Something Interesting In WWW2007

Yiqun Liu, Min Zhang

State Key Lab of Intelligent Tech. & Sys.

Tsinghua University

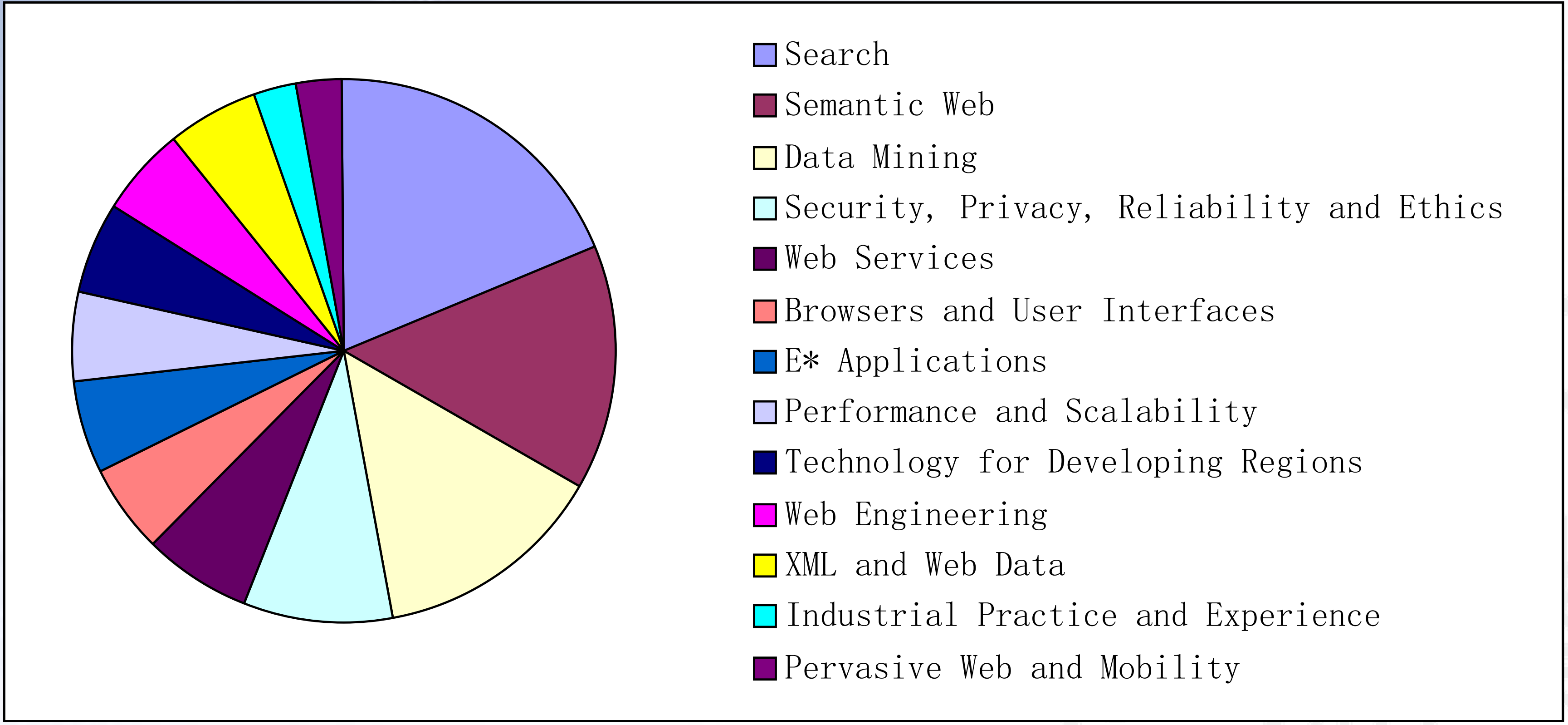
2007 / 06 / 06

General Information

- WWW2007
  - Organized by World Wide Web Consortium (W3C)
  - 12 tutorials, 8 workshops, 4 plenary speakers, 111 refereed papers, 119 posters, 7 panels, and 12 invited industry speakers.
  - The acceptance rate for refereed papers was 15%

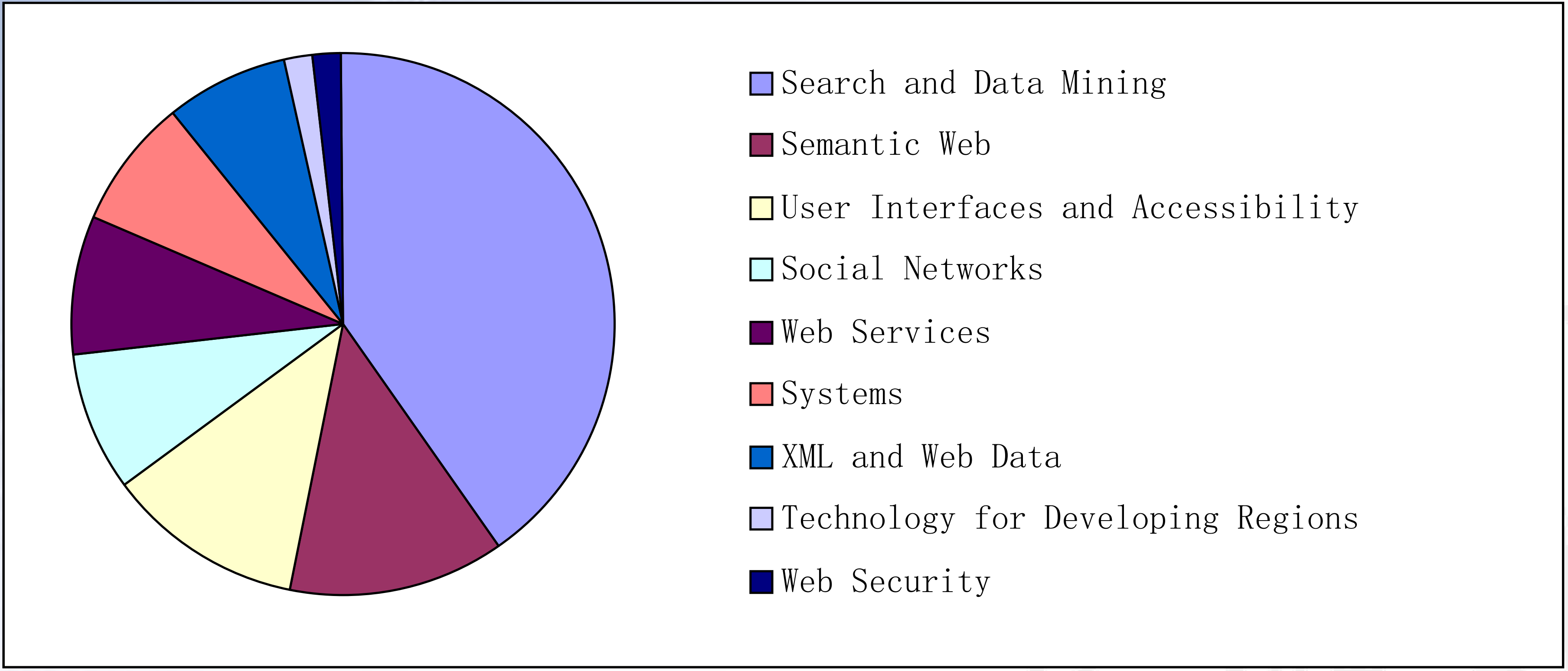
General Information

- Paper research fields



General Information

- Poster research fields



General Information

- Search, Data Mining & Semantic Web
  - Still hotspots
  - 46.85% in refereed papers
  - 52.94% in posters
  - Similar with WWW2005 (47% in refereed papers)
- User Interfaces
  - More and more attractive
- XML and Web Data
  - 12% in 2005, about 5% in 2007

General Information

- Some Interesting Topics
  - User behavior analysis
  - Evaluation issues
  - Spam identification
  - … …

User behavior analysis

- Workshop
  - *Query Logs Alone are not Enough*
- Papers
  - *Web N.0: What science will it take?*
  - *Google news personalization*
  - *Optimizing Web Search Using Social Annotations*
- Posters
  - *Identifying Ambiguous Queries in Web Search*

Query Logs Alone are not Enough

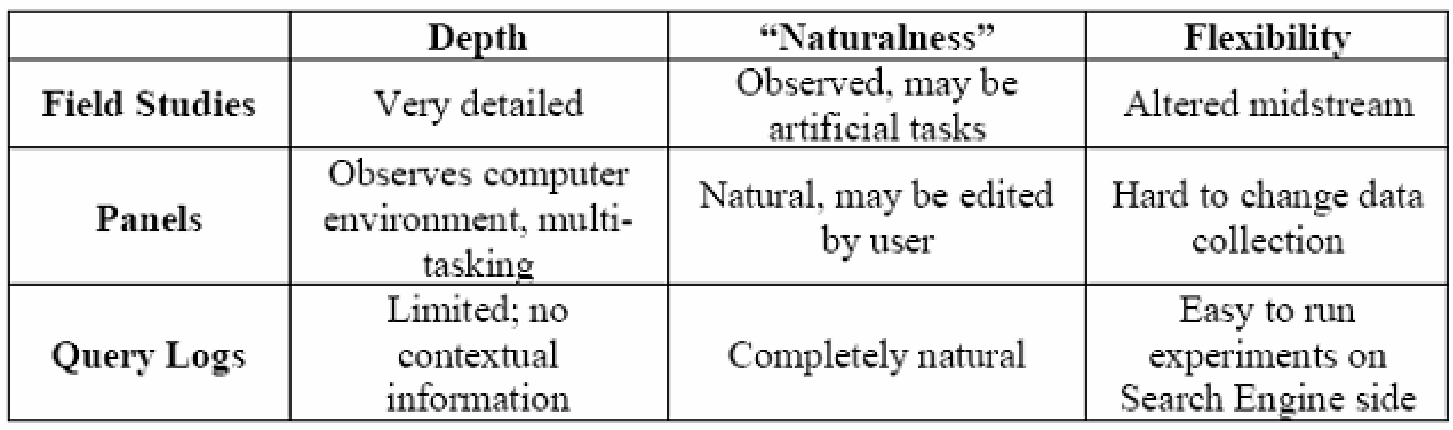
- Carrie Grimes et al, Google Inc.
- Do we really need query logs?
  - Query logs are one of the largest sources of data potentially available to a search engine.
  - Collecting query logs may cause problems
    - accidentally or intentionally released to the public
    - misused in some way internally
    - Potential threats to the privacy of the users
    - “AOL search-query scandal”: CTO Maureen Govern has decided to leave the company immediately.

Query Logs Alone are not Enough

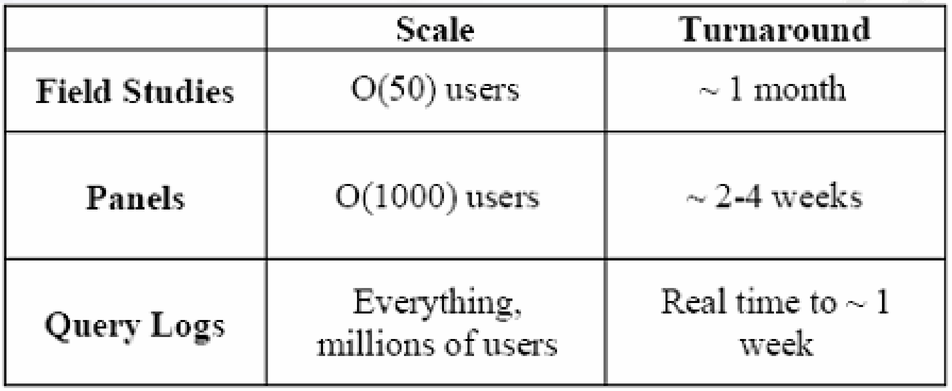
- In order to “understand” the intent of the search query, there are several information sources
  - Supervised lab studies (one on one interaction)
  - Instrumentation or other passive observation technology (web based application or software sending back data)
  - Query logs created by user transactions
- Conclusion: Query logs contain unique data that can significantly improve search engine performance

Query Logs Alone are not Enough

- Comparison



	Depth	"Naturalness"	Flexibility
Field Studies	Very detailed	Observed, may be artificial tasks	Altered midstream
Panels	Observes computer environment, multi-tasking	Natural, may be edited by user	Hard to change data collection
Query Logs	Limited; no contextual information	Completely natural	Easy to run experiments on Search Engine side



	Scale	Turnaround
Field Studies	O(50) users	~ 1 month
Panels	O(1000) users	~ 2-4 weeks
Query Logs	Everything, millions of users	Real time to ~ 1 week

Query Logs Alone are not Enough

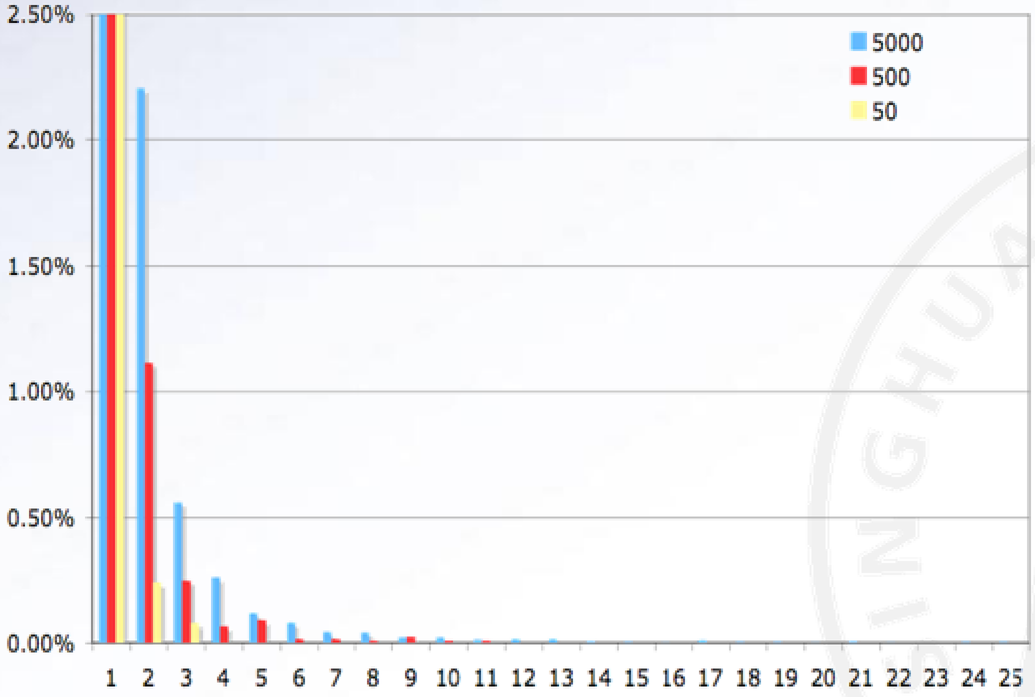
- Trade off between scale and detail
  - Loss of detail
    - Why a user does something, How to label the data
  - Higher scale => more noise / more diversity
- Trade off between automation and timeliness
  - Field / lab studies: manual, rarely longitudinal.
  - Query logs and instrumented panels: capture longitudinal trends

Query Logs Alone are not Enough

- Advantage of query logs
  - Scalable and easy to obtain
  - a diversity of tasks, queries and user experiences
  - measure users in the wild
- Disadvantage of query logs
  - only measure the how and the what, rather than the why
  - completely unlabeled except for the presence/absence of an event.
  - only measure the system being logged (20% search activities)
  - noisy, including robots, spam, data outages, recording errors, etc
  - don’t necessarily allow long-term studies of a single user

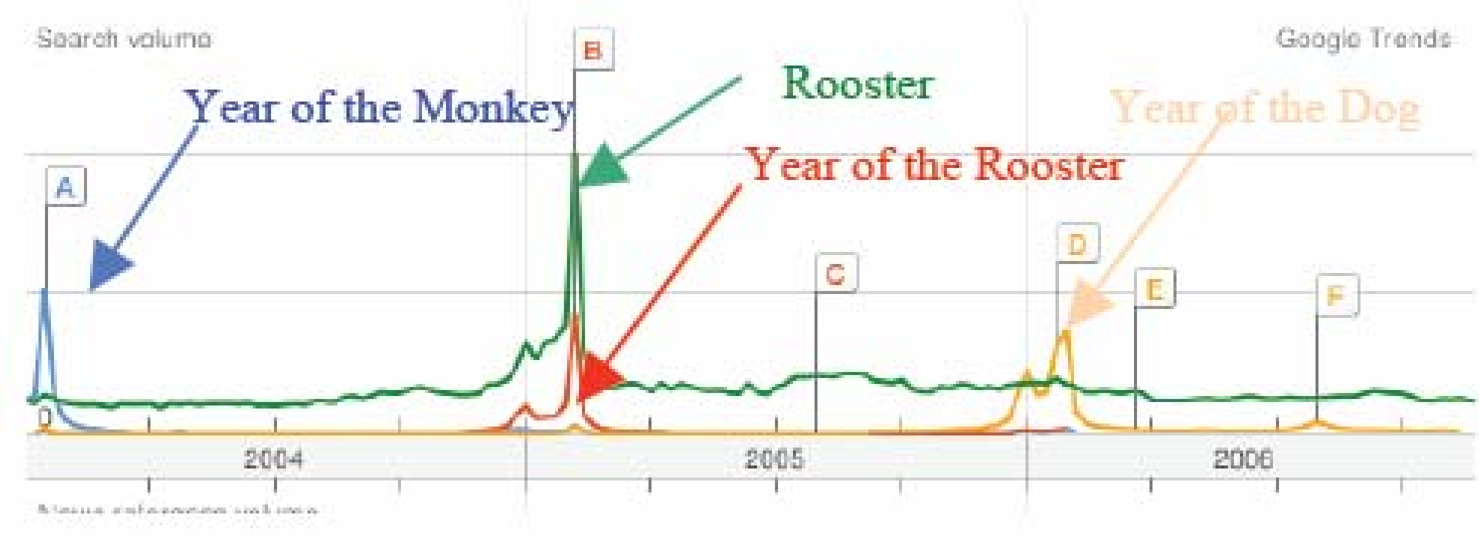
Query Logs Alone are not Enough

- Diversity of queries in search engine logs
  - a small- or medium-scale study would not approach the diversity observable from large scale query logs



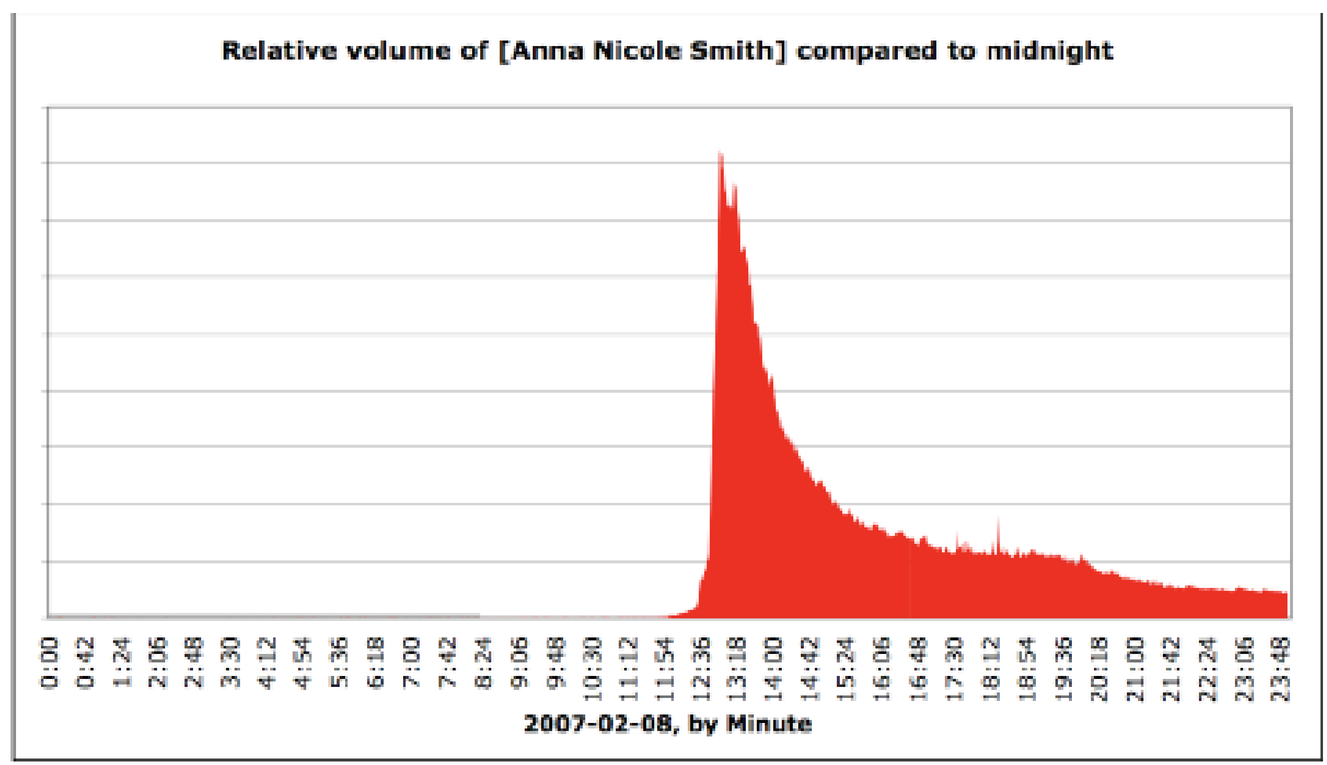
Query Logs Alone are not Enough

- Disambiguating Queries
  - Making use of related queries “year of rooster” and “rooster”



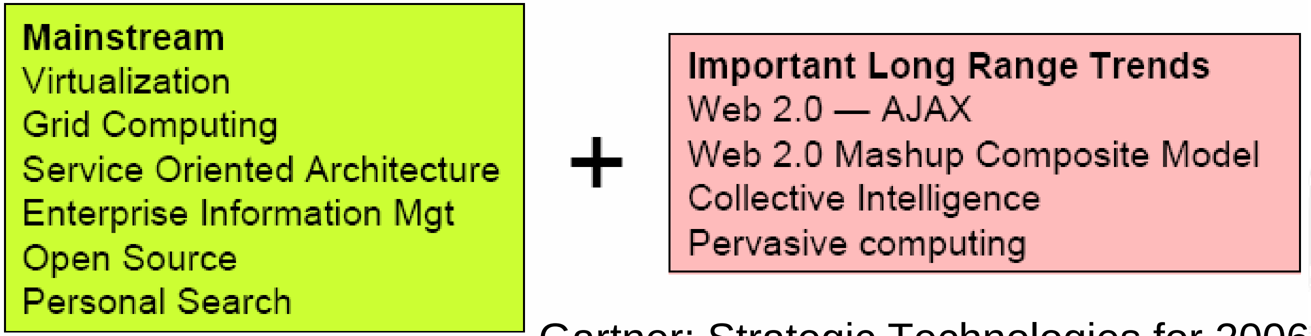
Query Logs Alone are not Enough

- Immediacy of Data
  - The death of “Anna Nicole Smith”



Web N.0: What science will it take?

- Prabhakar Raghavan, Yahoo! Research
  - Web 2.0的出现为WWW带来了深远的变化
  - Web的使用体验从单纯的人机交互转变为一个社会行为
  - 逐渐淡化计算机科学与社会科学（微观经济学、心理学、社会学）的界限



Mainstream
Virtualization
Grid Computing
Service Oriented Architecture
Enterprise Information Mgt
Open Source
Personal Search

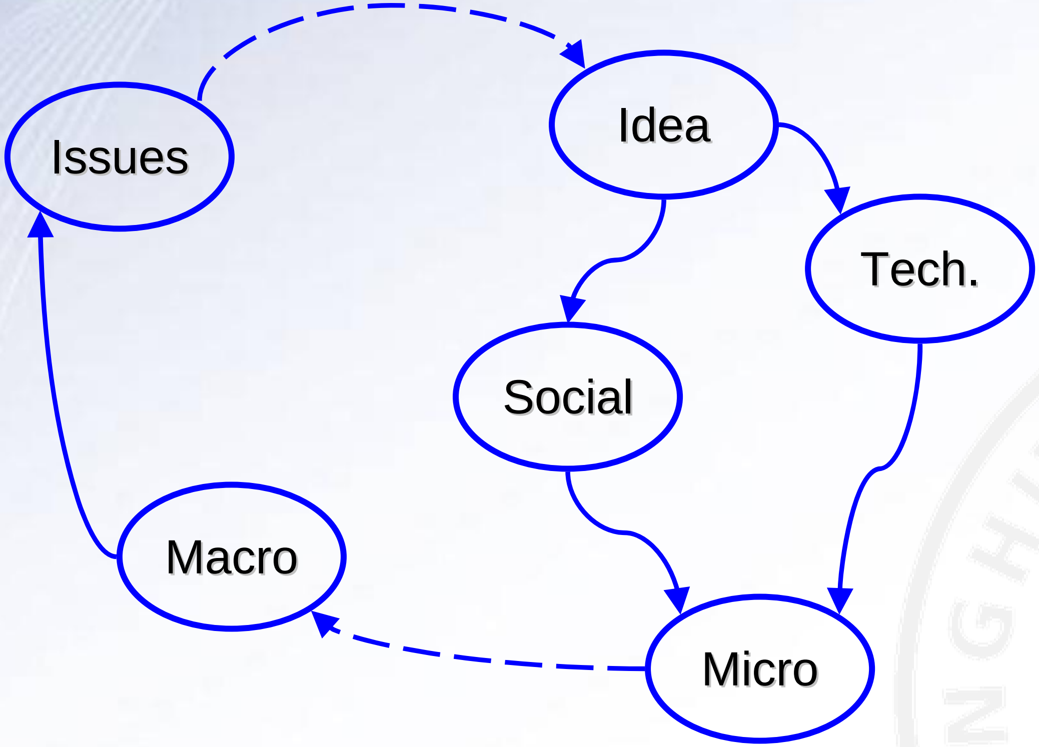
+

Important Long Range Trends
Web 2.0 — AJAX
Web 2.0 Mashup Composite Model
Collective Intelligence
Pervasive computing

Gartner: Strategic Technologies for 2006 and 2016

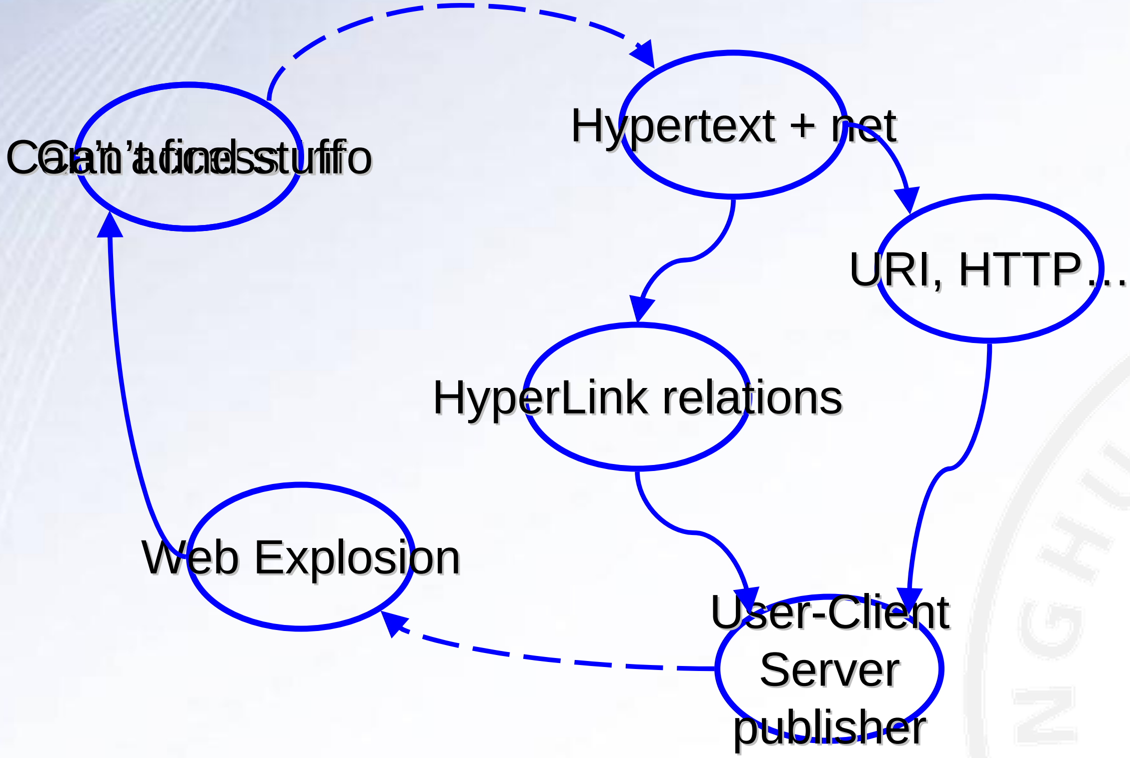
Web N.0: What science will it take?

- 问题产生-> 发展-> 解决-> 新问题产生



Web N.0: What science will it take?

- Cannot access Information



Web N.0: What science will it take?

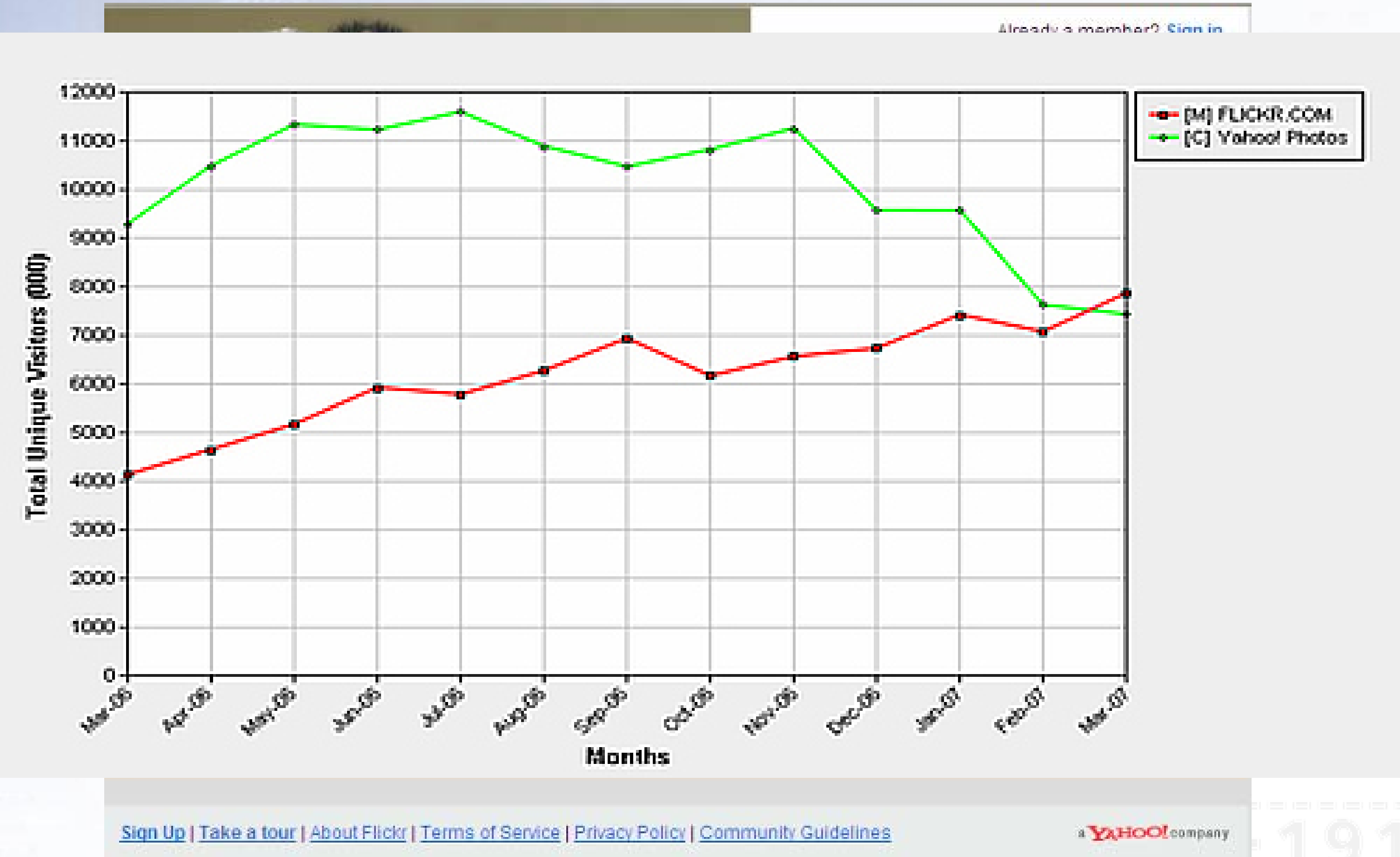
- Similar issues
  - – Can’t find stuff, Can’t write stuff, Can’t reuse Web data...
- Growth of content
  - – Published content: 3-4 Gb/day
  - – Professional Web content: 2Gb/day
  - – User generated content: 5-10Gb/day
  - – Private text content: 2Tb/day
  - – Upper bound on typed content: 140Tb/day

Web N.0: What science will it take?

- Estimate growth of metadata
  - Anchor text: 100Mb/day
  - Tags: 100Mb/day
- The power of social tagging
  - The wisdom of crowd can be used to search
  - The principle is not new: “anchor text” search
  - www.filckr.com

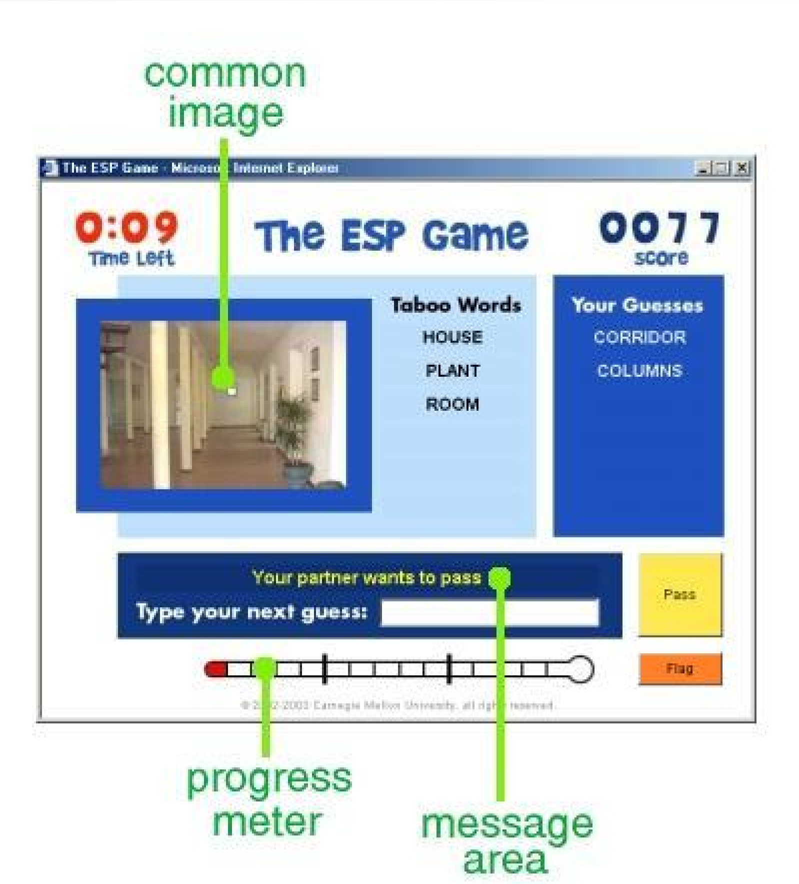
Web N.0: What science will it take?

- www.flickr.com (run by Yahoo!)



Web N.0: What science will it take?

- Challenge in tag-based search
  - How do we use the tags bette
  - How do we cope with spam (
- Where else can we explore
  - What are the incentive mecha
  - ESP game by CMU



0:09 Time Left

The ESP Game

0077 score

Taboo Words
HOUSE
PLANT
ROOM

Your Guesses
CORRIDOR
COLUMNS

Your partner wants to pass

Type your next guess:

© 2002-2003 Carnegie Mellon University, all rights reserved.

Web N.0: What science will it take?

- Other issues in Web science
  - – The science of online audience engagement
    - People interacting with people
    - Why people participate / creat
  - – How to measure audience engagement
    - For online advertising

$$\sum_{pageviews} repeat \times (time\_spent)^{\alpha} \times \log(user\_neighborhood)$$

Google news personalization

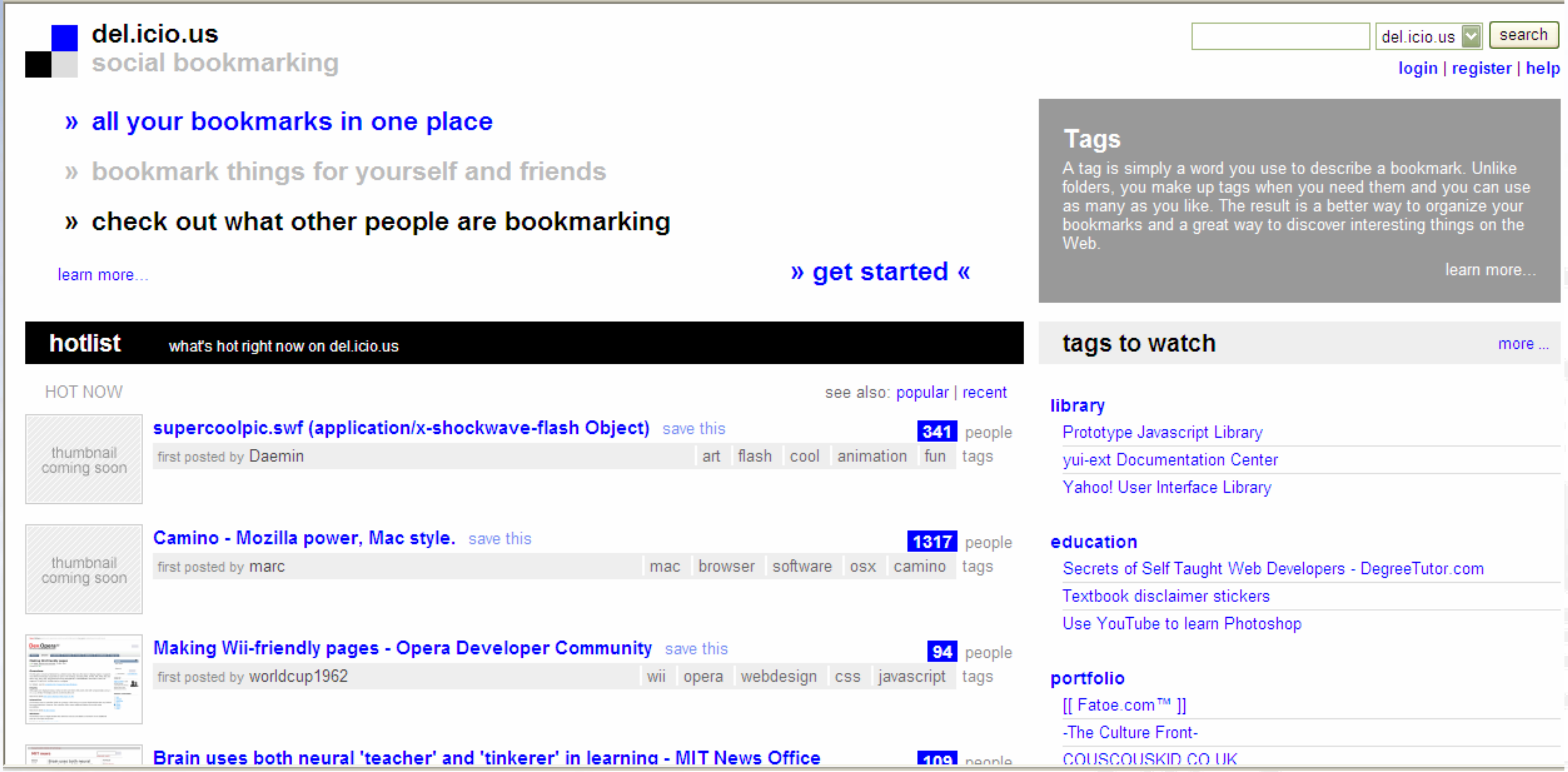
- Abhinandan Das, Google Inc.
- News recommendation system
- Challenges
  - Scale
    - Number of unique visitors in last 2 months, several million
    - Number of stories within last 2 months, several million
  - Item churn
    - News story changes every 10-15 min
  - Noisy ratings
    - Click treated as noisy positive vote

Google news personalization

- A method combined with Content-based and Collaborative filtering algorithms
  - User clustering algorithm
    - Minhash
    - Probabilistic Latent Semantic Indexing (PLSI)
  - Story-story co-visitation
    - for each story, store the co-visitation counts with other stories
  - Map-Reduce parallel frameworks

Optimizing Web Search Using Social Annotations

- Shenghua Bao (Shanghai Jiao Tong University)



del.icio.us
social bookmarking

del.icio.us search
login | register | help

» all your bookmarks in one place

» bookmark things for yourself and friends

» check out what other people are bookmarking

learn more...

» get started «

Tags

A tag is simply a word you use to describe a bookmark. Unlike folders, you make up tags when you need them and you can use as many as you like. The result is a better way to organize your bookmarks and a great way to discover interesting things on the Web.

learn more...

hotlist what's hot right now on del.icio.us

HOT NOW

see also: popular | recent

thumbnail coming soon
supercoolpic.swf (application/x-shockwave-flash Object) save this — 341 people
first posted by Daemin — art flash cool animation fun tags

thumbnail coming soon
Camino - Mozilla power, Mac style. save this — 1317 people
first posted by marc — mac browser software osx camino tags

Making Wii-friendly pages - Opera Developer Community save this — 94 people
first posted by worldcup1962 — wii opera webdesign css javascript tags

Brain uses both neural 'teacher' and 'tinkerer' in learning - MIT News Office — 109 people

tags to watch more ...

library
Prototype Javascript Library
yui-ext Documentation Center
Yahoo! User Interface Library

education
Secrets of Self Taught Web Developers - DegreeTutor.com
Textbook disclaimer stickers
Use YouTube to learn Photoshop

portfolio
[[Fatoe.com™]]
-The Culture Front-
COUSCOUSKID.CO.UK

Optimizing Web Search Using Social Annotations

- SocialSimRank (SSR)
  - calculates the similarity between social annotations and web queries;
  - to find the latent semantic association between queries and annotations
- SocialPageRank (SPR)
  - captures the popularity of web pages.
  - to measure the quality (popularity) of a web page from the web users' perspective

Identifying Ambiguous Queries in Web Search

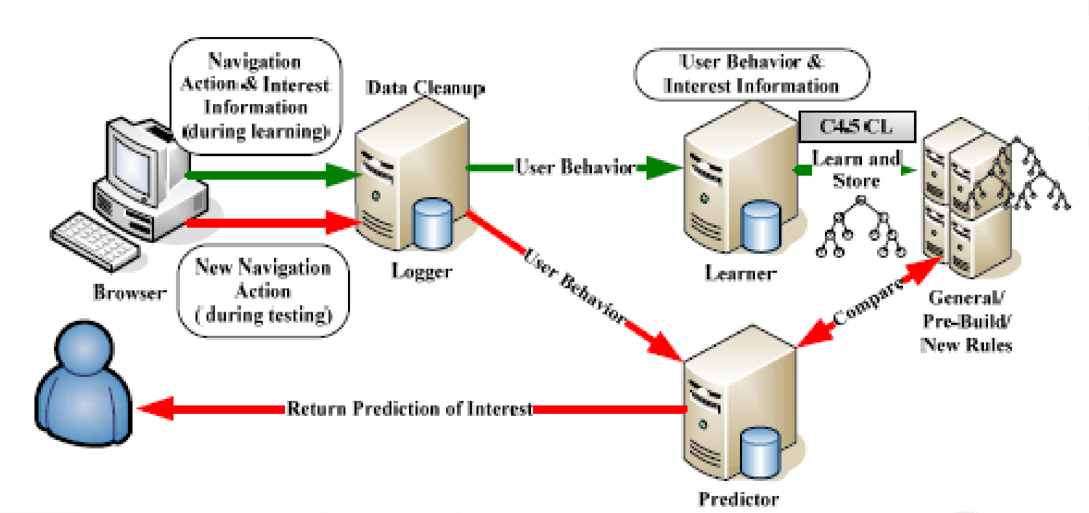
- Ruihua Song, MSRA and Shanghai Jiaotong Univeristy
- Taxonomy of Web search queries
  - Ambiguous queries: Apple, Gaint
  - Broad queries: Songs
  - Clear queries: Tsinghua University
  - 90% users agree with whether a query is ambiguous
  - Only 50% agree with whether it is “broad” or “clear”

Identifying Ambiguous Queries in Web Search

- Use Web search results to identify ambiguous queries
  - Classify result document D into categories defined by KDDcup 2005
  - an ambiguous query is that relevant documents probably belong to several different categories.
  - 12 features are derived to quantify the distribution of *D* ,
  - precision of 85.4%, recall of 80.9%
  - about 16% of all the queries are ambiguous

Can We Find Common Rules of Browsing Behavior

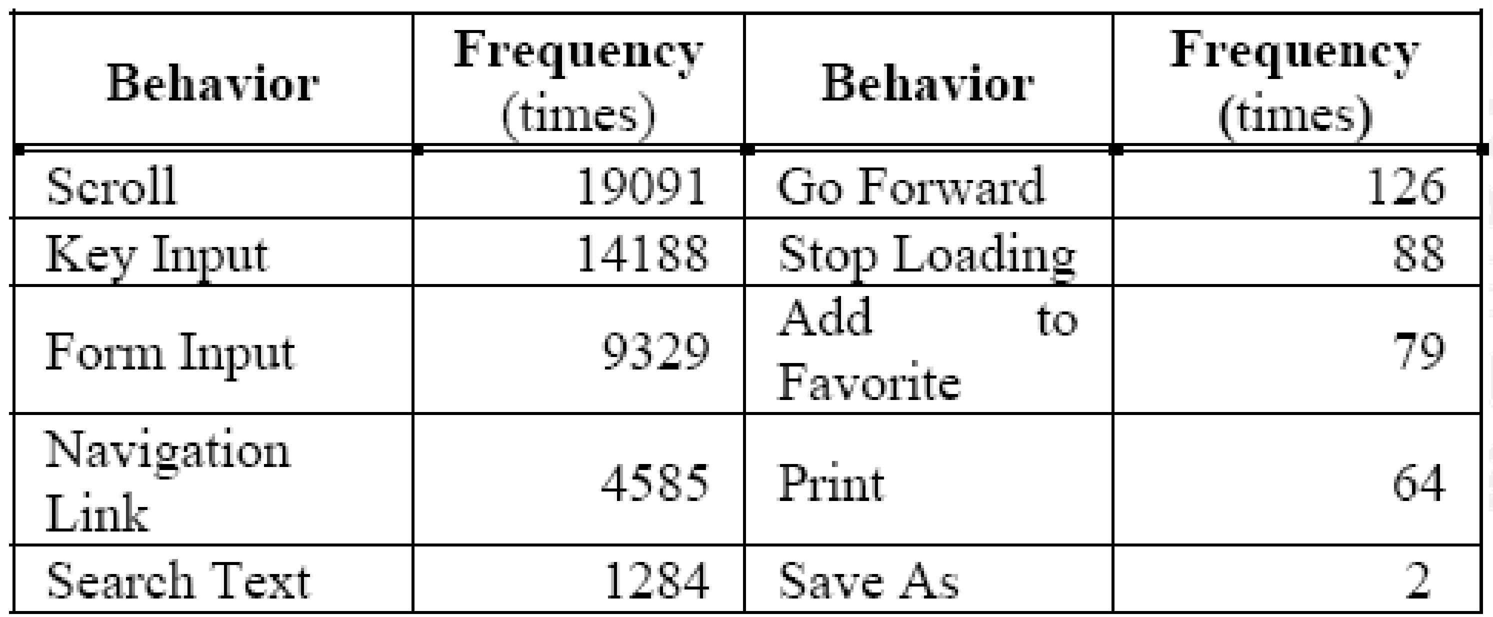
- Ganesan Velayathan, National Institute of Informatics, Japan
- Investigate factors in user's browsing behavior to automatically evaluate web pages that the user shows interest in.
- A client-side logging/analyzing tool: the GINIS Framework



Navigation Action & Interest Information (during learning)
Data Cleanup
User Behavior & Interest Information
C4.5 CL
Learn and Store
User Behavior
Browser
New Navigation Action (during testing)
Logger
Learner
User Behavior
Compare
General/ Pre-Build/ New Rules
Return Prediction of Interest
Predictor

Can We Find Common Rules of Browsing Behavior

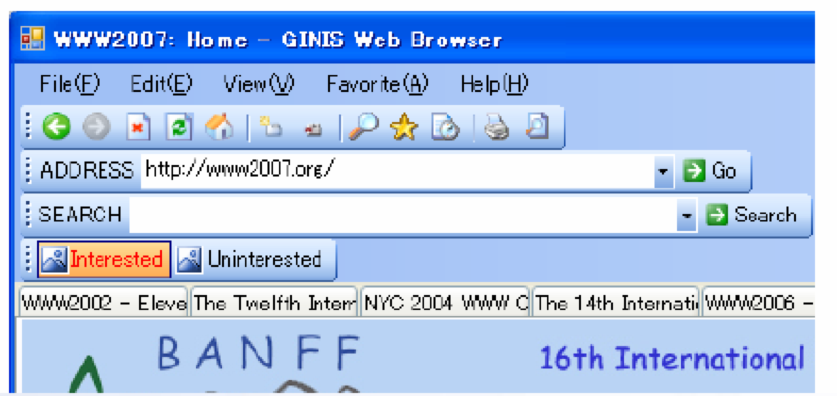
- Behavior Logging: 5 most and 5 least
  - Over 70 navigation actions and around 40 user behaviors were logged during this experiment.



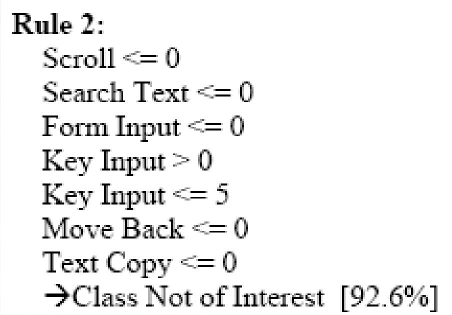
Behavior	Frequency (times)	Behavior	Frequency (times)
Scroll	19091	Go Forward	126
Key Input	14188	Stop Loading	88
Form Input	9329	Add to Favorite	79
Navigation Link	4585	Print	64
Search Text	1284	Save As	2

Can We Find Common Rules of Browsing Behavior

- Page Tagging

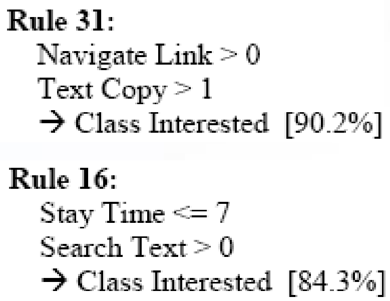


- Several rules found with C4.5



Rule 2:

Scroll <= 0
Search Text <= 0
Form Input <= 0
Key Input > 0
Key Input <= 5
Move Back <= 0
Text Copy <= 0
→Class Not of Interest [92.6%]



Rule 31:

Navigate Link > 0
Text Copy > 1
→ Class Interested [90.2%]

Rule 16:

Stay Time <= 7
Search Text > 0
→ Class Interested [84.3%]

User behavior analysis

- Also in this issue
  - – Comparing Click Logs and Editorial Labels for Training Query Rewriting
  - – Functional Faceted Web Query Analysis
  - – A Study of Mobile Search Queries in Japan
  - – …

Evaluation issues

- Workshop
  - *Web Search Engine Evaluation Using Clickthrough Data and a User Model*
- Papers
  - *Efficient Search Engine Measurements*
  - *The Discoverability of the Web*

Web Search Engine Evaluation Using Clickthrough Data and a User Model

- Georges Dupret (Yahoo! Research)
- How to rerank results with the help of click-through data?
- Click-through Information
  - Do have some information, but not reliable
  - Consideration: The user saw the document snippet
  - Attractivity: The document snippet is attractive

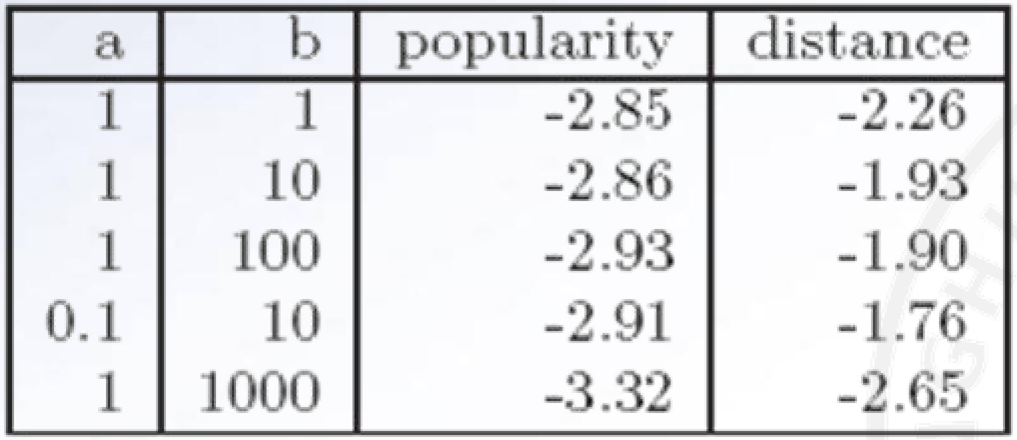
Web Search Engine Evaluation Using Clickthrough Data and a User Model

- Model and hypothesis:
  - users browse the result list sequentially
  - users select documents because they are considered and attractive
  - Attractivity depends on document snippet u and query q
  - Consideration depends on the distance d to the last selection

$$\mathrm{P}(\mathbf{s},\mathbf{a},\mathbf{c}|u,q,d) \quad = \quad \mathrm{P}(\mathbf{c}|d)\mathrm{P}(\mathbf{a}|u,q)$$

Web Search Engine Evaluation Using Clickthrough Data and a User Model

- The “popularity” model
  - Perseverence is constant (distance is not considered)
  - Worse than the distance model



a	b	popularity	distance
1	1	-2.85	-2.26
1	10	-2.86	-1.93
1	100	-2.93	-1.90
0.1	10	-2.91	-1.76
1	1000	-3.32	-2.65

Web Search Engine Evaluation Using Clickthrough Data and a User Model

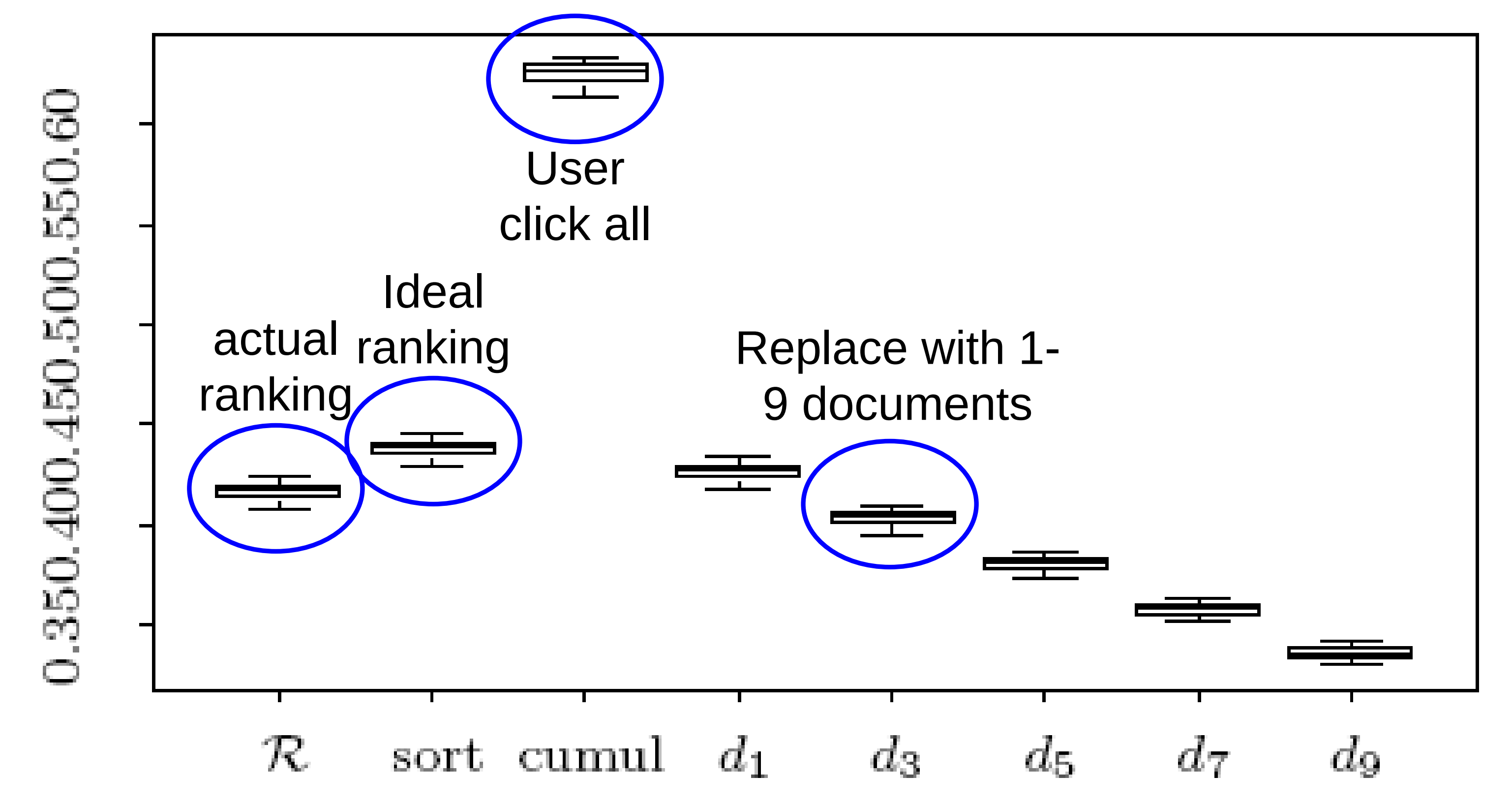
- Search engine evaluation
  - – The expected number of attractive documents that a user would see

$$\mathcal{R} = \sum_{q} \mathrm{P}(q) \sum_{o} \mathrm{P}(o|q) \sum_{\sigma} \mathrm{P}(\sigma|o,q)a(\sigma,o,q)$$

  - – o : a specific ordering
  - – σ : a sequence of selections in a result list

Web Search Engine Evaluation Using Clickthrough Data and a User Model

- Experiment results (with beta distribution as prior)



User click all

Ideal ranking

actual ranking

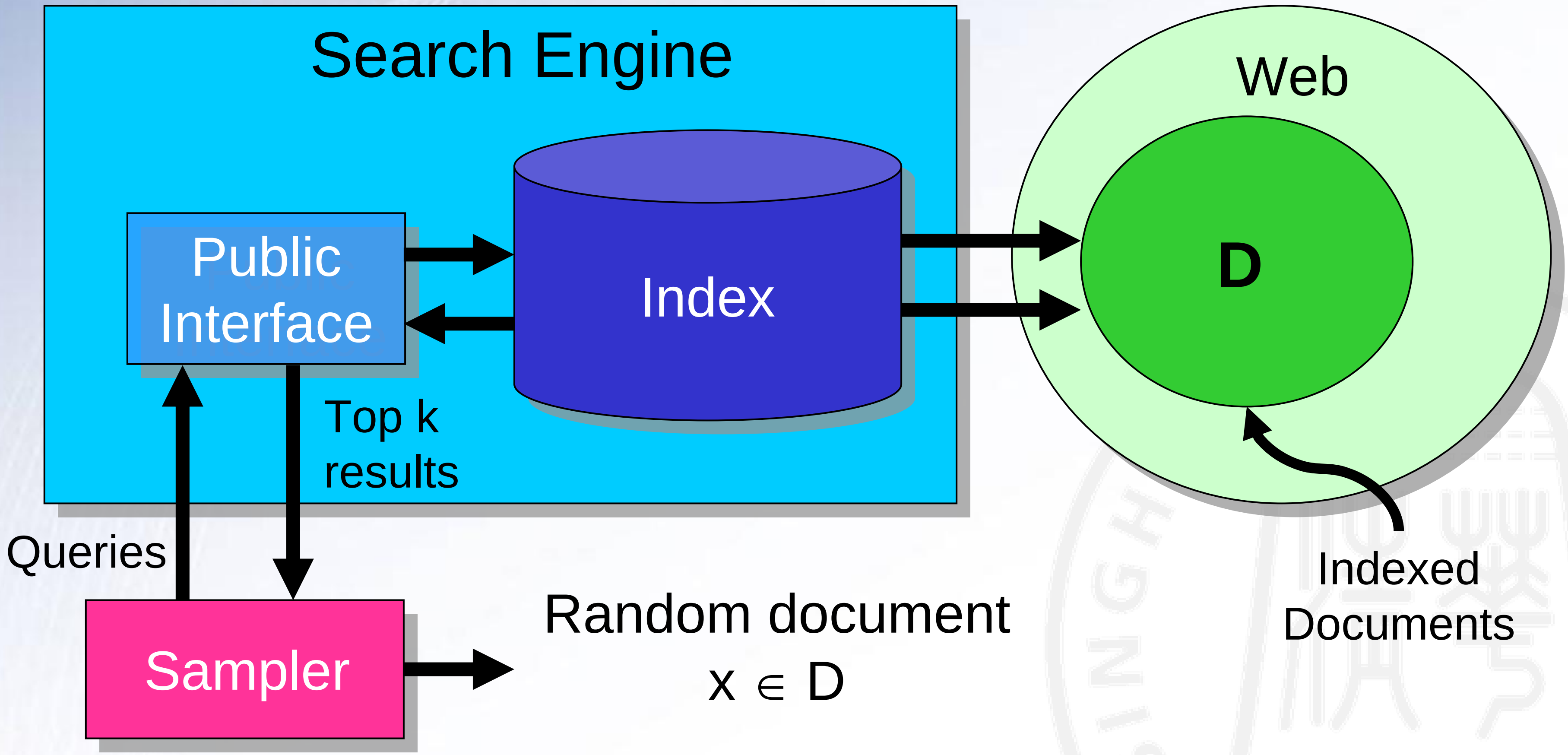
Replace with 1-9 documents

Efficient Search Engine Measurement

- Ziv BarYossef, Israel Institute of Technology and Google Haifa Engineering Center
- Estimate search engine from an objective, transparent way
  - corpus size
  - index freshness: average age of pages
  - number of unique pages: density of duplicates in the corpus
- Following the work of “Random Sampling from a Search Engine’s Index” (WWW2006 best paper)

Efficient Search Engine Measurement

- Search engine sampler



Random document

$x \in D$

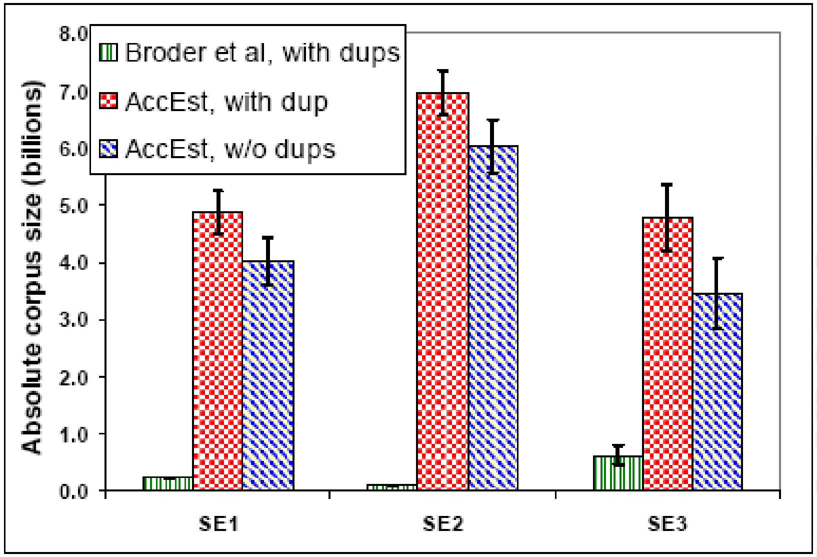
Estimate size of D

Efficient Search Engine Measurement

- Based on the computation of these two factors
  - Degree (q): how many results?
  - Degree (X): how many queries have such result?
- Problems
  - Cannot always compute query degrees (because of query overflow)
  - Cannot compute document degrees accurately (there may be q belonging to queries(x) that do not occur in x)(there may be q that occurs in x, but not included in queries(x) because of query overflow)

Efficient Search Engine Measurement

- Contributions
  - two new estimators that are able to overcome the bias introduced by approximate degrees.
  - Estimate with importance factors



Also in this issue…

- Papers
  - The Discoverability of the Web
- Posters
  - *Summary Attributes and Perceived Search Quality*
  - *Search Engine Retrieval of Changing Information*
  - *Behavior Based Web Page Evaluation*

Spam identification

- Workshop
  - *A Taxonomy of JavaScript Redirection Spam*
  - *A Large-Scale Study of Link Spam Detection by Graph Algorithms*
- Papers
  - *Spam Double-Funnel: Connecting Web Spammers with Advertisers*
- Posters
  - *Review Spam Detection*

Other issues

- All papers, posters, workshops and tutorials are available at http://www2007.org/proceedings.html
- All notes are available at ftp://166.111.138.76/incoming/paper/WWW07/
- WWW2008 in our city!

Thank you!

Questions or comments?