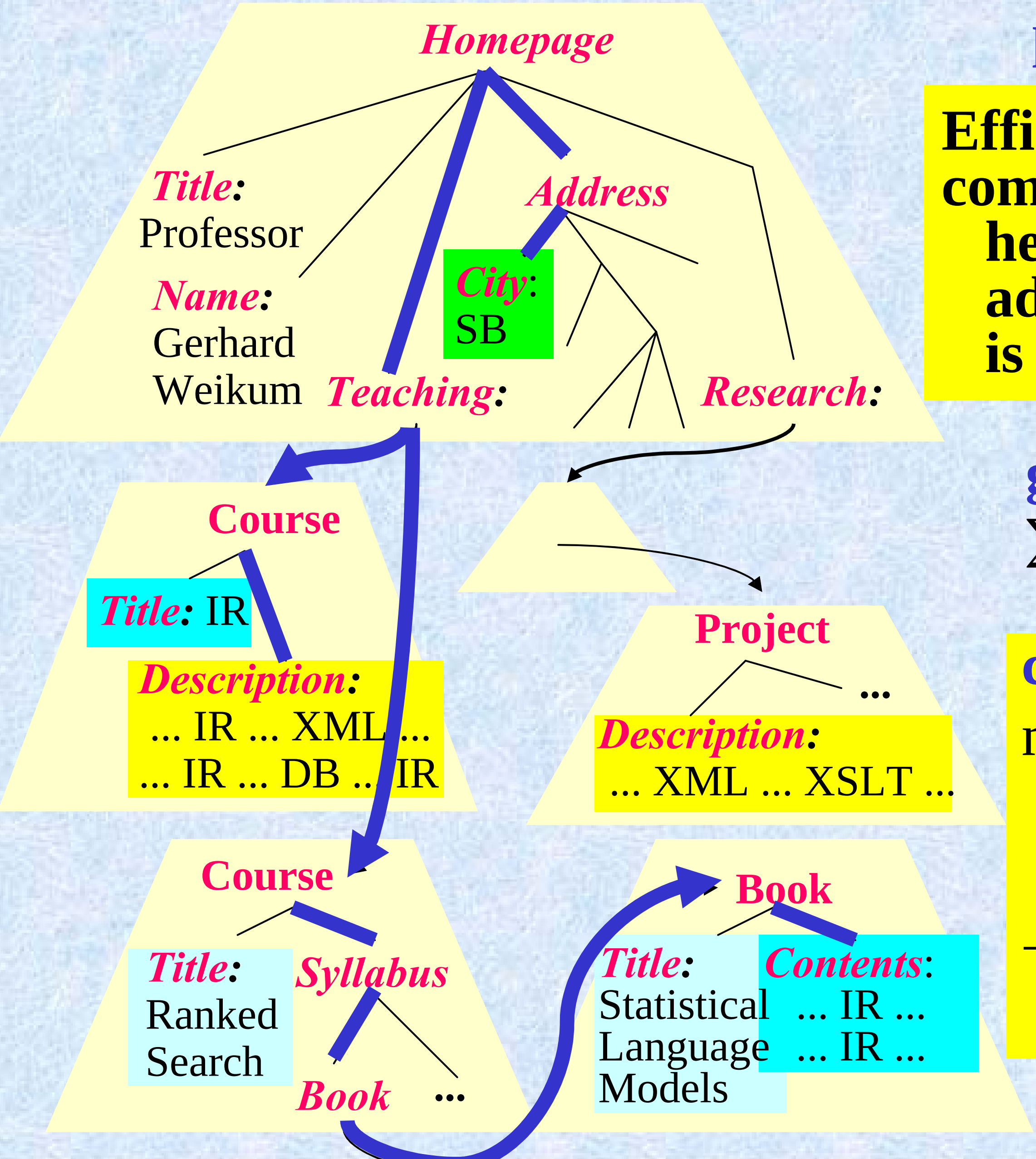


local score for elementary condition: based on tf*idf-style statistics for node or node context with score propagation

global score for query: $\sum$ local scores * compactness

compactness of result: max{$\sum$ node & edge weights | graph connecting matching nodes}
→ generalized MST (related to Steiner trees)

XML-IR Scoring Model



local score for

statistics

ext

Efficient score computation:
heuristics work;
advanced algorithms is open issue

global score for query:
$\sum$ local scores * compactness

compactness of result:
max{$\sum$ node & edge weights | graph connecting matching nodes}
$\rightarrow$ generalized MST (related to Steiner trees)

Outline

- ✓ Motivation and Challenges
- ✓ XXL & XXL-light: IR on XML Data
- Role of Ontologies
- Efficient Evaluation of Top-k Queries
- Ongoing and Future Work

On Thesauri and Ontologies

Taxonomy: classification of concepts into groups (and trees of groups)

Thesaurus: repository („treasure“) of synonyms
(and other relationships between words and concepts)

Ontology: metaphysical study of the nature of being & existence

Ontology (new definition): structured repository of knowledge
with a description of concepts and relationships,
possibly in the form of description logics formula

XML schemas, DTDs, namespaces:
syntactic conventions and standardized naming (plus typing info)

Gazetteer: (geographical) dictionary of names

Reasoning on Ontologies and Thesauri:

Professor $\subseteq$ Lecturer $\cap \exists$ hasStaff.Secretary
Teaching $\supseteq$ Cou
Professor $\subseteq$ Acac
Academician $\subseteq$ H
Human $\subseteq$ Carniv
...

poor man‘s ontology:
pragmatic, rich, efficient

→ logical inferences
with sub-FOL calculus

→ transitive closures,
shortest paths, etc.
along generalizations

Example WordNet

WordNet 1.6 Browser

File History Options Help

Search Word: woman

Redisplay Overview

Searches for woman: Noun

Senses:

1 of 4 senses of woman

Sense 1
woman, adult female -- (an adult female person (as opposed to a man); "the woman kept house while the man hunted")
    => Eve -- ((Old Testament) Adam's wife in Judeo-Christian mythology: the first woman and mother of the human race; God created Eve from Adam's rib and placed Ada
    => black woman -- (a woman who is Black)
    => white woman -- (a woman who is White)
    => yellow woman -- (offensive term for an Oriental woman)

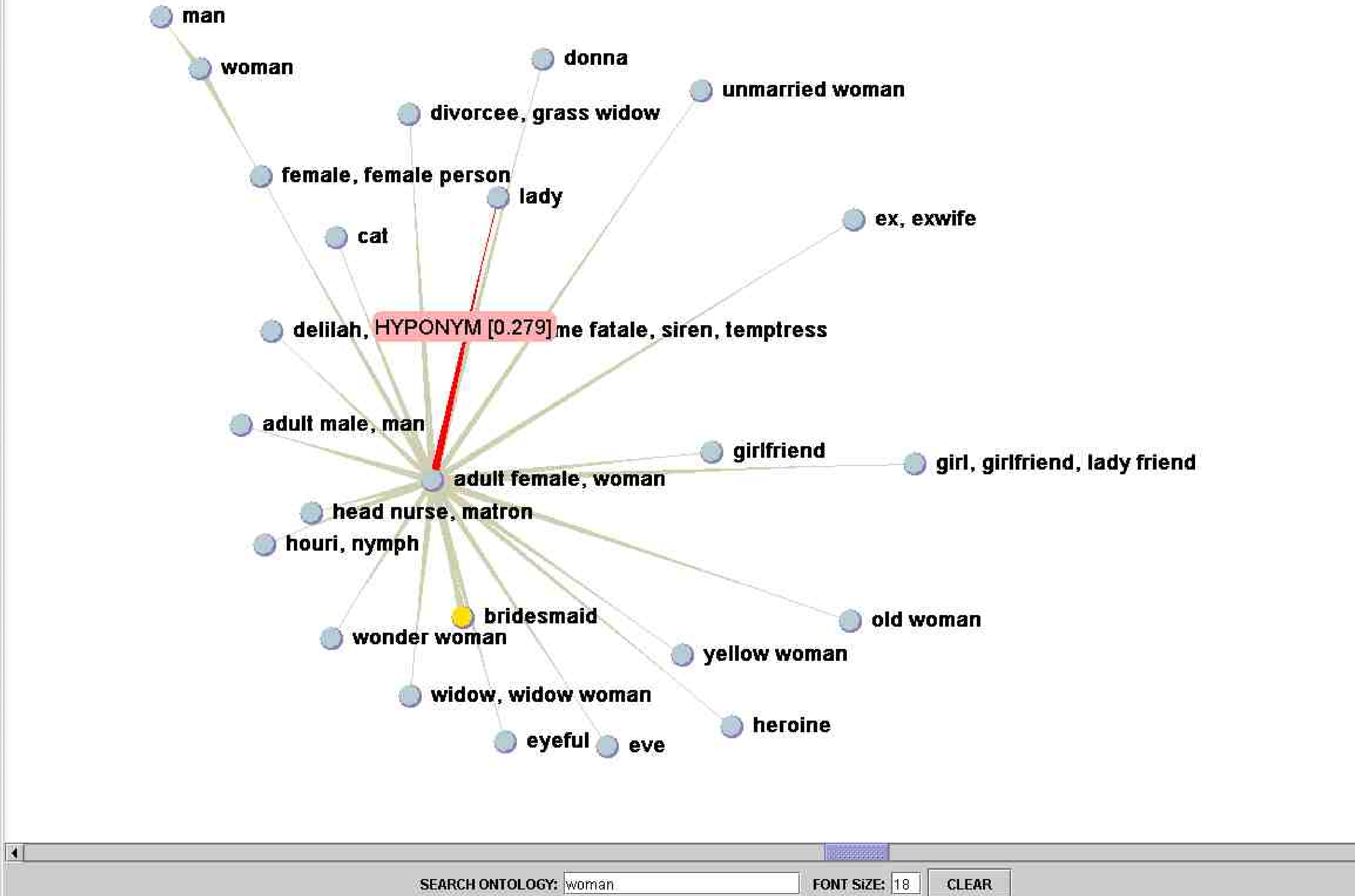
woman, adult female – (an adult female person)
=> amazon, virago – (a large strong and aggressive woman)
=> donna -- (an Italian woman of rank)
=> geisha, geisha girl -- (...)
=> lady (a polite name for any woman)
...
=> wife – (a married woman, a man's partner in marriage)

=> witch – (a being, usually female, imagined to have special powers derived from the devil)

    => maenad -- (an unnaturally frenzied or distraught woman)
    => matron, head nurse -- (a woman in charge of nursing in a medical institution)

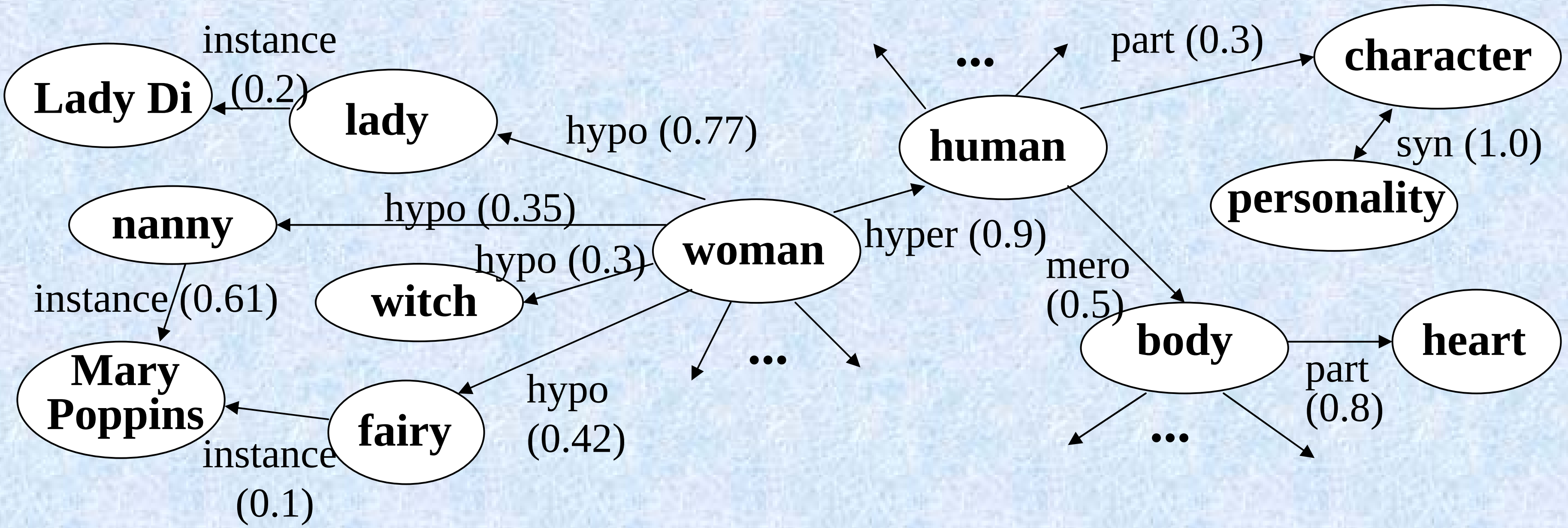
"Hyponyms (...is a kind of this), brief" search for noun "woman"

Ontology Visualization



Ontology Graph

An ontology graph is a directed graph with concepts (and their descriptions) as nodes and semantic relationships as edges (e.g., hypernyms).



Weighted edges capture strength of relationships
→ key for identifying closely related concepts

Statistics for Weighted Ontological Relations

Gather statistics from large corpus or by (focused) Web crawl

Various correlation measures for sim(c1, c2):

Dice coefficient:

$$\frac{2\left|\{\,docs\ with\ c1\,\}\cap\{\,docs\ with\ c2\,\}\right|}{\left|\{\,docs\ with\ c1\}\right|\ +\ \left|\{\,docs\ with\ c2\,\}\right|}$$

Jaccard coefficient:

$$\frac{\left|\{\,docs\ with\ c1\,\}\cap\{\,docs\ with\ c2\,\}\right|}{\left|\{\,docs\ with\ c1\}\right|\ +\ \left|\{\,docs\ with\ c2\}\right|-\left|\{\,docs\ with\ c1\ and\ c2\,\}\right|}$$

Conditional probabilites:

$$P[\,doc\ has\ c1\mid doc\ has\ c2\,]$$

Transitive similarity:

$$sim^*(c1,\,cn) = \ \ max\{\ \prod_{i=1..n-1} sim(c_i, c_{i+1}) \mid all\ \ paths\ \ from\ c1\ to\ cn\,\}$$

compute by (adaptation of) Dijkstra‘s shortest-path algorithm

Benefits from Ontology Service

Ontology service accessible via SOAP or RMI
Ontology filled with WordNet, geo gazetteer,
focused crawl results, extracted tables & forms

usefor for:

- Threshold-based query expansion
- Query keyword disambiguation
- Support for automatic tagging of HTML and enhanced XML tags
- Mapping of concept-value query conditions onto Deep-Web portals

Query Expansion

Threshold-based query expansion:
substitute ~w by (c_1 | ... | c_k) with all c_i for which $sim(w, c_i) \geq \delta$

„Old hat" in IR; highly disputed for danger of topic dilution

Approach to careful expansion:
- determine phrases from query or best initial query results (e.g., forming 3-grams and looking up ontology/thesaurus entries)
- if uniquely mapped to one concept then expand with synonyms and weighted hyponyms

Query Expansion Example

From TREC 2004 Robust Track:

Title: International Organized Crime

Description: Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.

Query = {international[0.145|1.00],
~META[1.00|1.00][{gangdom[1.00|1.00], gangland[0.742|1.00],
"organ[0.213|1.00] & crime[0
mafia[0.154|1.00], "sicilian[0.
"black[0.066|1.00] & hand[0.
organ[0.213|1.00], crime[0.31
columbian[0.686|0.20], cartel

Let us take, for example, the case of Medellin cartel's boss Pablo Escobar. Will the fact that he was eliminated change anything at all? No, it may perhaps have a psychological effect on other drug dealers but,
...

... for organizing the illicit export of metals and import of arms. It is extremely difficult for the law-enforcement organs to investigate and stamp out corruption among leading officials.
...

A parliamentary commission accused Swiss prosecutors today of doing little to stop drug and money-laundering international networks from pumping billions of dollars through Swiss companies.
...

135530 sorted accesses in 11.073

Results:

1. Interpol Chief on Fight Agai
2. Economic Counterintelligen
3. Supreme Procuratorate Wor
4. Crime and Punishment in th
5. SWITZERLAND CALLED

...

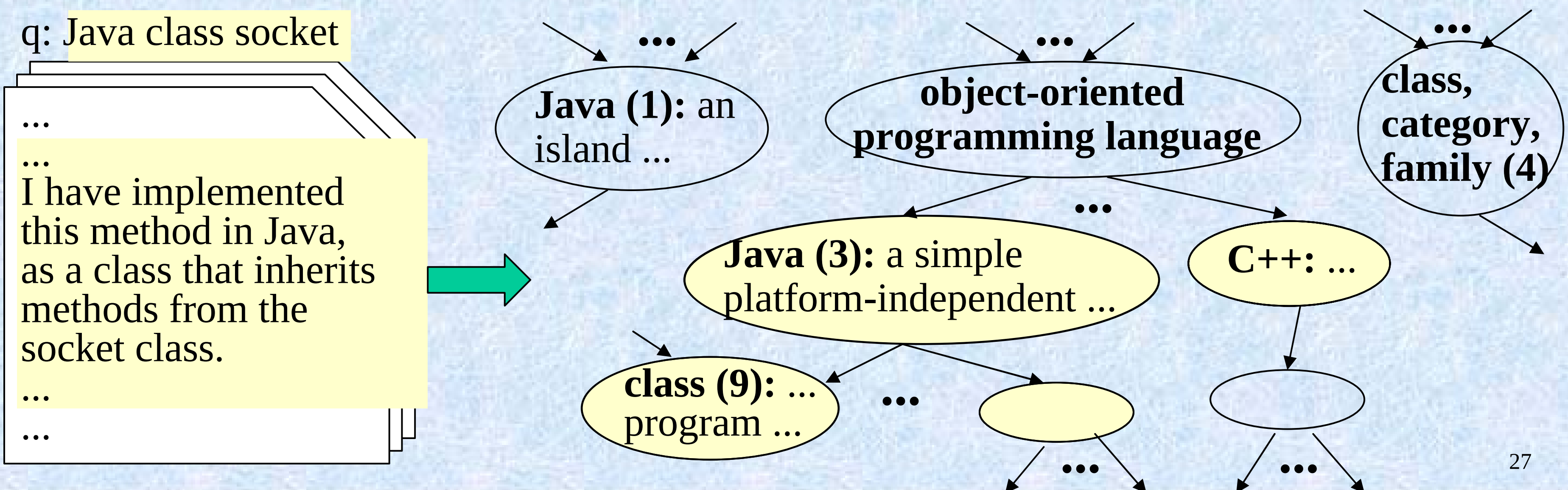
Keyword-to-Concept Mapping and Word Sense Disambiguation

Example: „Java class socket“ vs. „Java beach snorkeling“
Which concept should „Java“ be mapped to for query expansion?

Note: unlike in LSI or pLSI, concepts are explicit, not latent!

Approach for query keyword disambiguation:

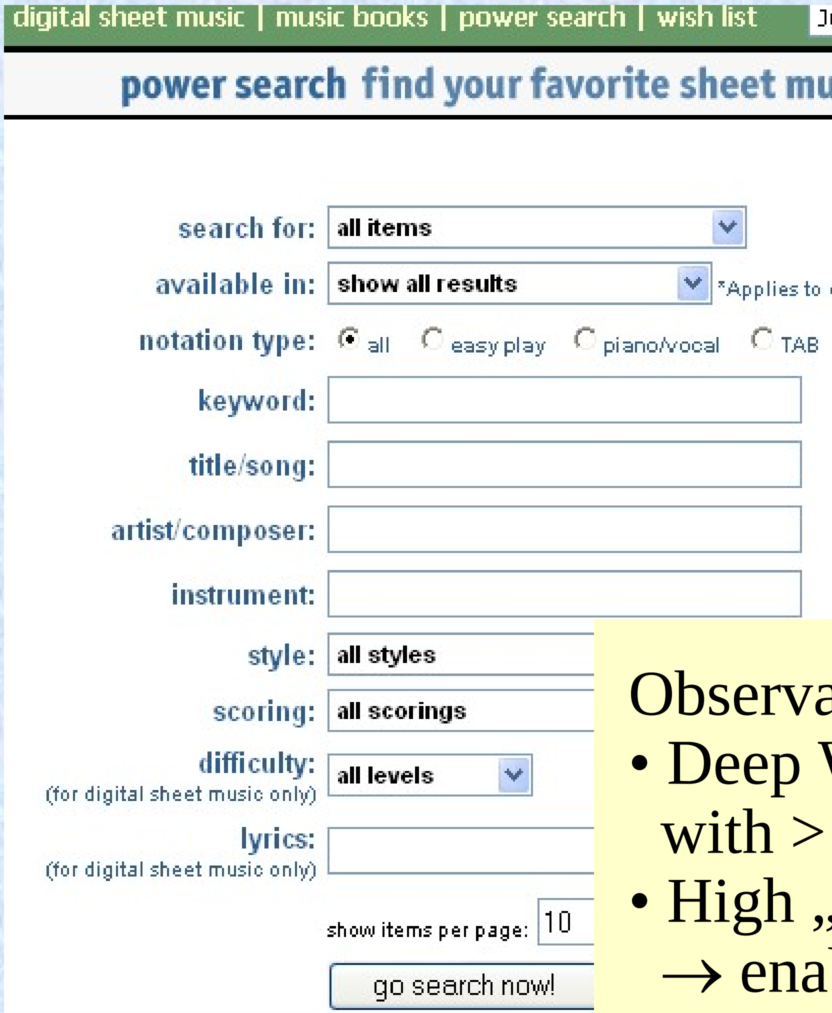
- form contexts con(w) and con(c_i) for keyword w and potential target concepts $c_i \in \{c_1, ..., c_k\}$
- bag-of-words similarity sim(con(w), con(c)) based on cos or KL diff
- choose concept argmax_c {sim(con(w), con(c))}



What About Deep Web and Web Services?

Mapping of concept-value query conditions onto Deep-Web portals:

~sheetmusic = „~flute“ → instrument = (flute | piccolo | recorder)
→ category = reeds
→ style = (classical | jazz | folk)




```
<element name="WSF_Form0Select0_Enum">
    <simpleType>
        <restriction base="string">
</restriction>
    </simpleType>
  </element>
  <simpleType name="WSF_Form0Select1_Enum">
    <restriction base="string">
        <enumeration value=„Alternative"/>
        <enumeration value=„Blues"/>
        <enumeration value=„Children‘s"/>
        <enumeration value=„Classical"/>
        <enumeration value=„Country"/>
```

Observations:

- Deep Web has > 500 000 hidden databases with > 500 billion ($5*10^{11}$) dynamic pages
- High „redundancy“ among query forms → enables exploitation of statistics

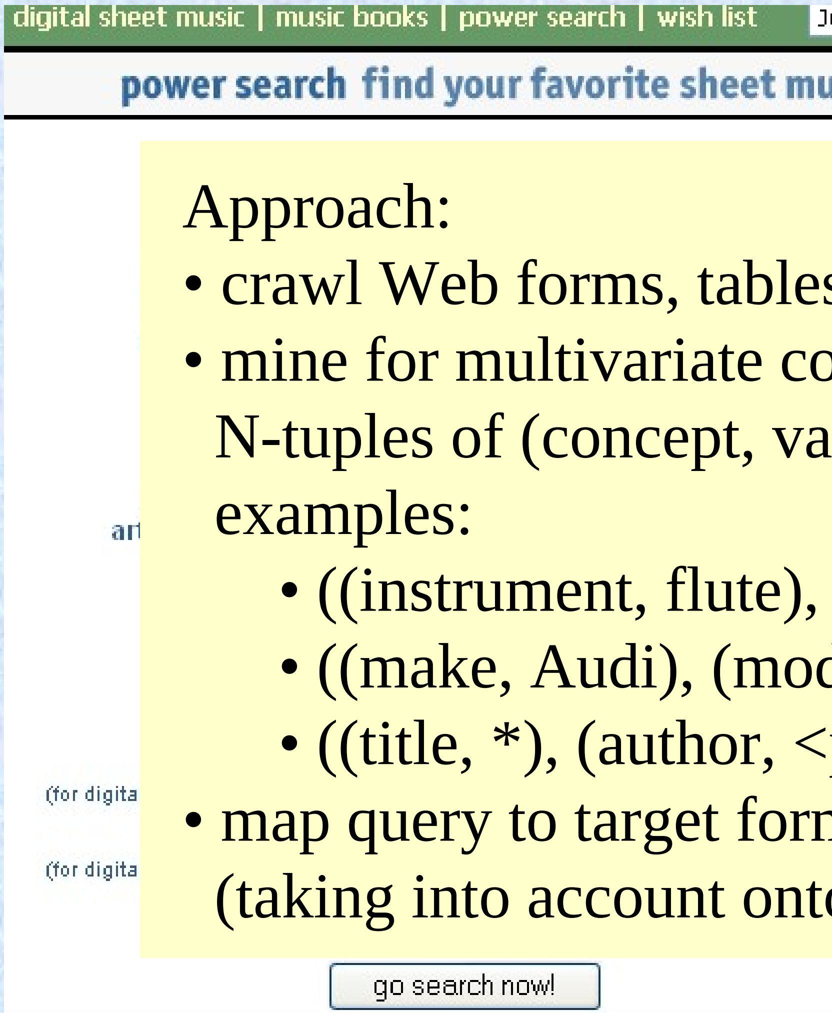
What About Deep Web and Web Services?

Mapping of concept-value query conditions onto Deep-Web portals:

~sheetmusic = „~flute“ → instrument = (flute | piccolo | recorder)

→ category = reeds

→ style = (classical | jazz | folk)



<element name="WSF_Form0Select0_Enum">
<simpleType>
<restriction base="string">
Enum">

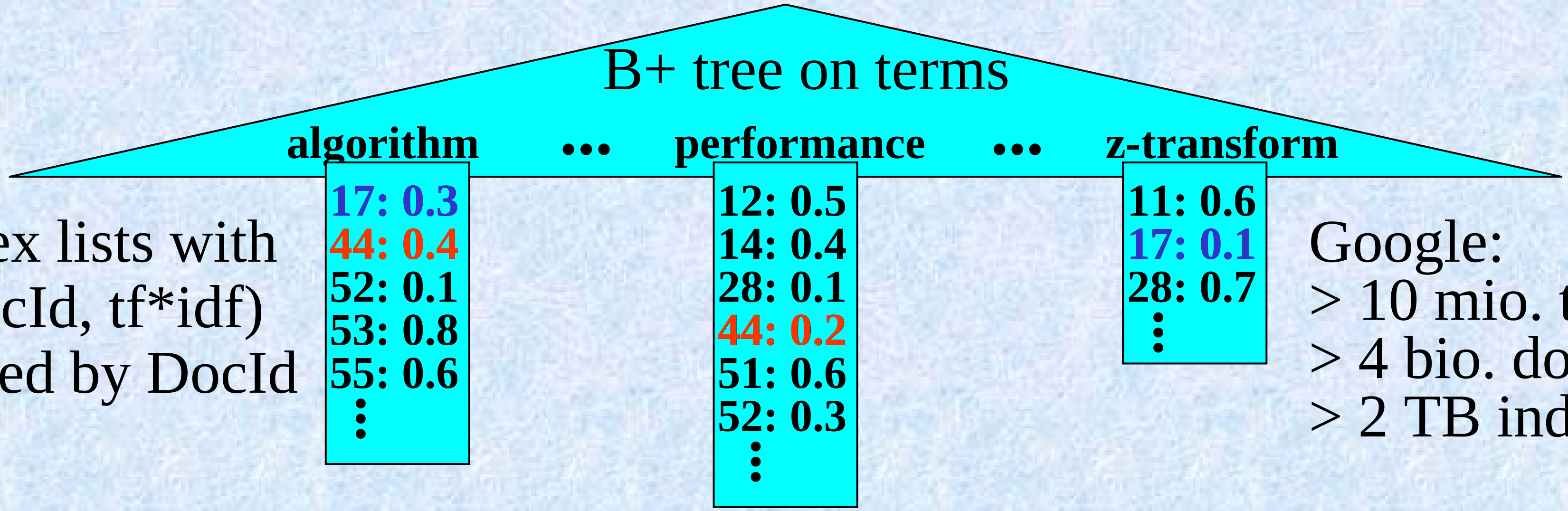
Approach:

- crawl Web forms, tables, WSDL
- mine for multivariate correlations among N-tuples of (concept, value) pairs, examples:
  - ((instrument, flute), (category, reeds))
  - ((make, Audi), (model, A4), (color, red))
  - ((title, *), (author, <person name>))
- map query to target form/WSDL for max likelihood (taking into account ontological similarities)

Outline

- ✓ Motivation and Challenges
- ✓ XXL & XXL-light: IR on XML Data
- ✓ Role of Ontologies
- Efficient Evaluation of Top-k Queries
- Ongoing and Future Work

Top-k Query Processing with Scoring



index lists with (DocId, tf*idf) sorted by DocId

Google:
> 10 mio. terms
> 4 bio. docs
> 2 TB index

Given: query q = t1 t2 ... tz with z (conjunctive) keywords
similarity scoring function score(q,d) for docs d∈D, e.g.: $\vec{q} \cdot \vec{d}$

Find: top k results with regard to score(q,d) (e.g.: $\Sigma_{i \in q}\, s_i(d)$)

Naive QP algorithm:

candidate-docs := ∅;
for i=1 to z do {
    candidate-docs := candidate-docs ∪ index-lookup(ti) };
for each dj ∈ candidate-docs do {compute score(q,dj)};
sort candidate-docs by score(q,dj) descending;

“Fagin’s TA” (Fagin’01; Güntzer/Kießling/Balke; Nepal et al.)

scan all lists L_i (i=1..m) in parallel:
consider dj at position pos_i in Li;
$high_i := s_i(dj)$;
if dj $\notin$ top-k then {
look up $s_\nu(dj)$ in all lists L_ν with $\nu \neq i$; // random access
compute $s(dj) := aggr \{s_\nu(dj) \mid \nu=1..m\}$;
if s(dj) > min score among top-k then
add dj to top-k and remove min-score d from top-k; };
threshold := aggr $\{high_\nu \mid \nu=1..m\}$;
if min score among top-k ≥ threshold then exit;

but random accesses are expensive !

m=3
aggr: sum
k=2

f: 0.5	a: 0.55	h: 0.35	**top-k:**
b: 0.4	b: 0.2	d: 0.35	
c: 0.35	f: 0.2	b: 0.2	~~f: 0.75~~
a: 0.3	g: 0.2	a: 0.1	a: 0.95
h: 0.1	c: 0.1	c: 0.05	b: 0.8
d: 0.1		f: 0.05	

applicable to XML data:
course ~ „Internet“ and ~topic = „performance“

TA-Sorted

scan index lists in parallel:

consider dj at position pos_i in Li;

E(dj) := E(dj) $\cup$ {i}; $high_i$:= si(q,dj);

bestscore(dj) := aggr{x1, ..., xm)

with xi := si(q,dj) for i$\in$E(dj), $high_i$ for i $\notin$E(dj);

worstscore(dj) := aggr{x1, ..., xm)

with xi := si(q,dj) for i$\in$E(dj), 0 for i $\notin$E(dj);

top-k := k docs with largest worstscore;

threshold := bestscore{d | d not in top-k};

if min worstscore among top-k $\geq$ threshold then exit;

m=3
aggr: sum
k=2

f: 0.5	a: 0.55	h: 0.35
b: 0.4	b: 0.2	d: 0.35
c: 0.35	f: 0.2	b: 0.2
a: 0.3	g: 0.2	a: 0.1
h: 0.1	c: 0.1	c: 0.05
d: 0.1		f: 0.05

top-k:

a: 0.95
b: 0.8

candidates:

~~f: 0.7 + ? $\leq$ 0.7 + 0.1~~
~~h: 0.45 + ? $\leq$ 0.45 + 0.2~~
~~c: 0.35 + ? $\leq$ 0.35 + 0.3~~
~~d: 0.35 + ? $\leq$ 0.35 + 0.3~~
~~g: 0.2 + ? $\leq$ 0.2 + 0.4~~

Top-k Queries with Probabilistic Guarantees

TA family of algorithms based on invariant (with sum as aggr)

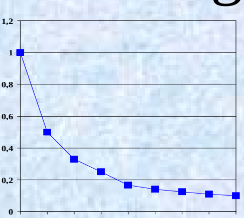
$$\sum_{i \in E(d)} s_i(d) \le s(d) \le \sum_{i \in E(d)} s_i(d) + \sum_{i \notin E(d)} high_i$$

Relaxed into probabilistic invariant

$$p(d) := P[s(d) > \delta] = P[\sum_{i \in E(d)} s_i(d) + \sum_{i \notin E(d)} S_i > threshold]$$

$$= P[\sum_{i \notin E(d)} S_i > threshold - \sum_{i \in E(d)} s_i(d)] =: P[\sum_{i \notin E(d)} S_i > \delta'] \le \varepsilon$$

where the RV S_i has some (postulated and/or estimated) distribution in the interval $(0, high_i]$



$high_i$

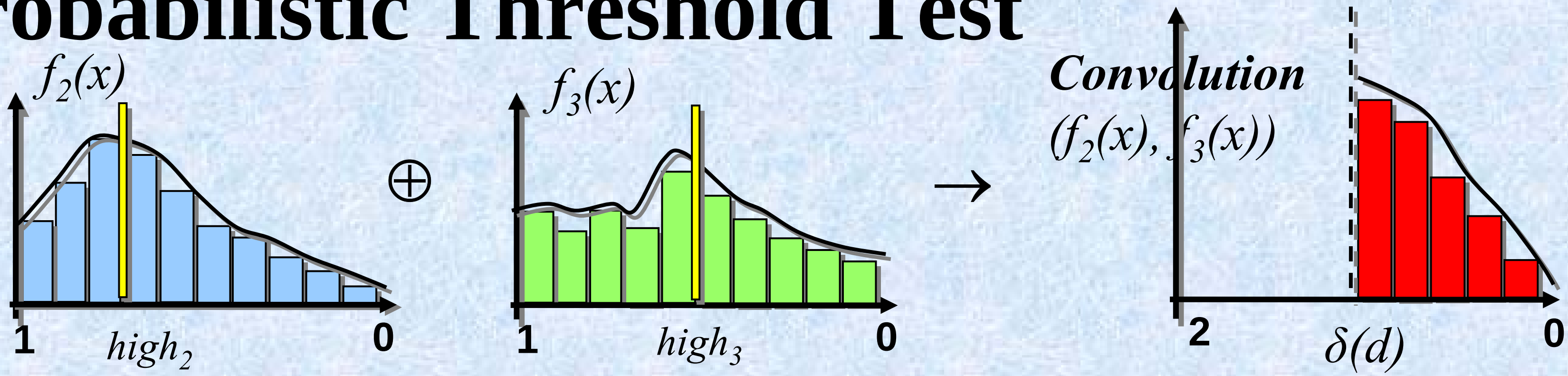
f: 0.5	a: 0.55	h: 0.35
b: 0.4	b: 0.2	d: 0.35
c: 0.35	f: 0.2	b: 0.2
a: 0.3	g: 0.2	a: 0.1
h: 0.1	c: 0.1	c: 0.05
d: 0.1		f: 0.05

S_1 (c: 0.35 … d: 0.1), S_2 (f: 0.2 … c: 0.1), S_3 (b: 0.2 … f: 0.05)

- ***Discard candidates with $p(d) \le \varepsilon$***
- **Exit index scan when candidate list empty**

Probabilistic Threshold Test

cand doc d with
$2 \notin E(d)$,
$3 \notin E(d)$



- **postulating *uniform or Zipf* score distribution in $[0, high_i]$**
  - **compute convolution using LSTs**
  - **use Chernoff-Hoeffding tail bounds or generalized bounds for correlated dimensions (Siegel 1995)**
- **fitting *Poisson* distribution (or Poisson mixture)**
  - **over equidistant values:** $P[d = v_j] = e^{-\alpha_i} \frac{\alpha_i^{\,j-1}}{(j-1)!}$
  - **easy and exact convolution**
- **distribution approximated by *histograms*:**
  - **precomputed for each dimension**
  - **dynamic convolution at query-execution time**

engineering-wise histograms work best!

with *independent* Si's or with *correlated* Si's

Prob-sorted Algorithm (Smart Variant)

Prob-sorted (RebuildPeriod r, QueueBound b):

...

scan all lists Li (i=1..m) in parallel:

…same code as TA-sorted…

// *<u>queue management</u>*

for all priority queues q for which d is relevant do

insert d into q with priority bestscore(d);

// *<u>periodic clean-up</u>*

if step-number mod r = 0 then

// *<u>rebuild; single bounded queue</u>*

if strategy = Smart then

for all queue elements e in q do

update bestscore(e) with current high_i values;

rebuild bounded queue with best b elements;

if prob[top(q) can qualify for top-k] < ε then exit;

if all queues are empty then exit;

Performance Results for .Gov Queries

on .GOV corpus from TREC-12 Web track:
1.25 Mio. docs (html, pdf, etc.)

50 keyword queries, e.g.:
- ***„Lewis Clark expedition“,***
- ***„juvenile delinquency“,***
- ***„legalization Marihuana“,***
- ***„air bag safety reducing injuries death facts“***

speedup by factor 10 at high precision/recall (relative to TA-sorted);

aggressive queue mgt. even yields factor 100 at 30-50 % prec./recall

	TA-sorted	Prob-sorted (smart)
#sorted accesses	2,263,652	527,980
elapsed time [s]	148.7	15.9
max queue size	10849	400
relative recall	1	0.69
rank distance	0	39.5
score error	0	0.031

.Gov Expanded Queries

on .GOV corpus with query expansion based on WordNet synonyms:

50 keyword queries, e.g.:

- **„juvenile delinquency *youth minor crime law jurisdiction offense prevention*“,**
- **„legalization marijuana *cannabis drug soft leaves plant smoked chewed euphoric abuse substance possession control pot grass dope weed smoke*“**

	TA-sorted	Prob-sorted (smart)
#sorted accesses	22,403,490	18,287,636
elapsed time [s]	7908	1066
max queue size	70896	400
relative recall	1	0.88
rank distance	0	14.5
score error	0	0.035

Performance Results for IMDB Queries

on IMDB corpus (Web site: Internet Movie Database):

375 000 movies, 1.2 Mio. persons (html/xml)

20 structured/text queries with Dice-coefficient-based similarities of categorical attributes Genre and Actor, e.g.:

- ***Genre*** $\supseteq$ ***{Western}*** $\wedge$ ***Actor*** $\supseteq$ ***{John Wayne, Katherine Hepburn}*** $\wedge$ ***Description*** $\supseteq$ ***{sheriff, marshall},***
- ***Genre*** $\supseteq$ ***{Thriller}*** $\wedge$ ***Actor*** $\supseteq$ ***{Arnold Schwarzenegger}*** $\wedge$ ***Description*** $\supseteq$ ***{robot}***

	TA-sorted	Prob-sorted (smart)
#sorted accesses	1,003,650	403,981
elapsed time [s]	201.9	12.7
max queue size	12628	400
relative recall	1	0.75
rank distance	0	126.7
score error	0	0.25

Outline

- ✓ Motivation and Challenges
- ✓ XXL & XXL-light: IR on XML Data
- ✓ Role of Ontologies
- ✓ Efficient Evaluation of Top-k Queries
- Ongoing and Future Work

Exploiting Collective Human Input for Collaborative Web Search

- Beyond Relevance Feedback and Beyond Google -

- href links are human endorsements → PageRank, etc.
- Opportunity: online analysis of human input & behavior may compensate deficiencies of search engine

Typical scenario for 3-keyword user query: a & b & c

→ top 10 results: user clicks on ranks 2, 5, 7

→ top 10 results: u

u

u

→ top 10 results: u

u

user asks friend for tips

query logs, bookmarks, etc. provide
- **human assessments & endorsements**
- **correlations among words & concepts and among documents**

Challenge: How can we use knowledge about the collective input of all users in a large community?

Concluding Remarks

long-term goal: **exploit the Web's potential for being the world's largest knowledge base**

- ***XML*** **and** ***Semantic Web*** **are key assets, but by themselves not sufficient; we need to cope with** ***diversity*****,** ***incompleteness,*** **and** ***uncertainty***
  $\rightarrow$ absolute need for ranked retrieval
- **view** ***information organization*** **and** ***information search*** **as dual views of the same problem**
- **combine techniques from** ***DBS*****,** ***IR*****,** ***CL*****,** ***AI*****, and** ***ML***
- **need better** ***theory*** **about quality/efficiency** ***tradeoffs*** **as well as** ***large-scale experiments***