

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)

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,)

Structured data (records) Unstructured data (documents)

Trend: quadrants getting blurred towards DB&IR technology integration



Gerhard Weikum, EDBT 2007 Summer School

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



XML IR on Heterogeneous Data

Union of heterogeneous sources without global schema Similarity-aware XPath: Which professors //~Professor [//* = "~Saarbruecken"] from Saarbruecken (SB) are teaching IR and have [//~Course [//* = "~IR"]] research projects on XML? [//~Research [//* = "~XML"]] alchemist Lecturer primadonna magician directo artist Scoring and ranking: investigator wizard • XML BM25 for content cond. intellectual RELATED (0.48) ontological similarity for relaxed tag condition professor researcher score aggregation with HYPONYM (0.749) probabilistic independence scientist lecturer scholar • extended TA for query exec. mentor academic. teacher statistical edge weighting by academician. Dice coeff.: 2 #(x,y) / (#x + #y) on Web faculty member max planck institut

Query Expansion with Incremental Merging

[M. Theobald et al.: SIGIR 2005]

relaxable query q: ~*professor research* with expansions $exp(i)=\{w \mid sim(i,w) \ge 0\}$ 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

alchemist magician artist wizard intellectual researcher scholar academic, academician, faculty member

B+ tree index on terms



ontology / meta-index



efficient, robust, self-tuning

max planck 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.

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





Keyword query: Greek art Paris





Semantic Search

Concept query: Person = "Max Planck" Concept query: "Greek art" & Location = "Paris"



Gerhard Weikum, EDBT 2007 Summer School

Entity Search: Example Google

Google[®]

Which politicians are also scientists ?

Search Advanced Search Preferences

Web

Science and Politics

Also: Why don't well-credentialed scientists get these TV gigs? This is an effort of the power elites: politicians, old-timer internet-illiterate ... sciencepolitics.blogspot.com/ - 9 Jul 2007 - Similar pages

NPR : When Science Drives Politics

And I expect my **politicians** to tell me the truth. That's the issue. ... And we **also** know that **scientists** immediately afterwards said, what 60 lines? ... www.npr.org/templates/story/story.php?storyId=5490898 - <u>Similar pages</u>

Issues in S and T, Fall 2004, Science, Politics, and U.S. Democracy

It follows that if either scientists or politicians so politicize their mutual ... However, the scientists must also understand that the officials being ... www.issues.org/21.1/branscomb.html - 36k - <u>Cached</u> - <u>Similar pages</u>

National Ledger - Global Warming Panic: Politicians Try to Stifle ...

Global Warming Panic: **Politicians** Try to Stifle **Scientists** ... "Hooey" is the term **also** used by Reid Bryson, the father of scientific climatology, ... www.nationalledger.com/artman/publish/article_272614277.shtml - 37k -<u>Cached</u> - <u>Similar pages</u>

Political Science Resources/United States Politics

Impeachment | Lobby Groups | Michigan Government | Minority **Politicians** ... Resources for the study of United States politics **also** appear throughout the ... www.lib.umich.edu/govdocs/psusp.html - 89k - <u>Cached</u> - <u>Similar pages</u>

news @ nature.com - UK civil servants accused of warping science ...

But the same message raised eyebrows on the other side of the Atlantic last week, when a UK parliament report suggested that **politicians** there **also** pick and ... www.nature.com/news/2006/061113/full/444252a.html - <u>Similar pages</u>

Political science - Wikipedia, the free encyclopedia

The antecedents of Western politics can also trace their roots back even earlier ... Political scientists may serve as advisers to specific politicians, ... en.wikipedia.org/wiki/Political_science - 61k - Cached - Similar pages



What is lacking?

- data is not knowledge
 - \rightarrow 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

Divesh Srivastava Mentions 1 - 10 out of 572 Wednesdey Wag 1, 2007 Colloquia page. Rutgers University 	sh	divesh						S	earch			Login	Help	The Cimple Pro	ject	Wiki
Wednesdey Aug 1, 2007 Colloquia page. Rutgers University Rozenbaum: RutgersDefense Committee: S. (Muthu) Muthukrishnan, Amelie Marian, Richard Martin, Divesh Srivastava (outside member) Login Return to DCIS Return to CS Department http://www.cs.rutgers.edu/news/colloquia/index.html Cached Annotated Selated Per Network Tuesday Jul 31, 2007 Session chair in VLDB 2007 Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/program/details_thursday.html Cached Annotated Related Top Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/program/details_thursday.html Cached Annotated Services Thursday Jul 26, 2007 "Publications" page, Himanshu Gupta's homepage on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. PDF H http://www.ds.unysb.edu/~hgupta/pubs.html Cached Annotated SiGMOD Publications	ri	Sriv	vas	stav	a 🗖			Men	tions 1	- 10 out of	526	<u>http://w</u> <u>Annot</u> AT&T, U	<u>ww.rese</u> ated JSA	arch.att.com/~di	vesh/	
Interstage Related Performance Related Performance Tuesday Session chair in VLDB 2007 Nick Kou Jul 31, 2007 Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/program/details_thursday.html Cached Annotated Related Top Saturday Jul 28, 2007 Session chair in VLDB 2007 * Xml Saturday Jul 28, 2007 Session chair in VLDB 2007 * Xml Saturday Jul 28, 2007 Session chair in VLDB 2007 * Xml Thursday Session chair in VLDB 2007 * Xml Saturday Session chair in VLDB 2007 * Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft * Xml Thursday * Thursday * Thursday.html Cached Annotated * Services Thursday * "Publications" page, Himanshu Gupta's homepage * VLDB 200 on Database Theory (ICDT), 1999, PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. * SiGMOD PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated * Publications <td colspan="6"><u>Colloquia page, Rutgers University</u> Rozenbaum: RutgersDefense Committee: S Amelie Marian, Richard Martin, <u>Divesh Srivasta</u> Return to DCIS <u>Return to CS Department</u></td> <td>S. (Mu <mark>ava</mark> (ou</td> <td>ithu) M itside m</td> <td>uthukrishn nember) Lo</td> <td>nan, ogin</td> <td colspan="5">528 total mentions occuring in 161 pag 0 new mentions found in the last 24 ho</td>	<u>Colloquia page, Rutgers University</u> Rozenbaum: RutgersDefense Committee: S Amelie Marian, Richard Martin, <u>Divesh Srivasta</u> Return to DCIS <u>Return to CS Department</u>						S. (Mu <mark>ava</mark> (ou	ithu) M itside m	uthukrishn nember) Lo	nan, ogin	528 total mentions occuring in 161 pag 0 new mentions found in the last 24 ho					
Tuesday Jul 31, 2007 Session chair in VLDB 2007 • Nick Kou Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/progrem/details_thursday.html Cached Annotated Related Tog Saturday Jul 28, 2007 Session chair in VLDB 2007 • xml Suddray Jul 28, 2007 Session chair in VLDB 2007 • xml Suddray Jul 28, 2007 Session chair in VLDB 2007 • xml Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/progrem/details_thursday.html Cached Annotated • xml Thursday Jul 26, 2007 • "Publications" page, Himanshu Gupta's homepage on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated Publications Publications Publications SigMOD more Publications Publications	http://www.cs.rutgers.edu/news/colloquia/index.html					ed Ann	otated			Relate	d Peop	le				
Natasayya (Wictosoft http://www.vidb2007.org/program/details_thursday.html Cached Annotated Related Top Saturday Session chair in VLDB 2007 Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vidb2007.org/program/details_thursday.html Cached Annotated • xml • logic Thursday Jul 28, 2007 "Publications" page, Himanshu Gupta's homepage • SiGMOD • VLDB 20 Thursday Jul 28, 2007 "Publications" page, Himanshu Gupta's homepage • SiGMOD • VLDB 20 Thursday Jul 28, 2007 "Publications" page, Himanshu Gupta's homepage • SiGMOD • VLDB 20 on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. • SiGMOD more PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated Publications	ch 27t m	sion chai pt 27th torium l base S	airin V n 09:0 Maxin System	LDB 20 0 - 10 num S ns: A	07 :30 10- Session Decade	Year E Chair of P	Best Pa : <mark>Dive</mark> rogress	aper A <mark>sh Sr</mark> s Sura	ward T <mark>ivastava</mark> jit Cha	alk Locat Self-Tur udhuri, Vi	ion: ning ivek	 <u>Nic</u> <u>H.1</u> <u>S.3</u> <u>Rat</u> <u>more</u> 	<u>k Kouda</u> /. Jagad Budarsh ghu Ran	i <u>s</u> ish an nakrishnan		
Saturday Jul 28, 2007 Session chair in VLDB 2007 Ioqic Saturday Jul 28, 2007 Sept 27th 09:00 - 10:30 10-Year Best Paper Award Talk Location: Auditorium Maximum Session Chair: Divesh Srivastava Self-Tuning Database Systems: A Decade of Progress Surajit Chaudhuri, Vivek Narasayya (Microsoft http://www.vldb2007.org/program/details_thursday.html Cached Annotated • xml Thursday Jul 28, 2007 "Publications" page, Himanshu Gupta's homepage on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated • xml Publications SigMOD Publications Publications PDF H Publicational Conference on Database Theory (ICDT), 1999. PDF H Publications	http://www.vldb2007.org/progra		g/program	n/details_1	thursday.html Cached Annotated						Related Topics					
Thursday "Publications" page, Himanshu Gupta's homepage Services Thursday "Publications" page, Himanshu Gupta's homepage • SIGMOD Jul 26, 2007 on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. PDF H • SIGMOD PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated • Publications	ch 27t m	sion chai pt 27th torium l base Sy	airin V 09:0 Maxin System Micros	LDB 20 0 - 10 num S ns: A	07 :30 10- Session Decade	Year E Chair of P	Best Pa Dive rogress	aper A <mark>sh Sr</mark> Sura	ward T <mark>ivastava</mark> jit Cha	alk Locat Self-Tur udhuri, Vi	ion: ning ivek	 <u>xml</u> <u>logi</u> <u>tran</u> <u>que</u> <u>more</u> 	l isaction ery proce	<u>s</u> essing		
Thursday "Publications" page, Himanshu Gupta's homepage • SIGMOD Jul 26, 2007 "Publications" page, Himanshu Gupta's homepage • VLDB 20 on Database Theory (ICDT), 1999. PDF [Over 150 citations (from scholar.google.com)]. H. Gupta and D. Srivastava. Data Warehouse of Newsgroups. International Conference on Database Theory (ICDT), 1999. • SIGMOD PDF H http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated	vld	www.vldb2	2007.or	g/program	n/details_t	thursday.	html C	ached /	Annotated			Service	es			
http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated Publication	tio ta loc	Dications Databa Iar.googl sgroups.	ase T gle.con . Inter	<u>e, Hima</u> Theory n)]. H. nationa	(ICDT), Gupta Confer	upta's h 1999. and D rence c	nomepa PDF). Sriva on Data	i <u>ge</u> [Over stava abase	150 ci Data \ Theory	tations (fi Varehouse (ICDT), 19	irom e of 999.	 SIG VLI VLI SIG more 	MOD 20 0B 2007 0B 2007 0B 2007 0D 20 20 2	107 (Session cha (Session chair) (Program chair) 107 (PC member	<u>iir)</u>)	
	http://www.cs.sunysb.edu/~hgupta/pubs.html Cached Annotated									Publica	ations					
Tuesday Program chair of VLDB 2007 and meta	cl	iram cha	air of V	/LDB 20	07							 Interaction and 	ensional 1 metada	associations be ata	tween	<u>data</u>

-

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:

max planck institut informatik





Information Extraction (IE): Text to Records

Person

Constant

Planck's constant

BirthDate

Person

Person

4/23, 1858 Kiel

3/14, 1879 Ulm

Max Planck Quantum Theory

Max Planck Albert Einstein

Max Planck Niels Bohr

extracted facts often

have confidence < 1

 \rightarrow DB with uncertainty

(probabilistic DB)

Collaborator

Organization

BirthPlace

ScientificResult

Max Planck



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 2, 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

max planck institut informatik

own set of units of measurement based on fundamental physical constance and and

year later, he discovered the law of heat radiation, which is named Planck' Person Radiation . This law became the basis of quantum theory &, which emerge Max Planck KWG / MPG later in cooperation with Albert Einstein & and Niels Bohr .

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



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

Assume entity e was seen in k records r₁, ..., r_k extracted from k pages $d_1, ..., d_k$ with accuracy $\alpha_1, ..., \alpha_k$

$$P[w | e] \sim \sum_{i} \alpha_{i} P[w | context(r_{i}, d_{i})] \quad \text{record-level LM}$$

$$\rightarrow s(e,q) = \sum_{w \in q} \lambda \left(\sum_{i} \alpha_{i} \frac{tf(w, context(r_{i}, d_{i}))}{|context(r_{i}, d_{i})|} \right) + (1-\lambda) \frac{\sum_{i} tf(w, r)}{|records|}$$

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

alternatively consider individual attributes e.a, with importance β_i extracted from page d_i with accuracy γ_{ii}

$$P[w | e] \sim \sum_{i} \alpha_{i} \sum_{j} \beta_{j} \gamma_{ij} P[w | context(r_{i} . a_{j} , d_{i})]$$

$$= \sum_{i} \alpha_{i} \sum_{j} \beta_{j} \gamma_{ij} P[w | context(r_{i} . a_{j} , d_{i})]$$

Gerhard Weikum, EDBT 2007 Summer School

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 \rightarrow entity if entity is extracted from page
- page1 \rightarrow page2 if there is hyperlink or implicit link between pages
- entity1 \rightarrow 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.:

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

- 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

<u>Use cases:</u>

- 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





YAGO Knowledge Representation



YAGO Enhancement by IE on Text Sources



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) \rightarrow patterns (in text) \rightarrow (extraction) rule \rightarrow (new) facts



Methods for Web-Scale Fact Extration



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

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



Gerhard Weikum, EDBT 2007 Summer School

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



Search Results Without Ranking

Yago A Core of Semantic Knowledge	Enter your Yago-query: Fisher isa scientist Fisher isa \$x	q: Fisher isa scientist Fisher isa \$x			
Yago		<pre>\$@Fisher = Ronald_Fisher \$@scientist = scientist_109871938 \$X = alumnus_109165182</pre>			
Yago is a huge semantic knowledge base. Currently, Yago knows over 900,000 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	Submit Query	<pre>\$@Fisher = Irving_Fisher \$@scientist = scientist_109871938</pre>			
mathematician_109635652	-subClassOf-> scie	\$X = social_scientist_109927304			
"Fisher" —familyNameOf Ronald_Fisher —type—>	<pre>\$@Fisher = James_Fisher \$@scientist = scientist_10981938 \$X = ornithologist_109711173 \$@Fisher = Ronald_Fisher</pre>				
"scientist" —means—> sci					
RDFS, MySQL, Oracle and Postgres (version: 2007-w21-2).		$\ensuremath{\$}^{\ensuremath{\varnothing}}$			
If you would like your results to be ranked by informativeness, check out our project NAGA.	mathematician_109635652 —subClassOf—> scientist_109871938	$X = $ theorist_110008610			
	Ahumni_of_Gonville_and_Caius_College,_Camb subClassOf—> ahumnus_109165182 "Eisher" —familyNameOf—> Ronald Eisher	\$@Fisher = Ronald_Fisher			
Queries	Ronald_Fisher —type—> Alumni_of_Gonville_and_Caius_College,_Cam	$\$$ Scientist = scientist_109871938 \$ - colleague 109301221			
A query has the form	Ronald_Fisher —type—> 20th_century_math "scientist" —means—> scientist_109871938	$\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ $			
ERE	mathematician 109635652				
		\$@scientist = scientist_109871938			
where the F's are entities (a g Einstein) and the P's are	 S@Fisher = Ronald_Fisher S@calentist = acientist = 100871028 	X - 0 000000000000000000000000000000000			
relations. The following relations are allowed: bornInYear.	 S@scientist = scientist_1098/1938 \$X = alumnus_109165182 	$\psi x = 019a11311 - 100003220$			
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 <u>Yago</u> and consists of millions of entities and relationships automatically extracted fromWeb-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 <u>Max-Plack-Institute Saarbrücken</u>. For details on NAGA, take a look at our technical report

 <u>Gjergiji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Mava Ramanath, Gerhard Weikum</u> "NAGA: Searching and Ranking Knowledge" (pdf, bib, slides) Techincal Report (MPII 2007)

NAGA Queries	Enter your NAGA-query:	
A query has the form E1 R1 E2 E3 R2 E4 where the Ei's are entities (e.g. Einstein) and the Rj's	Submit Query	
Score: 7.1844625 "Fisher" —family	521168058E-13 mathematic NameOf—> Ronald_Fishe	iar er

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

\$@scientist = scientist_109871938

\$@Fisher = Ronald Fisher

\$X = scientist 109871938

#Fisher _____tamilyNameOf ___> Ronald_Fisher _____ Ronald_Fisher ___type ___> 20th_century_mathemathematicians _____subClassOf ___> (% = producting_rectron to retron to retro to retron to retro to retron to retr

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

(try your own name...) • Albert_Einstein type subClassOf Sx

- Zidane isa Sx
- "Mostly Harmless"
- (entity names can b "Albert Einstein"))
- Bohr connect Einst
 Sx isa humanist
- Sx bornInYear 187

Score: 7.184462521168058E-13

→ statistical language model

for result graphs

max planck institut

informatik

Gerhard Weikum, EDBT 2007 Summer School

10

NAGA: Searching & Ranking Knowledge

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 <u>Yago</u> and consists of millions of entities and relationships automatically extracted fromWeb-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 <u>Max-Plack-Institute Saarbrücken</u>. For details on NAGA, take a look at our technical report

 <u>Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Mava Rama</u> "NAGA: Searching and Ranking Knowledge" (pdf, bib, slides) Techincal Report (MPII 2007)



A query has the form

```
E1 R1 E2
E3 R2 E4
```

where the Ei's are entities (e.g. Einstein) and the Rj's are relations. The following relations are allowed: means, type, subClassOf, isa, connect, domain, range, familyNameOf, givenNameOf, bornInYear, diedInYear, establishedInYear, hasWonPrize, writtenInYear, locatedIn, politicianOf, context, discoveredBy, discoveredInYear, before (for years), after(for years).

The R's can also be regular expressions over these relations. Both the relations and the entities can be variables (starting with \$). The two relations **connect** and isa are pseudo-relations that only work infix in a plain line.

Sample queries

- Planck familyNameOf \$x (try your own name...)
- Albert_Einstein type subClassOf Sx
- Zidane isa \$x
- "Mostly Harmless" writtenInYear Sx (entity names can be written in quotes (e.g. "Albert Einstein"))
- Bohr connect Einstein
 Sx isa humanist Sx bornInYear 1879

Inn. Mava Rame
(ndf, bid, sides)Carl_Sagan — hasWonPrize—> Pulitzer_Prize
Carl_Sagan — type—> Planetary_scientistsEnter your NAY
Sty contextPlanetary_scientistsSt 1st sciet
Sty contextPlanetary_scientistsSt 1st sciet
Sty contextStarl_Sagan

q: \$x isa Scientist

\$x hasWonPrize \$y

\$y context Literature

- E._O._Wilson —hasWonPrize—> Pulitzer_Prize
- submitQuer E._O._Wilson —type—> Evolutionary_biologists
- Ranked results
 Evolutionary_biologists subClassOf —> biologist_109855630

 Score: 1.69248
 biologist_109855630 subClassOf —> scientist_110560637

 "scientist" "
 \$X = E._O._Wilson
- Jewish-America Carl_Sagan —
- S@scient
SX = CaBertrand_Russell —hasWonPrize—> Nobel_Prize_in_Literature
Bertrand_Russell —type—> mathematicianScore: 1.69248Bertrand_Russell —type—> mathematician
mathematician —subClassOf—> scientist_110560637
\$X = Bertrand_Russell
 - \$@scien • \$X = Ca ■■■

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

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)

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

using
generative
mixture model
$$P[q | g] = \prod_{i=1}^{n} (1-\alpha) \cdot P[q_i | g_i] + \alpha \cdot P[q_i]$$

$$P[Goethe | bornIn, Frankfurt] = \frac{P[Goethe, bornIn, Frankfurt]}{P[bornIn, Frankfurt]}$$
based on
IE accuracy
and
authority
analysis
$$conf(e) = \sum_{i=1}^{n_e} acc(e, P_i) \cdot trust(P_i) \quad P(x|r,z) = \frac{P(x,r,z)}{P(r,z)} = \frac{P(x,r,z)}{\sum_{x} P(x',r,z)}$$
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 (Cld, Title, Location, Year)Journals (Jld, Title)CPublications (Pld, Title, Cld)JPublications (Pld, Title, Vol, No, Year)Authors (Pld, Person)Editors (Cld, 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

informatik

max planck institut

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



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={t₁, ..., t_m} find best tree (r, x₁, ..., 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



37/62

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

max planck institut informatik

- 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 ...

max planck institut informatik



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 \rightarrow \$ 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



Gerhard Weikum, EDBT 2007 Summer School

Social-Network Database

Simplified and cast into relational schema: Users (UId, Nickname, ...) **Docs** (<u>DId</u>, Author, PostingDate, ...) **Tags** (<u>TId</u>, String) Friendship (Uld1, Uld2, FScore) Content (Dld, Tld, Score) Rating (Uld, Dld, 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 (Uld, Tld, UTscore)

x.UTscore := Σ_d {s | (x.UId, x.TId, d, s) \in Ratings}

TagsDocs (Tld, Did, TDscore)

x.TDscore := Σ_u {s | (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]:

max planck institut informatik

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

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 informatik



Tag Mining from Social Networks

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



informatik

Tag Mining from Social Networks

<u>Given</u>: tag frequencies at daily resolution <u>Wanted</u>: "most interesting" tags for app-provided time intervals

Define **"interestingness"** of tag x for interval T

- <u>Requirements</u>
 - tag should be frequent in T and not so frequent at other times
 - tag with singular peaks in T should not dominate
- <u>Approach:</u>
 - interestingness (x, T) =

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

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 T₁, ..., T_m covering T, mutually disjoint, of max. length run TA over the lists for T₁, ..., 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:

max planck institut informatik

 $T_1 \dots T_k$ in desc freq order, $T_1^{\, \cdot} \dots T_l^{\, \cdot}$ in asc freq order

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

played against random, anonymous partner on Internet



<u>taboo:</u> 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

> max planck institut informatik



- 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

max planck institut informatik

- 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 !



Gerhard Weikum, EDBT 2007 Summer School

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



Gerhard Weikum, EDBT 2007 Summer School

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

