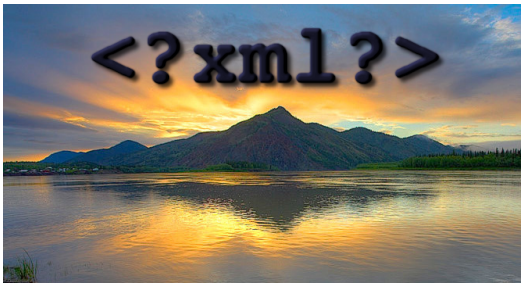
Lecture 24
Topic Review and
The Glorious Future



Administration

- Final Projects
  - We're available if you want more advice
  - CODE DEADLINE is 10:30AM on April 19
  - That's right before class; do not modify your code after that time
  - In class, you will have 4 minutes to demo your work on our laptop
  - If URL is different from Alpha, email us!
  - Short document description of your project due by 11:55PM April 20; this is to help us grade

2

Administration

- Midterm #2 available after class
- Final Exam
  - Format will be similar to midterms
  - A two-hour exam, but closer to MT#2 than to MT#1 in terms of "time-challenge"
  - All the content in the lectures, throughout class, is fair game
  - You are permitted two double-sided 8.5x11 sheets of handwritten notes
  - **April 28, 10:30AM-12:30PM**
  - Makeup: April 23, 1PM-3PM

3

Outline

- Overview and review of topics
- The Future Of The Web
  - *aka*, the most exciting new topics

4

Review

- We've covered three very broad areas
  - Part 1: Client/server data exchange
  - Part 2: Processing many users
  - Part 3: Large-scale system support
- Many technical details, but a few main topics to remember for each

5

Client/Server Data Exchange

- 1: HTTP & Client/Server model
- 2: Dynamic Content
  - Multi-tiers & Model-View-Controller model
  - JS on client & security concerns
- 3: Networking Basics
  - TCP/IP algs & HTTP-TCP interactions
- 4: Personalization
  - Sessions, cookies, logins

6

Client/Server Data Exchange

- 5-6: Security
  - Attacks & traditional encryption
  - Public key crypto, digital signatures
- 7-9: XML
  - Formatting, DTDs, XSLT
  - Schemas, XPath, Xquery
  - Web Services

7

Processing Many Users

- 10-11: Info Retrieval
  - Search architecture basics
  - IR scoring, vector space model, tf-idf
  - Precision & Recall, Kendall's Tau
- 12-13: Modern Search Engines
  - PageRank and other Link Analysis
  - Crawling, index construction, shingling
  - Distributed operation
- 14: Research Topics
  - No questions on this

8

Processing Many Users

- 15: Auctions
  - Auction types, bidder/seller motives

- 16: Recommendation Systems
  - Item-based vs User-based
  - Basic similarity measurement techniques
- 17: Logs & Data Mining
  - Classifiers, cross-validation, supervised vs unsupervised
  - Apriori

9

Large-Scale Systems

- 18: Domain Name System
  - DNS caching, Akamai
- 19, 21: Scaling & Caching
  - Replication vs Partitioning
  - Distributed writes and 2-phase commit
  - Proxies, HTTP caching techniques
- 22: MapReduce and GFS
  - Architecture, programming model
- 23: Datacenters

10

The Future

Client/Server	Many Users	Large Systems
HTTP, crypto, security, XML	IR, search, rec+auction, data mining	DNS, caches MapReduce, GFS

11

The Future

Client/Server	Many Users	Large Systems
HTTP, crypto, security, XML	IR, search, rec+auction, data mining	DNS, caches MapReduce, GFS
(HTML5) Zoetrope	DBPedia, Freebase, Wolfram Alpha	BigTable

12

HTML5

- HTML5 is a huge upgrade
  - 2D drawing on canvas
  - Local in-browser storage
  - Video and audio playback
- Way too big for single class
- Most issues are political at this point

13

Zoetrope

- Adar, Dontcheva, Fogarty, Weld (2008)
- Offers brand-new query interaction model with server-side data
- Current "now Web" has no history
  - Internet Archive captures some deltas
  - Captured pages are hard to explore
- Zoetrope users apply operations on *content streams*
  - Could also be useful for other stream types: sensors, weather, etc

14

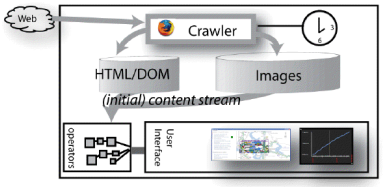
Zoetrope

- <video>

15

Zoetrope Internals

- Blends search, browsing. Employs:
  - Custom crawler, db, novel UI
- Content stream: series of <T_i, C_i> pairs
  - T_i is time crawled
  - C_i is content (both XML structure & render)



16

Zoetrope Internals

- Lenses are applied to tuple stream
  - Visual lenses crop a region of the screen; repeated ops on C_i info
  - Structural lenses also focus on region, but track content across time; ops on C_i
    - Relies on some amt of webpage stability
  - Textual lenses allow selecting a text-elt, not a spatial/structural one
- Filters also possible, usually on time

17

Zoetrope Crawler

- How does the Zoetrope crawler differ from the standard search crawler?

18

Zoetrope Summary

- Web currently has no memory at all
- Zoetrope both collects and exposes historical web info
- Taking advantage of history means more complicated queries

19

Linked Data

- A usable subset of the Semantic Web?
  - URIs identify real-world *things*
  - Following URI yields metadata
  - Metadata includes refs to other *things*

20

About: Iceland

An Entity of Type: populated place, in Data Space: dbpedia.org

Iceland is a European island country located in the North Atlantic Ocean. It has a population of about 320,000 and a total
Reykjavík, whose surrounding area is home to some two-thirds of the national population. Located on the Mid-Atlantic Ri
defines the landscape.

Property	Value
dbpedia-owl:abstract	Island er et land, der ligger i den nordlige del af Atlanterhavet mellem Grønland og befolkning på 325.000. Landets hovedstad og største by er Reykjavík. Grundet la geotermisk aktivitet på Island. De centrale dele af Island består af et plateau kara der flyder mod havet. Pga. den varme Golfstrøm har Island et relativt mildt klima beboet ca. siden 874, da den norske høvding Ingólfur Arnarson ifølge Landnámab tidligere og havde overvintret. I løbet af de næste århundreder bosatte folk af nord af først Norge og senere Danmark. I det 20. århundrede har Islands økonomi og v rigeste og mest udviklede i verden med en meget ekspansiv politik og risikofyldte økonomisk rutsjetur. I februar 2009 var 8,2 procent af arbejdsstyrken uden job. Is Iceland is a European island country located in the North Atlantic Ocean. It has a

21

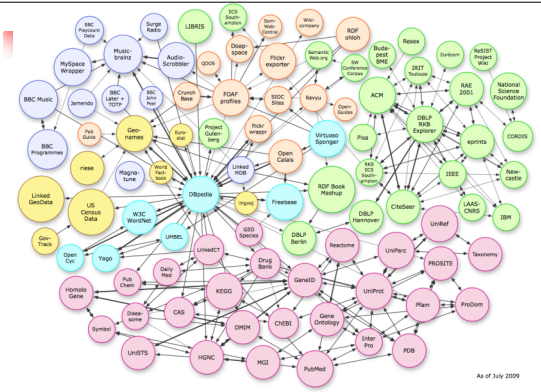
dbpedia-owl:anthem	dbpedia:Lofsöngur
dbpedia-owl:areaMetro	103001000000.000000 (xsd:double) 103003827148.062729 (xsd:double)
dbpedia-owl:capital	dbpedia:Reykjavík
dbpedia-owl:demonym	Icelander, Icelandic
dbpedia-owl:ethnicGroup	dbpedia:Icelanders dbpedia:Iceland#Demographics
dbpedia-owl:foundingDate	1944-06-17 (xsd:date)
dbpedia-owl:governmentType	dbpedia:Parliamentary_republic
dbpedia-owl:leaderName	dbpedia:Ólafur_Ragnar_Grímsson dbpedia:Jóhanna_Sigurðardóttir dbpedia:Ásta_Ragnheiður_Jóhannesdóttir
dbpedia-owl:leaderTitle	President Prime Minister Althing President
dbpedia-owl:percentageOfAreaWater	2.700000 (xsd:float)
dbpedia-owl:populationDensity	2.895766 (xsd:double) 3.100000 (xsd:double)
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/c/ce/Fl
dbpprop:areaKm	103001 (xsd:integer)
dbpprop:areaMagnitude	1 E11

22

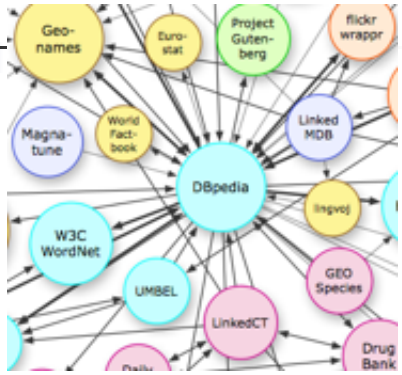
Linked Data

- URL-based interlinking and "alias" system makes it easy for sites to specialize in one kind of data

23



24



25

Open Linked Data FAQ

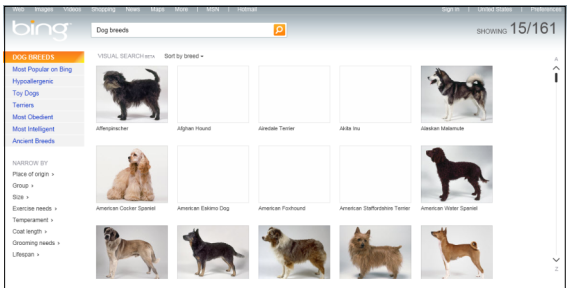
- Datasets contain fact triples
  - einstein invented relativity
  - limes contain vitamin-C
- (But for each obj and pred, use a unique and agreed-upon URL, as with DTDs in XML)
- RDF is triple format
- Usually, SPARQL is query lang
  - Freebase is similar, offers MQL
  - Wolfram Alpha offers only a textbox!!!

26

Future of Open Linked Data

- Right now, in the data-building phase
  - DBpedia has emerged as huge "hub" of other data resources
- 3.4 million things
  - 312K people
  - 413K places
  - 140K organizations
  - Etc, etc
- Total English Wikipedia has 3.2M (!)

27



28

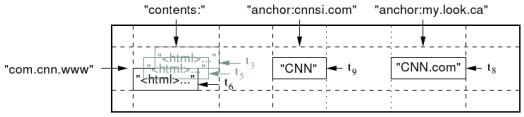
Linked Data Summary

- Very lightweight approach to building shared structured dataset
- "All items have URI", and "URI must be checked securely" in some amount of conflicting data elts
- DBPedia seems to be the king right now
  - Holds some of the promise of the semantic web

29

BigTable

- Traditional databases lack scale, failover
  - Should scale to petabytes, 1000s machines
  - Need updates not found in GFS
  - Tx semantics not needed
- Database-style storage built on cluster
  - Simple SQL queries
  - Sparse table format, w/timestamps



30

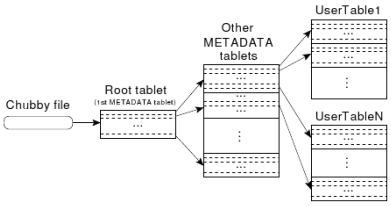
BT Details

- Sparse schema format
  - No need to preannounce schema
  - Empty cells are assumed
  - Columns can be stored in column families on-disk together
  - Column-oriented compression
  - Data values timestamped; BT keeps last *k*
- No locking across rows

31

BT Internals

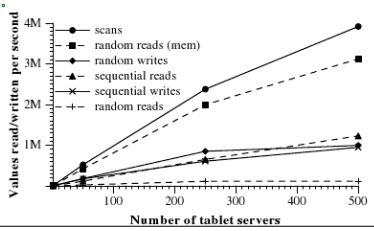
- Google File System stores data pieces
- MapReduce can implement "coprocessors" on data instead of query
- Tablet-finding critical piece



32

Scaling Challenges

- GFS provides failover
- BT responsible for loading/flushing tablets
- Expts show good scaling behavior



33

Conclusion

- New work on Web comes in a few flavors
  - User-specific (HTTP, Zoetrope)
  - Aggregating users (auctions, Linked Data)
  - Scaling systems (caching, BigTable)
- New datasets, applications, technology can spur advances in each or all

34

Conclusion

- Thanks.

35