Intelligente Suchmaschinen der Zukunft

Gerhard Weikum

weikum@mpi-inf.mpg.de
http://www.mpi-inf.mpg.de/~weikum/
Beyond Google: Knowledge Queries

Turn the Web, Web2.0, and Web3.0 into the world’s most comprehensive knowledge base („semantic DB“)!

Answer „knowledge queries“ such as:

- proteins that inhibit both proteases and some other enzymes
- neutron stars with Xray bursts > 10^{40} \text{ erg s}^{-1} & black holes in 10^{40} \text{ s}^{-1}
- differences in Rembetiko music from Greece and from Turkey
- connection between Thomas Mann and Goethe
- market impact of Web2.0 technology in December 2006
- sympathy or antipathy for Germany from May to August 2006
- Nobel laureate who survived both world wars and his children
- drama with three women making a prophecy to a British nobleman that he will become king
Don't Let Me Be Misunderstood

Keyword query: Max Planck

or

Entity query: Person = „Max Planck“

Keyword query: Greek art Paris

or

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

Keyword query: Madonna child

or

Entity-Relation query: Person x = „Madonna“ & HasChild (x,y)
What is lacking?

- data is not knowledge  
  → extraction and organization
- keywords cannot express advanced user intentions  
  → concepts, entities, properties, relations
Three Roads to Web Knowledge and Intelligent Search

Leibniz Approach: **Semantic Web** from Handcrafted Knowledge Sources and Search over Entities, Relations, Ontologies

Planck Approach: **Statistical Web** from Large-scale Information Extraction & Harvesting and Search over Uncertain Relations

Darwin Approach: **Social Wisdom** from Web 2.0 Communities and Search by Collaborative Recommendation
Outline

✓ Motivation

- Leibniz: Semantic Web
- Planck: Statistical Web
- Darwin: Social Web

- Conclusion
Leibniz Approach: Semantic Web

Handcrafted High-Quality Knowledge:
- Ontologies and other Lexical Sources
- Build on Rigorous Knowledge Atoms („Characteristica Universalis“)
- Semantic Search over Entities, Relations, Ontologies

Gottfried Wilhelm Leibniz (1646 - 1716)
High-Quality Knowledge Sources
General-purpose thesauri and concept networks: WordNet family

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)

---

=> pezzled -- (a shameless impudent scurrilous woman)
=> jilt -- (a woman who jits a lover)
=> lady -- (a polite name for any woman)
=> mephit -- (an unnaturally frenzied or distraught woman)
=> matron, head nurse -- (a woman in charge of nursing in a medical institution)
General-purpose thesauri and concept networks: WordNet family

200 000 concepts and relations; can be cast into
• description logics or
• graph, with weights for relation strengths (derived from co-occurrence statistics)

enzyme -- (any of several complex proteins that are produced by cells and act as catalysts)
  ➞ protein -- (any of a large group of nitrogenous organic compounds that are essential constituents of living cells; ...)
  ➞ macromolecule, supermolecule

  ➞ organic compound -- (any compound of carbon and another element or a radical)

  ➞ catalyst, accelerator -- ((chemistry) a substance that initiates or accelerates a chemical reaction without itself being affected)
  ➞ activator -- ((biology) any agency bringing about activation; ...)

max planeck institut informatik
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.

---

1 sense of organized crime

**Sense 1**

*organized crime*, gangland, gangdom -- (underworld organizations)

  => yakuza -- (organized crime in Japan; an alliance of criminal organizations and illegal enterprises)

  => Mafia, Maffia, Sicilian Mafia -- (a secret terrorist group in Sicily; originally opposed tyranny but evolved into a criminal organization in the middle of the 19th century)

  => Black Hand -- (a secret terrorist society in the United States early in the 20th century)

  => Camorra -- (a secret society in Naples notorious for violence and blackmail)

  => syndicate, crime syndicate, mob, family -- (a loose affiliation of gangsters in charge of organized criminal activities)
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], mafia[0.154|1.00], "sicilian[0.201|1.00],
"black[0.066|1.00] & hand[0.053|1.00],
organ[0.213|1.00], crime[0.312|1.00],
criminal[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

... 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.

... 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, ...
Exploit Hand-Crafted Knowledge

Wikipedia, WordNet, and other lexical sources

Max Karl Ernst Ludwig Planck (April 23, 1858 – October 4, 1947 in Göttingen, Germany) was a German physicist. He is considered to be the founder of quantum theory, and therefore one of the most important physicists of the twentieth century.

**Contents**

1 Life and work
   1.1 Early Childhood
   1.2 Education
   1.3 Academic career
   1.4 Family
   1.5 Professor at Berlin University
   1.6 Black-body radiation
   1.7 Einstein and the Theory of Relativity
   1.8 World War and Weimar Republic
   1.9 Quantum mechanics
   1.10 Nazi dictatorship and Second World War
2 Honours and medals
3 See also
4 Publications
5 Bibliography
6 External links
   6.1 Biographies
   6.2 Articles
7 Notes

**Life and work**

**Early Childhood**

Planck came from a traditional, intellectual family. His paternal great-grandfather and grandfather were both theology professors in Göttingen, his father was a law professor in Kiel and Munich, and his paternal uncle was a judge.

Planck was born in Kiel to Johann Julius Wilhelm Planck and his second wife, Emma Patzig. He was the sixth child in the family, though two of his siblings were from his father's first marriage. Among his earliest memories was the marching of Prussian and
Exploit Hand-Crafted Knowledge

Wikipedia, WordNet, and other lexical sources

{{Infobox_Scientist
 | name = Max Planck
 | birth_date = [[April 23]], [[1858]]
 | birth_place = [[Kiel]], [[Germany]]
 | death_date = [[October 4]], [[1947]]
 | death_place = [[Göttingen]], [[Germany]]
 | residence = [[Germany]]
 | nationality = [[Germany|German]]
 | field = [[Physicist]]
 | work_institution = [[University of Kiel]]
 | Humboldt-Universität zu Berlin]
 | Georg-August-Universität Göttingen]
 | alma_mater = [[Ludwig-Maximilians-Universität München]]
 | doctoral_advisor = [[Philipp von Jolly]]
 | doctoral_students =
 | Gustav Ludwig Hertz]]
 | known_for = [[Planck's constant]],
 | Quantum mechanics|quantum theory]]
 | prizes = [[Nobel Prize in Physics]] (1918)
 | ...}}
Exploit Hand-Crafted Knowledge

Wikipedia, WordNet, and other lexical sources

Nobel Prize in Physics: Laureates (1901-1925)


Categories: 1858 births | 1947 deaths | Cornell University faculty | German Nobel laureates | German physicists | Members of the Pontifical Academy of Sciences | Nobel laureates in Physics | Particle physics | People from Kiel | Quantum theory physicists | Thermodynamicists | Humboldt University of Berlin alumni | University of Munich alumni
YAGO Knowledge Base

<table>
<thead>
<tr>
<th>Entities</th>
<th>Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>KnowItAll</td>
<td>30 000</td>
</tr>
<tr>
<td>SUMO</td>
<td>20 000</td>
</tr>
<tr>
<td>WordNet</td>
<td>120 000</td>
</tr>
<tr>
<td>Cyc</td>
<td>300 000</td>
</tr>
<tr>
<td>TextRunner</td>
<td>n/a</td>
</tr>
<tr>
<td>YAGO</td>
<td>1.7 Mio.</td>
</tr>
<tr>
<td>DBpedia</td>
<td>1.9 Mio.</td>
</tr>
<tr>
<td></td>
<td>60 000</td>
</tr>
<tr>
<td></td>
<td>800 000</td>
</tr>
<tr>
<td></td>
<td>5 Mio.</td>
</tr>
<tr>
<td></td>
<td>8 Mio.</td>
</tr>
<tr>
<td></td>
<td>15 Mio.</td>
</tr>
<tr>
<td></td>
<td>103 Mio.</td>
</tr>
</tbody>
</table>

Accuracy ≈ 95%

Online access and download at [http://www.mpi-inf.mpg.de/~suchanek/yago/](http://www.mpi-inf.mpg.de/~suchanek/yago/)
NAGA: Graph IR on YAGO [G. Kasneci et al.: WWW’07, ICDE’08]

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

**Discovery queries**

- $x$ isa scientist
  - bornIn Kiel

**Connectedness queries**

- isa German novelist
  - Thomas Mann -> Goethe
  - isa Nobel prize
    - hasWon $x$
    - diedOn $a$
    - hasSon $b$
    - diedOn $b$

**Queries with regular expressions**

- $x$ isa scientist
  - hasFirstName | hasLastName
  - (coAuthor | advisor)*
    - worksFor Beng Chin Ooi
  - locatedIn* Zhejiang
Motivation

Leibniz: Semantic Web

- Planck: Statistical Web
- Darwin: Social Web

Conclusion
Planck Approach: Statistical Web

Information Extraction & Harvesting:
• Gather Entities, Relations, Facts
• Live with Uncertainty
• Search & Rank Knowledge Facts

Max Planck (1858 - 1947)
### 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 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 Radiation Law. This law became the basis of quantum theory, which emerged later in cooperation with Albert Einstein and Niels Bohr.

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

**Planck’s constant**

\[ h = 6.626 \times 10^{-23} \text{Js} \]

**Collaborator**

<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>

**Organization**

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

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

Combine NLP, pattern matching, lexicons, statistical learning
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:
- city (Seattle) in downtown Seattle in downtown X
- city (Seattle) Seattle and other towns X and other towns
- city (Las Vegas) Las Vegas and other towns
- plays (Zappa, guitar) playing guitar: … Zappa playing Y:
- plays (Davis, trumpet) Davis … blows trumpet X … blows Y
- in downtown Beijing Coltrane blows sax
- city (Beijing) plays (C., sax)
- old center of Beijing old center of X
- sax player Coltrane 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)
Performance of Web-IE

State-of-the-art precision/recall results:

<table>
<thead>
<tr>
<th>relation</th>
<th>precision</th>
<th>recall</th>
<th>corpus</th>
<th>systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>countries</td>
<td>80%</td>
<td>90%</td>
<td>Web</td>
<td>KnowItAll</td>
</tr>
<tr>
<td>cities</td>
<td>80%</td>
<td>???</td>
<td>Web</td>
<td>KnowItAll</td>
</tr>
<tr>
<td>scientists</td>
<td>60%</td>
<td>???</td>
<td>Web</td>
<td>KnowItAll</td>
</tr>
<tr>
<td>CEOs</td>
<td>80%</td>
<td>50%</td>
<td>News</td>
<td>Snowball, LEILA</td>
</tr>
<tr>
<td>birthdates</td>
<td>80%</td>
<td>70%</td>
<td>Wikipedia</td>
<td>LEILA</td>
</tr>
<tr>
<td>instanceof</td>
<td>40%</td>
<td>20%</td>
<td>Web</td>
<td>Text2Onto, LEILA</td>
</tr>
</tbody>
</table>

**precision value-chain:** entities 80%, attributes 70%, facts 60%, events 50%

Anecdotic evidence:

- invented (A.G. Bell, telephone)
- married (Hillary Clinton, Bill Clinton)
- isa (yoga, relaxation technique)
- isa (zearalenone, mycotoxin)
- contains (chocolate, theobromine)
- contains (Singapore sling, gin)

- invented (Johannes Kepler, logarithm tables)
- married (Segolene Royal, Francois Hollande)
- isa (yoga, excellent way)
- isa (your day, good one)
- contains (chocolate, raisins)
- plays (the liver, central role)
- makes (everybody, mistakes)
Beyond Surface Learning with LEILAK

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]**), …

Cologne lies on the banks of the Rhine

People in Cairo like wine from the Rhine valley

Paris was founded on an island in the Seine

We visited Paris last summer. It has many museums along the banks of the Seine.
YAGO Enhancement by IE on Text Sources

Capture confidence value for each fact

ongoing work: harvesting relations by IE tools like GATE, LEILIA, ...
(e.g.: which enzyme catalyzes which biochemical process, who discovered or invented what, ...)

max planck institut informatik
Yago

Yago is a huge semantic knowledge base. Currently, Yago knows over 500,000 entities (like persons, organizations, cities, etc.). It knows about 6 million facts about these entities. This Web-Interface allows users to pose questions to Yago in the RDF, MySQL, Oracle and Postgres (version 2007-21-2).

If you would like your results to be ranked by informativeness, check out our project NAGA.

Queries

A query has the form

E R E

where the E's are entities (e.g. Einstein) and the R's are relations. The following relations are allowed: bornInYear, diedInYear, establishedInYear, hasWonPrize,
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 novel scoring model, based on the principles of generative language models, which formalizes 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.

- Gerald Kainz, Fabian M. Suchanek, Georgina Rätse, Vyas Ramamurthy, Gerhard Weikum
  "NAGA: Searching and Ranking Knowledge" (pdf, bib, slides)
  Technical Report (MPII 2007)

### NAGA Queries

A query has the form:

\[ E1 \ Rel1 \ E2 \ Rel2 \ ... \]

where the \( E \)'s are entities (e.g. Einstein) and the \( Rel \)'s are relations. The following relations are allowed:

- the family relation \( familyNameOf \)
- the type relation \( type \)
- the means relation \( means \)
- the subClassOf relation \( subClassOf \)

### Score: 7.184462521168058E-13 mathematician_109635652

- Fisher → Ronald_Fisher
- scientist → scientist_109871938
- mathematician_109635652

→ statistical language model for result graphs

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

**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)
```

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

q: isa ($x, vegetarian)
isa (Einstein, vegetarian)
isa (Al Nobody, vegetarian)
```
NAGA Ranking Model

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

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

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

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

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

based on IE accuracy and authority analysis

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

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

for $q_i = (x, r, z)$ with variable $x$ estimated by correlation statistics

---

max planck institut informatik
Query: $x$ isa politician
$x$ isa scientist

Results:
Benjamin Franklin
Paul Wolfowitz
Angela Merkel

Entity Search: Example NAGA
Outline

✓ Motivation

✓ Leibniz: Semantic Web

✓ Planck: Statistical Web

• Darwin: Social Web

• Conclusion
Darwin Approach (Social Web)

Social Wisdom & Natural Selection:
- Evolution of (Web 2.0) species
- Survival of the fittest
- Search & rank by collaborative recommendation

Charles Darwin (1809 - 1882)
“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.
Social Tagging: Example Flickr (1)

Tags / happiness

Sort by:
Most recent • Most interesting

happiness clusters
Explore and refine happiness photos with our clustery goodness!

Find similar images on Yahoo! image search

Ads from Yahoo!
AdPinnet Improve Your Life
Discover the inner secrets to attain success, prosperity and happiness.

www.adp.net

Double happiness Jade Pendant
Buy 14kt gold Chinese double happiness jade pendant from Paradise.
Social Tagging: Example Flickr (2)

Tags / happiness / clusters

- smile, love, happy, joy, portrait, girl, woman, fun, bw, friends
- children, child, kids, boys, kid, girls
- wedding, marriage, bride
- brasil, brazil

These are the most recent photos tagged with happiness. See more...
Social Tagging: Example Flickr (3)
Social-Tagging Community

> 20 Mio. users
> 2 Bio. photos
> 10 Bio. tags
30% monthly growth

Source: www.flickr.com
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)

Congratulations! You scored 1 point!

taboo:
pyramid
Louvre museum
Paris

Mitterand
Mona Lisa
metro lignes 7, 14
Da Vinci code
Dark Side of Social Wisdom

• **Spam** (Web 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: PageRank

**PageRank (PR):** links are endorsements & increase page authority; authority is higher if links come from high-authority pages

\[
PR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p, q)
\]

with \( t(p, q) = 1 / \text{outdegree}(p) \)

and \( j(q) = 1 / N \)

equivalent to principal eigenvector:

\[
p_{n \times 1} = A_{n \times n} \times p_{n \times 1}
\]

**Social Ranking**

**Authority (page q) = stationary prob. of visiting q**

**random walk:** uniformly random choice of links + random jumps; add **bias** to transitions and jumps for personal PR, **TrustRank**, etc.
Wisdom of Crowds: Graphs are Everywhere

http://www.flickr.com/photos/lukemontague/14038129/

http://www.flickr.com/photos/shopping2null/395271855/

http://datamining.typepad.com/gallery/core.png

http://datamining.typepad.com/gallery/newblog-crop.png
Wisdom of Crowds: Beyond PR

Typed graphs: data items, users, friends, groups, postings, ratings, queries, clicks, …

with weighted edges → spectral analysis of various graphs
Outline

✓ Motivation

✓ Leibniz: Semantic Web
✓ Planck: Statistical Web
✓ Darwin: Social Web

• Conclusion
Summary

Harvesting **knowledge** & organizing in **semantic DB / graph** for

- scholarly Web,
- digital libraries,
- enterprise know-how,
- online communities, etc.

Enable **semantic search**, with ranking of uncertain facts

Three roads to knowledge and intelligent search:

- **Leibniz / Semantic Web**: ontologies, encyclopedia, etc.
- **Planck / Statistical Web**: large-scale IE from text, speech, etc.
- **Darwin / Social Web**: wisdom of crowds, tagging, folksonomies
Thank you!