

Web Search and Mining “**WMa**”

CS635 Autumn 2019

Mon Thu 5:30—7:00pm

SIC201

Motivation

- Few areas of computer science have had as much impact on our lives in the last 15 years as the Internet, the Web, search engines, e-commerce, and social media
- Rapid rise of some of the largest and most accomplished companies on earth
- Draws heavily from data structures, algorithms, databases, probability and statistics, (deep!) machine learning, parallel computing, economics and game theory, social sciences, ...

Highlights

1970

0

1990

0

1995

1995

5

1999

2000

2004

4

2006

2009

9

Arthur C. Clarke predicts that satellites would "bring the accumulated knowledge of the world to your fingertips" using a console combining the functionality of the photocopier, telephone, television and a small computer, allowing data transfer and video conferencing around the globe.

Tim Berners-Lee & co build first prototype at CERN

Mosaic web browser

developed

NorthernLight, Lycos.

Alta Vista launched by Digital Equipment Corp

Goto.com introduces paid search

Google launches

Facebook founded

Twitter

founded

Bing launched by

Infrastructure

Crawling

Indexing

PageRank

De-

spamming

Auctions

Personalization

Query+click log mining

Social media

Collaborative filtering

Local search

Text mining

CS635 and CS728

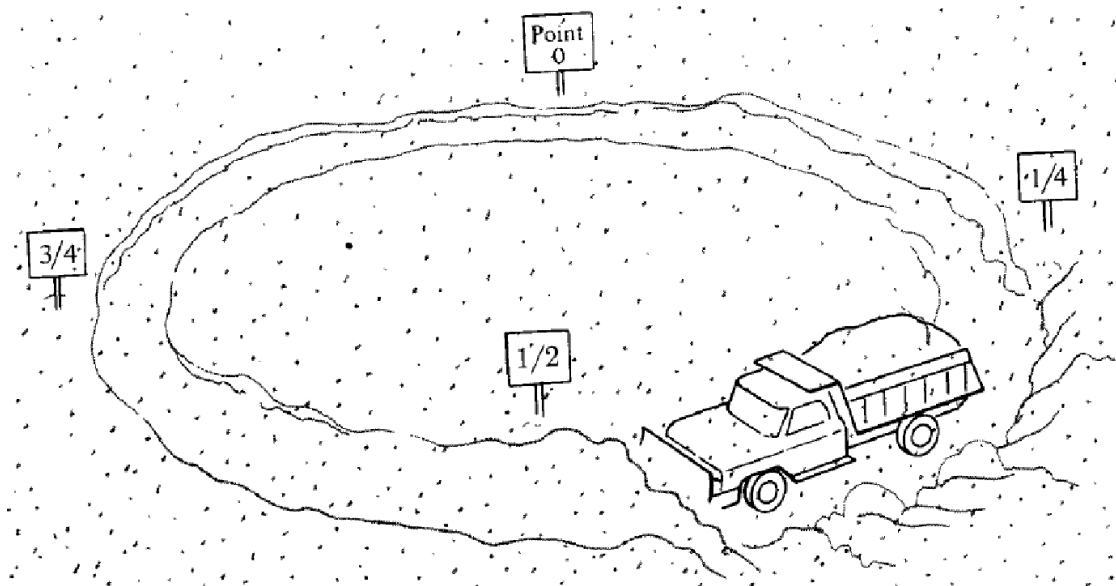
- CS635 = Web search and mining = first course on Web information retrieval
 - Indexing, query processing, ranking
 - Corpus and document models, text mining
 - Hyperlinks and social network analysis
 - Large scale data analysis: “big data”
- CS728 = Organization of Web information = second course with more recent topics
 - Bootstrapping knowledge bases
 - Named entity and relation recognition
 - Open domain knowledge extraction
 - Semantic annotations and search

Administrivia

- Course material will be on Moodle
- Schedule goog.gl/GxDW2g
- Meant primarily for Mtech**1**, Btech**3**, **PhD**
- Mtech2, Btech4 ok but too late for research?
- Weak prerequisites: prob+stat, databases
- If you can read and understand a bit of math, you will need ~10 days of self-study to pick up the relevant pieces from basic machine learning

Course design: books

- Basic material from a few books
 - [Manning-Raghavan-Schutze online book](#)
 - Baeza-Yates and Ribeiro-Neto
 - Managing Gigabytes (plumbing tips)
 - MTW second edition in progress, notes will be available for many segments



Course design: research papers

- Papers written much faster than you can write a book
 - In fact, much faster than anyone can read
- Digesting papers: critical survival skill
- They sound like there's nothing left to do
- You get credit for pointing out loopholes, flaws, and further ideas to follow from existing papers

Evaluation

- Homework
 - All homeworks and exams from all earlier offerings are effectively (ungraded) homeworks (many with solutions) — take advantage!
 - Hands-on work (coding, simulations and measurements)
- Exams
 - Limited time, in-class
 - Problems “served on a platter”
 - Hard to pose and solve real systems issues
- Optional extra-credit projects; auditing

Multidisciplinary synthesis

- Content: hypermedia, markup standards, text and semistructured data models
- Activity: linking, blogging, selling, spamming
- Algorithms: graphs, indexing, compression, string processing, ranking
- Statistics: models for text and link graph, catching (link) spam, profiling queries
- Language: tagging, extraction
- Plumbing: networking, storage

CS635 syllabus

- Text indexing, search
- Document and corpus models
 - Word embeddings
- Ranking, learning to rank
- Whole-document labeling/classification
- Measuring and modeling (social) networks
 - Embeddings for knowledge graphs
- Hyperlink assisted search and mining
- Web sampling, crawling, monitoring

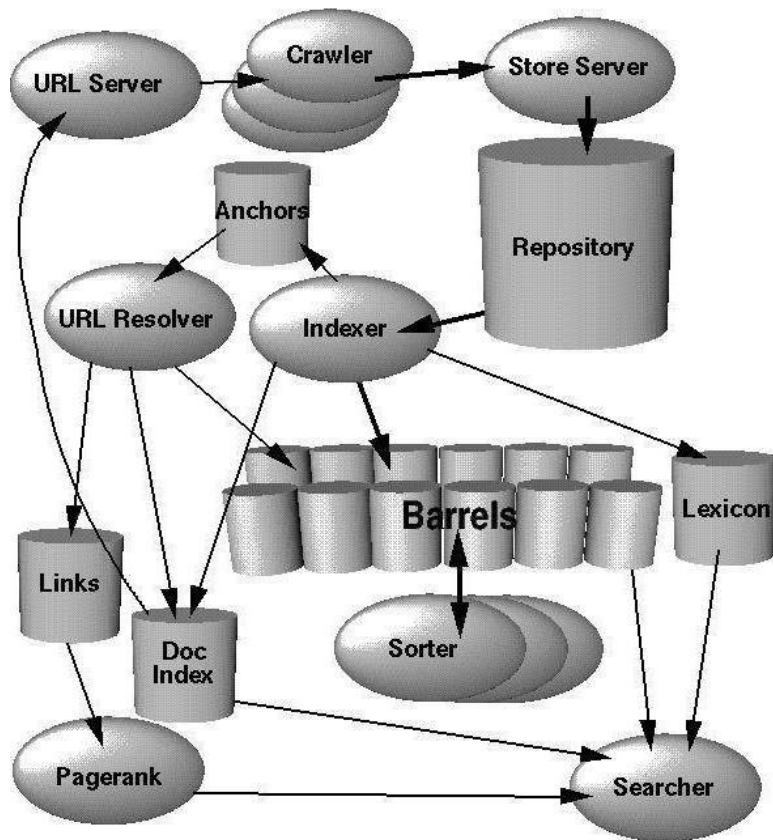
CS728 syllabus

- Fine-grained front-end text analysis
 - Natural language, lists, tables, site layout, ...
- Entity, type and relation annotation
 - Using various embeddings
- Information extraction, integration, coref
- Bootstrapping Web knowledge bases
 - “Open Information Extraction” or OpenIE
- Question answering; query interpretation
 - Adding graph structure to text search

Text indexing, search and ranking

- Boolean queries involving word occurrence in documents
- Inverted index design, construction, compression, updates; query processing
- Vector space model, relevance ranking, recall, precision, F1, break-even
- Probabilistic ranking, belief networks
- Fast top-k search
- Similarity search: minhash, random projections, locality-preserving hash

Industry-strength search



- Distributed crawling
- Corpus storage and indexing
- Scalable query processing
- Query correct and suggest
- Log collection and processing
- Verticals structured callouts

Document and corpus models

- Token, compound, phrase, stems, stopwords
- Language and typography issues
- Practical computer representations
- Bag of words, corpus, term-document matrix
- Probabilistic generative models
- Modeling multi-topic corpus and documents
- Statistical notions of semantic similarity
- Text clustering applications

Diversity

Mahatma Gandhi (MG) Road (Bengaluru) - Top Tips Befo

<https://www.tripadvisor.in> > ... > Bengaluru > Places to visit in Bengaluru ▼

★★★★★ Rating: 4 - 919 reviews

Mahatma Gandhi (MG) Road, Bengaluru: See 919 reviews, articles, and 43 phc (MG) Road, ranked No.20 on TripAdvisor among 326 ...

Suggested restaurants near MG Road - Pune - Zomato

<https://www.zomato.com> > India > Pune ▼

Suggested restaurants near MG Road. ... special corporate packages and offer Crazy Frog at DP Road. Come join us for a lovely evening!

M. G. Road, Fort - Mumbai - Wikimapia

wikimapia.org/street/134414/M-G-Road-Fort ▼

M. G. Road, Fort. ... M. G. Road, Fort, related objects. Police Station / Church cities: leadbull - Lead Generation Company in Mumbai, ...

Book Hotels near MG Road Bangalore, Tariff @ ₹1201| F

<https://www.oyorooms.com/hotels-near-mg-road-bangalore/> ▼

Book Hotels near MG Road Bangalore NOW!! And get ✓Free WiFi ✓AC Room Cancellation, Spotless linen and clean wash rooms all at ...

MG Road - Burrp

www.burrrp.com/mumbai/restaurants/mg-road-area ▼

Restaurants in Mg Road, Mg Road Restaurants, Mg Road Restaurants addre Restaurants phone numbers, Mg Road Restaurants reviews, ...

Hotels in MG Road , Matheran - Book from 8 Hotels & Ge

https://www.makemytrip.com/hotels/hotels-in-mg_road-matheran.html ▼

★★★★★ Rating: 4.9 - Review by Siddhartha Kathpalia ...

Get the best Offers on Hotels in MG Road, Matheran . Compare from 1 availabl Use coupon code & Get Upto 70% OFF instantly on MG ...

McCallum (TV Series 1995–1998) - IMDb

www.imdb.com/title/tt0112067/ ▼

★★★★★ Rating: 7.8/10 - 280 votes

Drama · From deep within the morgue at St. Patrick's Hospital Dr. Iain McCallum and Dr. Angela Moloney along with a team

McCallum Bagpipes Ltd

www.mccallumbagpipes.com/ ▼

McCallum Bagpipes offers an unrivalled range of bagpipes, p chanter reeds and smallpipes. We also stock an extensive s

McCallum Theatre - Palm Desert, California - www.mc

www.mccallumtheatre.com/ ▼

McCallum Theatre in Palm Desert, California. Learn more at mccallumtheatre.com.

Images for mccallum



→ More images for mccallum

| McCallum Australia |

<https://www.mccallumhinges.com.au/> ▼

McCallum Australia are market leaders in high quality archite 1938, we has been introducing innovative products specialisi

McCallum

www.mccallum.org.au/ ▼

WELCOME TO McCALLUM DISABILITY SERVICES. Since grown into one of Western Victoria's most comprehensive org

Today's Best Music, SL100

Today's Best Music on the radio. WNSL services the Laurel - Hattiesburg area of Mississippi.
www.sl100.com/ - 53k - 3 Jan 2007 - [Cached](#) - [Similar pages](#)

Nortel: Programs - Developer Program -Compatible with Meridian SL100

A list of Developer Program products tested compatible in a laboratory environment with Meridian **SL100**.

www.nortel.com/prd/dpp/product/sl100.html - 15k - [Cached](#) - [Similar pages](#)

Nortel: Programs - Developer Program Compatibility Certificate for ...

Meridian **SL100**. OPERATING SYSTEMS: Windows Server 2003. DEV. PRODUCT RLS LEVEL: 3.012, NORTEL RELEASE LEVEL: SE06. S/W Patch Release: Not Applicable ...

www.nortel.com/prd/dpp/product/prodpages/certs/cert1193.html - 11k -

[Cached](#) - [Similar pages](#)

See results for. [sl100 transistor](#)

Query
suggestion

[IndustryCommunity.com] Re: SL100 transistor specs and datasheet ...

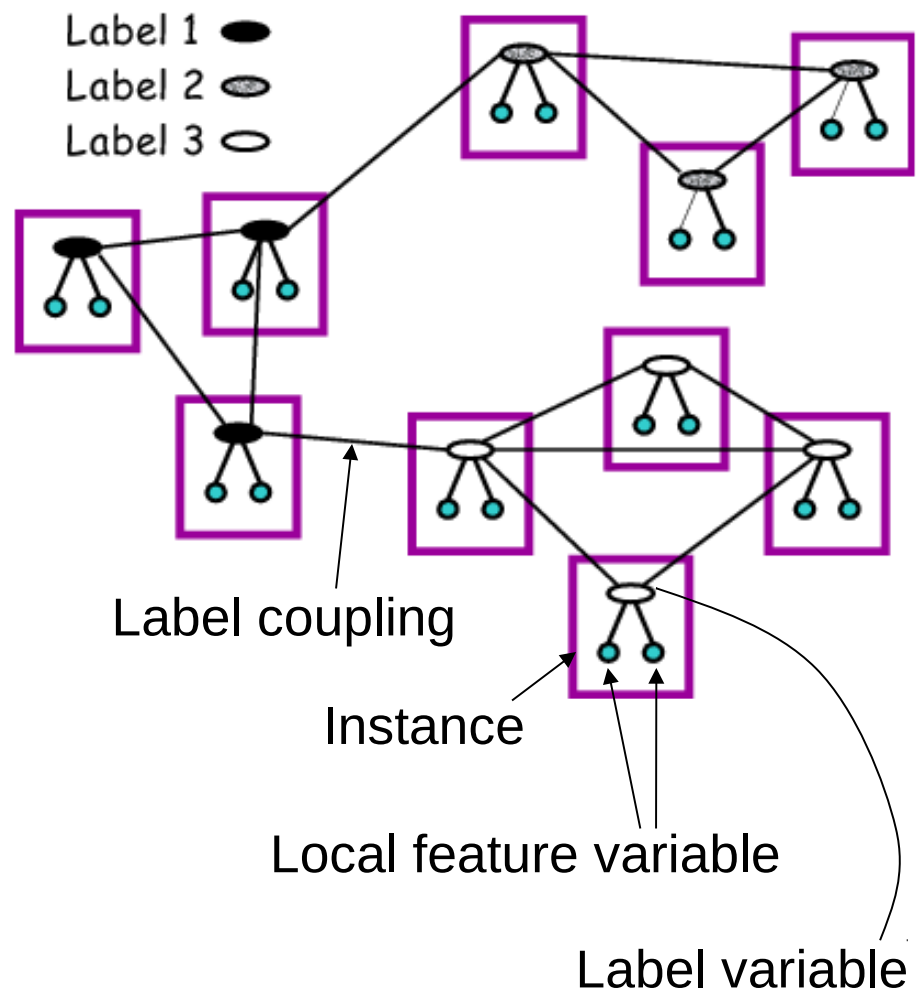
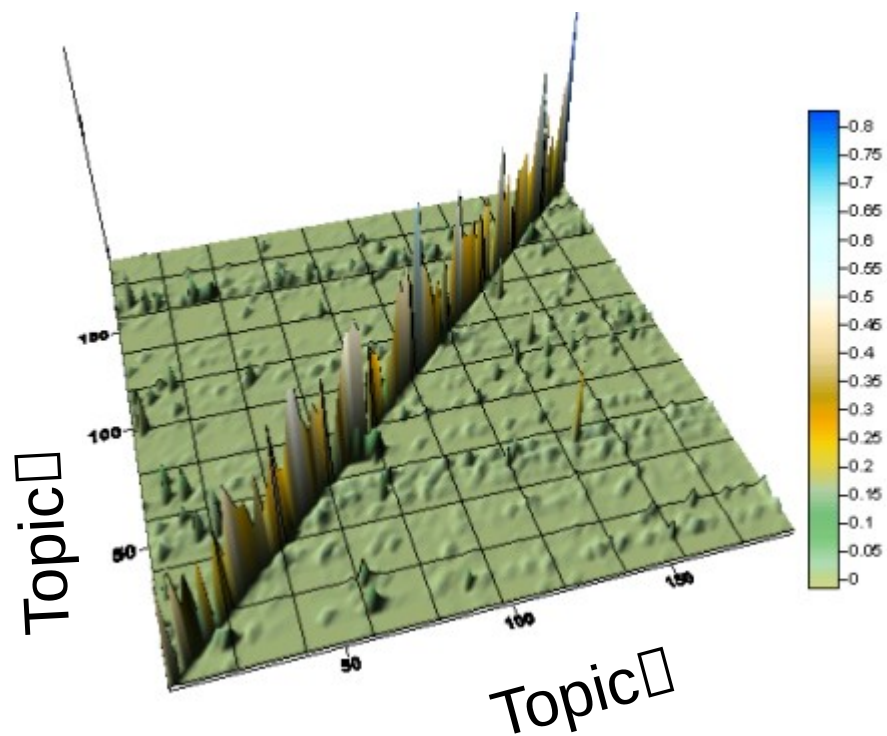
In Reply to: Re: **SL100 transistor** specs needed posted by shailendra mishra on ... **SL100 &Sk100 transistor** specs and datasheet needed T.M.KAREEMULLAH Posted ...

www.industrycommunity.com/myforum/john_dunn_next6/messages/462.html

Query = SL100

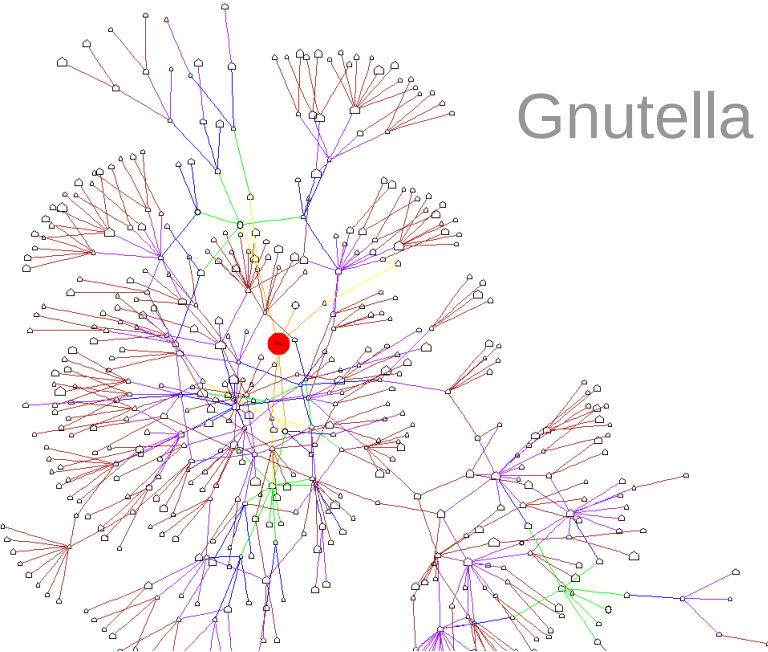
Whole-document labeling/classification

- Topics on Yahoo!, spam vs. non-spam, etc.
- Bayesian classification using generative corpus models
- Conditional probabilistic classification
- Discriminative classification
- Transductive, semi-supervised and active labeling
- Exploiting graph connectivity for improved labeling accuracy
- Machine learning needed here

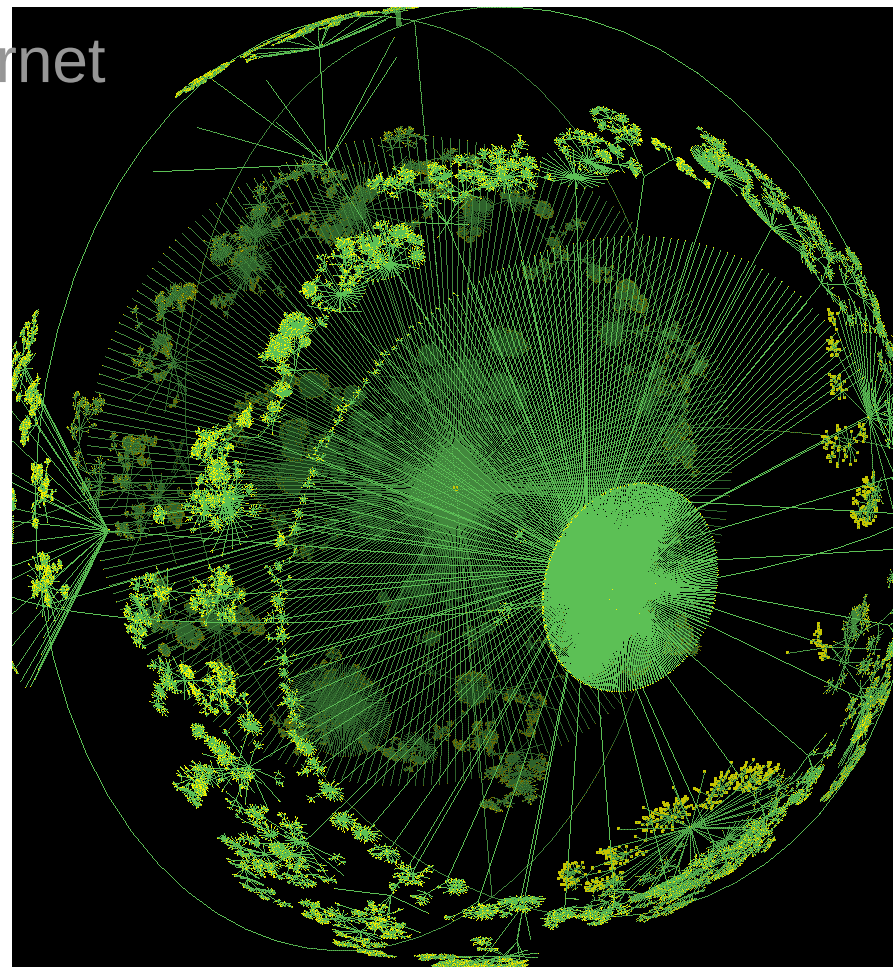


Measuring and modeling social networks

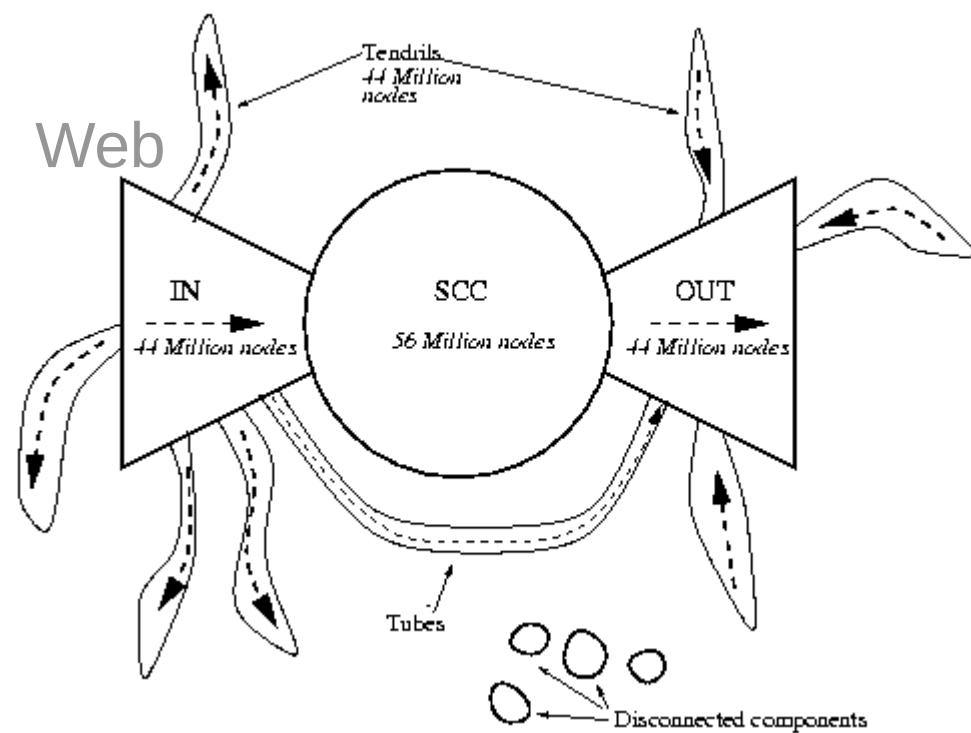
- Prestige, centrality, co-citation
- Web graph: degree, diameter, dense subgraphs, giant bow-tie
- Generative models: preferential attachment, copying links, “Googlearchy”, aging
- Link locality and content locality
- Generating synthetic social networks
- Compressing and indexing large social networks, reference compression, connectivity server, reachability index



Internet



Web



Hyperlink assisted search and mining

- Review of spectral graph theory
- Hyperlink induced topic search (HITS) and Google's Pagerank; computational issues
- Stability, topology sensitivity, spam resilience
- Other random walks: SALSA, PHITS, maxent walks, absorbing walks, SimRank, ...
- Personalized and topic-sensitive Pagerank
- Viral marketing, social networks

Web sampling, crawling, monitoring

- Plumbing: DNS, TCP/IP, HTTP, HTML
- Large-scale crawling issues:
concurrency, network load, shared
work pool, spider traps, politeness
- Setting crawl priorities using graph
properties and page contents
- Sampling Web pages using random
walks
- Monitoring change and refreshing
crawls
 - Driven by query workload
 - Driven by search and advertising

Froogle

Results **1 - 10** of about **12** confirmed / **17** total results for **digo**. (0.22 seconds)

View

> **List view**

[Grid view](#)

Sort By

> **Best match**

[Price: low to](#)

[high](#)

[Price: high to](#)

[low](#)

Price Range

\$ to

\$

Group By

[Store](#)

> **Show All**

Products

Search within

> **All Categories**



[Sony DSC-P72 Cybershot Digital Camera 3.2M Pixel](#)

\$287.95 - [Compare prices](#)

SONY DSC-P72 CYBERSHOT DIGITAL CAMERA 3.2M PIXEL

ATACOM: [3.2 / 5](#)



[Sony DSC-F717 Digital Still Camera 5M Pixel](#)

\$698.95

SONY DSC-F717 DIGITAL STILL CAMERA 5M PIXEL

ATACOM: [3.2 / 5](#)



[Sony DSC-U60 Cybershot Digital Camera 2M Pixel](#)

\$318.95

SONY DSC-U60 CYBERSHOT DIGITAL CAMERA 2M PIXEL

ATACOM: [3.2 / 5](#)

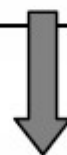


[Sony DSC-V1 Cybershot Digital Camera Optical Zoom](#)

\$549.95 - [Compare prices](#)

SONY DSC-V1 CYBERSHOT DIGITAL CAMERA OPTICAL ZOOM

ATACOM: [3.2 / 5](#)



[Sony DSC-V1 Cybershot Digital Camera Optical Zoom](#)

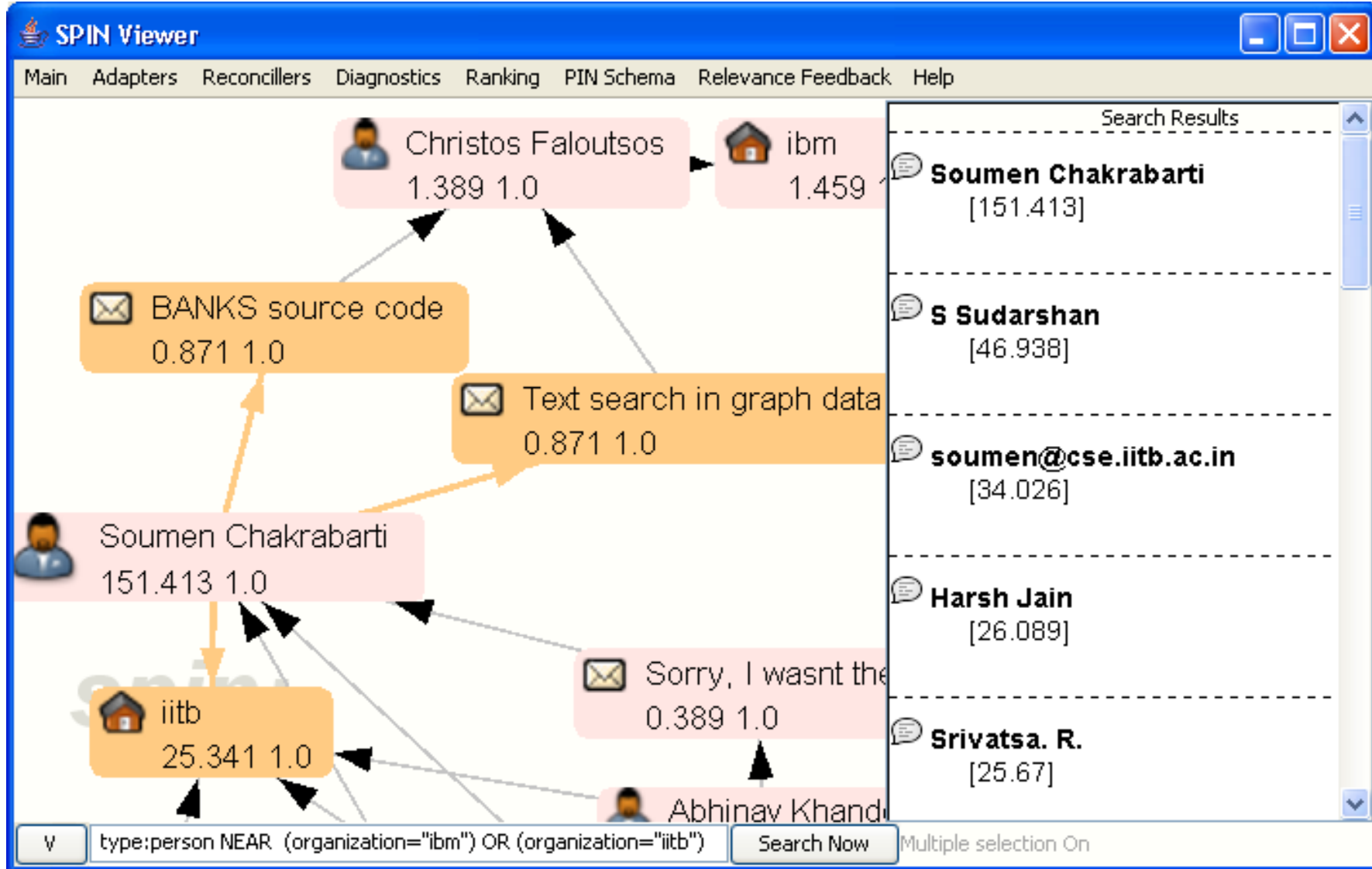
\$549.95 - [Compare prices](#)

SONY DSC-V1 CYBERSHOT DIGITAL CAMERA OPTICAL ZOOM

ATACOM: [3.2 / 5](#)

Adding graph structure to text search

- XML and related (largely) tree data models
- Typed entity-relationship networks
- Path expressions, integrating word queries with path matches; indexing, query processing
- Spreading activation queries: find entity of specified type “near” matching predicates
- Steiner query: explain why two or more entities (or words) are closely related



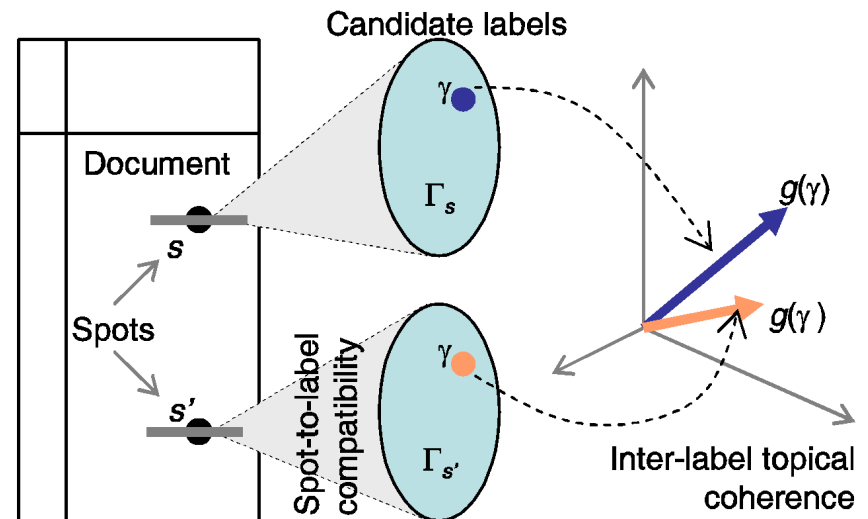
“Find a person near IBM and IITB”

Entity annotation in text

[http://en.wikipedia.org/wiki/Training_\(meteorology\)](http://en.wikipedia.org/wiki/Training_(meteorology))

In meteorology, training is when a successive series of showers or thunderstorms moves repeatedly over the same area, usually causing some form of flooding, especially flash floods. Often, this happens when a line of rain or storms forms along a stationary front, and moves down the length of the front, while the front is stalled. It is named so because this is similar to the way train cars from your training sessions, the nutrients and supplements that you consume after you've a huge impact on how you'll be rewarded for the work you did while you were there. Post-exercise Nutrition During intense exercise, our bodies use glycogen, amino acids and fluids at a rapid rate, what is often referred to as a catabolic state. Our goal with your post-workout nutrition is to return the body to an anabolic state as soon as we can once your session is over. This will help you recover from the training and allow you to improve and conditioning at a faster rate. Let's take a look at some general guidelines here as effectively as possible. Carbohydrates

- Need a label catalog
- Michael Jordan, Stuart Russell
- Collective annotation



Bridging Structured-Unstructured Gap

Name a **physicist** who searched for intelligent life in the cosmos

□ type=**physicist** NEAR “cosmos”...

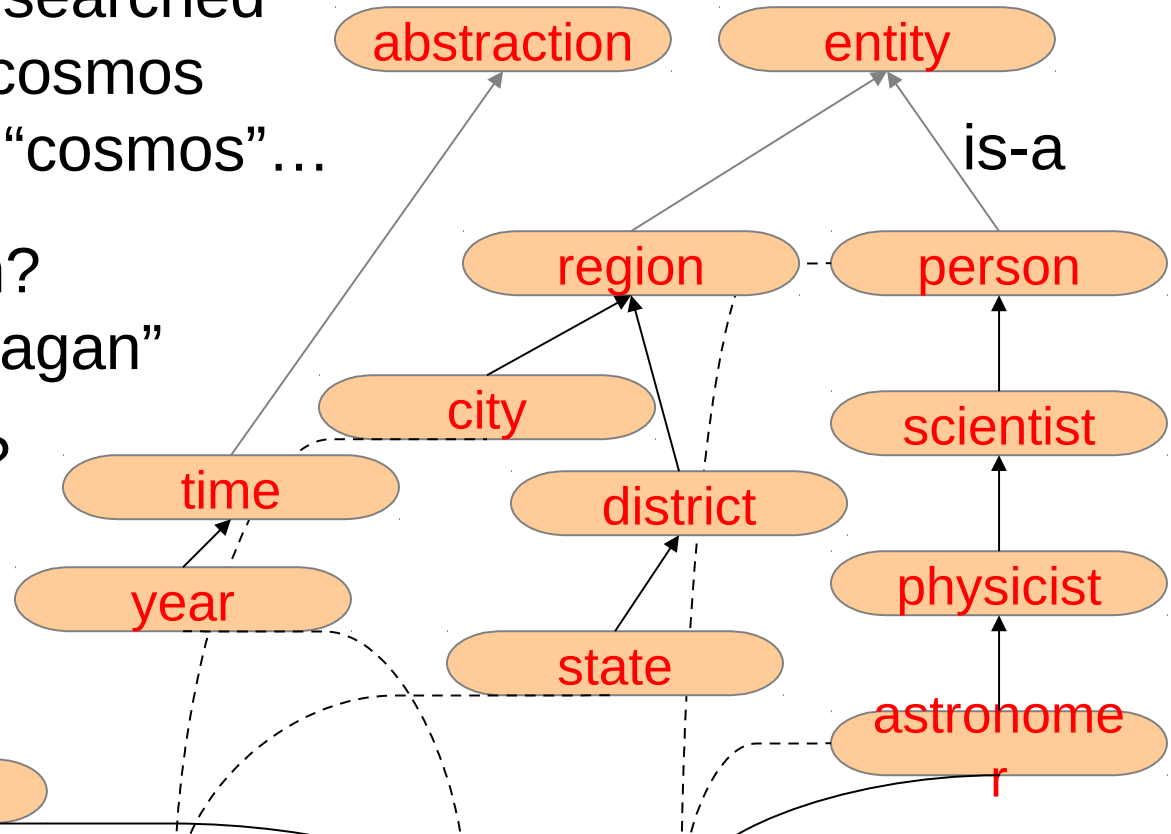
Where was Sagan born?

□ type=**region** NEAR “Sagan”

When was Sagan born?

□ type=**time**

pattern=**isDDDD** NEAR
“Sagan” “born”



Born in New York in 1934, Sagan was a noted astronomer whose lifelong passion was searching for intelligent life in the cosmos.

Mining the Web as a noisy database

Target Entity Set: ?distance

Matching conditions: {Mumbai}, {Vadodara}, distance, far

Wikipedia entities

String tokens

Search

Example query in basic query language

Target Entity Set: ?{Category: Scientist}

Matching conditions: {Category: Musical Instrument}, played

Wikipedia categories

String tokens

Search

Example query involving Wikipedia categories

Query1

Target Entity Set: ?{Category: French Film} ?{Number}

Matching conditions: academy, award, won

Search

Query2

Target Entity Set: ?{Category: French Film} ?{money:USD}

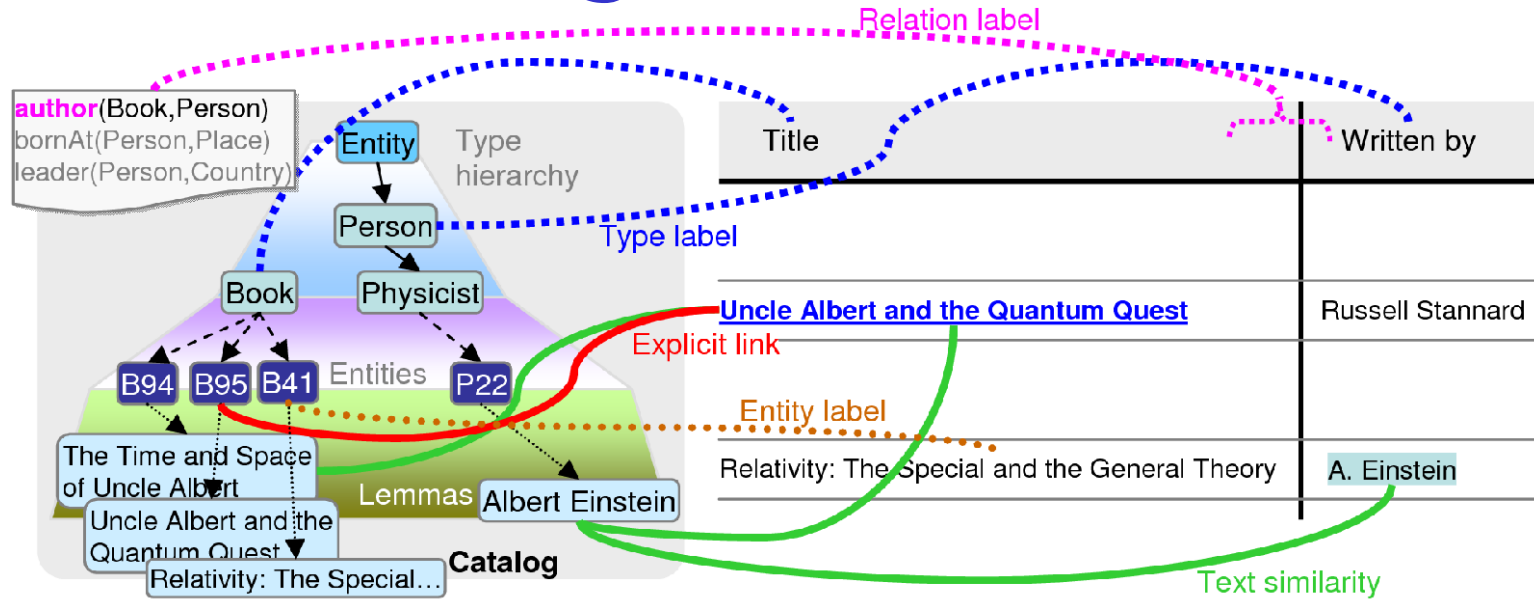
Matching conditions: production, cost, budget

Search

Result table with French films, number of academy awards won and production cost for each

Compile a table from multiple queries

Mining Web tables



- “Organically grown” Web tables
- Much relational information
- No coordinated schema
- Use social entity and type catalogs to label

Small answer subgraph search

BanKS Nick Roussopoulos Christos Faloutsos in using

[Search](#) [Browse](#) [Templates](#) [Query](#)

Searched DBLP [Complete] for **Nick Roussopoulos Christos Faloutsos** Results **1 - 10**. Search took **14.033** seconds.

Keyword(s) [nick](#) matches 161 nodes; [roussopoulos](#) matches 3 nodes; [christos](#) matches 81 nodes; [faloutsos](#) matches 4 nodes; Click on keywords to select or filter nodes. Time Profile: 1:651:13381[dbLoad:dbLookup:Expansion]

Rank: 1 **Score: 0.17376289** (es=0.17445762, ns=0.17101157) **Seqnum: 3** **Time: 1748** [[Similar Results](#)]

Table: writes Prestige=2.56348E-7, EdgeCost=0.0
name=Nick Roussopoulos, **paperid=conf/Vldb/SellisRF87**,

Table: paper Prestige=1.08929E-6, EdgeCost=1.0
paperid=conf/Vldb/SellisRF87, **title=The R+-Tree: A Dynamic Index for Multi-Dimensional Objects.**, **year=1987**,

Table: writes Prestige=2.52925E-7, EdgeCost=1.7320508
name=Christos Faloutsos, **paperid=conf/Vldb/SellisRF87**,

Table: author Prestige=1.35053E-5, EdgeCost=1.0
name=Christos Faloutsos, **url=**,

Table: author Prestige=1.04098E-5, EdgeCost=1.0
name=Nick Roussopoulos, **url=**,

<http://www.cse.iitb.ac.in/banks/>

Bootstrapping Web knowledge bases

- Hearst, 1992; KnowItAll (Etzioni+ 2004)
 - T such as x, x and other Ts, x or other Ts, T
x, x is a T, x is the only T, ...
- Google sets

<u>Cat</u>	<u>cat</u>	<u>England</u>	<u>Japan</u>
<u>Dog</u>	<u>more</u>	<u>France</u>	<u>China</u>
<u>Horse</u>	<u>ls</u>	<u>Germany</u>	<u>India</u>
<u>Fish</u>	<u>rm</u>	<u>Italy</u>	<u>Indonesia</u>
<u>Bird</u>	<u>mv</u>	<u>Ireland</u>	<u>Malaysia</u>
<u>Rabbit</u>	<u>cd</u>	<u>Spain</u>	<u>Korea</u>
<u>Cattle</u>	<u>cp</u>	<u>Scotland</u>	<u>Taiwan</u>
<u>Rat</u>	<u>mkdir</u>	<u>Belgium</u>	<u>Thailand</u>
<u>Livestock</u>	<u>man</u>	<u>Canada</u>	<u>Singapore</u>
<u>Mouse</u>	<u>tail</u>	<u>Austria</u>	<u>Australia</u>
<u>Human</u>	<u>pwd</u>	<u>Australia</u>	<u>Bangladesh</u>

Information carnivores at work

KO :: India Pakistan Cricket Series

A web site by Khalid Omar, sort of live from Karachi, **Pakistan**.

Probe	Word	Phrase
Khalid	1.3M	0
Omar	6.63M	0
sort	130M	0
Karachi	2.51M	629
Pakistan	50.5M	1

“cities such as [probe]”

“[probe] and other cities”, “[probe] is a city”, etc.

- “Garth Brooks is a country” [singer],
“gift such as wall” [clock]
- “person like Paris” [Hilton],
“researchers like Michael Jordan” (which one?)

Sample output

- author; “Harry Potter”
 - J K Rowling, Ron
- person; “Eiffel Tower”
 - Gustave, (Eiffel), Paris
- director; Swades movie
 - Ashutosh Gowariker, Ashutosh Gowarikar
- What can search engines do to help?
 - Cluster mentions and assign IDs
 - Allow queries for IDs — expensive!
 - “Harry Potter” context in “Ron is an author”

Ambiguity and
extremely skewed
Web popularity

Summary

- Large scale text analysis and search
- Interface between unstructured text and structured knowledge
- Mix of theory and practice
- Applies algorithms, statistics, machine learning to the text + structured data domains