Web Search and Mining
CS610 Spring 2007
TuF 6:35—8pm
SIC301

Soumen Chakrabarti

TAs to be decided

Administrivia

- http://www.cse.iitb.ac.in/~soumen/teach/cs610s2007/
- Newsgroup iitb.courses.cs610 on jeeves.cse.iitb.ac.in
- Meant for Mtech1, Btech3, Mtech2, Btech4
- CS705 in Fall 2006 is strongly recommended as a prerequisite
- If you can read and understand a fair bit of math, you will need ~10 days of self-study to pick up the relevant pieces

Multidisciplinary synthesis

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

Course design

- Basic material from a few books
- Research papers
  - You get credit for pointing out loopholes, flaws, and further ideas to follow from existing papers
- Homeworks
  - Light on paperwork, slightly heavier on hands-on work (coding, simulations and measurements)
  - Exploring possibility of switching one or few homework assignments with larger projects
- Exams
  - “Served on a platter” than real systems issues

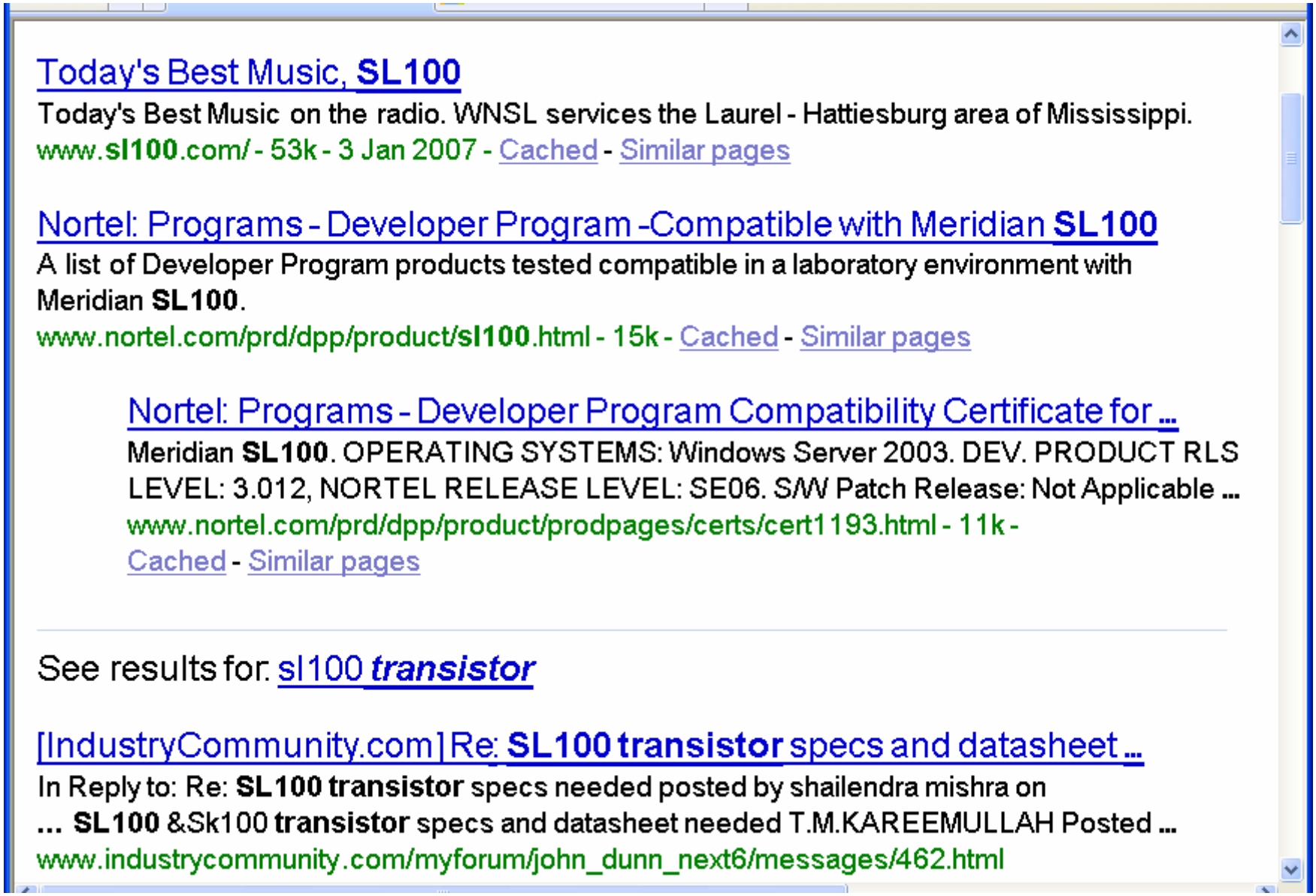
Syllabus at a high level

- Document and corpus models
- Text indexing, search and ranking
- Whole-document labeling/classification
- Measuring and modeling social networks
- Hyperlink assisted search and mining
- Adding (XML) graph structure to text search
- Bootstrapping Web knowledge bases
- Web sampling, crawling, monitoring

Now, drilling down a bit, …

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


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
[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

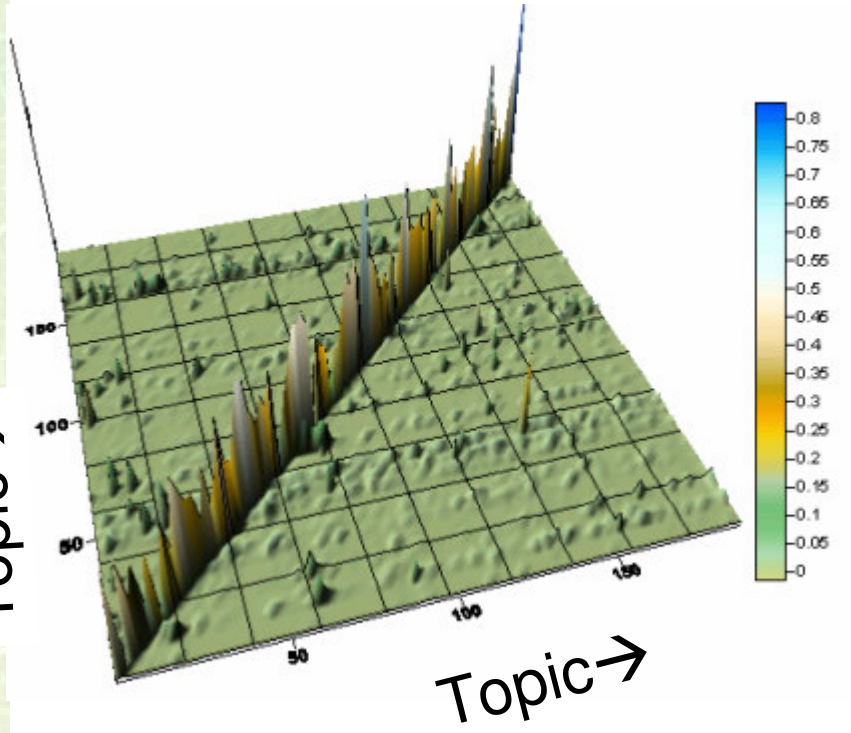
Text indexing, search and ranking

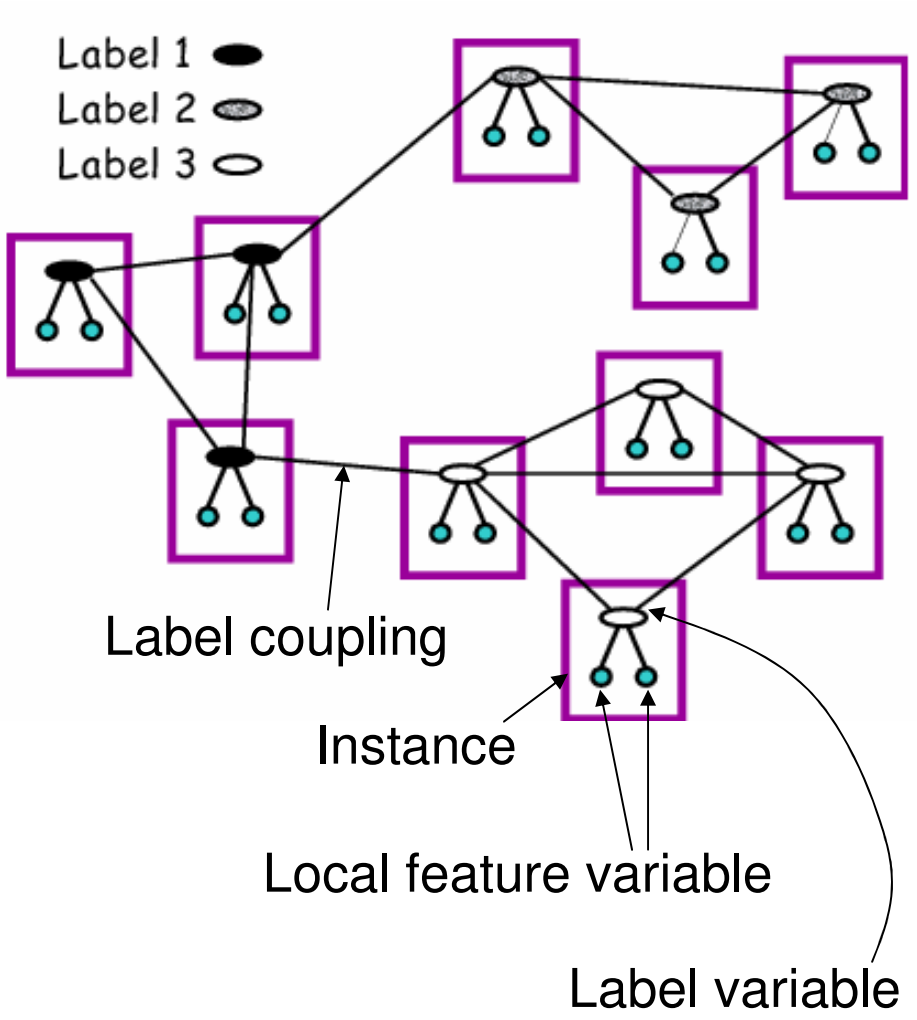
- Boolean queries involving word occurrence in documents
- Inverted index design, construction, compression, updates; query processing
- Vector space model, relevance ranking, recall, precision, F1, break-even
- Probabilistic ranking, belief networks
- Fast top-k search
- Similarity search: minhash,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

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

CS705 needed here

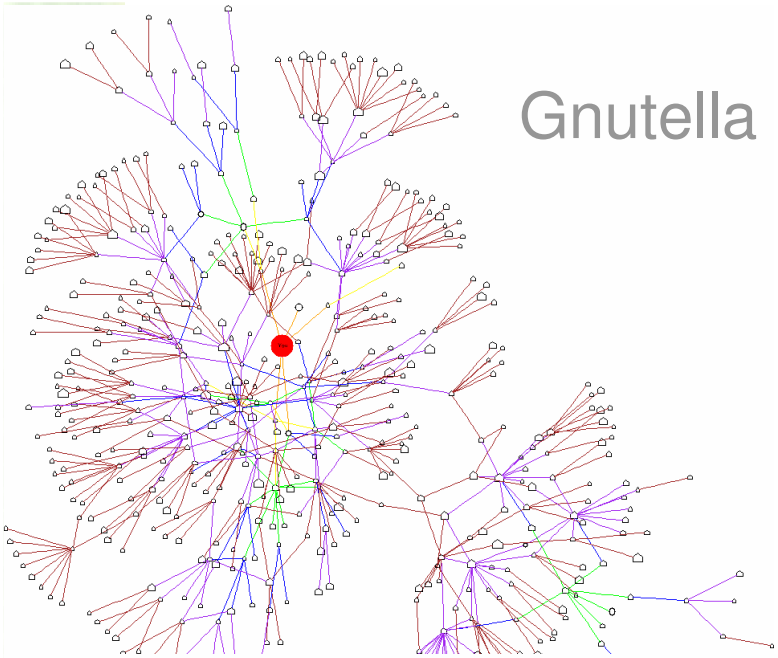

Topic→
Topic→
150
100
50
50
100
150
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0


Label 1
Label 2
Label 3
Label coupling
Instance
Local feature variable
Label variable

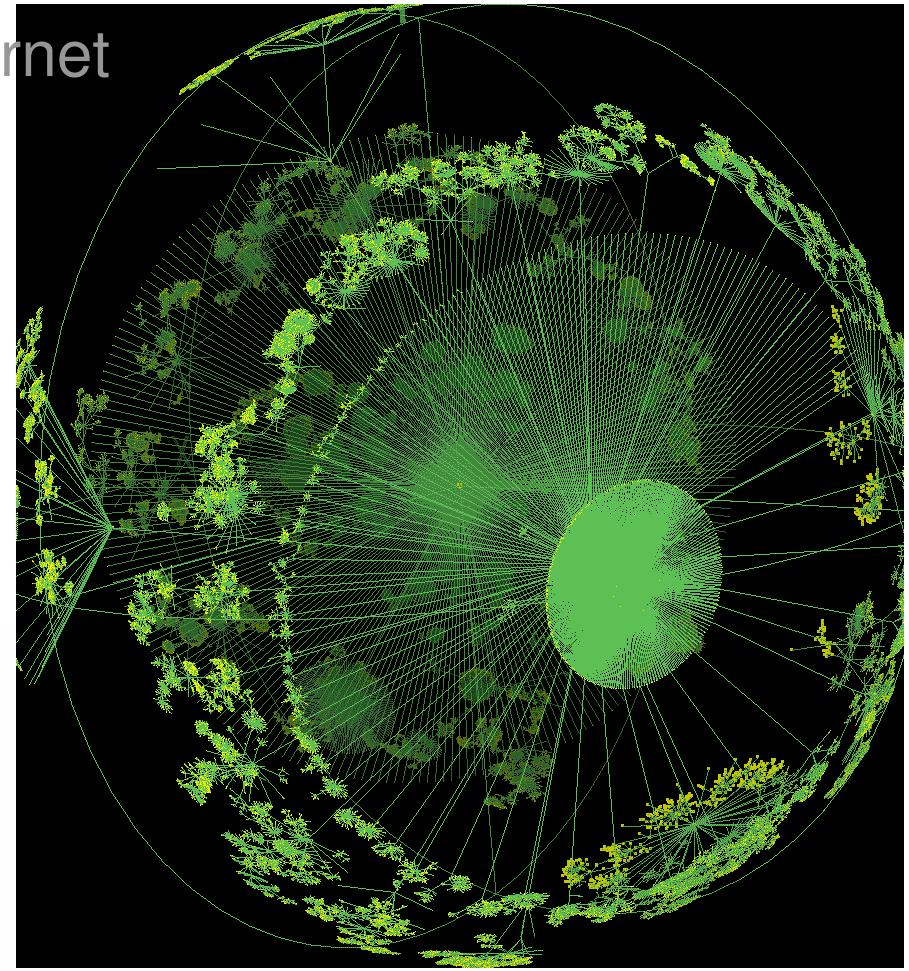
Measuring and modeling social networks

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

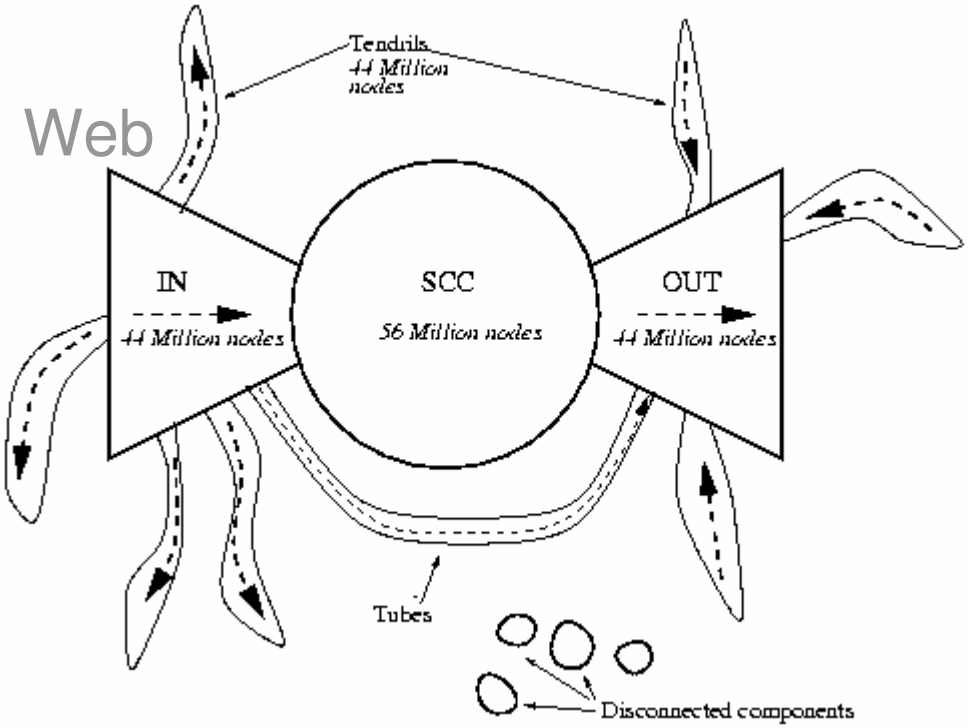
Gnutella



Internet

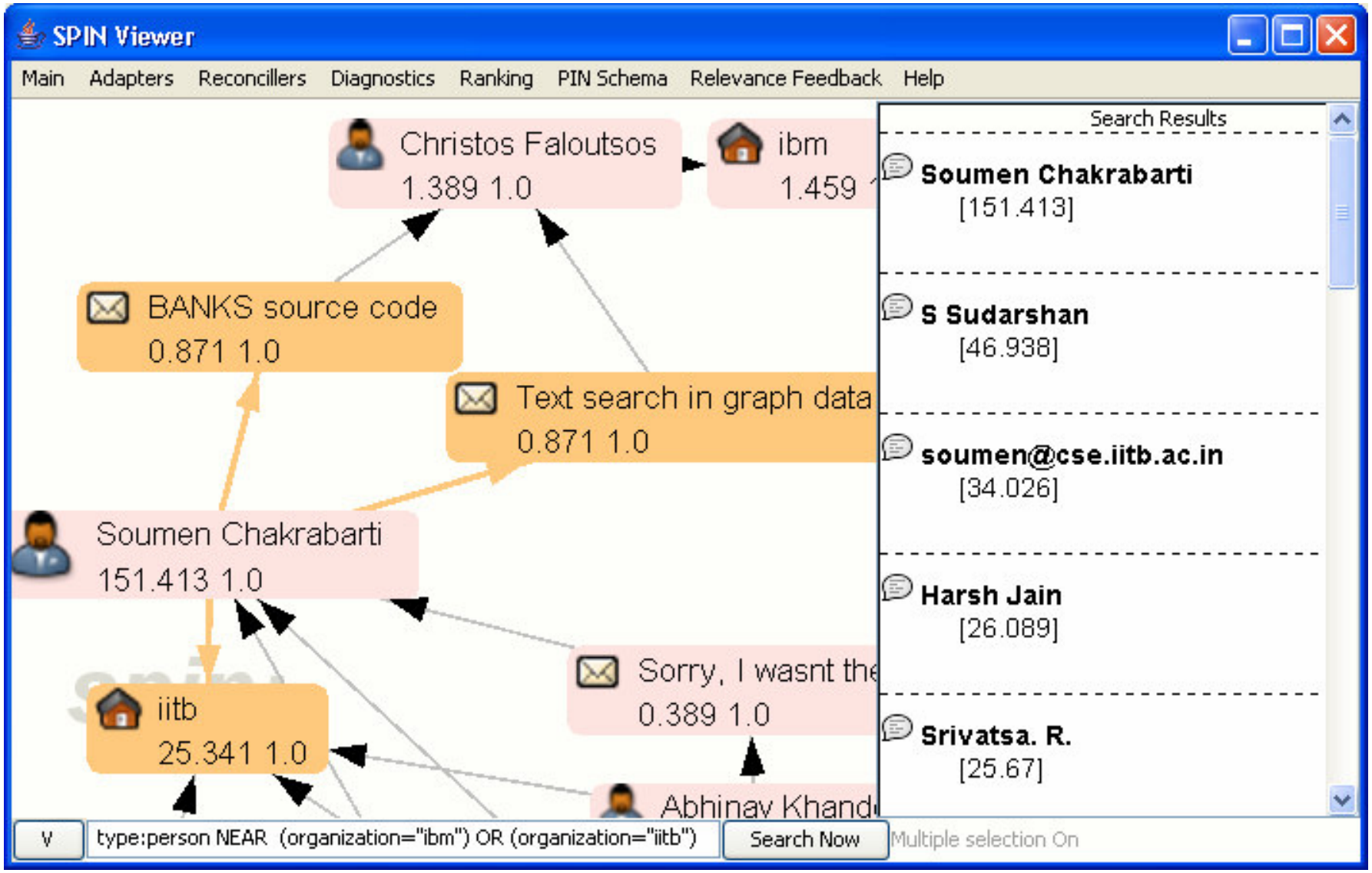


Web



Hyperlink assisted search and mining

- Review of spectral graph theory CS705
- Hyperlink induced topic search (HITS) and Google's Pagerank
- Large-scale computational issues
- Stability, topology sensitivity, spam resilience
- Other random walks: SALSA, PHITS, maxent walks, absorbing walks, SimRank, …
- Personalized and topic-sensitive Pagerank
- Viral marketing in social networks

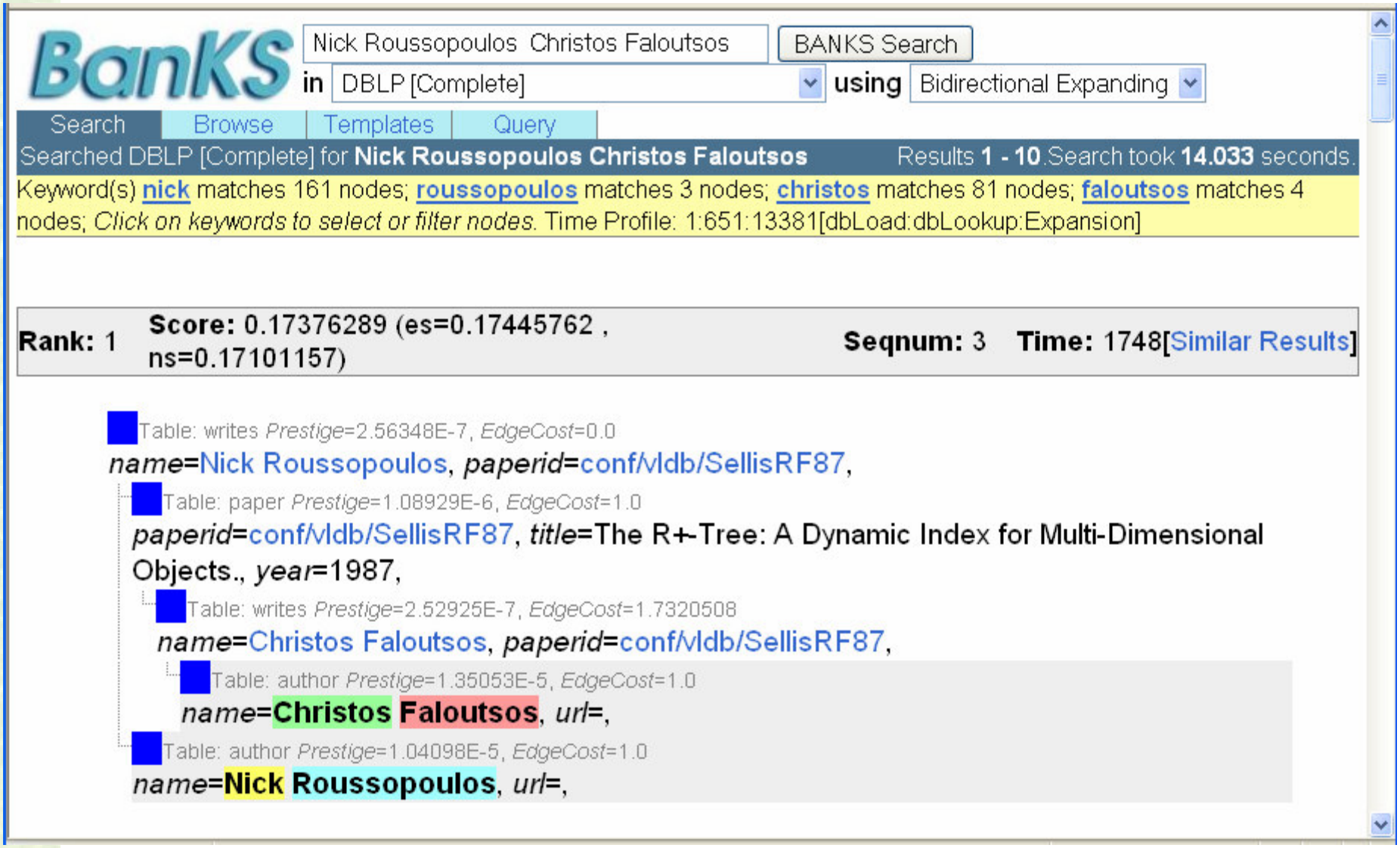


"Find a person near IBM and IITB"

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

Small answer subgraph search



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

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

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

Froogle BETA
Web Images Groups News Froogle more »
digc
Search Froogle
Search the Web
Advanced Fro
Preferences
Froogle
Results 1 - 10 of about 12 confirmed / 17 total results for digc. (0.22 seconds)
View
> List view
Grid view
Sort By
> Best match
Price: low to high
Price: high to low
Price Range
$ to
$ Go
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

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