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)

<table>
<thead>
<tr>
<th>Unstructured search (keywords)</th>
<th>Structured search (SQL,XQuery)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword Search on Relational Graphs (BANKS, Discover, DBexplorer, …) + Web 2.0</td>
<td>IR Systems Search Engines + Digital Libraries + Enterprise Search</td>
</tr>
<tr>
<td>DB Systems + Text + Relax. &amp; Approx. + Ranking</td>
<td>Querying entities &amp; relations from IE (Libra, ExDB, NAGA, … )</td>
</tr>
</tbody>
</table>

Structured data (records) Unstructured data (documents)

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?

- **Professor**
  - Name: Gerhard Weikum
  - Address: Max-Planck Institute for Informatics, Germany
  - Teaching:
    - Course: IR
    - Description: Information retrieval ...
    - Syllabus: ...
    - Book: ...
    - Article: ...
  - City: SB
  - Country: Germany
  - Research:
    - Title: Intelligent Search of Heterogeneous XML Data
  - Project:
    - Title: INEX task coordinator (Initiative for the Evaluation of XML ...)
    - Contents: Ranked retrieval ...
    - Literature: ...
    - Funding: EU
  - Activities:
    - Seminar
    - Scientific
    - Other
  - Lecturer: Ralf Schenkel
  - Name: Ralf Schenkel
  - Address: Max-Planck Institute for Informatics, Germany
  - Activities:
    - Seminar
    - Scientific
    - Other
  - Sponsor: EU
  - Name: EU
  - Sponsorship: EU
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?

**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 \( q: \sim \text{professor research} \)
with expansions \( \exp(i) = \{ w \mid \text{sim}(i, w) \geq \theta \} \)
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

<table>
<thead>
<tr>
<th>research</th>
<th>professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>57: 0.6</td>
<td>44: 0.4</td>
</tr>
<tr>
<td>44: 0.5</td>
<td>33: 0.3</td>
</tr>
<tr>
<td>52: 0.4</td>
<td>75: 0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lecturer</th>
<th>scholar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>57: 0.9</td>
</tr>
<tr>
<td>44: 0.8</td>
</tr>
<tr>
<td>22: 0.7</td>
</tr>
<tr>
<td>23: 0.6</td>
</tr>
<tr>
<td>51: 0.6</td>
</tr>
<tr>
<td>52: 0.6</td>
</tr>
<tr>
<td>92: 0.9</td>
</tr>
<tr>
<td>67: 0.9</td>
</tr>
<tr>
<td>52: 0.9</td>
</tr>
<tr>
<td>44: 0.8</td>
</tr>
<tr>
<td>55: 0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scholar</th>
</tr>
</thead>
<tbody>
<tr>
<td>67: 0.9</td>
</tr>
<tr>
<td>52: 0.9</td>
</tr>
<tr>
<td>44: 0.8</td>
</tr>
<tr>
<td>55: 0.8</td>
</tr>
</tbody>
</table>

ontology / meta-index

<table>
<thead>
<tr>
<th>professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>lecturer: 0.7</td>
</tr>
<tr>
<td>scholar: 0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>academic</th>
</tr>
</thead>
<tbody>
<tr>
<td>53: 0.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scientist</th>
</tr>
</thead>
<tbody>
<tr>
<td>44: 0.5</td>
</tr>
</tbody>
</table>

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**

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

Query:
$x$ isa politician
$x$ isa scientist

Results:
Benjamin Franklin
Paul Wolfowitz
Angela Merkel

Score: 6.247457440558111E-7
'scientist' — meanings -> scientist_110560637
Benjamin_Franklin — type -> Massachusetts_politicians
'politician' — meanings -> politician_110451263
American_scientists — subClassOf -> scientist_110560637
Benjamin_Franklin — type -> American_scientists
Massachusetts_politicians — subClassOf -> politician_110451263

- $x$@politician = politician_110451263
- $x$@scientist = scientist_110560637
- $\forall x =$ Benjamin_Franklin

Score: 3.185850362140424E-7
'scientist' — meanings -> scientist_110560637
'politician' — meanings -> 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

- $x$@politician = politician_110451263
- $x$@scientist = scientist_110560637
- $\forall x =$ Paul_Wolfowitz

Score: 1.121658976926192E-7
Angela_Merkel — type -> German_scientists
'scientist' — meanings -> scientist_110560637
German_Christian_Democrat_politicians — subClassOf ->
politician_110451263
Angela_Merkel — type -> German_Christian_Democrat_politicians
'politician' — meanings -> politician_110451263
German_scientists — subClassOf -> scientist_110560637

- $x$@politician = politician_110451263
- $x$@scientist = scientist_110560637
- $\forall x =$ Angela_Merkel
Entity Search: Example DBLife

http://dblife.cs.wisc.edu

Divesh Srivastava

Mentions 1 - 10 out of 526

http://www.research.att.com/~divesh/

Announced

AT&T, USA

528 total mentions occurring in 161 pages. 0 new mentions found in the last 24 hours.

Related People
- Nick Koukas
- H. V. Jagadish
- S. Sudarshan
- Raghu Ramakrishnan
more

Related Topics
- xml
- logic
- transactions
- query processing
more

Services
- SIGMOD 2007 (Session chair)
- VLDB 2007 (Session chair)
- VLDB 2007 (Program chair)
- SIGMOD 2007 (PC member)
more

Publications
- Intensional associations between data and metadata
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:
- focused crawling & Deep-Web crawling
- record extraction (named entity, attributes)
- record linkage & aggregation (entity matching)
- keyword / record search (faceted GUI)
- entity ranking
Information Extraction (IE): Text to Records

<table>
<thead>
<tr>
<th>Person</th>
<th>BirthDate</th>
<th>BirthPlace</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>4/23, 1858</td>
<td>Kiel</td>
<td></td>
</tr>
<tr>
<td>Albert Einstein</td>
<td>3/14, 1879</td>
<td>Ulm</td>
<td></td>
</tr>
<tr>
<td>Mahatma Gandhi</td>
<td>10/2, 1869</td>
<td>Porbandar</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>ScientificResult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>Quantum Theory</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Collaborator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>Albert Einstein</td>
</tr>
<tr>
<td>Max Planck</td>
<td>Niels Bohr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td>KWG / MPG</td>
</tr>
</tbody>
</table>

Max 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 developed his 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 emerged later in cooperation with Albert Einstein and Niels Bohr.

Combined NLP, pattern matching, lexicons, statistical learning

extracted facts often have confidence < 1
→ DB with uncertainty (probabilistic DB)
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!**, …

Ian Foster, father of the Grid, talks at the **GES conference** in **Germany** on **05/02/07**.
Entity-Search Ranking with LM


Standard LM for docs with background model (smoothing):
\[
s(d, q) = \lambda P[q | d] + (1 - \lambda) P[q] \quad P[q | d] \sim \sum_{w \in q} \log \frac{tf(w,d)}{\sum_u tf(u,d)}
\]

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

\[
P[w | e] \sim \sum_i \alpha_i P[w | \text{context}(r_i, d_i)] \quad \text{record-level LM}
\]

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

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

Alternatively consider individual attributes \( e.a_j \) with importance \( \beta_j \) extracted from page \( d_i \) with accuracy \( \gamma_{ij} \)

\[
P[w | e] \sim \sum_i \alpha_i \sum_j \beta_j \gamma_{ij} P[w | \text{context}(r_i.a_j, d_i)]
\]
Entity Authority (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 Ranking by Link Analysis (2)

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

\[
\tilde{r}_e \sim \alpha M_{e\rightarrow e} \times \tilde{r}_e + (1 - \alpha) M_{p\rightarrow e} \times \tilde{r}_p
\]

\[
\tilde{r}_p \sim \beta M_{p\rightarrow p} \times \tilde{r}_p + (1 - \beta) M_{e\rightarrow p} \times \tilde{r}_e
\]

• 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

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th># Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>KnowItAll</td>
<td>30 000</td>
</tr>
<tr>
<td>SUMO</td>
<td>60 000</td>
</tr>
<tr>
<td>WordNet</td>
<td>200 000</td>
</tr>
<tr>
<td>OpenCyc</td>
<td>300 000</td>
</tr>
<tr>
<td>Cyc</td>
<td>5 000 000</td>
</tr>
<tr>
<td>YAGO</td>
<td>6 000 000</td>
</tr>
</tbody>
</table>

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

Accuracy ≈ 97%

Gerhard Weikum, EDBT 2007 Summer School
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, ...)

Gerhard Weikum, EDBT 2007 Summer School
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 Extraction

seeds → text → rules → new facts

Example:

in downtown Seattle
Seattle and other towns
City (Las Vegas) Las Vegas and other towns
plays (Zappa, guitar) playing guitar: Zappa
plays (Davis, trumpet) Davis blows trumpet

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

We visited Paris last summer. It has many museums along the banks of the Seine.
NAGA: Graph IR on YAGO

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

**Discovery queries**

```
Kiel bornIn $x isa scientist
```

**Connectedness queries**

```
German novelist isa Thomas Mann
```

**Queries with regular expressions**

```
Ling hasFirstName | hasLastName
```

```
Beng Chin Ooi
```

```
Zhejiang
```

```
Nobel prize hasWon $x
```

```
Thomas Mann hasSon $y
```

```
Thomas Mann diedOn $a
```

```
Kiel bornIn $x isa scientist
```

```
Thomas Mann diedOn $y
```

```
Beng Chin Ooi worksFor $y
```

```
Beng Chin Ooi locatedIn $y
```

```
Beng Chin Ooi (coAuthor | advisor) $x
```

```
Nobel prize hasWon $x
```

```
Nobel prize hasWon $b
```

```
Goethe isa Thomas Mann
```

```
Goethe isa German novelist
```

```
Thomas Mann isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Thomas Mann isa German novelist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```

```
Beng Chin Ooi isa scientist
```
Search Results Without Ranking

$q$: Fisher isa scientist

Fisher isa $x$

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

$X = 
alumnus_109165182$

$@Fisher = Irving_Fisher$
$@scientist = scientist_109871938$

$X = 
social_scientist_109927304$

$@Fisher = James_Fisher$
$@scientist = scientist_10981938$

$X = 
ornithologist_109711173$

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

$X = 
thorist_110008610$

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

$X = 
colleague_109301221$

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

$X = 
organism_100003226$

...
Ranking with Statistical Language Model

NAGA: Searching and Ranking Knowledge

NAGA is a new semantic search engine. It uses a knowledge base, which is organized as a graph with typed edges. This knowledge base is a projection of Yago and consists of millions of entities and relationships automatically extracted from Web-based corpora. Our query language is capable of expressing keyword search for the casual user as well as graph queries with regular expressions for the expert user. Furthermore, it enables the formulation of queries with additional semantic information. The results are ranked due to a novelty scoring model, based on the principles of generative language models, which formulates several notions like confidence, informativeness and compactness. NAGA is being developed at the Max-Planck-Institute Saarbrücken. For details on NAGA, take a look at our technical report.

- Gitak Kasneci, Fabian M. Strohmeier, Georgios Stin, Marco Ramerstorfer, Gerhard Weikum
  "NAGA: Searching and Ranking Knowledge" (pdf, bib, slides)
  Technical Report (MPII 2007)

A query has the form

\[
ERI_1 R ERI_2
\]

where the ERI's are entities (e.g. Einstein) and the RI's are relations. The following relations are allowed:

- `familyNameOf` relation
- `type` relation
- `means` relation
- `subClassOf` relation

Score: \(7.184462521168058 \times 10^{-13}\) mathematician_109635652

q: Fisher isa scientist
Fisher isa \(x\)

\[@Fisher = Ronald_Fisher\]

\[@scientist = scientist_109871938\]

\(X = mathematician_109635652\)

\[@Fisher = Ronald_Fisher\]

\[@scientist = scientist_109871938\]

\(X = statistician_109958989\)

\[@Fisher = Ronald_Fisher\]

\[@scientist = scientist_109871938\]

\(X = president_109787431\)

\[@Fisher = Ronald_Fisher\]

\[@scientist = scientist_109871938\]

\(X = geneticist_109475749\)

\[@Fisher = Ronald_Fisher\]

\[@scientist = scientist_109871938\]

\(X = scientist_109871938\)

...
NAGA: Searching & Ranking Knowledge

NAGA is a new semantic search engine. It uses a knowledge base, which is organized as a graph with typed edges. This knowledge base is a projection of YAGO and consists of millions of entities and relationships automatically extracted from Web-based corpora. Our query language is capable of expressing keyword search for both users and as graph queries with regular expressions for the expert user. Furthermore, it enables the formulation of queries with additional semantic information. The results are ranked due to a novel scoring model, based on the principle of conservative language models, which formulates several notions like confidence, informativeness and compactness. NAGA is being developed at the Max-Planck-Institute Saarbrücken. For details on NAGA, take a look at our technical report.

- Gitež Kafanci, Fabian M. Suchanek, Georgios Stin, Marco Ramaz
  "NAGA: Searching and Ranking Knowledge" (pdf, bib, slides)
  Technical Report (MPII 2007)

A query has the form

```
$X isa Scientist  
$X hasWonPrize $y  
$y context Literature
```

Carl_Sagan —hasWonPrize—> Pulitzer_Prize

E._O._Wilson —hasWonPrize—> Pulitzer_Prize

Bertrand_Russell —hasWonPrize—> Nobel_Prize_in_Literature

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

Gerhard Weikum, EDBT 2007 Summer School
**Ranking Factors**

**Confidence:** 
Prefer results that are likely to be correct  
- Certainty of IE  
- Authenticity and Authority of Sources

**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

**Compactness:** 
Prefer results that are tightly connected  
- Size of answer graph

---

- `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)

**Example Queries:**

- `q: isa (Einstein, $y)`
  - `isa (Einstein, scientist)`
  - `isa (Einstein, vegetarian)`

- `q: isa ($x, vegetarian)`
  - `isa (Einstein, vegetarian)`
  - `isa (Al Nobody, vegetarian)`

---

![Diagram showing relationships between concepts]

- `isa (vegetarian`)
- `isa (Tom Cruise)`
- `isa (Einstein)`
- `isa (Bohr)`
- `isa (Nobel Prize)`
- `bornIn (Max Planck, Kiel)`
- `livesIn (Elvis Presley, Mars)`
- `won (Einstein, Nobel Prize)`
- `diedIn (Bohr)`
- `1962`
NAGA Ranking Model

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

For query $q$ with fact templates $q_1 \ldots q_n$ rank result graphs $g$ with facts $g_1 \ldots g_n$ by decreasing likelihoods:

using generative mixture model

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

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

based on IE accuracy and authority analysis

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

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

estimated by correlation statistics

For $q_i = (x, r, z)$ with variable $x$
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)  Journals (JId, Title)
CPublications (PId, Title, CId)  JPublications (PId, Title, Vol, No, Year)
Authors (PId, Person)  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:**

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

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
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 \( S = \sum_{i=1}^{m} S_{content}(x_i, t_i) + S_{path}(r, x_i) \)

(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
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 …
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 $\rightarrow$ spectral analysis of various graphs

Evolving over time $\rightarrow$ 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** \((UId, TId, UTscore)\)
  \[ x.UTscore := \sum_d \{ s \mid (x.UId, x.TId, d, s) \in \text{Ratings} \} \]

  **TagsDocs** \((TId, Did, TDscore)\)
  \[ x.TDscore := \sum_u \{ s \mid (u, x.TId, x.DId, s) \in \text{Ratings} \} \]

  **DocsUsers** \((DId, UId, DUscore)\)
  \[ x.DUscore := \sum_t \{ s \mid (x.UId, t, x.DId, s) \in \text{Ratings} \} \]
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** [Hothen 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(t₁, t₂) \sim \text{aggr} \{\text{sim}(d₁, d₂) | (t₁, d₁), (t₂, d₂) \in \text{Tagging}\}
sim(d₁, d₂) \sim \text{aggr} \{\text{sim}(t₁, t₂) | (t₁, d₁), (t₂, d₂) \in \text{Tagging}\}

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

\[ s(q, d, u) = \sum_{t \in q} \sum_{c \in \text{SimTags}(t)} \sum_{f \in \text{Friends}(u)} T\text{Score}(f, c, d) \cdot T\text{Sim}(t, c) \cdot F\text{Score}(u, f) \cdot \text{UR}(f) \cdot \text{PR}(d) \]

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:

\[
\text{interestingness} (x, T) = \frac{\sum_{t \in T} \text{freq}(x,t)}{(C + \text{freq}(x,[0,\infty)))}
\]

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:
  - \( T_1 \ldots T_k \) in desc freq order, \( T_{1'} \ldots T_{l'} \) in asc freq order
Human Computing: ESP Game  
[Luis von Ahn et al. 2004 ]

played against random, anonymous partner on Internet

- 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:

- 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
- M. Hearst: Clustering versus faceted categories for information exploration. CACM 49(4), 2006


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, 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
• J. Stoyanovich, S. Bedathur, K. Berberich, G. Weikum: EntityAuthority Semantically Enriched Graph-Based Authority Propagation. WebDB 2007
• E. Agichtein, S. Sarawagi: Scalable Information Extraction and Integration. Tutorial. KDD 2006
• A. Doan et al.: Managing Information Extraction, Tutorial. SIGMOD 2006
• 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:**

- B. Kimelfeld, Y. Sagiv: Finding and Approximating Top-k Answers in Keyword Proximity Search. PODS 2006
- B. Kimelfeld, Y. Sagiv: Combining Incompleteness and Ranking in Tree Queries. ICDT 2007
- V. Kacholian et al.: Bidirectional Expansion For Keyword Search on Graph Databases. VLDB 2005
- B. Ding et al.: Finding Top-k Min-Cost Connected Trees in Databases. ICDE 2007
- 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
social IR:
- N. Bansal, N. Koudas, Searching the Blogosphere. WebDB 2007