Welcome to the class of

Advanced Topics in Information Retrieval !

Min ZHANG (张敏)
z-m@tsinghua.edu.cn

Tea Time

Workshop Schedule

Timing and Scoring

- Release your demo online test URL by 8am of the Tuesday morning in your presentation week on http://learn.tsinghua.edu.cn "课程讨论" section.
- Presentation time: 17 minutes for each project
  - 13 minutes presentation
    - E.g. 10 minutes design and introduction + 3 minutes demo
  - 4 minutes QA
- Every student is required to ask at least 1 questions on each workshop day

3

Scoring sheets and paper submission

Scoring

Present. (1)	Design (1.5)	Tech. (1.5)	Pre-test (1)	Live Demo (2)	QA (2)	Timing (1)	Total (10)

- Paper :
  - Submit your paper (~6 pages, A4, single space, body text font: not larger than 10pt) on the project by 23:59, June 20th on the homework section of learn.tsinghua.

Workshop Schedule – Day 1

- May 31, Friday

Workshop要求说明		
文献搜索与管理系统	邵韵秋	王晨阳
NBA球员搜索引擎	曲建波	赵东杰
大众新闻搜索引擎	石朋	林一夫
养生搜索引擎	王振	杨俊

Workshop Schedule – Day 2

- June 10, Monday (**online demo by Thur. Jun 6 morning**)

论坛搜索引擎优化: 以虎扑为例	邹昊	李昊阳
美食搜索引擎	石岩松	袁浩禹
音乐搜索引擎	宋鑫	石灵奇
美食搜索引擎	树扬	陈彦肇
机票搜索引擎	邓志杰	王鑫

Workshop Schedule – Day 3

- June 14, Friday

旅游攻略搜索引擎	吴至婧	余文梦
音乐领域搜索引擎	陈佳	张景辉
体育新闻搜索引擎	王淮	尹誉衡
？？	殷敏	/
Best Project Award		

Part II. Search Engine Techniques

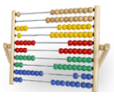
- Link-based analysis
- Challenging Topics
- User Modeling

CHALLENGES:

II. SCALE

Challenges II – Scale

- 2006, 400,000,000 queries
- July 2008, 7.5 Billion queries*
- Nov. 2004, docs：8,000,000,000
- July 25, 2008，1 trillion docs 1,000,000,000,000



- 2017: 20 Petabytes data per day
- Around 10-15 Exabytes in total+

1Exabyte = 1,000Petabyte

* Searchengine watch http://searchenginewatch.com/3630718
\+ https://www.heshmore.com/how-much-data-does-google-handle/

Twitter Moves 300 Petabytes to Google Cloud – That's a Lot of Covfefe

NEWS ANALYSIS
MITCH WAGNER, Executive Editor, Light Reading
9/18/2018

COMMENT (0)

Login

50% 50%

Like 19
Tweet
Share

Twitter is using tools running on Google Cloud to find out more about how people use its service, including monitoring abuse and mapping how conversations develop.

"We operate at a massive scale with a relatively small team," Twitter CTO Parag Agrawal said during a brief presentation at Google's annual cloud conference in San Francisco in July.

That small team needs to wrangle a lot of data, with over 300 petabytes migrated to Google Cloud, primarily for cold storage and ad hoc analysis, Agrawal said.

Twitter selected Google Cloud because the service could deliver performance. Google's architecture allows Twitter to scale storage and compute independently, and run multiple data analyses simultaneously.

Google also provides cost efficiency, using custom machine types to keep costs down. And Twitter liked Google's culture, including its open source commitment and focus on security, which gives Twitter confidence in the future of the partnership, Agrawal said.

Twitter made a long-term bet that Google is investing more in big data than Twitter can, said Derek Lyon, Twitter director of engineering and data infrastructure, in a later session at the conference.

The analytics move is part of a larger migration to Google cloud for multiple Twitter infrastructure platforms, including real-time analytics, NoSQL, messaging, object store, general computing, batch computing and Hadoop, Lyon said.

Twitter began its migration two years ago. Before migrating, Twitter took two to three months to evaluate multiple providers, Lyon said. Evaluation included a ten-year financial time horizon -- ten years being the life of a data center. Twitter compared the costs and merits of running analytics on-premises and in multiple different cloud scenarios. The evaluation factored in both migration costs and long-term operations.

Over the course of its evaluation, Twitter determined that a go-slow migration approach is best. "An immediate all-in migration at Twitter scale is expensive, distracting and risky," Lyon said.

Additionally, Twitter had concerns about whether Google could serve a company at Twitter's scale. "As a test, we asked for five petabytes of flash [storage] and Google was able to turn it around quickly," Lyon said. "This was a test of Google's capability to fill demand, and Google turned it around."

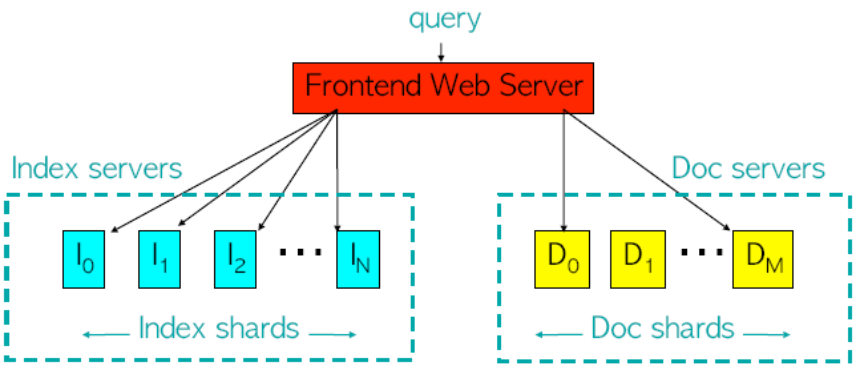
Watch the Twitter Hadoop presentation here:

https://www.lightreading.com/enterprise-cloud/data-strategy-and-analytics/twitter-moves-300-petabytes-to-google-cloud---thats-a-lot-of-covfefe/d/d-id/746167

Challenges II – Scale

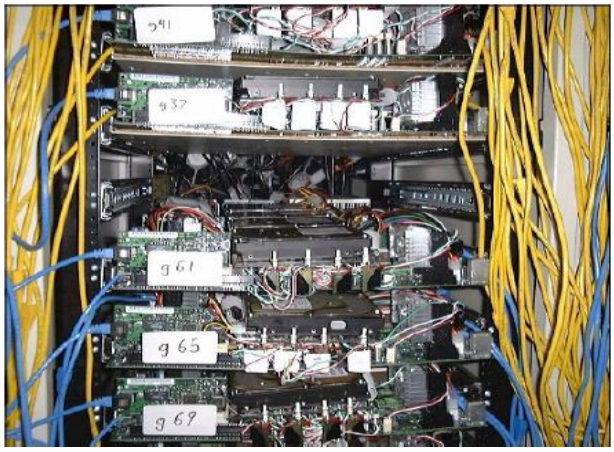
- Peak of google.stanford.edu ~ 1997

Research Project, circa 1997

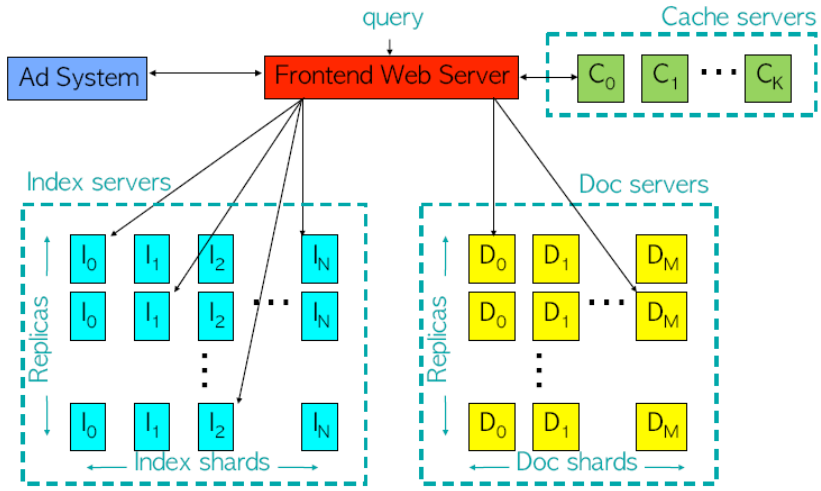


Challenges II – Scale

- "Corkboards" Google servers 1999

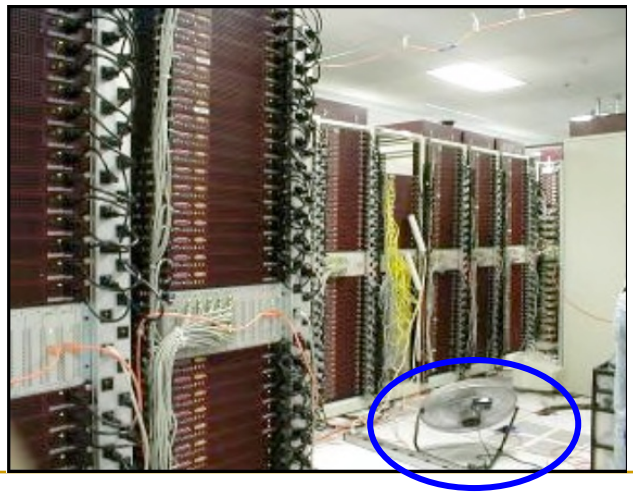


Serving System, circa 1999



Challenges II – Scale

- Google Datacenter 2000

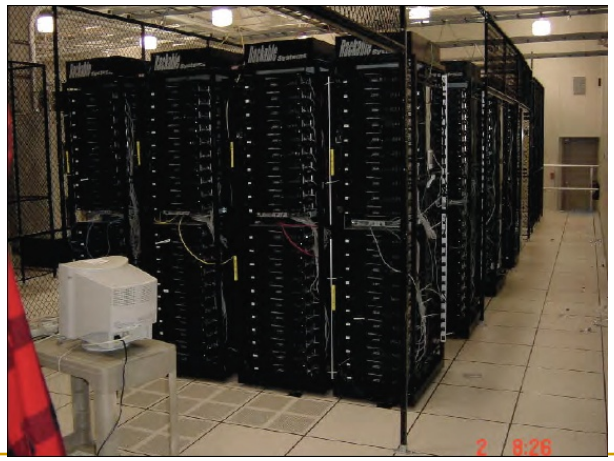


Challenges II – Scale

- Google new datacenter 2001

Challenges II – Scale

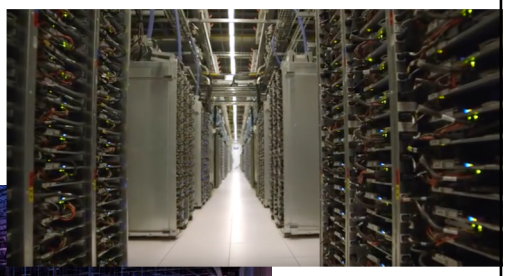
- Google new datacenter 2001 (3 days later)

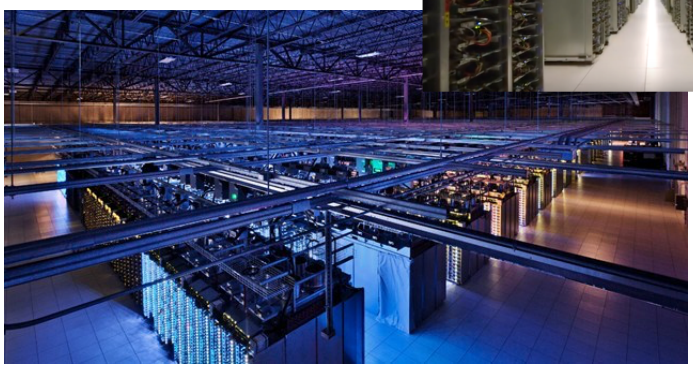


Challenges II – Scale

- Around 2015

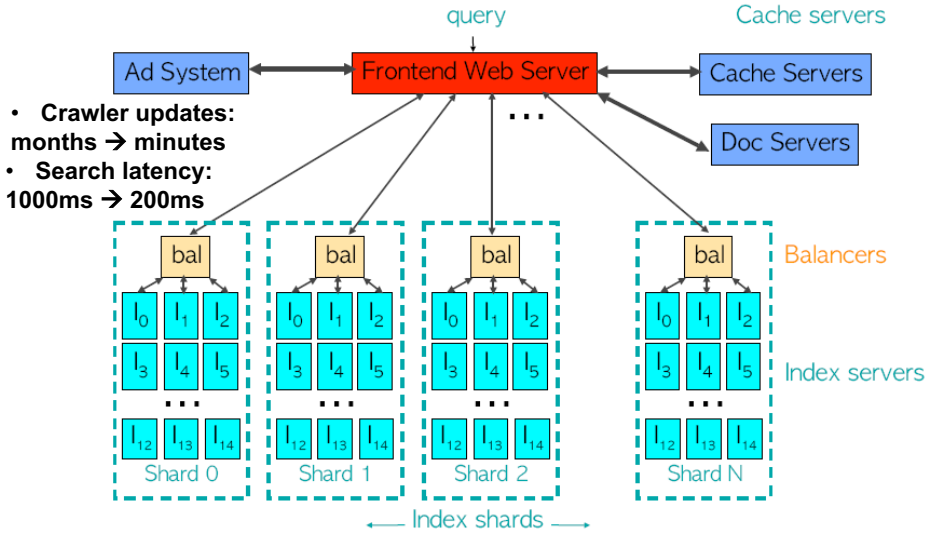
https://www.youtube.com/watch?v=XZmGGAbHqa0
Inside a Google data center





Joe Kava,
VP of Google's
Data Center
Operations
Dec. 16, 2014

Early 2001, in-memory index



Holding the complete search index in memory: resulting in **the use of 1000 machines** to handle a single query **compared to just 12 previously**

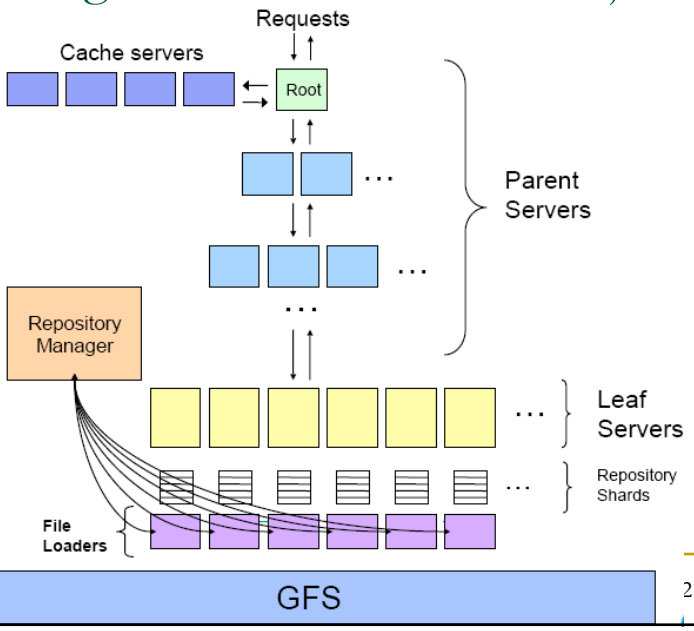
Challenges II – Scale

- Current large scale computing

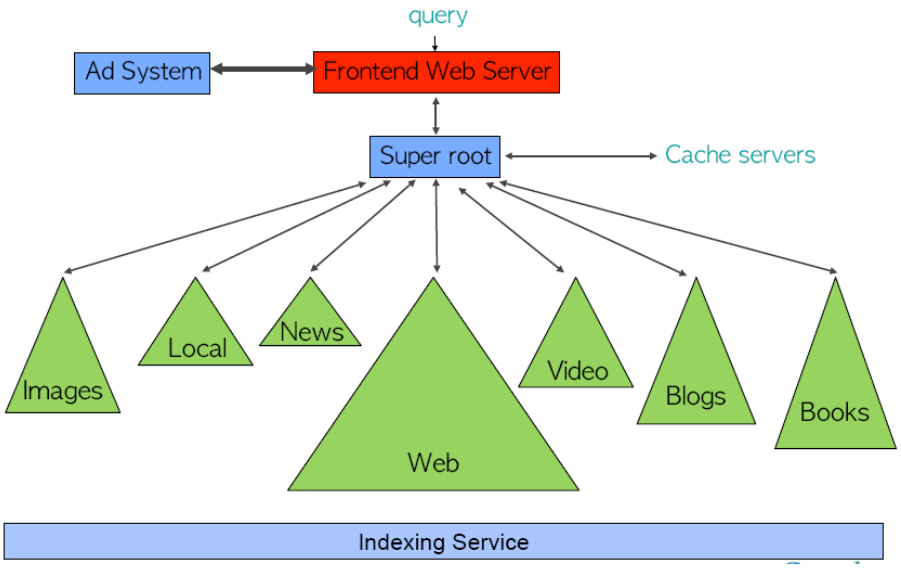
Server design 2004: Manhattan Project

Feature a 12-volt battery on each server in lieu of a centralized uninterruptible power supply.

Efficiency is raised from 92-95% for a large UPS to over 99.9% for the individual battery design.



2007 universal search



Late 2009: 'Caffeine' engine

Google Caffeine

- 11, Aug, 2009 "Google unveils new "Caffeine" search engine."
  - Released right after the partnership between Yahoo！ and Bing
  - To make the search process much faster, more intuitive and overall produce results with greater accuracy for the user.
  - The new architecture is said to include size, indexing, speed, accuracy, crawling and ranking changes.
- Some conclusions:
  - Returns results at over double the speed to the current Google.
  - Places more reliance on keyword strings rather than exact phrases.
  - Gives greater weight to authority domains and social media, e.g. blogs
  - Gives increased weight on domain names
  - Have more accurate real time results in an effort to match Facebook and Twitter's real time search engines.

Challenges II – Scale

Beyond system architecture:

- IR methods need significant rethinking
  - *idf*
    - How meaningful is idf in a heterogeneous collection of this size?
    - All words have very high idf
  - *Stemming*
    - Often used as a recall tool
    - Do you really need recall when you have 1 trillion docs?
  - *Query expansion*
    - Is it at all needed? Does it work? What about query drift?
    - How fast can you do it? What is the effect of slowdown on user?

Challenges II – Scale

Search Engine	Reported index Size	Page Depth
Google	8.1 billion (Dec. 2004)	101K
MSN	5.0 billion	150K
Yahoo!	19.2 billion (Aug. 2005)	500K
Ask Jeeves	2.5 billion	101K+

Search engine size war

ɔm Danny Sullivan, archEngineWatch b site

Challenges II – Scale, and quality

Search Engine	Reported index Size	Page Depth
Google		
MSN		
Yahoo!		
Ask Jeeves		
All the Web		
All the Surface Web		

"As in the past, we are not disclosing the size of our index for competitive reasons. That said, we believe our index is highly competitive. Search quality is comprised of a variety of factors including freshness, relevance etc. and we continue to deliver high quality results for our consumers to ensure that they are able to find the best results for what they are looking for,"
-- said spokesperson **Stephanie Iwamasa, Yahoo!.**

Search engine size war

From Danny Sullivan, SearchEngineWatch web site

2002.12

"Absolute numbers are no longer useful" -- by Google, Sep 27, 2005

- How big is the Web? – 5 billion is enough?

-- by Kenneth Church (Microsoft Research), Jan, 2008

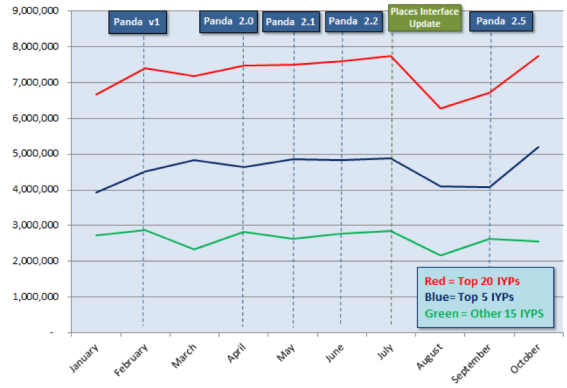
Feb. 2011 Google Panda

- First released in February 2011.
  - Rolled out about once a month for first 2 years
  - Released a "slow rollout" of Panda 4.2 starting on July 18, 2015
- The change aimed to lower the rank of "low-quality sites" or "thin sites"
- And return higher-quality sites near the top of the search results.
- The name "Panda" comes from Google engineer Navneet Panda, who developed the technology
- Impacts:
  - A surge in the rankings of news websites and social networking sites
  - A drop in rankings for sites containing large amounts of advertising.
  - Reportedly affected the rankings of ~12% of all search results
  - Complaints of scrapers/copyright infringers getting better rankings than sites with original content
- Creates a ratio with a site's inbound links and reference queries, search queries for the site's brand.
  - Sitewide modification factor and hence page modification factors

Panda Creates A Rollercoaster Ride For IYP Traffic

(IYP: Internet Yellow Page)

- A clear correlation: Google Panda vs. the traffic received by the IYPs.



- Visitors to IYPs have grown 39% between January-October
- Panda benefited larger IYPs but not smaller IYPs

This data is taken from Google AdPlanner*. It shows *average daily visitors* to the *Top 20 IYP sites within the US*.

http://searchengineland.com/how-google-panda-places-updates-created-a-rollercoaster-ride-for-iyp-traffic-101683

2015 Google's Mobilegeddon

- Mobile + Armageddon
- On Tuesday, April 21, 2015 Google mage a major update to its mobile search algorithm.
- An algorithm to favor sites that are "mobile-friendly."
- Ones with large text, easy-to-click links, and that resize to fit whatever screen they're viewed on and ranking them higher in search.
- Websites that aren't mobile-friendly will get demoted.
- Page level but not site level
- Detected by Mobile Crawler
- This will affect all mobile searches: 40 - 60% of search traffic.

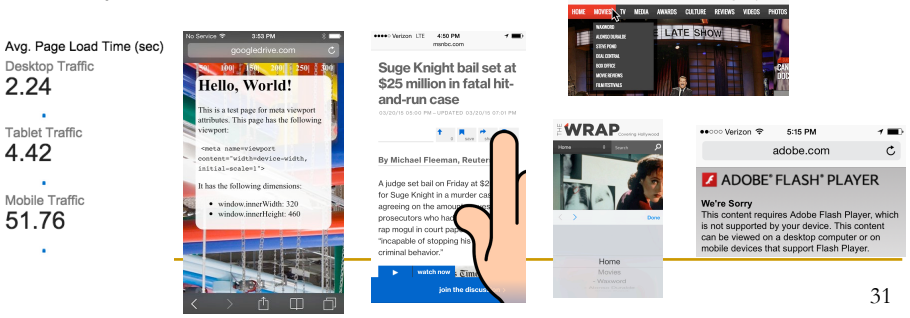
- WHY?
  - More Google searches take place on mobile devices than on computers in 10 countries including the US and Japan
  - About 60% of online traffic now comes from mobile

（http://adwords.blogspot.sg/2015/05/building-for-next-moment.html）

30

Surviving Google's Mobilegeddon

- Tip #1: Render content in <1 second on a mobile device
- Tip #2: Use a font size that's easily readable on a phone
- Tip #3: Create a "viewport" container for content, and size all content to that viewport; Fit all Text, Images, Videos to the Viewport
- Tip #4: Size and Space your Buttons For Mobile
- Tip #5: Don't Use Hover Menus on Mobile
- Tip #6: Don't use flash



31

CHALLENGES:

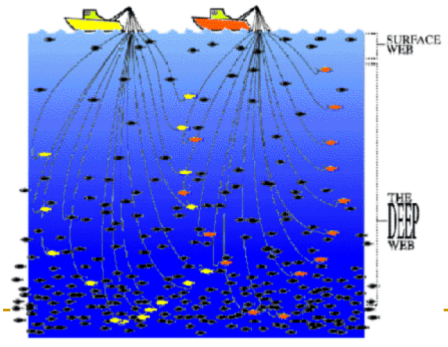
III. MULTI-SOURCES FUSION

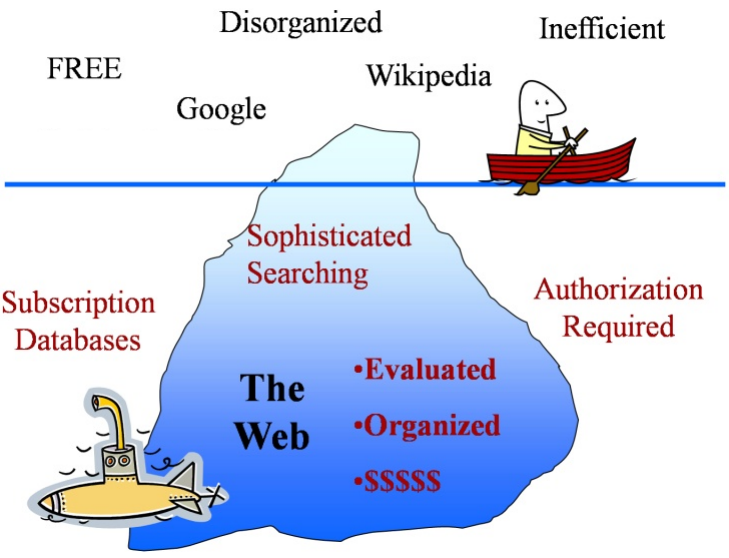
Challenges III – multi sources fusion

- Multi sources on the Web
  - Pure text file: txt, codes, etc
  - Semi-structured data: HTML, XML, emails, etc
  - Structured data: tables, databases, etc
  - Binary files: sounds, music, video, software, images …

Challenges III – multi sources fusion

- Difficulties
  - Acquisition of deep web data
    - Data: 400~500 times v.s. surface web
    - #visits: 1.5 times v.s. surface web
    - Higher quality of data (structured, manually revised)

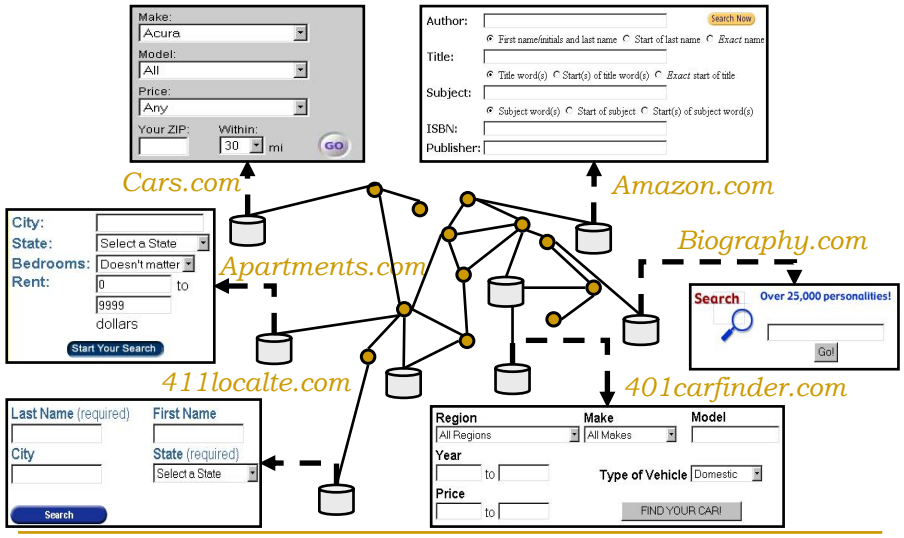




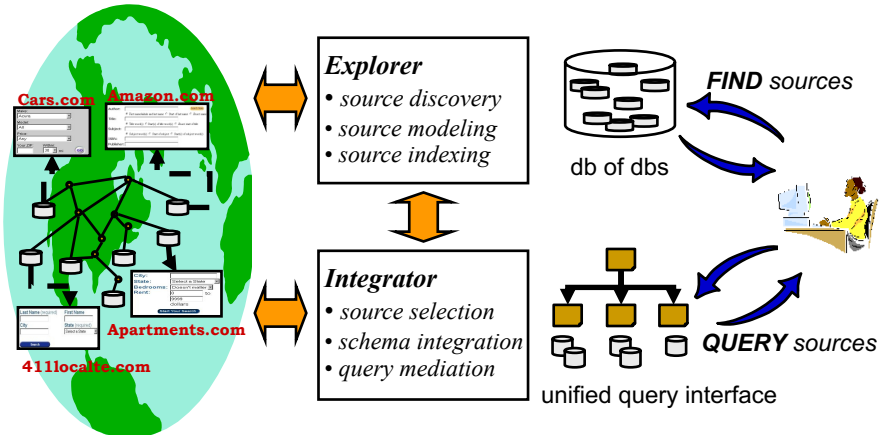
Classification of the deep web

- The Opaque Web
  - Can be crawled, but not be indexed. E.g. images, scripts, …
- The Private Web (e.g. Anti-robots)
- The Proprietary Web
  - Resources with registration info, domain/IP limitation
  - Information residing on an Intranet
- The Truly "Invisible" Web
  - Searchable Database (that require information be typed in) [airline schedules, etc]
    - Pages don't exist until requested
  - A lot of the real time data [stock quotes, sports scores, weather, election results, etc]
    - Next time same query used, a modified page may be generated

How to enable *effective access* to the deep Web?



Exploring and **integrating** deep Web



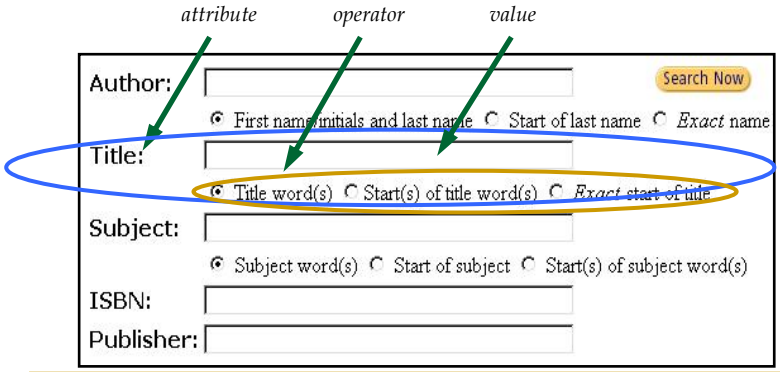
Key issues (I) – semantics

How to deal with "deep" semantics across different resources?

- How to understand a query interface?
  - Where is the first condition? What's its attribute?
- How to match query interfaces?
  - What does "author" on this source match on that?
- How to translate queries?
  - How to ask this query on that source?

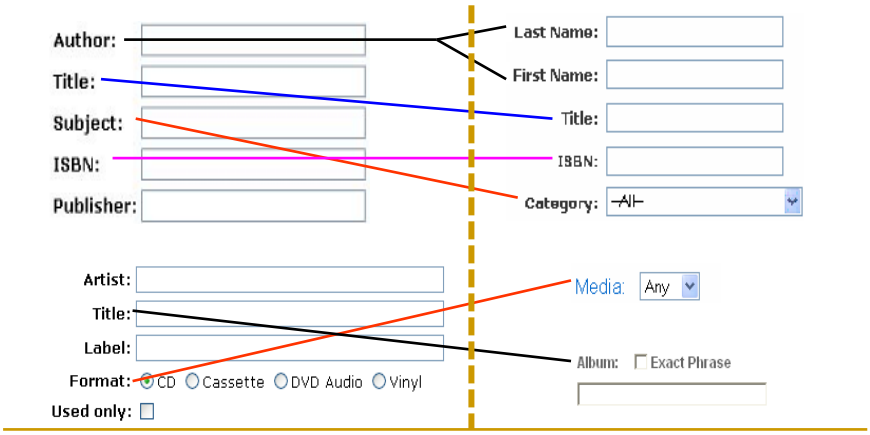
Key issues (I) – semantics(1) example: interface understanding

- Query Interface Understanding
  - Hidden-syntax parsing



Key issues (I) – semantics (2) & (3) example: interface matching and translation

- Matching Query Interfaces (Hidden-model discovery)



Key issues (II) – topic-specific searching

- Domain-based integration
  - Sources in the same domain are simpler to integrate
  - Such sources are useful to integrate
- Semi-transparent integration
  - Bring users to the right sources
  - Help users to interact as automatically as possible

An example

O 共2条，共1页，第1页

From China-pub

[刷新] [上一页] [下一页] [首页] [尾页]

外国古建筑二十讲:插图珍藏本
作者: 陈志华
市场价: ￥69.00
会员价：￥62.10(4-5星会员) ￥64.17(1-3星会员) ￥65.55(普通会员)
出版社: 三联书店 ISBN：7-108-01633-8
出版日期:2002-5-1 丛书:

加入购物车 团体购书 加入藏书阁

前言
第一讲公民的胜利
第二讲把人体美赋予建筑
第三讲面包和马戏
第四讲拱券革命
第五讲西方不亮东方亮
第六讲无情世界的感情
第七讲文明地域扩大
第八讲畸形的珍珠
……
下篇
第十六讲从自然神崇拜到皇帝崇拜
第二七讲四达之地
第十八讲宗教的象征
第十九讲敬慎者的家园
第二十讲敢于输入也勇于创造
后记
...

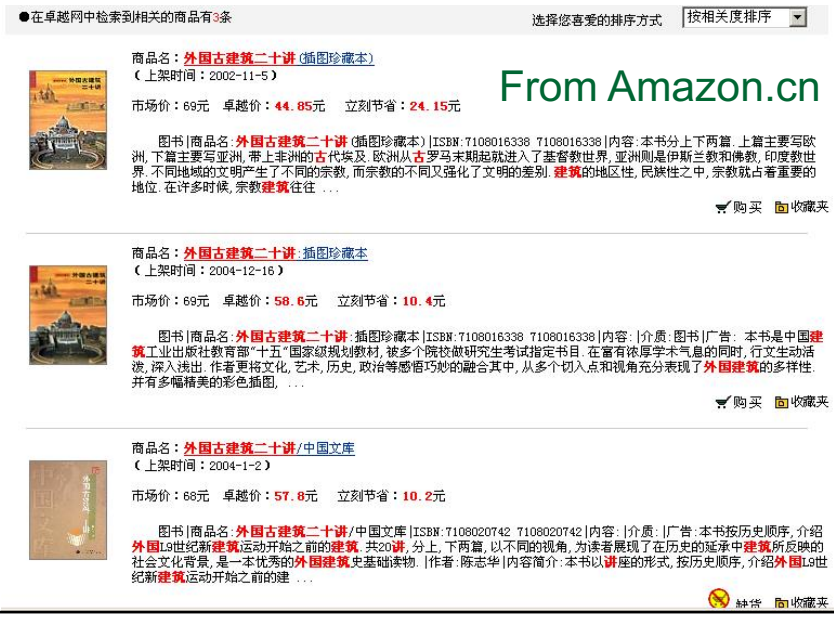
An example

外国古建筑二十讲 (插图珍藏本)
原价：~~￥69.00~~
当当价：￥53.90 购买
内容提要：本书以讲座的形式，基本按历史顺序，介如19世纪新建筑运动开始之前的外国建筑。全书共20讲，分上下两篇，不是完整的建筑史，而是择重点而述...
所属分类：图书->文化->世界文化

From dangdang

外国古建筑二十讲——中国文库
原价：~~￥68.00~~
当当价：￥57.80 购买
内容提要： 本书以讲座的形式，按历史顺序，介绍外国19世纪新建筑运动开始之前的建筑。共20讲，分上、下两篇，不是完整的建筑史，而是择重点而述...
所属分类：图书->艺术->建筑艺术

An example



From Amazon.cn

45

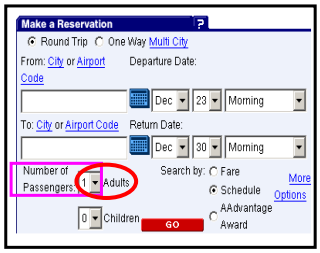
An example – Integration

- Integration
  - Name：20 lectures on Foreign Historic Buildings(Illustrated collector's Edition); 20 lectures on Foreign Historic Buildings
  - Picture

  - Author: Zhihua Chen(C)
  - Abstract
  - Price: Original ：69，68
    - Cut-off: 62.1, 64.17, 65.55(C); 53.9, 57.8(D);44.85, 58.6, 57.8(A)
  - ISBN: 7-108-01633-8(C); 7108016338(A), 7108020742(A)
  - Publisher, Publish date: San Lian Publisher，2002-05-01 (C)
  - Online date: 2002-11-5, 2004-12-16, 2004-1-2 (A)
  - Classification: Book→Culture→World Culture, Book→Art→Architectural art(D)
  - Available: out of stock (A)

46

Key issues (III) – schema matching



Result of extraction:

6	adults	Equal	1
7	children	Equal	0
8	number of passengers	Equal	1

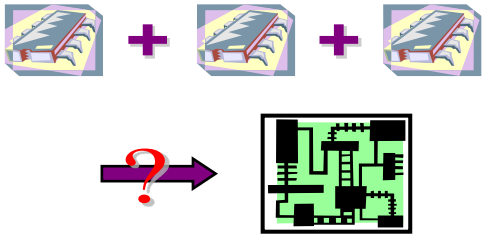
- New challenges
  - How will errors in automatic form extraction impact the subsequent schema matching?

Lessons people have learned:

System integration of an integration system is non-trivial.

"Putting together" may not be that shortest section in your paper…

"system" research often ends up with "components in isolation"



Current research idea highlight

- Manual construction/annotation
  - Easy
  - High accuracy
  - Difficult to generalization
- Automatic learning
  - Power of generalization
  - Really difficult!!
  - Poor accuracy
- Learning by active queries (annotated samples)
  - Good tradeoff between the two strategies
- Learning from user behavior (similar queries, co-clicks, similar resources)
- Assist user in identifying best sources for a given search

The "invisible" web will never disappear!