EECS 395/495: Web Information Retrieval and Extraction

Spring 2011

Outline

- Introductions
- Course goals and logistics
- Why take this class?
- What is Web Search?
  - What will it be in five years? Ten years?

Introductions

- Professor: Doug Downey
- TA: Vaibhav Rastogi

Goals

- Learn:
  - How does Web search work?
  - What's the future of Web search?
- Learn how to:
  - Analyze, discuss, and present research papers
  - Do projects at the frontier of Web search and information extraction

Logistics

- Tuesdays: lectures
- Thursdays: debates on research papers
  - Offense and Defense (2 on each side)
    - Defense (in favor of paper) starts with a 10 minute presentation on the key ideas
      - Then, offense/defense debate, with participation from all
      - Debaters graded by prof, TA, and peers
    - See Web page and sign up (e-mail TA) this week

Logistics

- Grading
  - Participation (30%)
    - During lectures/discussions (5%)
    - Leading a debate (15%)
    - Synopses of papers (10%)
  - Projects (70%)
    - Project grade, based on individual contribution (60%)
    - Review of another team's project (10%)

Paper synopses

- About two paragraphs, due each Thursday
  - Incl. this Thursday – paper TBA this evening
  - Turn-in via e-mail to prof and TA
- Include:
  - Brief summary of what the paper says and its main contributions
  - One potential flaw in the paper
  - One suggested “next step” for the paper

Projects

- Groups of 2-4 *(not necessarily = debate groups)*
- Examples:
  - Read 3-5 recent research papers and summarize the state of the art in some area of IR/IE
  - Implement an IR/IE system and report on the results
  - Answer theoretical questions
  - Etc.

Specific Examples

- Systems
  - Create a search “knob” for specifying reading level
  - Execute relevant background searches as I write an e-mail
  - Automatically monitor bias in search engines
- Theoretical questions
  - Read a paper on search ad auctions or PageRank and attempt to extend the results
- Summary of Research
  - Read 5 papers on automated question answering; summarize the field and suggest future directions

Project Milestones

- **April 11**: Proposal (~1 page)
  - Meetings with me to finalize **April 12/13**

…lots of progress…

- **May 9**: Report preliminary results (~2 pages)
  - **May 11**: Review group provides feedback (~1 page)
- **June 1**: Final Report (~4 pages)
- **June 2/June 8**: Final Presentations
  (~8 mins + 3 min for Q&A)

Outline

- Introductions
- Course goals and logistics
- **Why take this class?**
- What is Web Search?
  - What will it be in five years? Ten years?

Who cares about search?

- "The most important application for the foreseeable future...is search.“
  - Steve Ballmer, CEO of Microsoft
    Financial Times http://news.cnet.com/8301-10784_3-9973650-7.html
- Why?
  - Search’s utility scales as the Web scales
  - People use it all the time
  - Control
  - Profit

Why should *you* care about search?

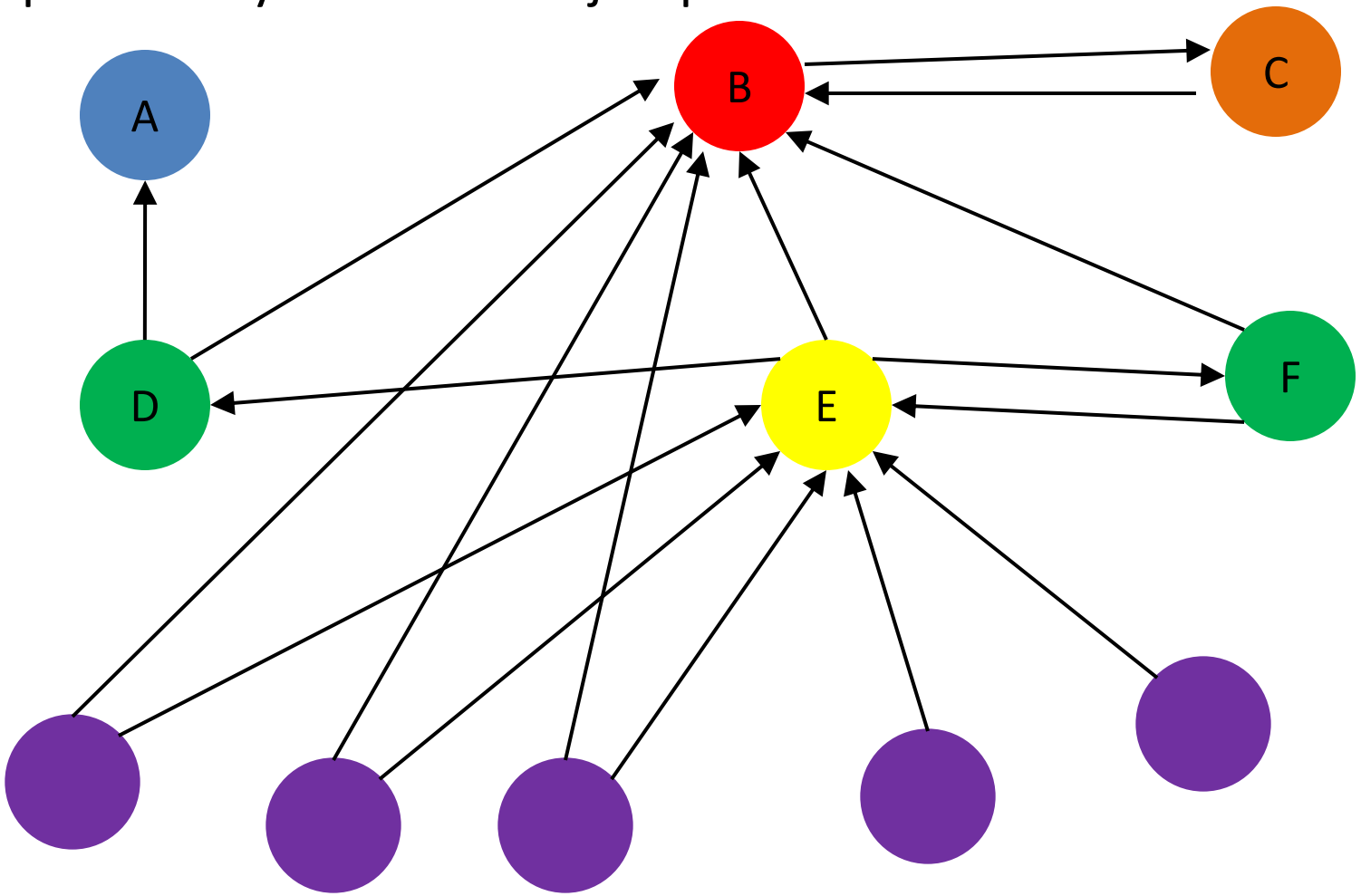
- Opportunity
- Fascinating & important enabling technologies:
  - Scaling
  - Machine Learning/Data Mining
  - Graph-based algorithms 
  - Language Understanding
  - Auction theory
  - User Interfaces
  - …

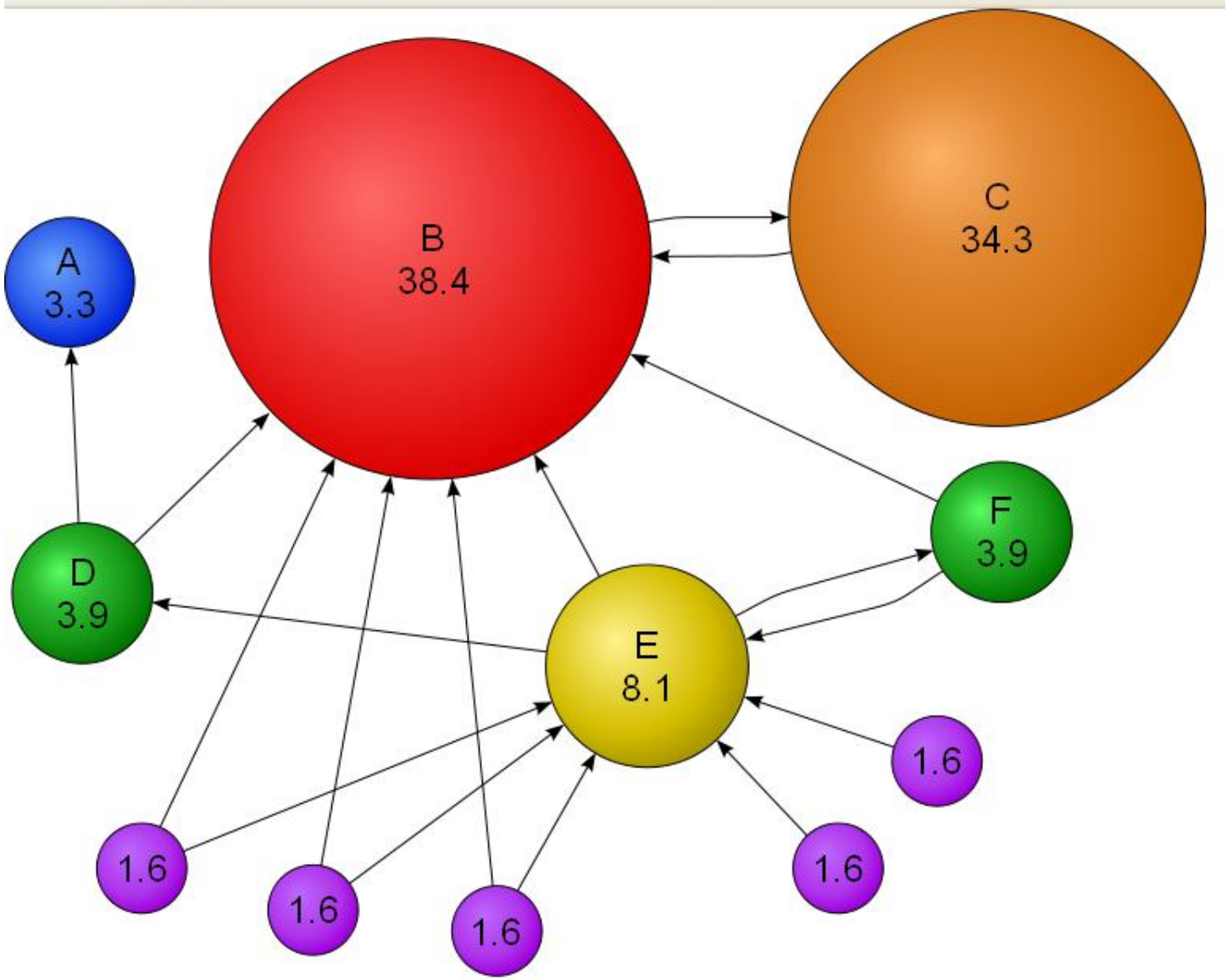
Graph-based Algorithms

- Example: PageRank
  - Google's original claim to fame
  - Idea: Quality of *p* is proportional to the aggregate quality of the pages linking to *p*
- PageRank(*p*) = probability that a "random surfer" lands on *p*
  - Pick a starting page at random
  - Follow links uniformly at random
    - Every now and then, jump to a random page
  - PageRank(*p*) = proportion of visits to *p*

PageRank example

15% probability of a random jump





Wikipedia.com

How to compute PageRank

- Simulate a random surfer?
  - On 20 billion pages and 400 billion hyperlinks…
- It can be done
  - You'll learn how
- What about:
  - Personalized PageRank? Link spam?

What do you need to take this class?

- EECS 311
  - basic understanding of algorithms and data structures
- Helpful:
  - Basics of linear algebra and probability theory
- Willingness to participate

Outline

- Introductions
- Course goals and logistics
- Why take this class?
- **What is Web Search?**
  - **What will it be in five years? Ten years?**

Web Search today…

- Performs an **easy** task
- Extremely **quickly**
- At massive **scale**
- Relatively **well**

Easy?

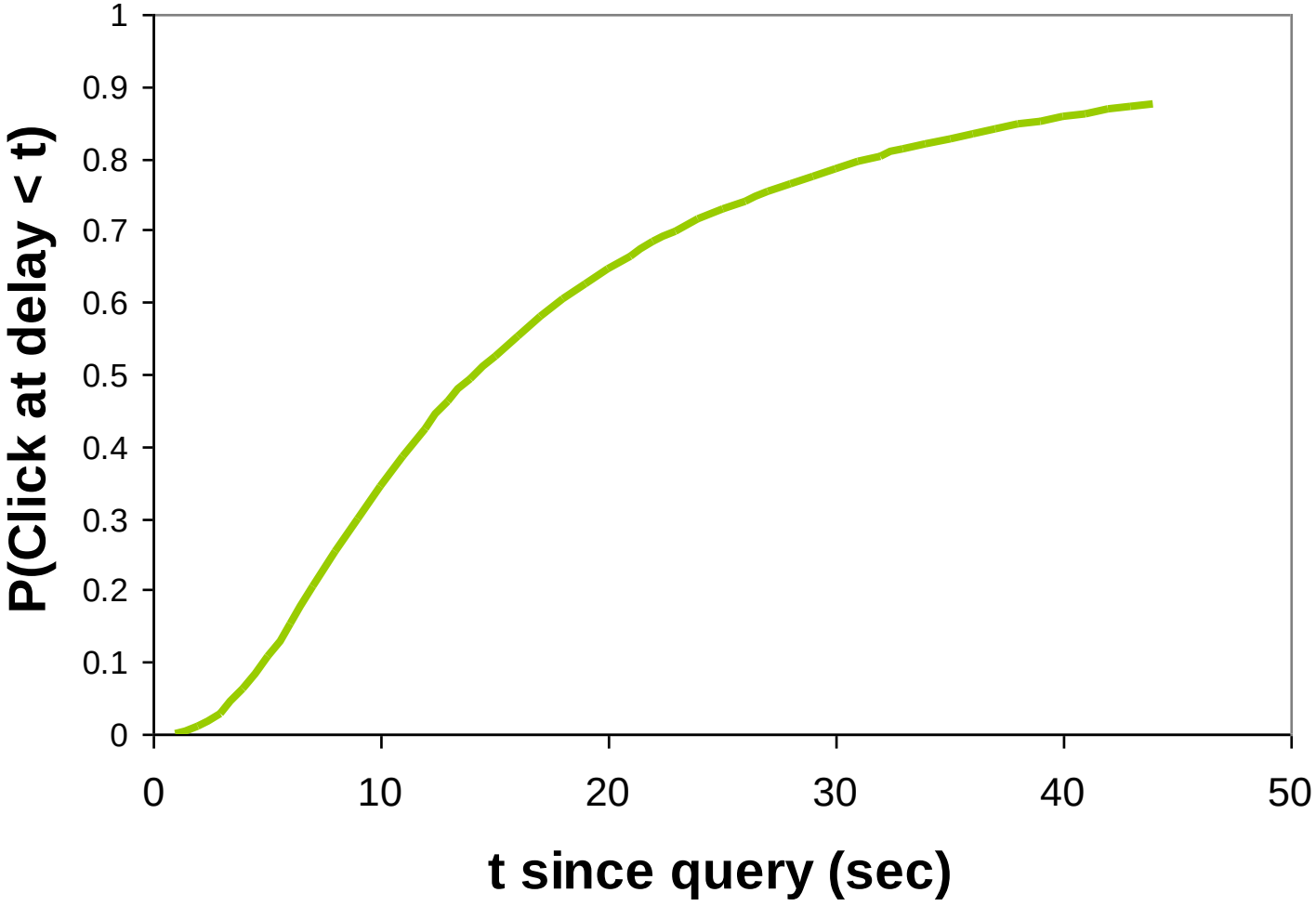
- **Belief**: Web Search engines have to understand my query and find a needle in a haystack of 20 billion documents!
- **Reality**: *Most* search queries are
  - **short** (avg. ~2.5 words [2005])
  - satisfied by pages from a **small subset** of the Web
    - Millions rather than billions [Mei et al., 2008]

=> More like finding a *pencil* in a haystack!

Extremely Quickly

- Results returned in < 1 sec
  - For any query
  - For any engine
- It was not always thus!
  - 3-4 seconds in early days ([Chu & Rosenthal, 1996; Garratt et al. 2001])
- How? Inverted indices, enormous data centers, clever algorithms…

Users are quick too…



[Downey, Dumais, & Horvitz 2007]

At Massive Scale

Search Engine	February 2008	Searches per Day February 2008	February 2009	Searches per Day February 2009
Total	9882	341	13104	468
Google	5855	202	8293	296
Yahoo!	2136	74	2696	96
Microsoft	953	33	1073	38
Ask	488	17	536	19
AOL	450	16	507	18

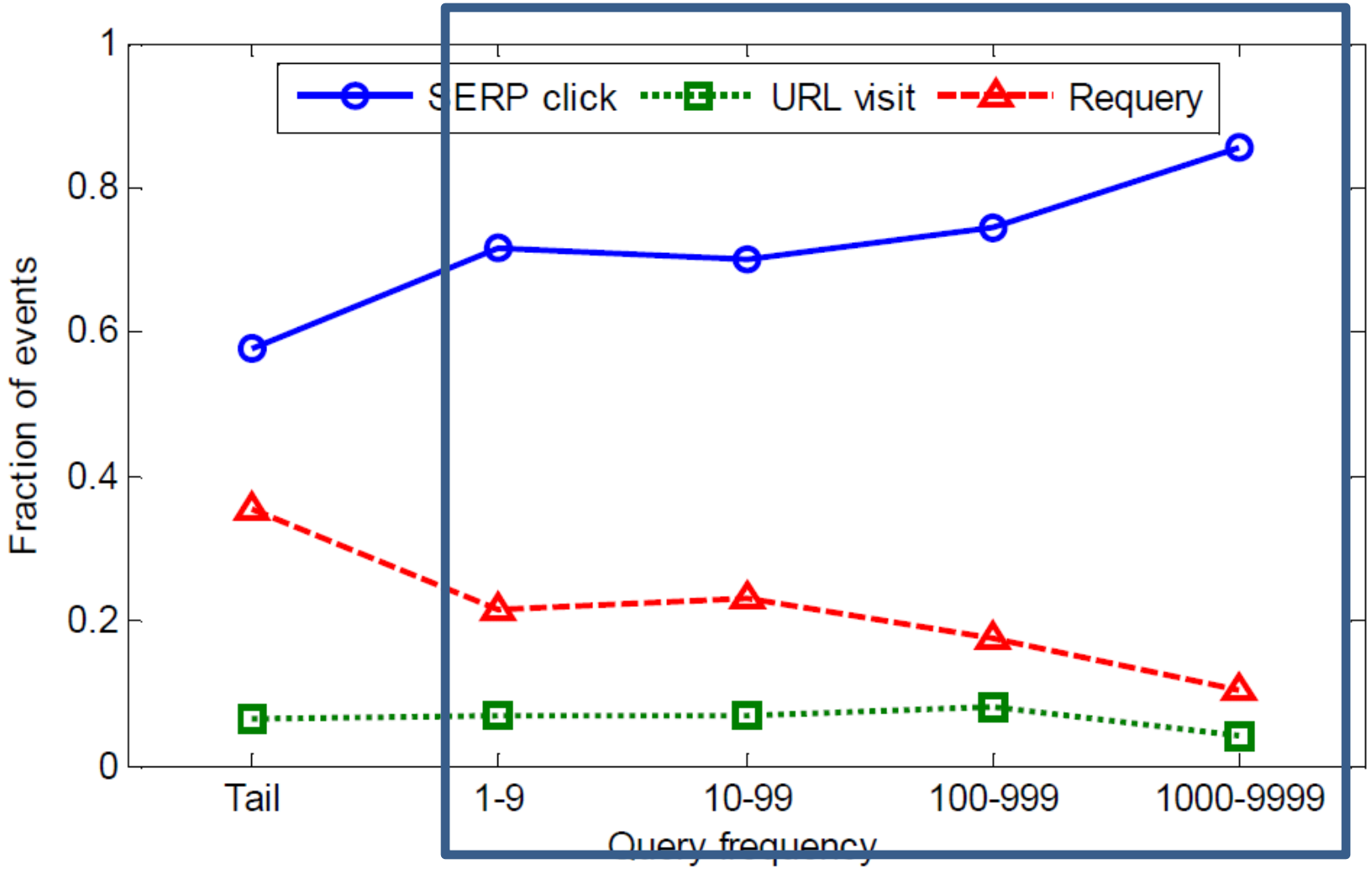
(in millions)

Jan 2011: **18.5 Billion**

How? Inverted indices, enormous data centers, clever algorithms…

http://blog.searchenginewatch.com/comscoresearchsharefeb2009_0309.jpg

Relatively Well



http://www.seoresearcher.com/distribution-of-clicks-on-googles-serps-and-eye-tracking-analysis.htm

Web Search Engines today…

- For the most part:
  - Perform an **easy** task
  - Extremely **quickly**
  - At massive **scale**
  - Relatively **well**
- The future of Web search is in **more difficult tasks**

Trend toward more difficult queries

- Search queries are getting longer (sort of)

- *Note: To see data this slide and next, search for "Percentage of US clicks by number of keywords"*

Trend toward more difficult queries

- Search queries are getting longer? Sort of

Things you can't do with search today

- Query by *description* rather than *content*
  - `Humorous anecdotes about frogs`
- Extracting and synthesizing over multiple pages
  - `Nanotechnology companies hiring on the West Coast`
  - `Substances the FDA has banned`
- Organizing bodies of documents
  - `Show me the most compelling cases for or against QE2`
  - `Who says drinking from aluminum soda cans increases my Alzheimer's risk? Should I believe them?`

2009 Image Search for **famous pink building**

Bing and Decide!

- Google “owns” algorithmic document search

Competitive Landscape

Core Search Entity	1/10: Search Share (%) 1/11-2-11: Explicit Core Search Share (%)		
	Jan-10	Jan-11	Feb-11
Google Sites	65.4%	65.6%	65.4%
Yahoo! Sites	17.0%	16.1%	16.1%
Microsoft Sites	11.3%	13.1%	13.6%
Ask Network	3.8%	3.4%	3.2%
AOL, Inc.	2.5%	1.7%	1.7%

Source: ComScore, Inc.

Bing and Decide!

- Google “owns” algorithmic document search

- Bing’s strategy:
  - Searchers intend to complete **tasks**
  - Today: search engines are the **first** step
  - Goal: Have engine do the first ***k*** steps

Reminder

- 2-paragraph paper summary due Thursday
- Sign-up for papers this week
- Look for e-mail on these items this evening