max planck institut informatik

Web, Semantic, and Social Information Retrieval

Gerhard Weikum

weikum@mpi-inf.mpg.de
http://www.mpi-inf.mpg.de/~weikum/

EDBT 2007 Summer School, Bolzano, Italy, September 3, 2007

Adding Semantics to IR (or Adding Ranking to DB)

	Structured data (records)	Unstructured data (documents)
Unstructured search (keywords)	**Keyword Search on Relational Graphs** (BANKS, Discover, DBexplorer, …) + Web 2.0	**IR Systems Search Engines** + Digital Libraries + Enterprise Search
Structured search (SQL,XQuery)	**DB Systems** + Text + Relax. & Approx. + Ranking	**Querying entities & relations from IE** (Libra, ExDB, NAGA, …)

Trend: quadrants getting blurred
towards DB&IR technology integration

Overview

- **Part 1: Web IR**
  - State of the Art
  - Scalability Challenge
  - Quality Challenge
  - Personalization
  - Research Opportunities
- **Part 2: Semantic & Social IR**
  - Ontologies in XML IR
  - Entity Search and Ranking
  - Graph IR
  - Web 2.0 Search and Mining
  - Research Opportunities

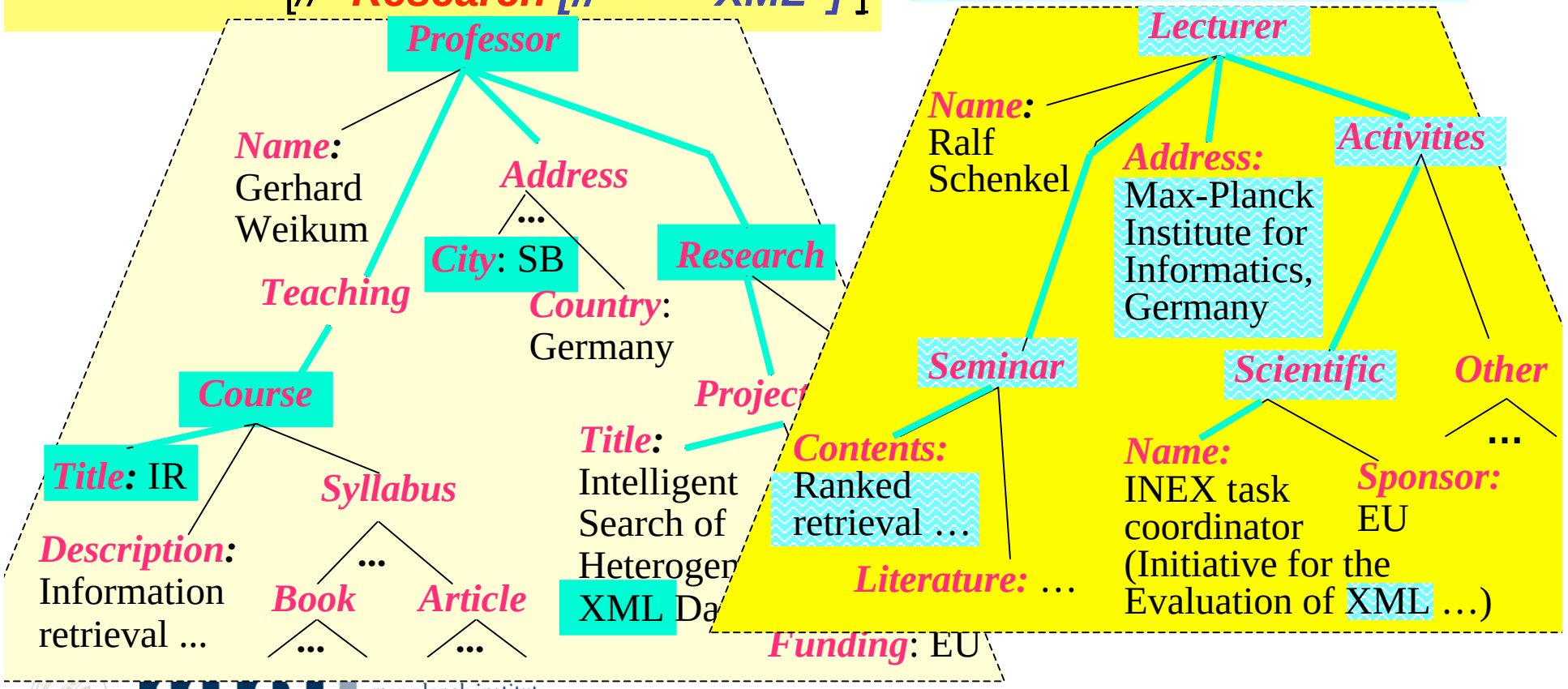
XML IR on Heterogeneous Data

Union of **heterogeneous** sources **without global schema**

Similarity-aware XPath:

```
//~Professor [//* = "~SB"]
           [//~Course [//* = "~IR"] ]
           [//~Research [//* = "~XML"] ]
```

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



XML IR on Heterogeneous Data

Union of **heterogeneous** sources **without global schema**

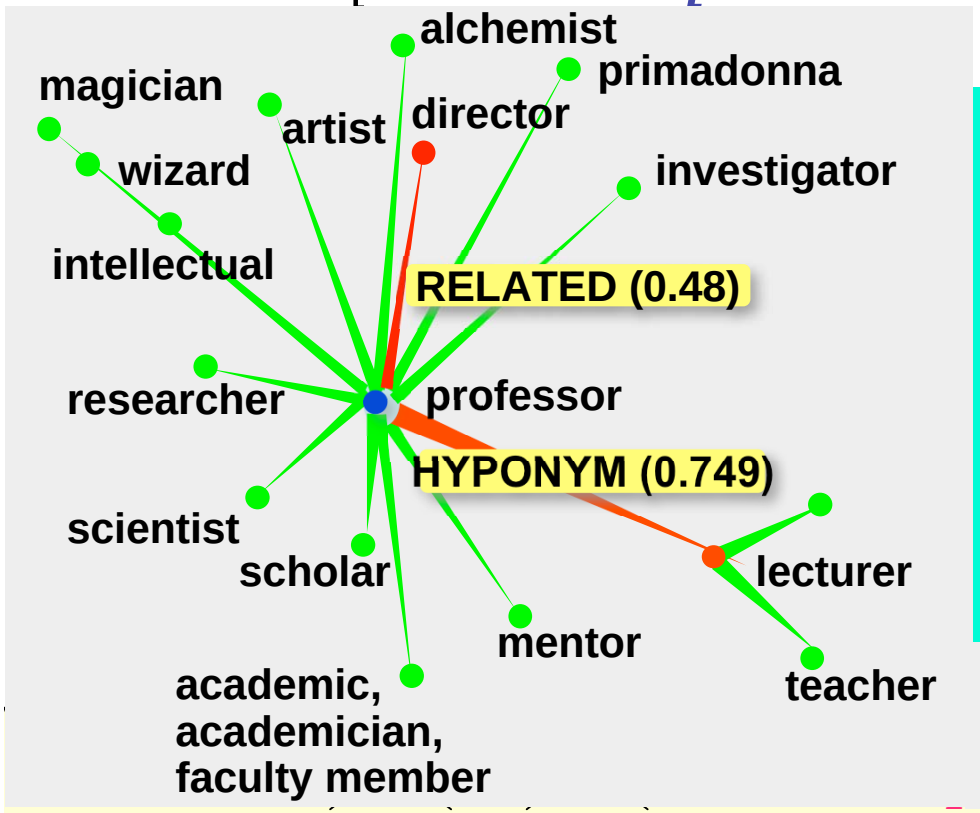
Similarity-aware XPath:

//~Professor [// = "~Saarbruecken"]*

[//~Course [// = "~IR"]*]

[//~Research [// = "~XML"]*]

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



Lecturer

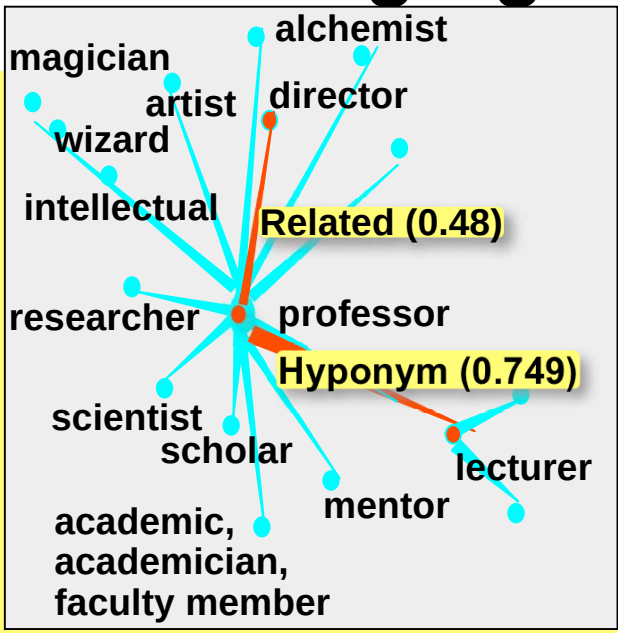
Scoring and ranking:

- **XML BM25** for **content** cond.
- **ontological** similarity for **relaxed tag** condition
- score aggregation with **probabilistic independence**
- extended TA for query exec.

statistical edge weighting by Dice coeff.: 2 #(x,y) / (#x + #y) on Web

Query Expansion with Incremental Merging

[M. Theobald et al.: SIGIR 2005]

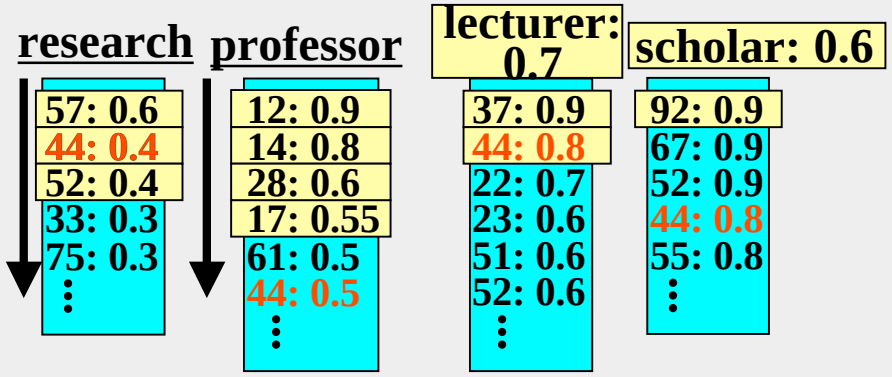


relaxable query q: ~*professor research*
with expansions exp(i)={w | sim(i,w) ≥ θ}
based on ontology relatedness modulating monotonic score aggregation

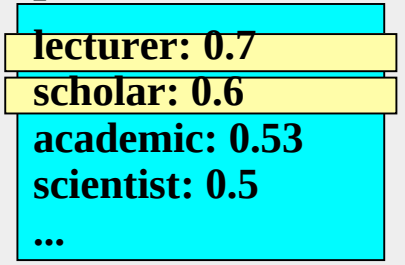
TA scans of index lists for $\bigcup_{i \in q}$ exp(i)

Better: dynamic query expansion with incremental merging of additional index lists

B+ tree index on terms



ontology / meta-index



efficient, robust, self-tuning

Query Expansion Example

From TREC 2004 Robust Track Benchmark:

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[0.312|1.00]", camorra[0.254|1.00], maffia[0.318|1.00],
mafia[0.154|1.00], "sicilian[0.201|1.00] & mafia[0.154|1.00]",
"black[0.066|1.00] & hand[0.053|1.00]", mob[0.123|1.00], syndicate[0.093|1.00]}],
organ[0.213|1.00], crime[0.312|1.00], collabor[0.415|0.20],
columbian[0.686|0.20], cartel[0.466|0.20], ...}}

135530 sorted accesses in 11.073s.

Results:

1. Interpol Chief on Fight Against Narcotics
2. Economic Counterintelligence Tasks Viewed
3. Dresden Conference Views Growth of Organized Crime in Europe
4. Report on Drug, Weapons Seizures in Southwest Border Region
5. SWITZERLAND CALLED SOFT ON CRIME

...

Overview

- **Part 1: Web IR**
  - State of the Art
  - Scalability Challenge
  - Quality Challenge
  - Personalization
  - Research Opportunities
- **Part 2: Semantic & Social IR**
  - ✓ Ontologies in XML IR
  - Entity Search and Ranking
  - Graph IR
  - Web 2.0 Search and Mining
  - Research Opportunities

Don‘t Let Me Be Misunderstood

Keyword query: Max Planck

or

Keyword query: Greek art Paris

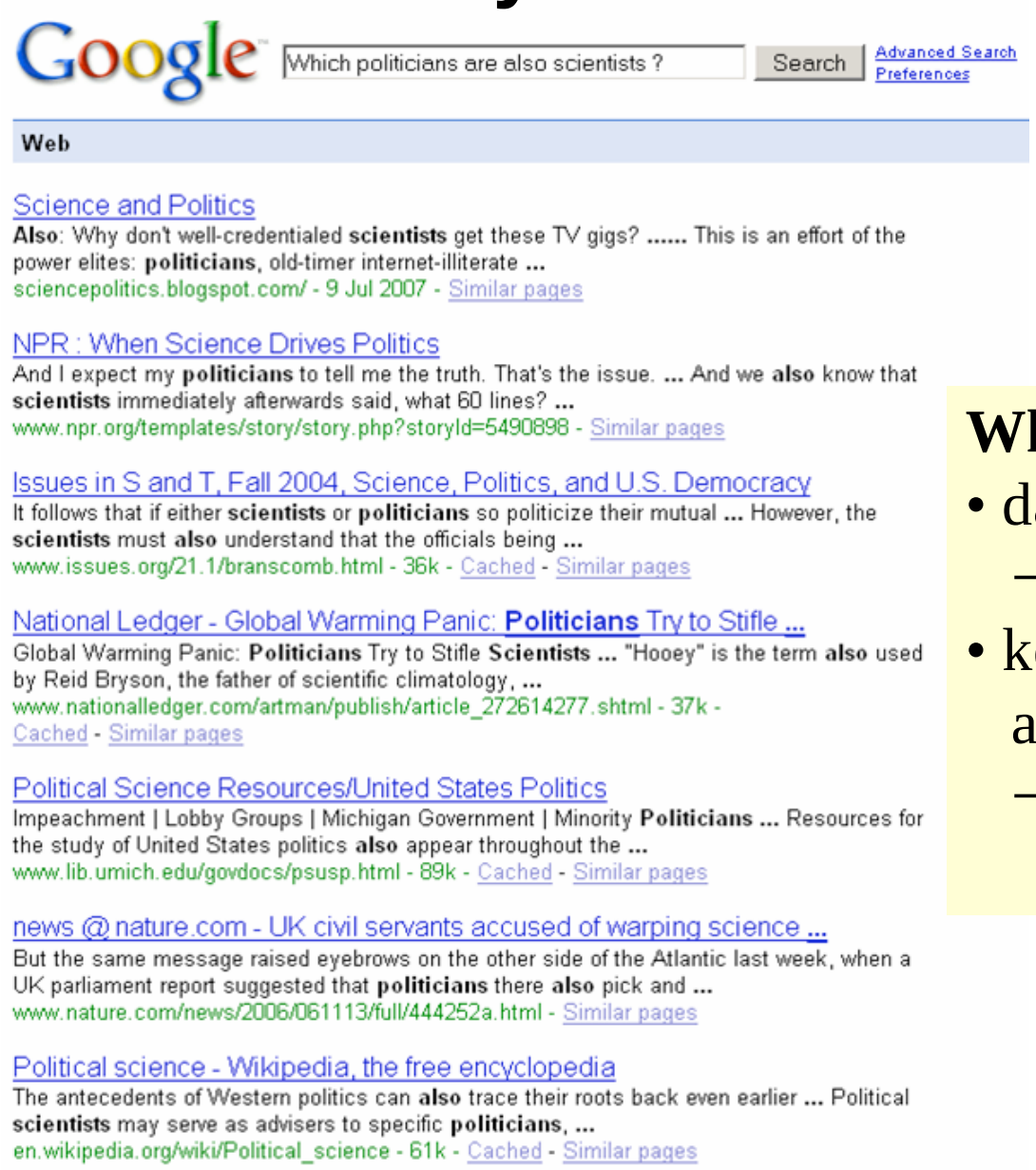
or

Semantic Search

Concept query:
Person = „Max Planck“

Concept query:
„Greek art“
& Location = „Paris“

Entity Search: Example Google



What is lacking?

- data is not knowledge
  → extraction and organization
- keywords cannot express advanced user intentions
  → concepts, entities, properties, relations

Entity Search: Example NAGA

Score: 6.247457440558111E-7

"scientist" —means—> scientist_110560637
Benjamin_Franklin —type—> Massachusetts_politicians
"politician" —means—> politician_110451263
American_scientists —subClassOf—> scientist_110560637
Benjamin_Franklin —type—> American_scientists
Massachusetts_politicians —subClassOf—> politician_110451263

- $@politician = politician_110451263
- $@scientist = scientist_110560637
- $X = Benjamin_Franklin

Score: 3.185850362140424E-7

"scientist" —means—> scientist_110560637
"politician" —means—> politician_110451263
Paul_Wolfowitz —type—> American_political_scientists
American_political_scientists —subClassOf—> scientist_110560637
Paul_Wolfowitz —type—> Jewish-American_politicians
Jewish-American_politicians —subClassOf—> politician_110451263

- $@politician = politician_110451263
- $@scientist = scientist_110560637
- $X = Paul_Wolfowitz

Score: 1.121658976926192E-7

Angela_Merkel —type—> German_scientists
"scientist" —means—> scientist_110560637
German_Christian_Democrat_politicians —subClassOf—>
politician_110451263
Angela_Merkel —type—> German_Christian_Democrat_politicians
"politician" —means—> politician_110451263
German_scientists —subClassOf—> scientist_110560637

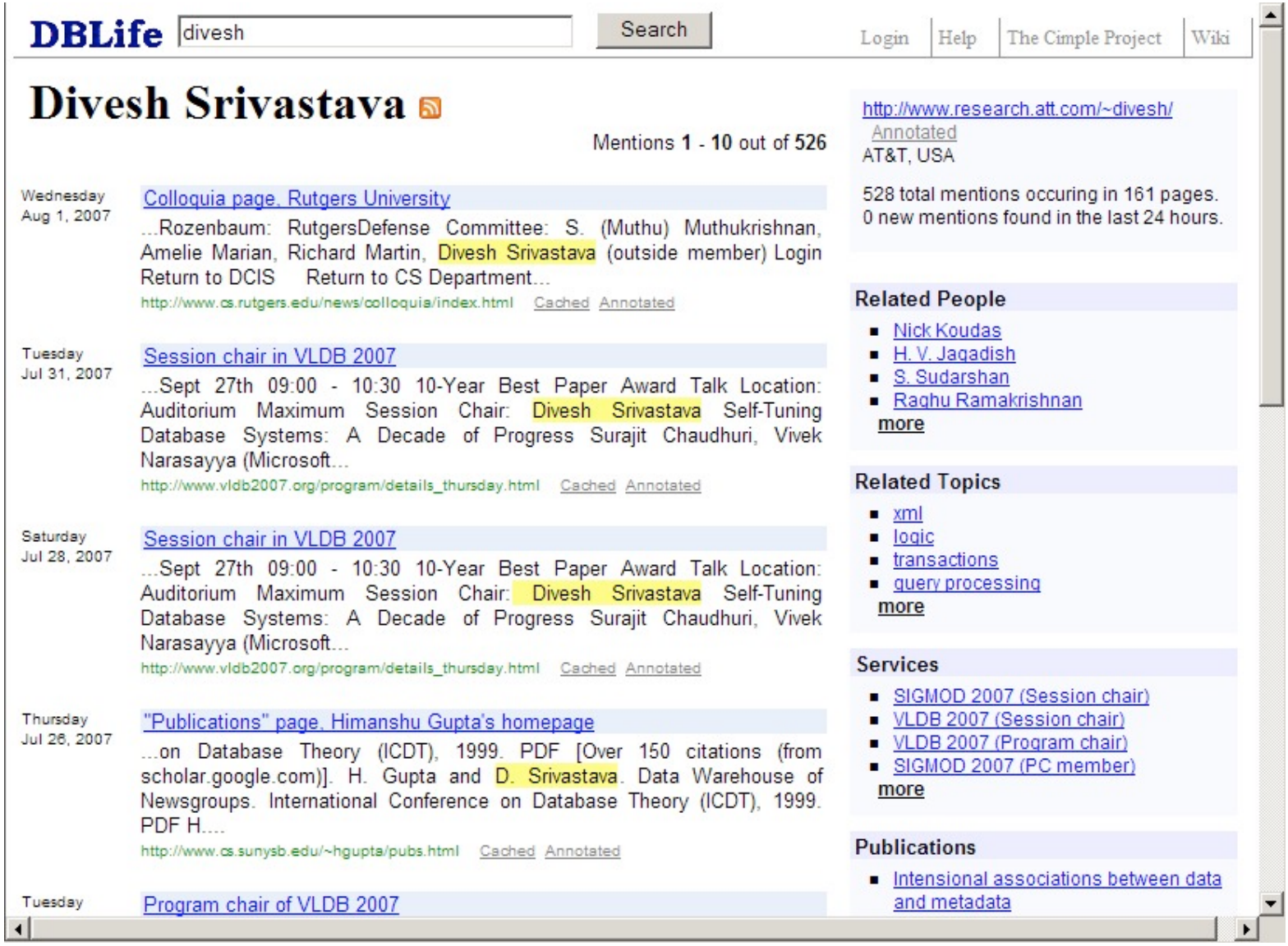
- $@politician = politician_110451263
- $@scientist = scientist_110560637
- $X = Angela_Merkel

Query:
$x isa politician
$x isa scientist

Results:
Benjamin Franklin
Paul Wolfowitz
Angela Merkel
…

Entity Search: Example DBLife

http://dblife.cs.wisc.edu



Entity Search

Instead of „interpreting“ text with background knowledge, extract facts and search entities, attributes, and relations

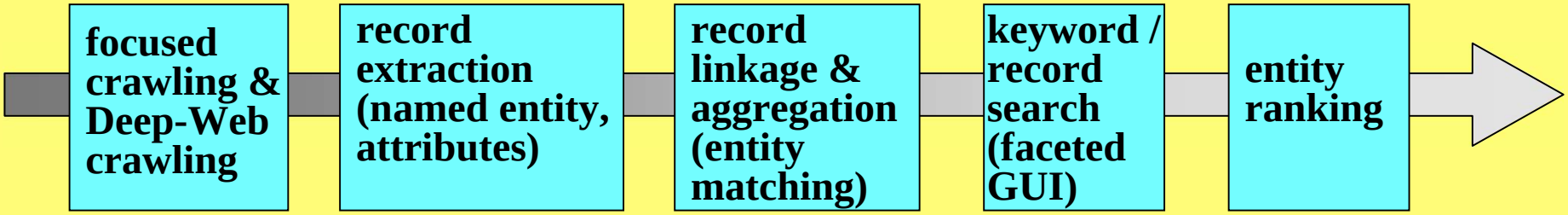
Motivation and Applications:

- Web search for vertical domains
  (products, traveling, entertainment, scholarly publications, intelligence agencies, etc.)
- preparation for natural-language QA
- step towards better Deep-Web search, digital libraries, e-science

Example systems:

- Libra (MSR), EntityRank (UIUC), ExDB (UW Seattle), NAGA (MPII), …
- probably all commercial search engines have some support for entities

Typical system architecture:



Information Extraction (IE): Text to Records

Max Planck

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947) was a German physicist who is considered to be the inventor of quantum theory.

Born in Kiel, Planck started his physics studies at Munich University in 1874, graduating in 1879 in Berlin. He returned to München in 1880 to teach at the university, and moved to Kiel in 1885. There he married Marie Merck in 1886. In 1889, he moved to Berlin, where from 1892 on he held the chair of theoretical physics.

In 1899, he discovered a new fundamental constant, which is named Planck's constant, and is, for example, used to calculate the energy of a photon. Also that year, he de... own set of units of measurement based on fundamental physical constants. One year later, he discovered the law of heat radiation, which is named Planck's Law of Radiation. This law became the basis of quantum theory, which emerg... later in cooperation with Albert Einstein and Niels Bohr.

Person	BirthDate	BirthPlace	...
Max Planck	4/23, 1858	Kiel	
Albert Einstein	3/14, 1879	Ulm	
Mahatma Gandhi	10/2, 1869	Porbandar	

Person	ScientificResult
Max Planck	Quantum Theory

Constant
Planck's constant

extracted facts often have confidence < 1 → DB with uncertainty (probabilistic DB)

Person	Collaborator
Max Planck	Albert Einstein
Max Planck	Niels Bohr

Person	Organization
Max Planck	KWG / MPG

combine NLP, pattern matching, lexicons, statistical learning

IE Technology: Rules, Patterns, Learning

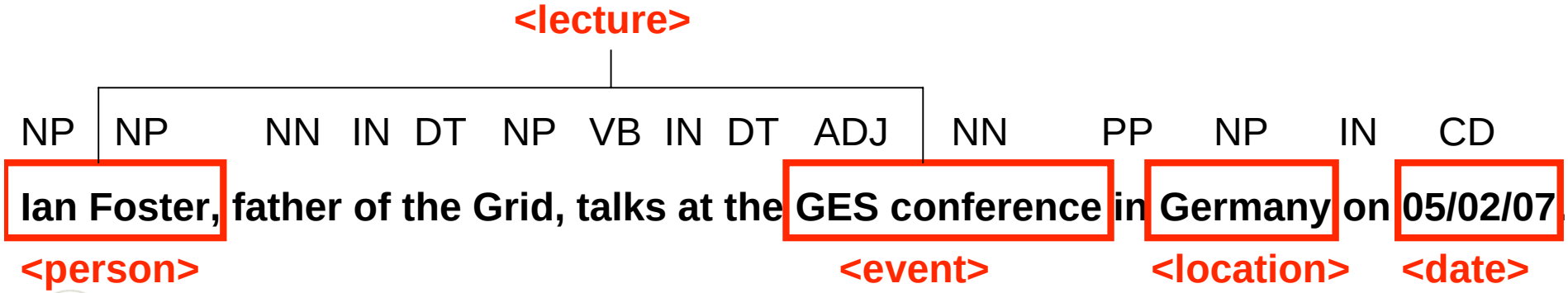
For heterogeneous sources and for natural-language text:

- **NLP techniques** (parser, PoS tagging) for tokenization
- **identify patterns** (regular expressions) as features
- **train statistical learners** for segmentation and labeling (HMM, CRF, SVM, etc.), augmented with lexicons
- use learned model to **automatically tag** newly seen input

Training data:

The WWW conference takes place in Banff in Canada.
Today's keynote speaker is Dr. Berners-Lee from W3C.
The panel in Edinburgh, chaired by Ron Brachman from Yahoo!, …
…

<location>
<organization>
<person>
<event>



Entity-Search Ranking with LM

[Z. Nie et al.: WWW 2007; cf. also T. Cheng: VLDB 2007]

Standard LM for docs with background model (smoothing):

$$s(d,q) = \lambda P[q \mid d] + (1-\lambda) P[q] \qquad P[q \mid d] \sim \sum_{w \in q} \log \frac{tf(w,d)}{\sum_u tf(u,d)}$$

Assume **entity e** was seen in **k records** $r_1, \ldots, r_k$
extracted from k pages $d_1, \ldots, d_k$ with **accuracy** $\alpha_1, \ldots, \alpha_k$

$$P[w \mid e] \sim \sum_i \alpha_i \, P[w \mid context(r_i, d_i)]$$ **record-level LM**

$$\rightarrow \quad s(e,q) = \sum_{w \in q} \lambda \left(\sum_i \alpha_i \frac{tf(w, context(r_i,d_i))}{|context(r_i,d_i)|} \right) + (1-\lambda) \frac{\sum_r tf(w,r)}{\#records}$$

with context window around r_i in d_i (default: only r_i itself)

alternatively consider individual attributes $e.a_j$ with importance β_j
extracted from page d_i with accuracy γ_{ij}

$$P[w \mid e] \sim \sum_i \alpha_i \sum_j \beta_j \gamma_{ij} \, P[w \mid context(r_i.a_j, d_i)]$$

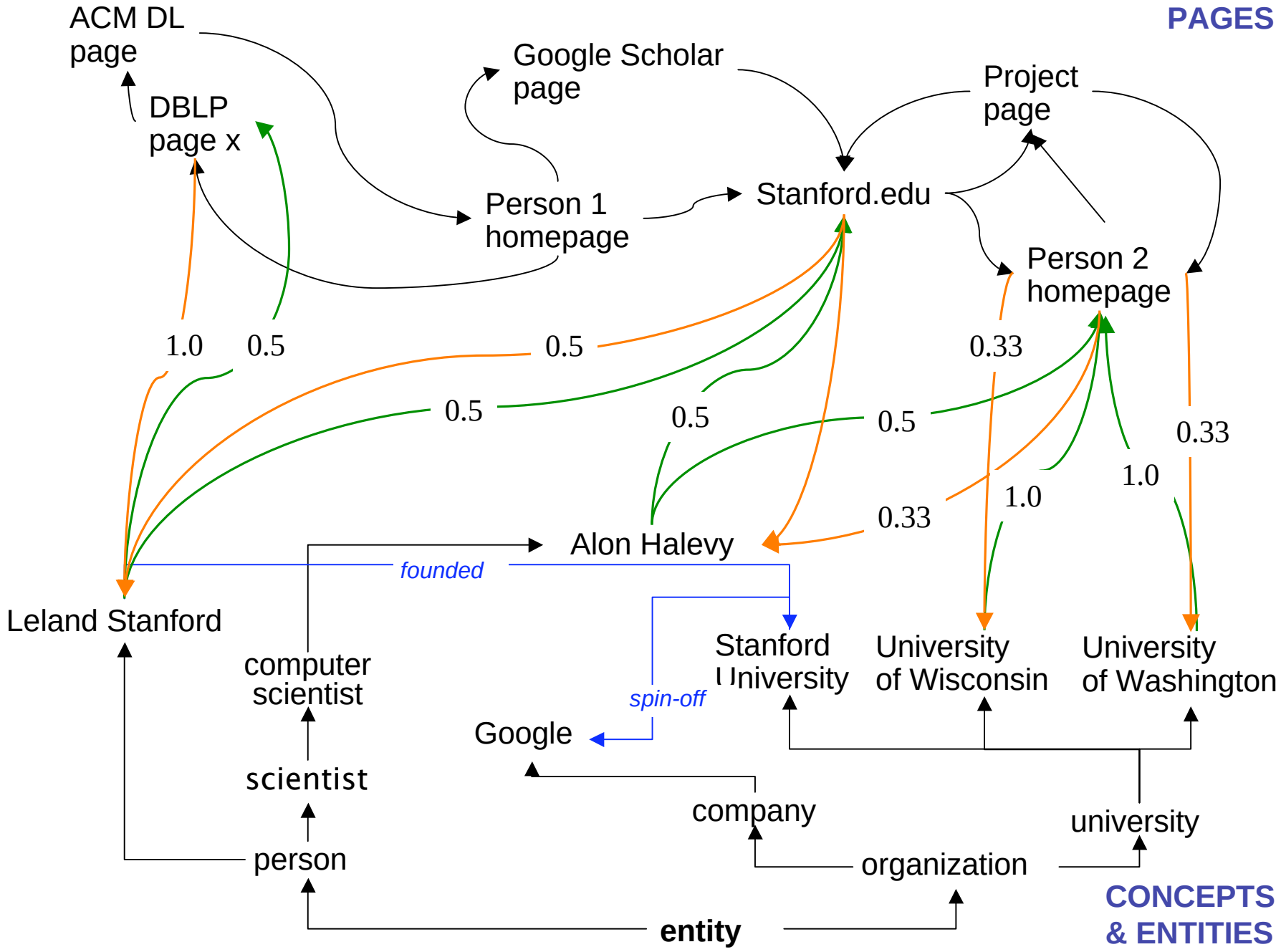
Entity-Search Ranking by Link Analysis (1)

[A. Balmin et al. 2004, Nie et al. 2005, Chakrabarti 2007, J. Stoyanovich 2007]

EntityAuthority (EVA; similar to ObjectRank, PopRank, HubRank):

- define **authority transfer graph** among entities and pages with edges:
  - entity → page if entity appears in page
  - page → entity if entity is extracted from page
  - page1 → page2 if there is hyperlink or implicit link between pages
  - entity1 → entity2 if there is a semantic relation between entities
- edges can be typed and (degree- or weight-) normalized and are weighted by confidence and type-importance
- also applicable to graph of DB records with foreign-key relations (e.g. bibliography with different weights of publisher vs. location for conference record)
- compared to standard Web graph, ER graphs of this kind have higher variation of edge weights

Entity-Search Authority Transfer Graph



Entity-Search Ranking by Link Analysis (2)

[A. Balmin et al. 2004, Nie et al. 2005, Chakrabarti 2007, J. Stoyanovich 2007]

- perform PR- or PPR- or HITS-style spectral analysis on query-time subgraph, e.g.:

$$\vec{r}_e \sim \alpha M_{e \to e} \times \vec{r}_e + (1-\alpha) M_{p \to e} \times \vec{r}_p$$

$$\vec{r}_p \sim \beta M_{p \to p} \times \vec{r}_p + (1-\beta) M_{e \to p} \times \vec{r}_e$$

- small-scale experiment: query *„Serbia basketball“* on Wikipedia subset with extraction of persons, organizations, locations (+ YAGO ontology)
  - top result pages with PR: *1977, Greece, Belgrade*
  - top result pages with EVA: *Basketball in Yugoslavia, Vlade Divac*
  - top result entities with EVA: *Michael Jordan, LA Lakers, Vlade Divac*

- for query-time efficiency, node scores may be precomputed for individual keywords or important queries based on query log

Overview

- **Part 1: Web IR**
  - State of the Art
  - Scalability Challenge
  - Quality Challenge
  - Personalization
  - Research Opportunities
- **Part 2: Semantic & Social IR**
  - ✓ Ontologies in XML IR
  - ✓ Entity Search and Ranking
  - Graph IR
  - Web 2.0 Search and Mining
  - Research Opportunities

Graph IR

graph (V, E) with
- V: data items (records, elements, docs, passages, entities, …)
- E: (semantic) relations as edges

set of keyword conditions or
more expressive (node-evaluable) conditions

Use cases:
- contextual multi-page Web search
- relational DBs
- XML beyond trees
- RDF graphs
- ER graphs (e.g. from IE)
- ontology / knowledge graphs
- social networks
- biological networks

YAGO: Yet Another Great Ontology

[F. Suchanek, G. Kasneci, G. Weikum: WWW 2007]

- Turn Wikipedia into explicit **knowledge base** (semantic DB)
- Exploit hand-crafted **categories** and **templates**
- Represent facts as explicit **knowledge triples**:

  relation (entity1, entity2)

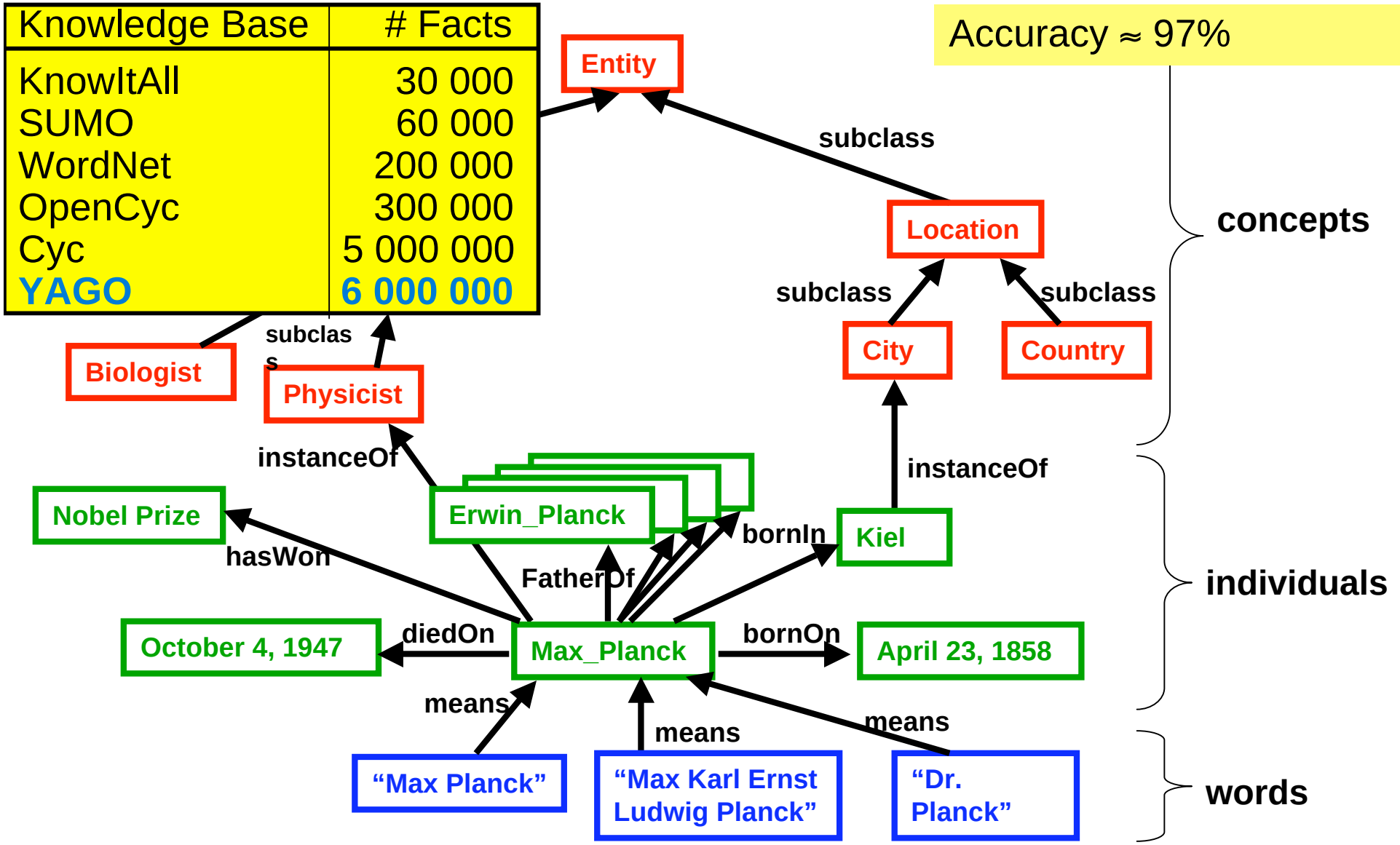
  (in 1st-order logic, compatible with RDF, OWL-lite, XML, etc.)
- Map (and disambiguate) relations into **WordNet** concept DAG

Examples:

YAGO Knowledge Representation

Knowledge Base	# Facts
KnowItAll	30 000
SUMO	60 000
WordNet	200 000
OpenCyc	300 000
Cyc	5 000 000
YAGO	**6 000 000**

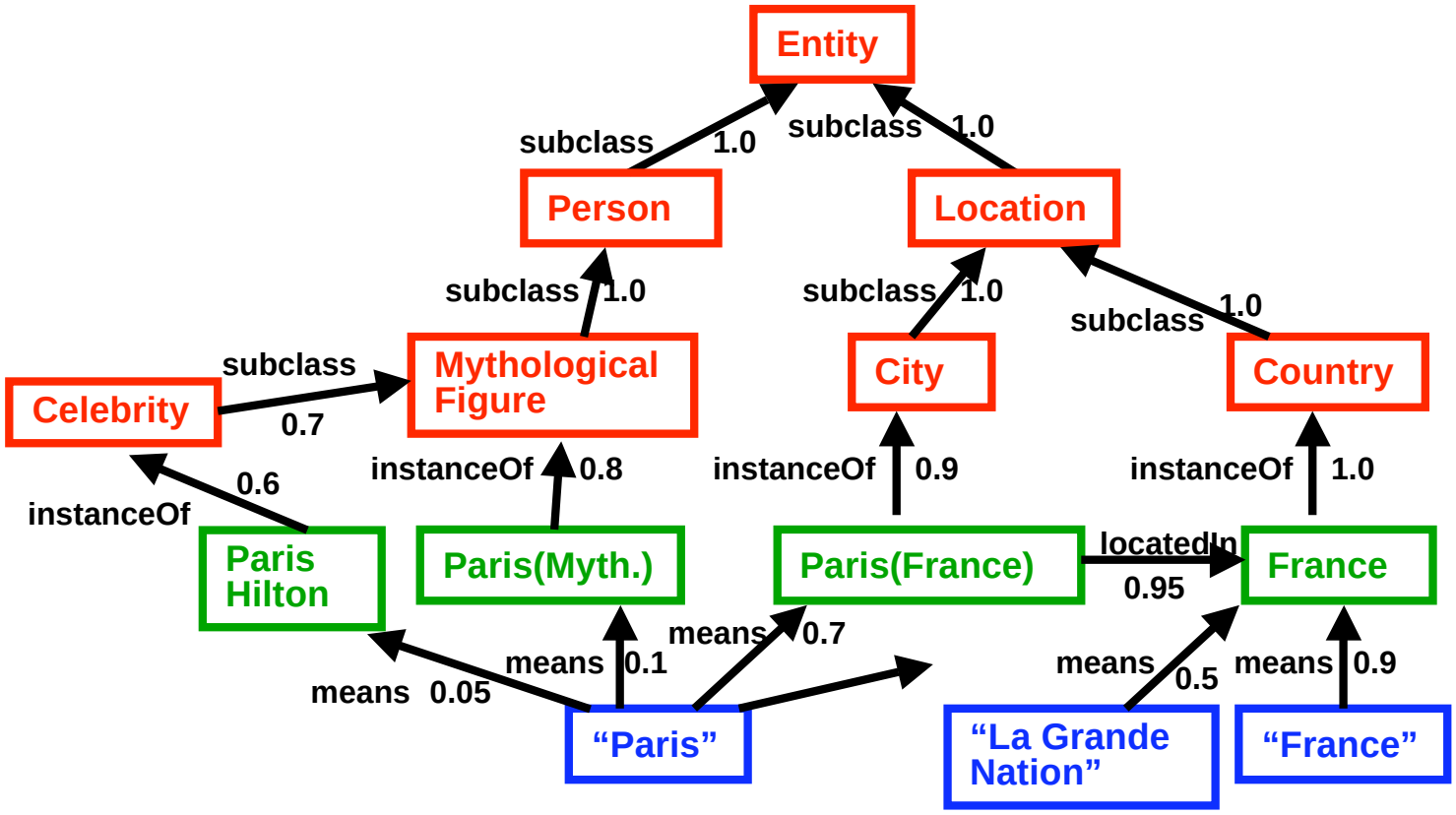
Accuracy ≈ 97%



Online access and download at http://www.mpi-inf.mpg.de/~suchanek/yago/

YAGO Enhancement by IE on Text Sources

Capture confidence value for each fact



ongoing work: harvesting relations by IE tools like GATE, LEILA, ...

(e.g.: which enzyme catalyzes which biochemical process, who discovered or invented what, ...)

Knowledge Acquisition from the Web

Learn Semantic Relations from Entire Corpora at Large Scale
(as exhaustively as possible but with high accuracy)

Examples:
- **all cities, all basketball players, all composers**
- **headquarters of companies, CEOs of companies, synonyms of proteins**
- **birthdates of people, capitals of countries, rivers in cities**
- **which musician plays which instruments**
- **who discovered or invented what**
- **which enzyme catalyzes which biochemical reaction**

Existing approaches and tools
(Snowball [Gravano et al. 2000], KnowItAll [Etzioni et al. 2004], …):
almost-unsupervised pattern matching and learning:
seeds (known facts) → patterns (in text) → (extraction) rule → (new) facts

Methods for Web-Scale Fact Extration

Example:

in downtown X
X and other towns

X … blows Y

in downtown Beijing
Coltrane blows sax

city(Beijing)
plays(C., sax)

city(Beijing)
plays(Coltrane, sax)

old center of Beijing
sax player Coltrane

old center of X
Y player X

Assessment of facts & generation of rules based on statistics

Rules can be more sophisticated:

playing NN: (ADJ|ADV)* NP & class(NN)=instrument & class(head(NP))=person
→ plays(head(NP), NN)

Beyond Surface Learning with LEILA

Learning to Extract Information by Linguistic Analysis [F. Suchanek et al.: KDD'06]

Limitation of surface patterns:

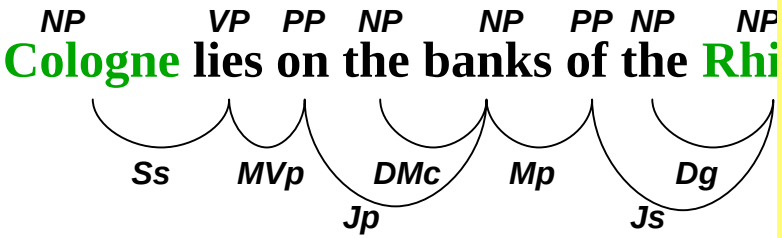
who discovered or invented what

"Tesla's work formed the basis of AC electric power"

"Al Gore funded more work for a better basis of the Internet"

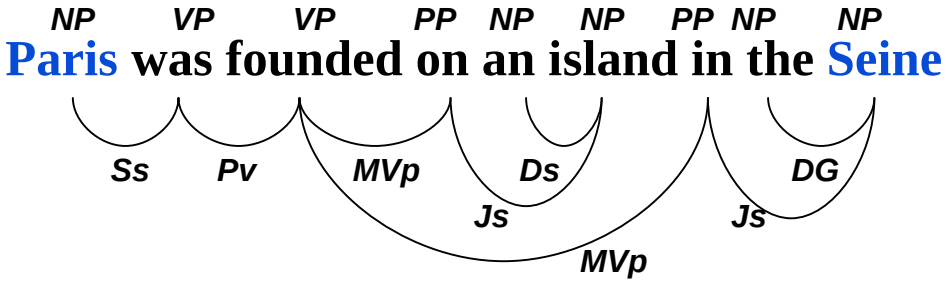
Almost-unsupervised Statistical Learning with Dependency Parsing

(Cologne, Rhine), (Cairo, Nile), … ***(Cairo, Rhine), (Rome, 0911), (*, *[0..9]**), …***



LEILA outperforms other Web-IE methods in terms of precision, recall, F1, but:

- **dependency parser is slow**
- **one relation at a time**



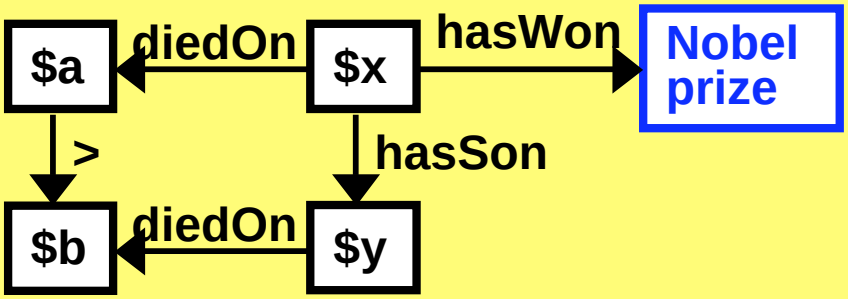
(Paris, Seine)

We visited Paris last summer. It has many museums along the banks of the Seine.

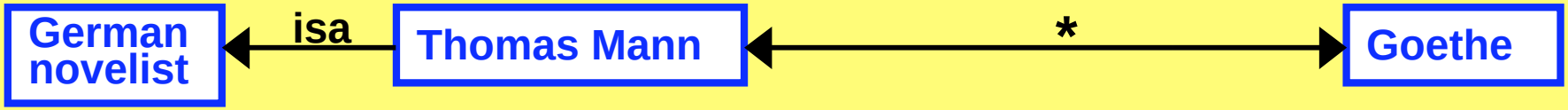
NAGA: Graph IR on YAGO [G. Kasneci et al.: WWW'07]

Graph-based search on YAGO-style knowledge bases
with built-in **ranking** based on **confidence** and **informativeness**

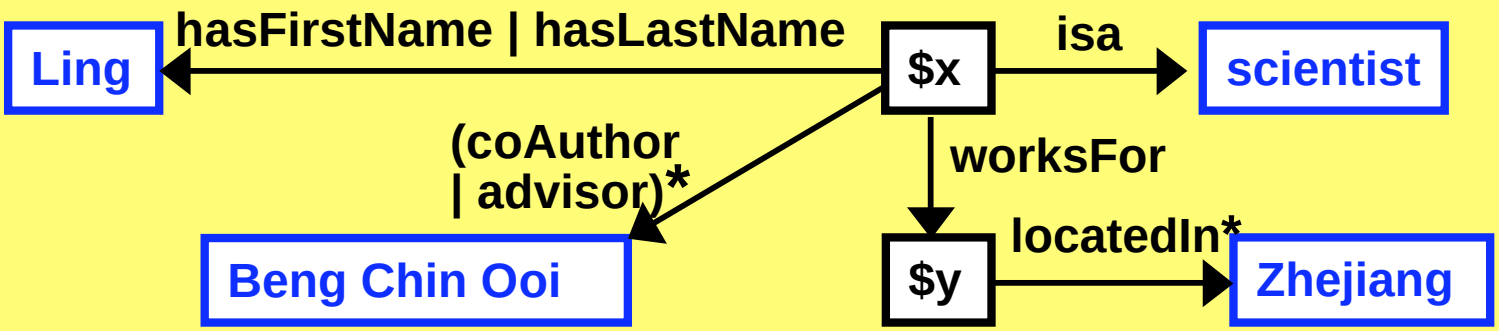
discovery queries



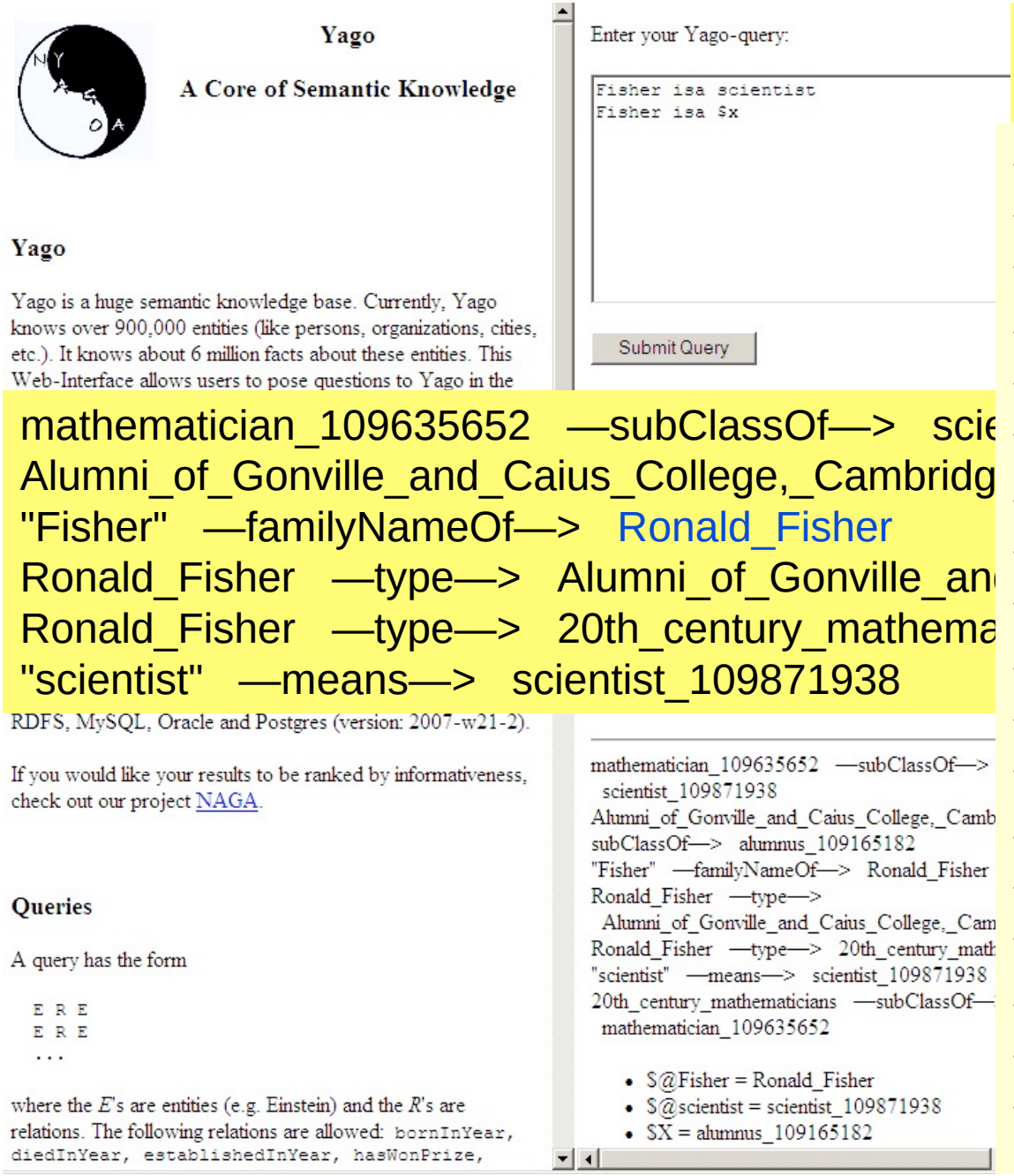
connectedness queries



queries with regular expressions



Search Results Without Ranking



q: Fisher isa scientist
Fisher isa $x

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = alumnus_109165182

$@Fisher = Irving_Fisher
$@scientist = scientist_109871938
$X = social_scientist_109927304

$@Fisher = James_Fisher
$@scientist = scientist_10981938
$X = ornithologist_109711173

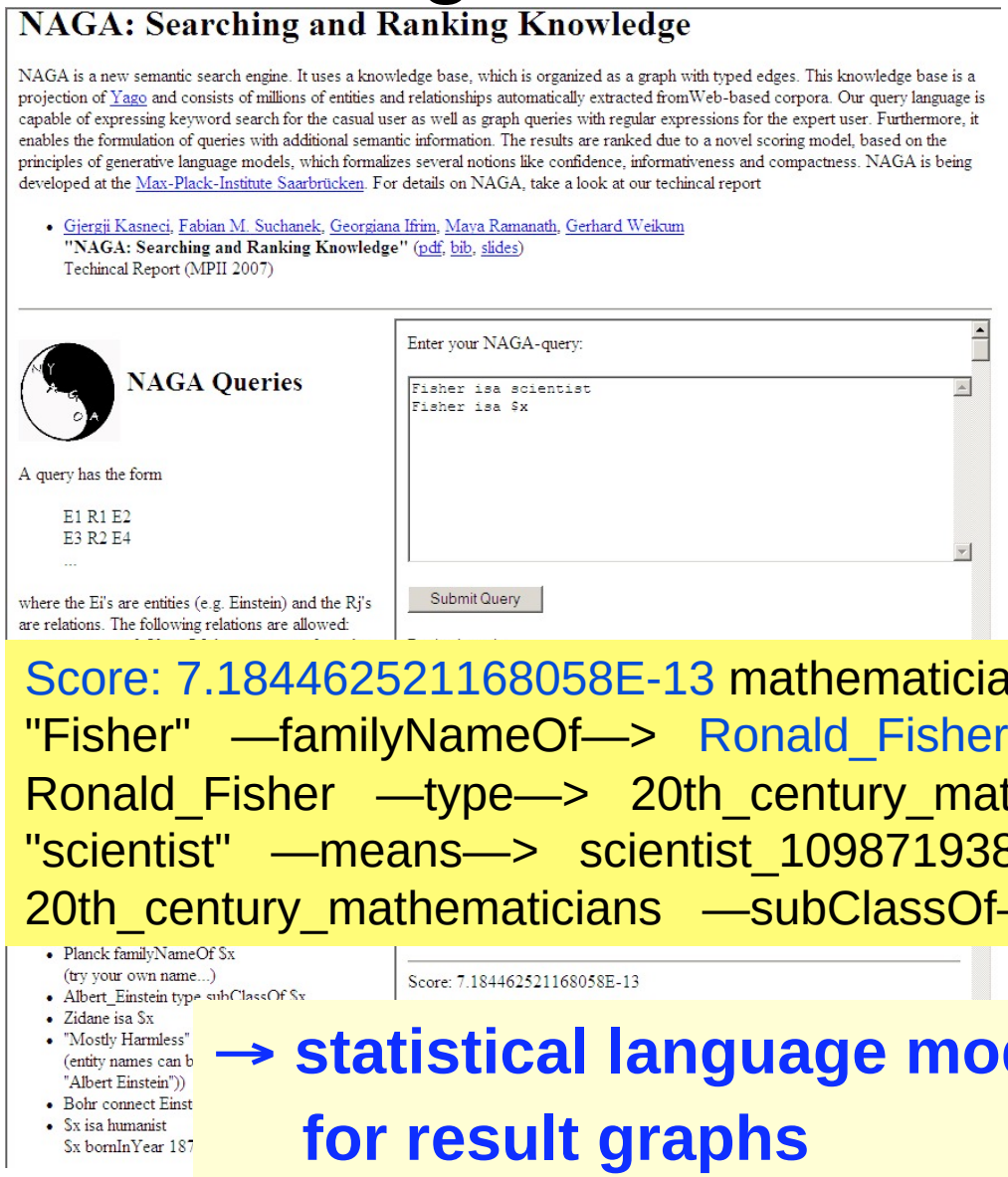
$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = theorist_110008610

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = colleague_109301221

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = organism_100003226

...

Ranking with Statistical Language Model



q: Fisher isa scientist
Fisher isa $x

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = mathematician_109635652

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = statistician_109958989

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = president_109787431

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = geneticist_109475749

$@Fisher = Ronald_Fisher
$@scientist = scientist_109871938
$X = scientist_109871938

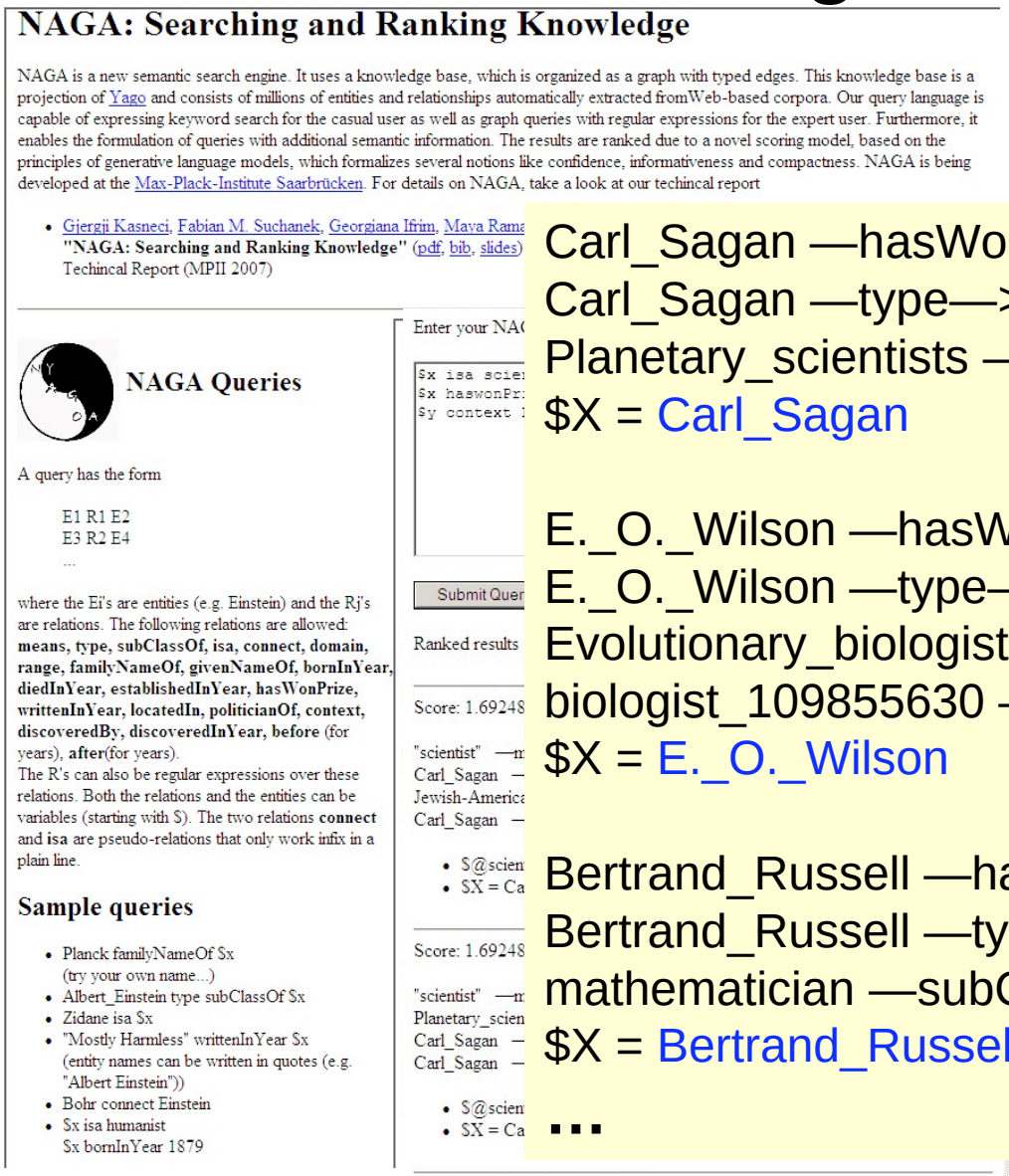
...

Score: 7.184462521168058E-13 mathematician_10
"Fisher" —familyNameOf—> Ronald_Fisher
Ronald_Fisher —type—> 20th_century_mathema
"scientist" —means—> scientist_109871938
20th_century_mathematicians —subClassOf—> r

→ statistical language model for result graphs

Online access at http://www.mpi-inf.mpg.de/~kasneci/naga/

NAGA: Searching & Ranking Knowledge



q: $x isa Scientist
$x hasWonPrize $y
$y context Literature

Carl_Sagan —hasWonPrize—> Pulitzer_Prize
Carl_Sagan —type—> Planetary_scientists
Planetary_scientists —subClassOf—> scientist_110560637
$X = Carl_Sagan

E._O._Wilson —hasWonPrize—> Pulitzer_Prize
E._O._Wilson —type—> Evolutionary_biologists
Evolutionary_biologists —subClassOf—> biologist_109855630
biologist_109855630 —subClassOf—> scientist_110560637
$X = E._O._Wilson

Bertrand_Russell —hasWonPrize—> Nobel_Prize_in_Literature
Bertrand_Russell —type—> mathematician
mathematician —subClassOf—> scientist_110560637
$X = Bertrand_Russell

...

Online access at http://www.mpi-inf.mpg.de/~kasneci/naga/

Ranking Factors

Confidence:

Prefer results that are likely to be correct

- Certainty of IE
- Authenticity and Authority of Sources

bornIn (Max Planck, Kiel) from „Max Planck was born in Kiel“ (Wikipedia)

livesIn (Elvis Presley, Mars) from „They believe Elvis hides on Mars“ (Martian Bloggeria)

Informativeness:

Prefer results that are likely important

May prefer results that are likely new to user

- Frequency in answer
- Frequency in corpus (e.g. Web)
- Frequency in query log

q: isa (Einstein, $y)

isa (Einstein, scientist)
isa (Einstein, vegetarian)

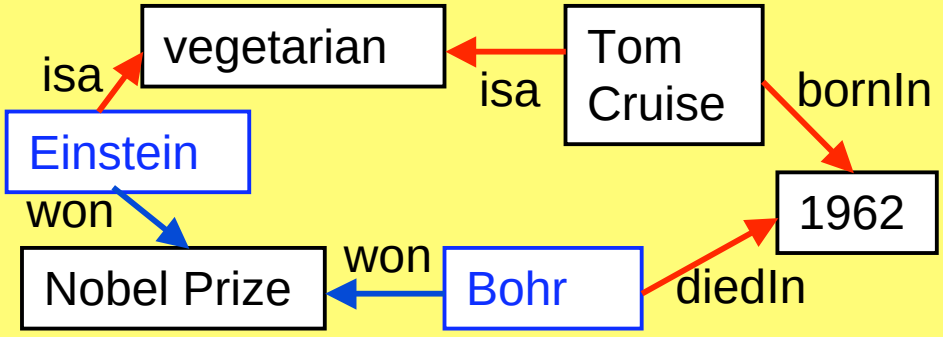
q: isa ($x, vegetarian)

isa (Einstein, vegetarian)
isa (Al Nobody, vegetarian)

Compactness:

Prefer results that are tightly connected

- Size of answer graph



NAGA Ranking Model

Following the paradigm of ***statistical language models*** (used in speech recognition and modern IR)

For query q with fact templates $q_1 \ldots q_n$ ex.: bornIn ($x, Frankfurt)
rank result graphs g with facts $g_1 \ldots g_n$ ex.: bornIn (Goethe, Frankfurt)
by **decreasing likelihoods**:

using generative mixture model

$$P[q \mid g] = \prod_{i=1}^{n} (1-\alpha) \cdot P[q_i \mid g_i] + \alpha \cdot P[q_i]$$

$$\beta \cdot P_{\text{conf}}[q_i \mid g_i] + (1-\beta) \cdot P_{\text{inform}}[q_i \mid g_i]$$

$$P[Goethe \mid bornIn, Frankfurt] = \frac{P[Goethe, bornIn, Frankfurt]}{P[bornIn, Frankfurt]}$$

based on IE accuracy and authority analysis

$$conf(e) = \sum_{i=1}^{n_e} acc(e, P_i) \cdot trust(P_i)$$

$$P(x \mid r, z) = \frac{P(x,r,z)}{P(r,z)} = \frac{P(x,r,z)}{\sum_{x'} P(x',r,z)}$$

for q_i = (x, r, z) with variable x

estimated by correlation statistics

Keyword Search on Graphs

[BANKS, Discover, DBExplorer, KPS, SphereSearch, BLINKS]

Schema-agnostic **keyword search** over **multiple tables**:
graph of tuples with foreign-key relationships as edges

Example:

Conferences (CId, Title, Location, Year)
CPublications (PId, Title, CId)
Authors (PId, Person)

Journals (JId, Title)
JPublications (PId, Title, Vol, No, Year)
Editors (CId, Person)

Select * From * Where * Contains "*Gray, DeWitt, XML, Performance*" And Year > 95

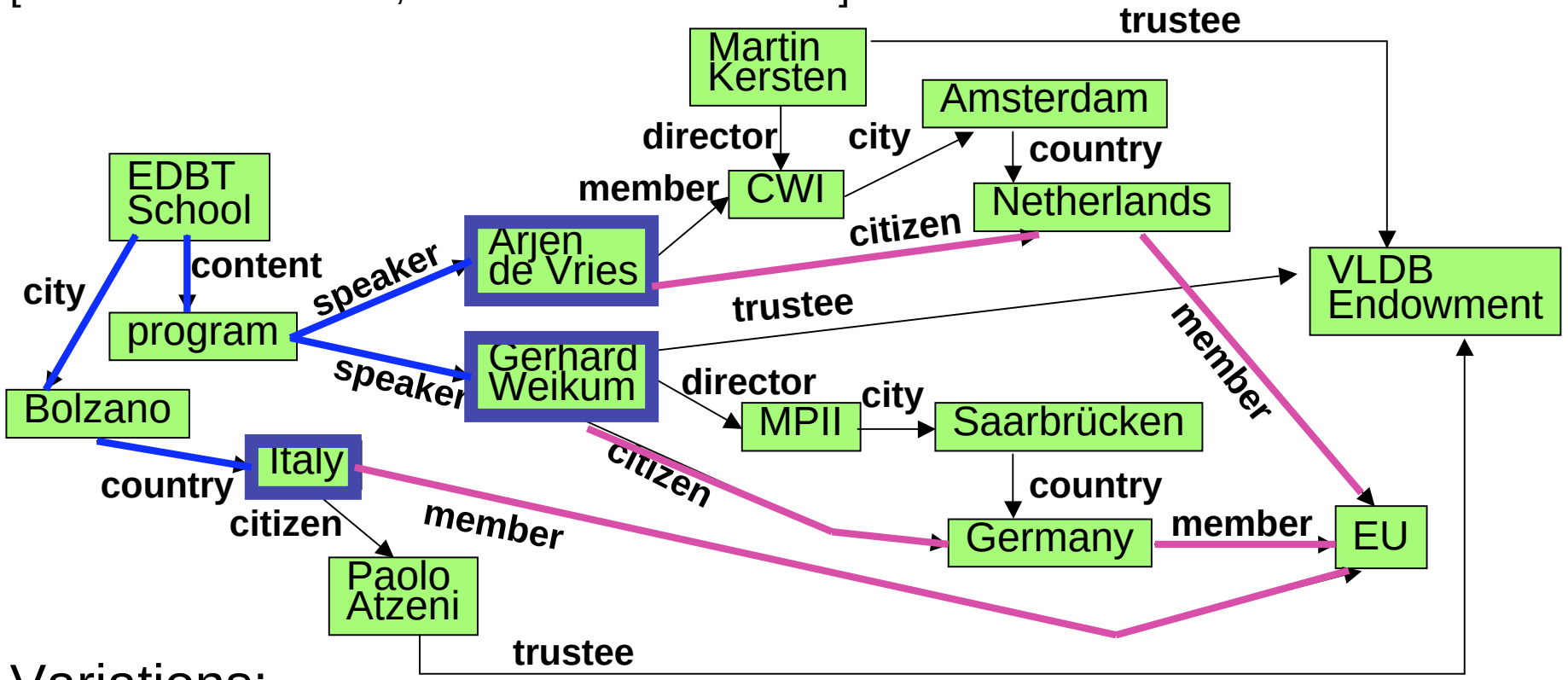
Result is **connected tree** with nodes that together contain all query keywords (or as many as possible)

QP approach for search over relational DB:
exploit schema, generate meaningful join trees (up to size limit)

Keyword Search on Graphs: Semantics

Subtleties of Interconnection Semantics

[S. Cohen et al. 2005, B. Kimelfeld et al. 2007]



Variations:

- directed vs. undirected graphs, strict vs. relaxed
- conditions on nodes, conditions on edges (node pairs)
- all conditions mandatory or some optional
- dependencies among conditions

Keyword Search on Graphs: Ranking (1)

Result is **connected tree** with nodes that contain as many query keywords as possible

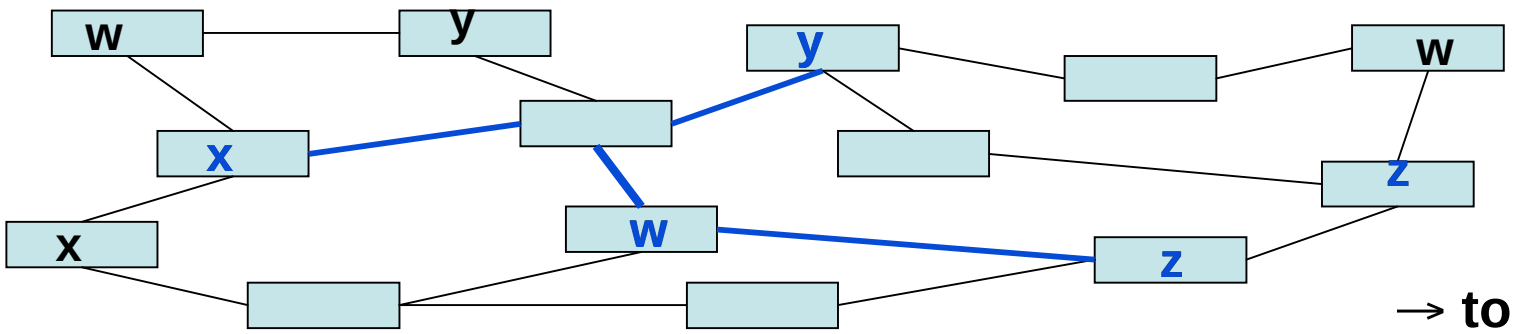
Ranking:

$$s(tree,q) = \alpha \cdot \sum_{nodes\ n} nodeScore(n,q) + (1-\alpha) \cdot \left(1 + \sum_{edges\ e} edgeScore(e)\right)^{-1}$$

with **nodeScore** based on tf*idf or prob. IR
and **edgeScore** reflecting importance of relationships (or confidence, authority, etc.)

Top-k querying: compute best trees, e.g. Steiner trees (NP-hard)

Example: keyword search „w x y z“ on relational-DB graph

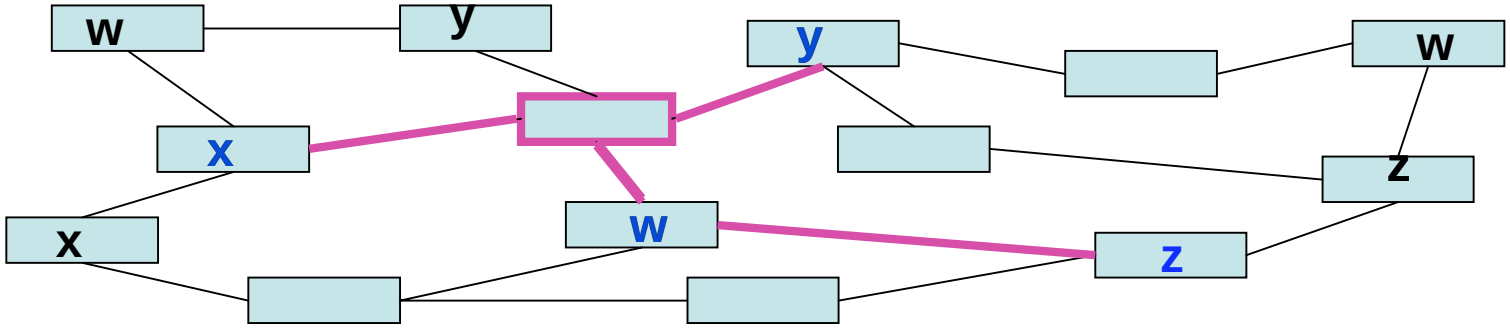


→ **top-k Steiner trees**

Keyword Search on Graphs: Ranking (2)

Define aggregation function to be **distributive**
rather than holistic [Kacholia et al. 2005, He et al. 2007]:
for $q=\{t_1, ..., t_m\}$ find best tree $(r, x_1, ..., x_m)$ rooted at r
according to $S = \Sigma_{i=1..m} S_{content}(x_i, t_i) + S_{path}(r, x_i)$
(aggregating shortest paths of matching nodes to root)

Example: keyword search „w x y z“ on relational-DB graph



Keyword Search on Graphs: Top-k QP (1)

[Graupmann et al.: VLDB 2005]

given: query with node conditions t1, …, tm

precompute
- inverted index **IX (term, node, nodescore)**
- shortest paths **SP (node1, node2, pathscore)**

to compute best Steiner tree use
2-approximation by **MST (minimum spanning tree):**
- evaluate t1, …, tm on IX:
  form m groups of candidate nodes in desc nodescore order
- compute MSTs for m-tuples from groups
- or better:
  - run **TA on m groups**
  - merge same-node entries from different groups
  - test **connectivity** and look up **pathscore in SP**
  - use additional thresholding heuristics

Keyword Search on Graphs: Top-k QP (2)

[Bhalotia et al.: ICDE 2002, Kacholia et al.: VLDB 2005]

Use distributive scoring model (with aggr. of shortest paths)
inverted index IX (term, node, nodescore)
simple neighbor index **NEIX (node1, node2, edgescore)**

- evaluate t1, …, tm on IX:
  form m groups of candidate nodes in desc nodescore order
- iterate over candidate nodes and candidate trees:
  - for each candidate node **backward-expand** its predecessor set, running shortest-path algorithm on NEIX
  - **combine nodes** into result-candidate tree when their **predecessor sets intersect**

- highly depends on expansion strategy (heuristics)
- extend with **forward-expansions** from result-candidate roots
- consider using degree-distribution statistics …

Keyword Search on Graphs: Top-k QP (3)

[He et al.: SIGMOD 2007]

Use distributive scoring model (with aggr. of shortest paths)
inverted index IX (term, node, nodescore)
+ keyword-path index **KPX (n1, k2, n2, pathscore)**
with shortest path from n1 to n2 containing k2

- evaluate t1, …, tm on IX:
  form m groups of cand. nodes in desc nodescore order
- iterate over candidate nodes and candidate trees:
  - run **backward expansion**, **forward expansion**, and evaluate KPX for candidate nodes and trees the **nearest matches of other keywords** using KPX
  - judiciously choose expansion nodes (various strategies)
  - use **TA-style threshold test** for pruning & stopping
  - actually use **bilevel index** instead of full KPX:
    - run graph partitioning on full data graph
    - precompute KPX for inter-partition graph and all partitions

Summary: Semantic IR

- **variety of „semantics“:** text + ontologies; relaxable XML; faceted data; vertical domains in Web; ER graphs;
- semantic enrichment facilitated by **info extraction & harvesting**
- **entity ranking** leverages & extends link analysis methods
- **graph IR** faces semantic subtleties and algorithmic complexity,
- needs **principled ranking models** and **efficient top-k QP**
- research trends: from keyword matching to **knowledge queries**; natural-language QA

Overview

- **Part 1: Web IR**
  - State of the Art
  - Scalability Challenge
  - Quality Challenge
  - Personalization
  - Research Opportunities
- **Part 2: Semantic & Social IR**
  - ✓ Ontologies in XML IR
  - ✓ Entity Search and Ranking
  - ✓ Graph IR
  - Web 2.0 Search and Mining
  - Research Opportunities

„Wisdom of Crowds“ at Work on Web 2.0

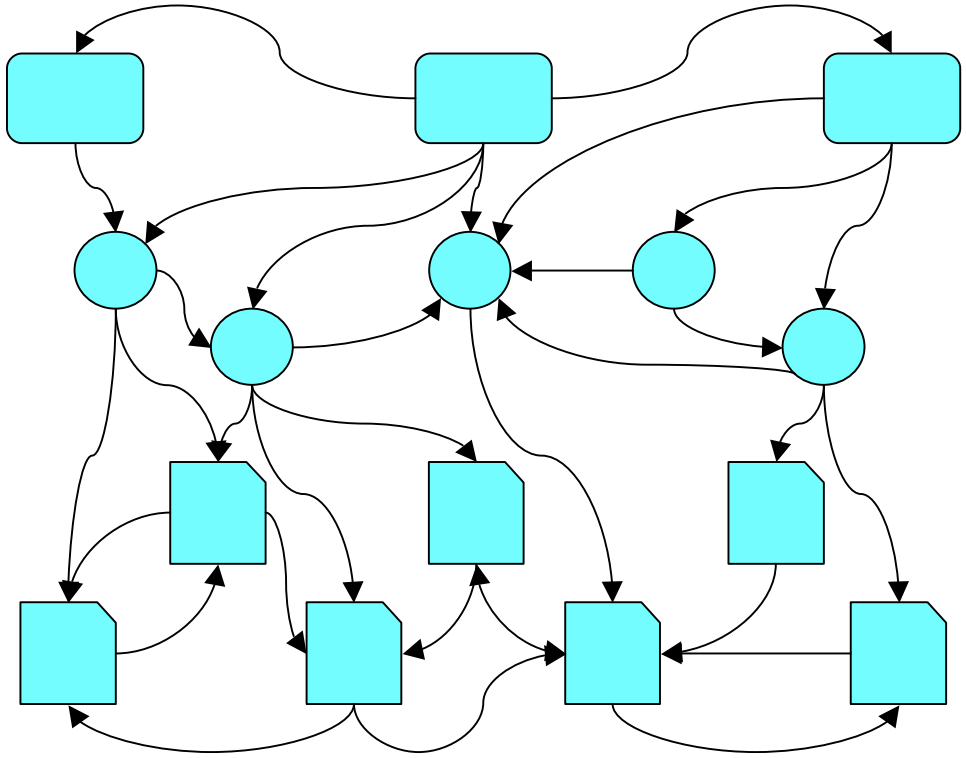
Information enrichment & knowledge extraction **by humans**:

- **Collaborative Recommendations & QA**
  - Amazon (product ratings & reviews, recommended products)
  - Netflix: movie DVD rentals → $ 1 Mio. Challenge
  - answers.yahoo, iknow.baidu, etc.
- **Social Tagging and Folksonomies**
  - del.icio.us: Web bookmarks and tags
  - flickr: photo annotation, categorization, rating
  - YouTube: same for video
- **Human Computing in Game Form**
  - ESP and Google Image Labeler: image tagging
  - Peekaboom: image segmenting and tagging
  - Verbosity: facts from natural-language sentences
- **Online Communities**
  - dblife.cs.wisc.edu for database research
  - www.lt-world.org for language technology
  - Yahoo! Groups, Myspace, Facebook, etc. etc.

Dark Side of Social Wisdom

- **Spam (Web & blog spam – not just for email anymore):**
  - lucky online casino, easy MBA diploma, cheap V!-4-gra, etc.;
  - law suits about „appropriate Google rank“
- **Truthiness:**
  - degree to which something is truthy (not necessarily facty);
  - truthy := property of something you know from your guts
- **Disputes:**
  - editorial fights over critical Wikipedia articles;
  - Citizendium: new endeavor with "gentle expert oversight"
- **Dishonesty, Bias, …**

The Wisdom of Crowds: Beyond PR



Typed graphs: data items, users, friends, groups,
postings, ratings, queries, clicks, …
with **weighted edges** → **spectral analysis** of various graphs

Evolving over time → tensor analysis

Social-Network Database

Simplified and cast into relational schema:

Users (UId, Nickname, …)

Docs (DId, Author, PostingDate, …)

Tags (TId, String)

Friendship (UId1, UId2, FScore)

Content (DId, TId, Score)

Rating (UId, DId, RScore)

Tagging (UId, TId, DId, TScore)

TagSim (TId1, TId2, TSim)

- Actually several kinds of „Friends“: same group, fan & star, true friend, etc.
- Tags could be typed or explicitly organized in hierarchies
- Numeric values for FScore, RScore, TScore, TSim
  may be explicitly specified or derived from co-occurrence statistics

Social-Network Graphs

Tagging relation is central:

- ternary relationship between users, tags, docs
- could be represented as hypergraph or tensor
- or (lossfully) decomposed into 3 binary projections (graphs):

UsersTags (UId, TId, UTscore)

$x.UTscore := \Sigma_d \{s \mid (x.UId, x.TId, d, s) \in Ratings\}$

TagsDocs (TId, DId, TDscore)

$x.TDscore := \Sigma_u \{s \mid (u, x.TId, x.DId, s) \in Ratings\}$

DocsUsers (DId, UId, DUscore)

$x.DUscore := \Sigma_t \{s \mid (x.UId, t, x.DId, s) \in Ratings\}$

Authority in Social Networks

Apply link analysis (PR etc.) to appropriately defined matrices

- **SocialPageRank** [Bao et al.: WWW 2007]:

  Let M_{UT}, M_{TD}, M_{DU} be the matrices corresponding to relations UsersTags, TagsDocs, DocsUsers

  Compute iteratively: $\vec{r}_U = M'_{DU} \times \vec{r}_D$

  $$\vec{r}_D = M'_{TD} \times \vec{r}_T$$

  $$\vec{r}_T = M'_{UT} \times \vec{r}_U$$

- **FolkRank** [Hotho et al.: ESWC 2006]:

  Define **graph G as union of graphs** UsersTags, TagsDocs, DocsUsers

  Assume each user has personal preference vector $\vec{p}$

  Compute iteratively: $\vec{r}_D = \alpha \vec{r}_D + \beta M_G \times \vec{r}_D + \gamma \vec{p}$

  FolkRank vector of docs is: $\vec{r}_D|_{\gamma>0} - \vec{r}_D|_{\gamma=0}$

Search & Ranking with Social Relations

Web search (or search in social network) can benefit from the "taste", "expertise", "experience", "recommendations" of friends

Naive method:
Look up your best friend's bookmarks or search with her tags

Combine content scoring with FolkRank, SocialPR, etc.

Additionally exploit tag co-occurrences in social network
[Bao et al.: WWW 2007, see also Jeh/Widom: KDD 2002]:

$sim(t_1, t_2) \sim aggr \{sim(d_1,d_2) \mid (t_1,d_1), (t_2,d_2) \in Tagging\}$

$sim(d_1, d_2) \sim aggr \{sim(t_1,t_2) \mid (t_1,d_1), (t_2,d_2) \in Tagging\}$

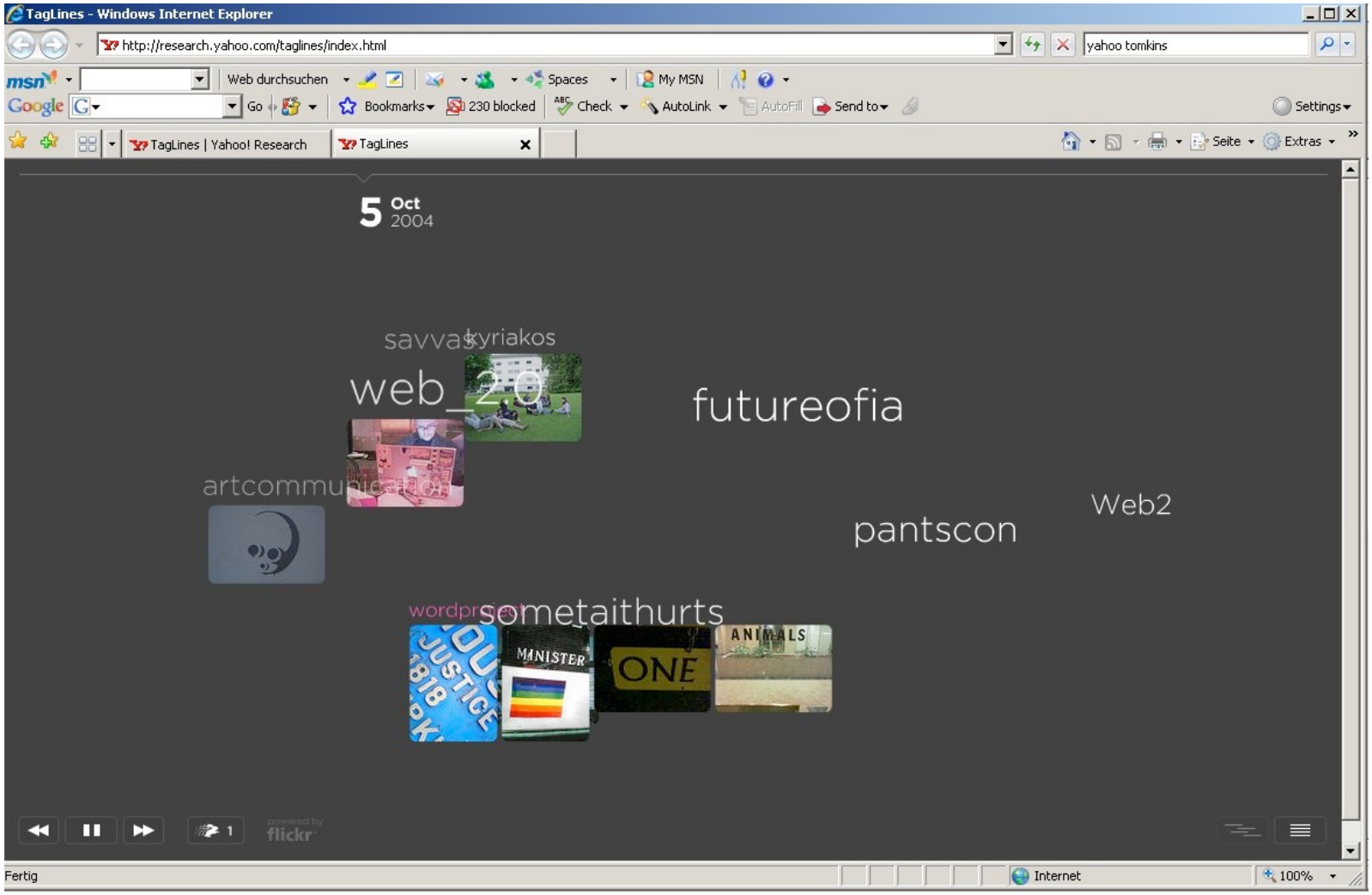
Integrate friendship strengths, tag similarities, user&page PR, e.g.:

$$s(q,d,u) = \sum_{t \in q} \sum_{c \in SimTags(t)} \sum_{f \in Friends(u)} TScore(f,c,d) \cdot TSim(t,c) \cdot FScore(u,f) \cdot UR(f) \cdot PR(d)$$

But: ranking models mostly ad hoc
efficient QP widely open

Tag Mining from Social Networks

Taglines [Dubinko/Kumar/Magnani/Novak/Raghavan/Tomkins WWW 2006]
http://research.yahoo.com/taglines



Tag Mining from Social Networks

Given: tag frequencies at daily resolution

Wanted: „most interesting“ tags for app-provided time intervals

Define **„interestingness“** of tag x for interval T

- Requirements
  - tag should be frequent in T and not so frequent at other times
  - tag with singular peaks in T should not dominate
- Approach:

  interestingness (x, T) =

  $\Sigma_{t \in T}$ **freq(x,t) / (C + freq(x,[0,∞)))**

  with regularization constant C

Tag Mining from Social Networks

Naive algorithm:

run TA over lists for all t in specified T, aggregating freq(x,t)

Additive algorithm:

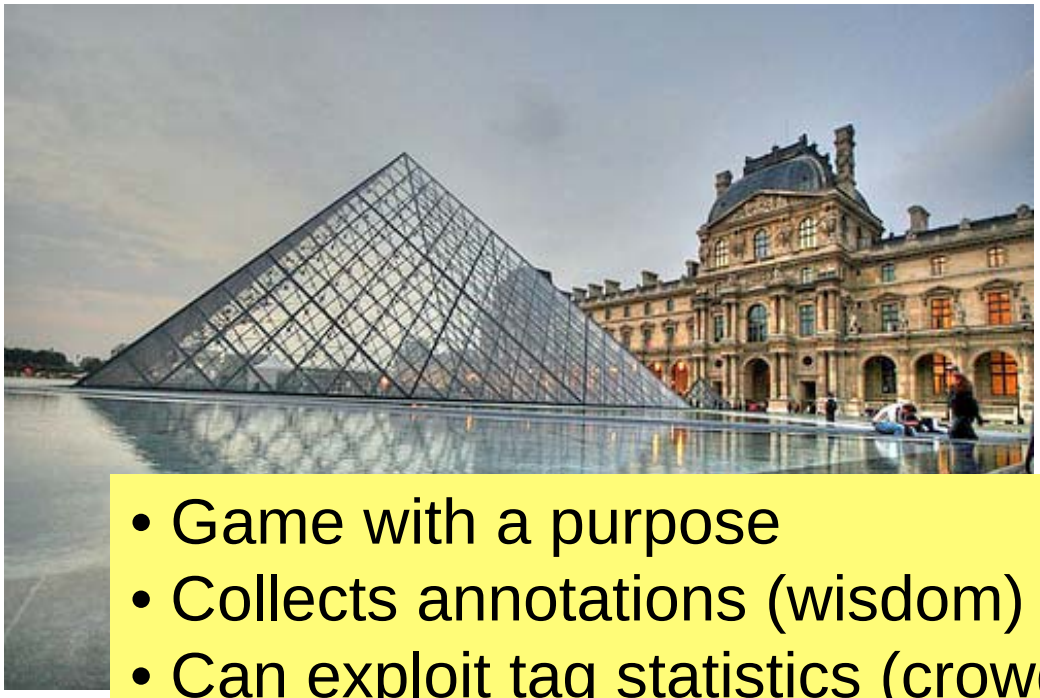
- **precompute** aggregated freq values for time intervals that start at and have lengths of powers of 2:
  [0,2), [2,3), …, [0,4), [4,8), …, [0, 8), [8, 16), …
- **decompose** query-specified T into intervals T_1, …, T_m covering T, mutually disjoint, of max. length
  run TA over the lists for T_1, …, T_m

Smart algorithm:

- represent query-specified T as **union and diff of intervals**
  $T = T_1 \cup ... \cup T_k - T_1' - ... - T_l'$ $(k+l < m)$
- **run TA** over these lists:
  T_1 ... T_k in desc freq order, T_1' … T_l' in asc freq order

Human Computing: ESP Game [Luis von Ahn et al. 2004]

played against random, anonymous partner on Internet



taboo:
pyramid
Louvre
museum
Paris
art

- Game with a purpose
- Collects annotations (wisdom)
- Can exploit tag statistics (crowds)
- Attracts people, fun to play, some play hours
- ESP game collected > 10 Mio. tags from > 20000 users
- 5000 people could tag all photos on the Web in 4 weeks (human computing)

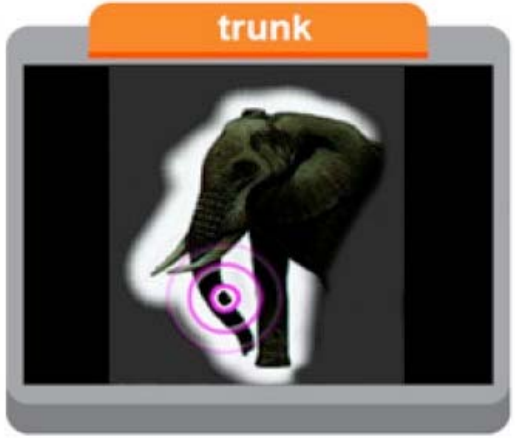
More Human Computing

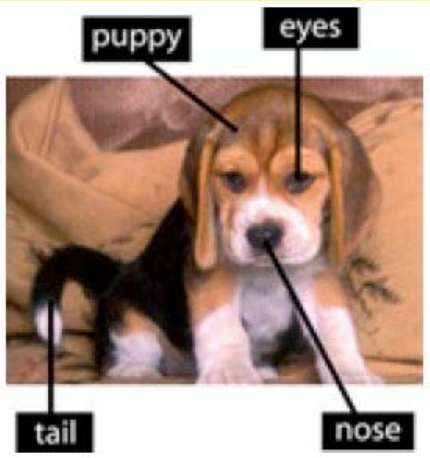
Verbosity [von Ahn 2006]:

- Collect common-knowledge facts (relation instances)
- 2 players: Narrator (N) and Guessor (G)
  N gives stylized clues:

  is a kind of …, is used for …, is typically near/in/on …, is the opposite of …, …
- random pairing for independence,
  can build statistics over many games for same concept

Peekaboom, Phetch, etc.:
locating & tagging objects





- incentives to play ?
- game design for moving up the value-chain ?

Summary: Social IR

- Great potential for leveraging **social networks** and **human computing**
- **Spectral analysis** methods applicable to ranking, but ranking models still not well understood
- Search result scoring should exploit **social tags** & **friendships**, but scoring models still not well understood
- **Query processing** becomes more difficult
- Managing very large **online-community sites** is difficult
- **Spam** occurs also in social networks („splog“)
- **Truthiness** (user-user correlations) and **temporal evolution** will be important issues
- Robust **reputation** and **trust** models will be crucial

Overview

- **Part 1: Web IR**
  - State of the Art
  - Scalability Challenge
  - Quality Challenge
  - Personalization
  - Research Opportunities
- **Part 2: Semantic & Social IR**
  - ✓ Ontologies in XML IR
  - ✓ Entity Search and Ranking
  - ✓ Graph IR
  - ✓ Web 2.0 Search and Mining
  - Research Opportunities

Semantic & Social IR: Research Opportunities

- large-scale **ontologies** and robust query expansion
- large-scale, **almost-unsupervised IE**; uncertain facts in QP
- **principled ranking** models and **efficient top-k QP**
  for knowledge queries on **ER graphs** (built by IE)
- general-purpose **Deep Web** search (without data integration)
- **principled models** for exploiting **social tagging & friendships**
- models for **reputation** and **trust**, robustness to **misbehavior**

- not covered in talk, but would be glad to discuss:
  data sets & usage logs, **experimental methodology**
- beyond scope, but relevant: **HCI, cognitive models, NLP**

Thank You !

Literature on Semantic & Social IR (1)

search with ontologies, facets, heterogeneity:

- S. Liu, F. Liu, C.T. Yu, W. Meng: An effective approach to document retrieval via utilizing WordNet and recognizing phrases. SIGIR 2004
- M. Theobald, R. Schenkel, G. Weikum: Efficient and self-tuning incremental query expansion for top-k query processing. SIGIR 2005
- W.W. Cohen: Data integration using similarity joins and a word-based information representation language. ACM Trans. Inf. Syst. 18(3), 2000
- S. Amer-Yahia et al.: Report on the DB/IR panel at SIGMOD 2005. ACM Sigmod Record 2005
- X. Zhou et al.: Query Relaxation Using Malleable Schemas. SIGMOD 2007
- K.C. Chang: Large-scale Deep Web Integration: Exploiting and Querying Structured Data on the Deep Web, Tutorial. SIGMOD 2006
- D. Suciu (Ed.): Special Issue Web-scale Data, Systems, Semantics. Data Eng. Bull. 31(4), 2006
- M. Hearst: Clustering versus faceted categories for information exploration. CACM 49(4), 2006
- J. Diederich, W.-T. Balke: The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. ECDL 2007
- H. Bast, A. Chitea, F. Suchanek, I. Weber: ESTER: Efficient Search on Text, Entities, and Relations. SIGIR 2007
- F. Suchanek, G. Kasneci, G. Weikum: YAGO: a Core of Semantic Knowledge Unifying WordNet and Wikipedia. WWW 2007
- A. Broder, M. Fontoura, V. Josifovski, L. Riedel: A Semantic Approach to Contextual Advertising. SIGIR 2007

Literature on Semantic & Social IR (2)

entity search, info extraction:

- S. Chakrabarti: Breaking Through the Syntax Barrier: Searching with Entities and Relations. ECML 2004.
- Z. Nie, J. Wen, W. Ma: Object-level Vertical Search. CIDR 2007.
- Z. Nie, Y. JMa, S. Shi, J. Wen, W. Ma: Web Object Retrieval. WWW 2007
- Z. Nie, Y. Zhang, J. Wen, W. Ma: Object-Level Ranking. WWW 2005
- A. Balmin, V. Hristidis, Y. Papakonstantinou: ObjectRank: Authority-based Keyword Search in Databases. VLDB 2004
- S. Chakrabarti: Dynamic Personalized Pagerank in Entity-Relation Graphs. WWW 2007
- J. Stoyanovich, S. Bedathur, K. Berberich, G. Weikum: EntityAuthority Semantically Enriched Graph-Based Authority Propagation. WebDB 2007
- T. Cheng, X. Yan, K. C.-C. Chang: EntityRank: Searching Entities Directly and Holistically. VLDB 2007
- E. Agichtein, S. Sarawagi: Scalable Information Extraction and Integration. Tutorial. KDD 2006
- A. Doan et al.: Managing Information Extraction, Tutorial. SIGMOD 2006
- W.W. Cohen: Information Extraction, Tutorial. http://www.cs.cmu.edu/~wcohen/ie-survey.ppt
- H. Cunningham: An Introduction to Information Extraction. Encyclopedia of Lang. & Ling. 2005
- O.Etzioni et al.: Unsupervised Named-Entity Extraction from the Web. Artif. Intell. 165(1), 2005
- M. Banko et al.: Open Information Extraction from the Web. IJCAI 2007
- F.M. Suchanek et al.: Combining linguistic and statistical analysis to extract relations from web documents. KDD 2006

Literature on Semantic & Social IR (3)

knowledge search, graph IR:

- M.J. Cafarella, C. Re, D. Suciu, O. Etzioni: Structured Querying of Web Text Data. CIDR 2007
- K. Anyanwu, A. Maduko, A. Sheth, SPARQ2L: Towards Support For Subgraph Extraction Queries in RDF Databases. WWW 2007.
- G. Kasneci et al.: NAGA: Searching and Ranking Knowledge. MPII Technical Report, 2007.
- B. Kimelfeld, Y. Sagiv: Finding and Approximating Top-k Answers in Keyword Proximity Search. PODS 2006
- S. Cohen, Yaron Kanza, Benny Kimelfeld, Yehoshua Sagiv: Interconnection Semantics for Keyword Search in XML. CIKM 2005.
- B. Kimelfeld, Y.Sagiv: Combining Incompleteness and Ranking in Tree Queries. ICDT 2007
- V. Kacholia et al.: Bidirectional Expansion For Keyword Search on Graph Databases. VLDB 2005
- H.He, H.Wang, J.Yang, P.Yu: BLINKS: Ranked Keyword Searches on Graphs. SIGMOD 2007
- B. Ding et al.: Finding Top-k Min-Cost Connected Trees in Databases. ICDE 2007
- J. Graupmann, R. Schenkel, G. Weikum: The SphereSearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents. VLDB 2005.
- S. Agrawal, S. Chaudhuri, G. Das: DBXplorer: A System for Keyword-Based Search over Relational Databases. ICDE 2002.
- V. Hristidis, Y. Papakonstantinou: DISCOVER: Keyword Search in Relational Databases. VLDB 2002
- G. Bhalotia et al.: Keyword Searching and Browsing in Databases using BANKS. ICDE 2002

Literature on Semantic & Social IR (4)

social IR:

- N. Koudas (Ed.): Special issue on Data Management Issues in Social Sciences. IEEE Data Engineering Bulletin 30(2), 2007
- N. Bansal, N. Koudas, Searching the Blogosphere. WebDB 2007
- L. von Ahn: Games with a Purpose. IEEE Computer 39(6), 2006
- L.von Ahn, M.Kedia, M.Blum: Verbosity: a game for collecting common-sense facts. CHI 2006
- A. Hotho, R. Jäschke, C. Schmitz, G. Stumme: Information Retrieval in Folksonomies: Search and Ranking. ESWC 2006
- S. Bao, X. Wu, B. Fei, G. Xue, Z. Su, Y. Yu: Optimizing Web Search Using Social Annotation. WWW 2007
- S. Marti, P. Ganesan, H. Garcia-Molina: DHT Routing using Social Links. IPTPS 2004
- A. Mislove, K. Gummadi, P. Druschel: Exploiting Social Networks for Internet Search. HotNets 2006
- M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, A. Tomkins: Visualizing tags over time. WWW 2006
- S. Golder, B.A. Huberman: Usage Patterns of Collaborative Tagging Systems. Journal of Information Science 32(2), 2006
- R. Ramakrishnan: Community Systems: The World Online. Keynote Slides. CIDR 2007