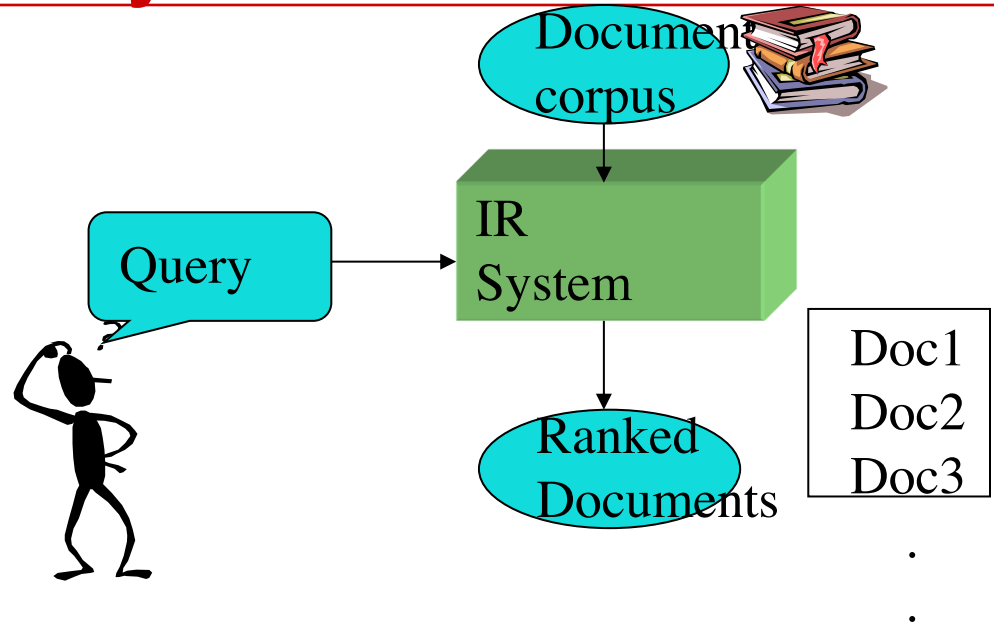
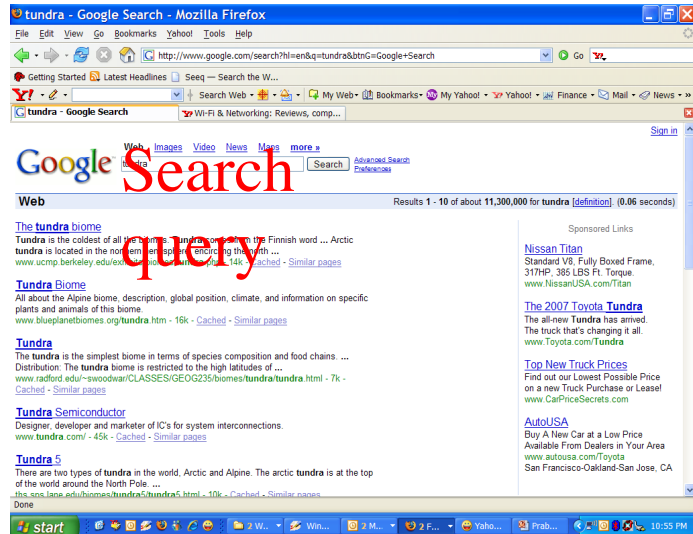


Introduction to Information Retrieval and Web Search

Tao Yang

UCSB CS293S, 2023

A Narrow View of Information Retrieval/Web Search Systems

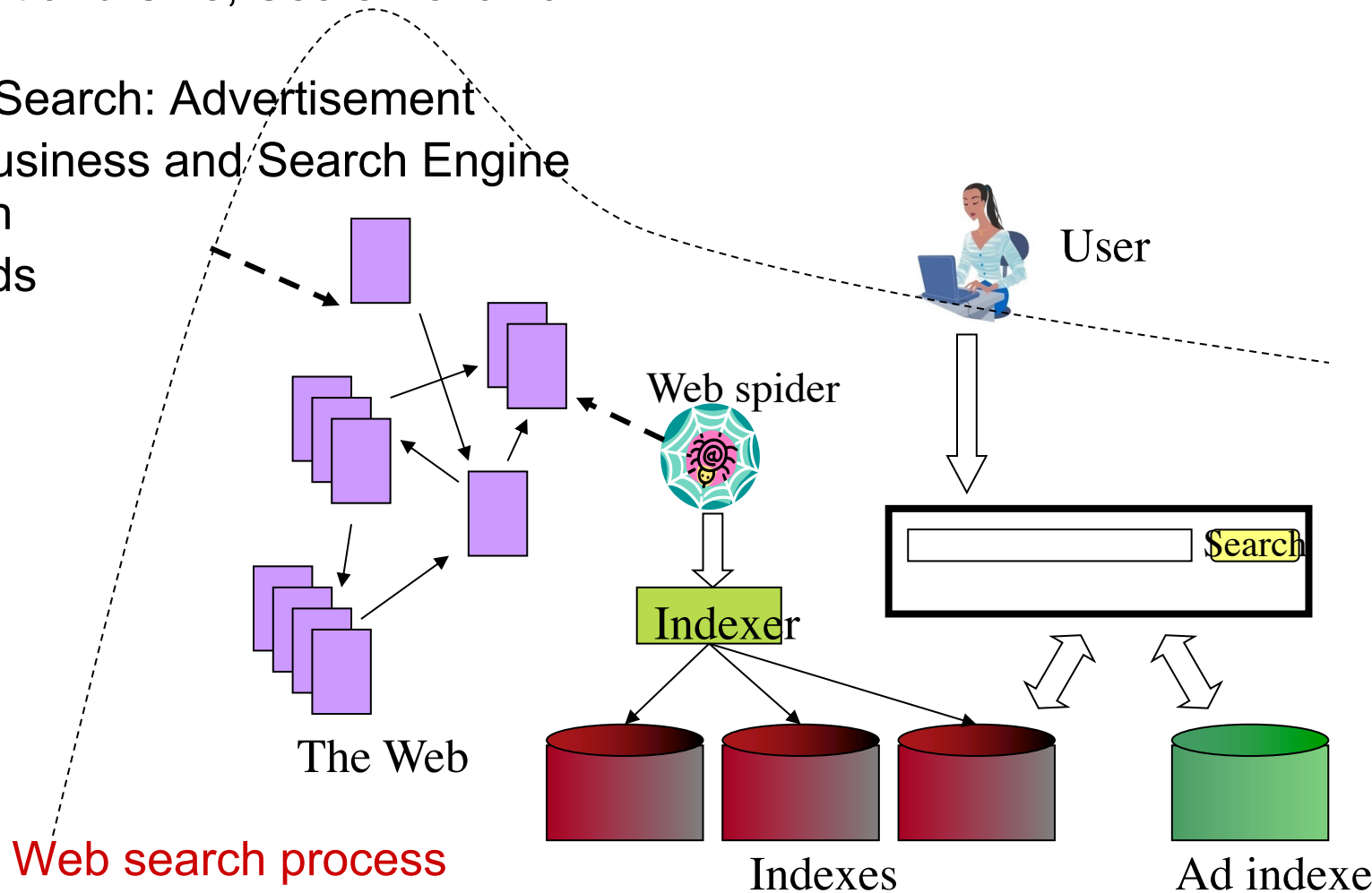


Why is IR/Search important?

- Too much information on the web and social networks
 - Everybody uses IR/search systems to select
- Search system is important for business
 - Most of web visit traffic is directed from search engines

Table of Content

- Information Retrieval& Search Engine Architecture
- Web Content and Size, Users Behavior in Search
- Sponsored Search: Advertisement
- Impact to Business and Search Engine Optimization
- Related fields
- This course

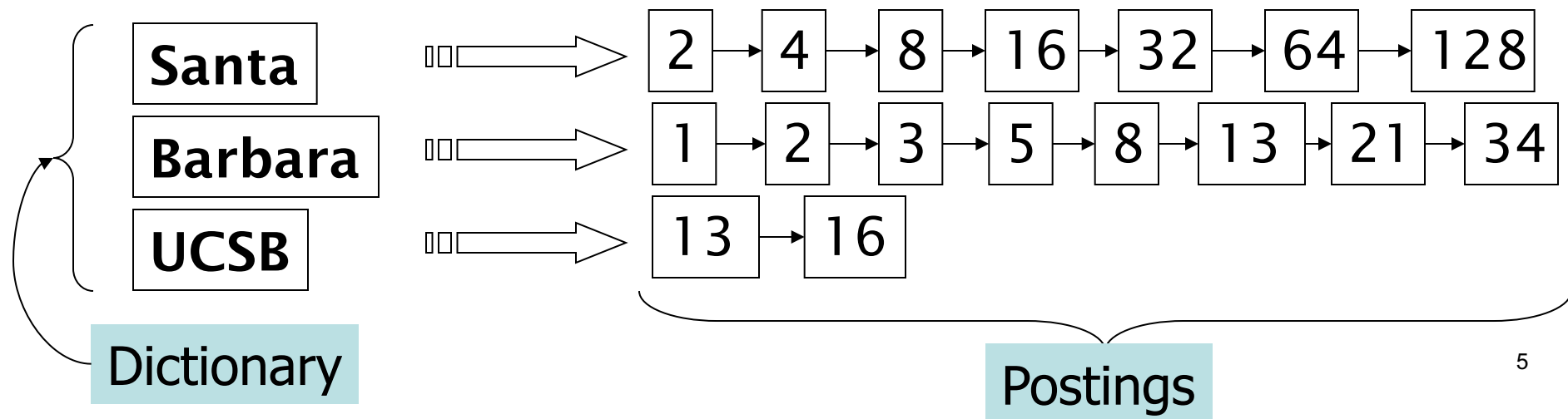


Search engine architecture: key pieces

- **Spider (a.k.a. crawler/robot) – builds corpus**
 - Collects web pages recursively
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - Additional pages from direct submissions & other sources
- **Indexer and offline text mining**
 - create inverted indexes so online system can search
 - Enrich knowledge on things and their relationship (e.g. names and events) and documents through data mining and learning
- **Online query process – serves query results**
 - Front end – query reformulation, word processing
 - Back end – finds matching documents and ranks them

Document Representation in Search Index

- A document is represented as a vector of features
 - As a bag of terms → a sparse vector with many zeros. Each element is a term weight
 - As a dense vector with neural features
- A sparse vector can be implemented as **inverted index**
 - A dictionary with a set of terms
 - Each term points to a list of postings with document IDs that contain such term feature



Indexing Process with Mining

- **Text acquisition**

- Gather data

- **Text transformation**

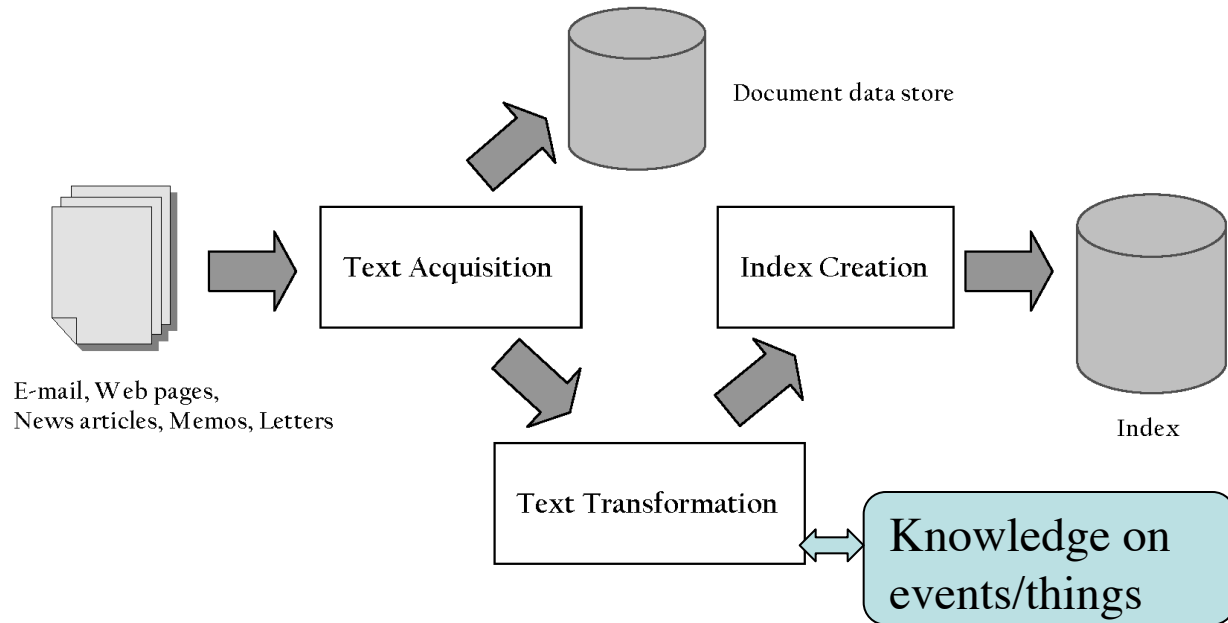
- transforms documents into *index terms* or *features*

- **Index creation**

- takes index terms and creates data structures (*indexes*) to support fast searching

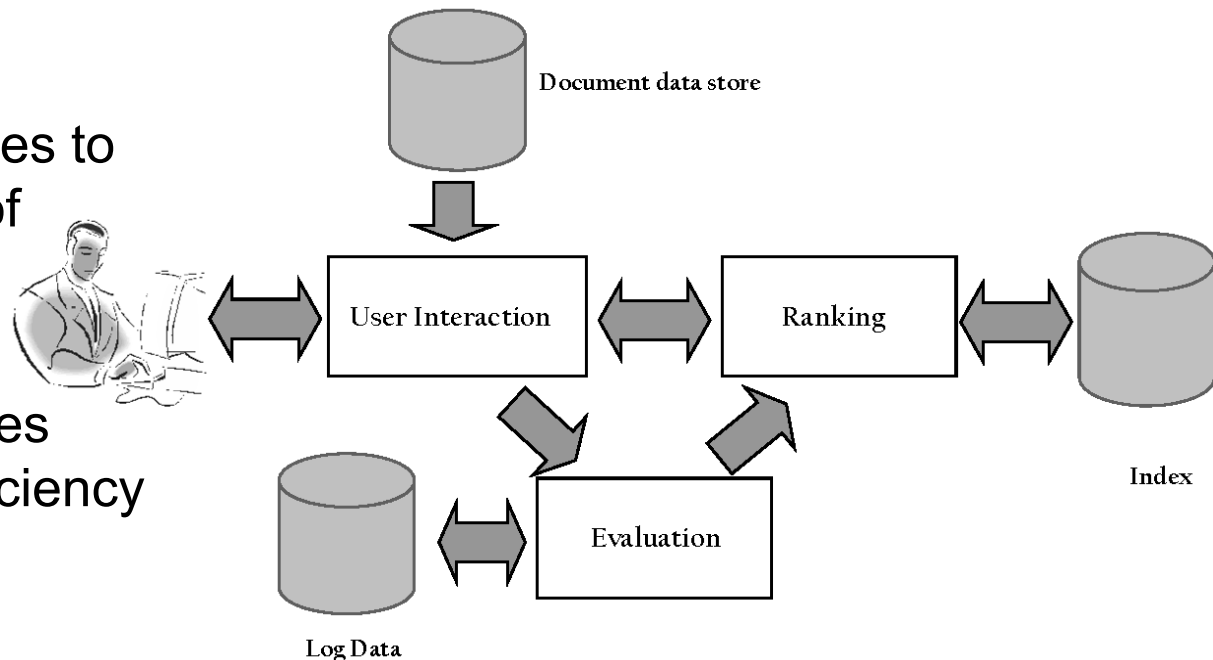
- **Data mining**

- Knowledge learning on entities (people name, organization, etc) and their relationship (knowledge graphs)

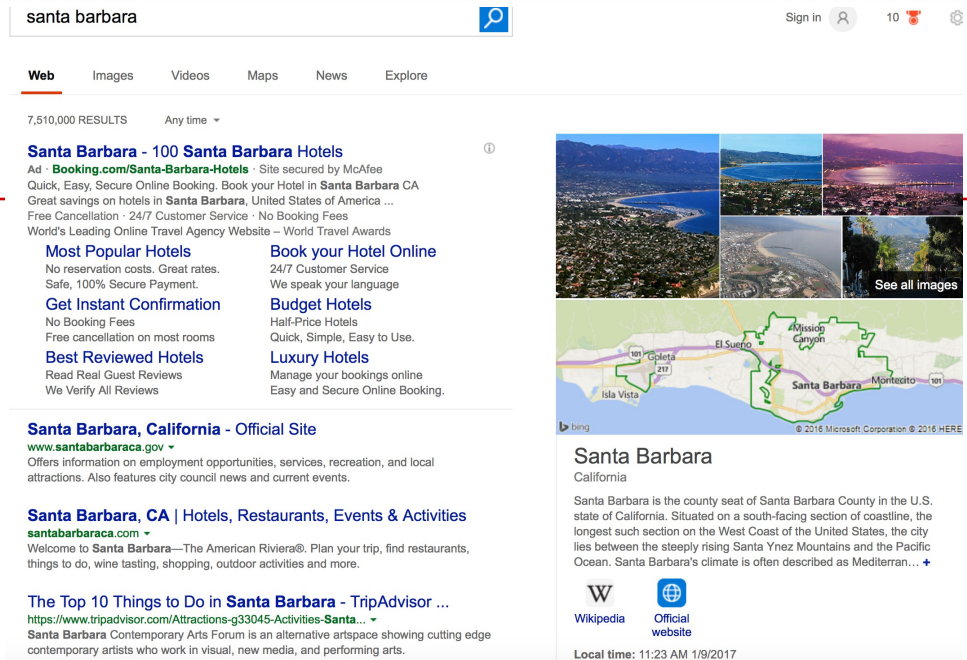


Query Process

- **User interaction**
 - supports creation and refinement of query, display of results
- **Ranking**
 - uses query and indexes to generate ranked list of documents
- **Evaluation**
 - monitors and measures effectiveness and efficiency (primarily offline)



User Interaction



- **Results output**
 - Constructs the display of ranked documents for a query
 - Merge results from multiple channels
 - Retrieves appropriate *advertising*
 - Generates *snippets (dynamic description)* to show how queries match documents
 - *Highlights* important words and passages
 - May provide *clustering* and other visualization tools

User Interaction

- **Query transformation**
 - Improves initial query,
 - Stopword removal, spell correction, long query trimming
 - marriot hotel at golet
 - *Spell checking suggestion* and *query suggestion* provide alternatives to original query
 - Did you mean “Marriott hotel at Goleta”?
 - *Query transformation or expansion* modifies the original query possibly with additional terms
 - *UC santa babara admission rate*

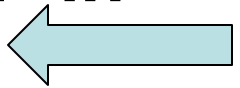
Online System Support

- **Performance optimization**
 - Designing matching & ranking algorithms for efficient processing
 - *Safe vs. unsafe* optimizations
- **Parallel/distributed computing. Caching**
 - Processing queries in a distributed environment
 - *Query broker* distributes queries and assembles results
 - *Caching of* intermediate or final results

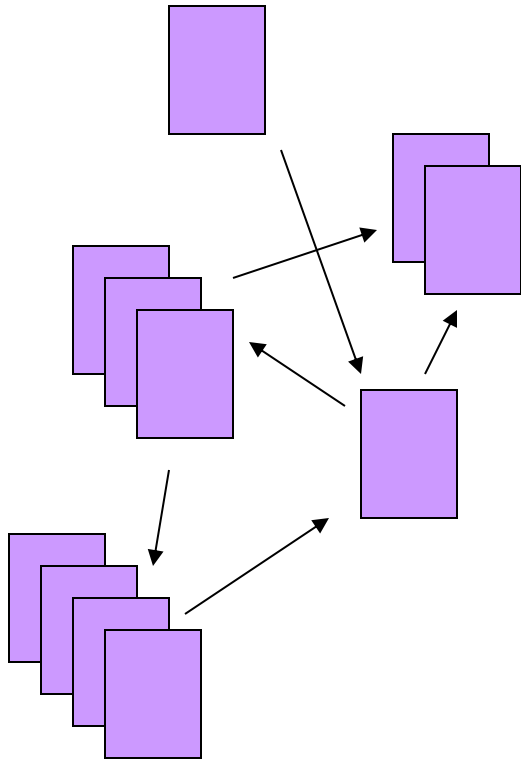
Evaluation

- **Logging**
 - Logging user queries and interaction is crucial for improving search effectiveness and efficiency
 - *Query logs and clickthrough data*
 - used for query suggestion, spell checking, query caching, ranking, advertising search, and other components
- **Ranking analysis**
 - Measuring and tuning ranking effectiveness
 - Use some metrics such as NDCG
- **Performance analysis**
 - Measuring and tuning system efficiency

Table of Content

- **Information Retrieval/Search Engine Architecture and Process**
- **Web Content and Size. Users Behavior in Search** 
- **Advertisement and Impact to Business**
 - Search Engine Optimization
- **Related Fields**

Characteristics of Web Content



The Web

- **No design/co-ordination**
- **Distributed content creation, linking**
- **Diverse content** includes truth, lies, obsolete information, contradictions ...
- **Structured (databases), semi-structured ...**
- **Scale -- huge**
- **Growth** – slowed down from initial “volume doubling every few months”
- **Content can be *dynamically generated***

General web search: identify relevant information with a horizontal/exhaustive view of the world.

Vertical search: Focus on specific segment of web content

The user



- **Diverse in access methodology**
 - Increasingly, high bandwidth connectivity
 - Growing segment of mobile users: limitations of form factor
 - keyboard, display
- **Diverse in search methodology**
 - Search, search + browse, filter by attribute ...
 - Average query length ~ 2.5 terms
- **Poor comprehension of syntax**
 - Early engines surfaced rich syntax – Boolean, phrase, etc.
 - Current engines hide these
- **Mobile users**
 - Bias towards shorter queries
 - Much higher location-based activity through map app


Web Search: How do users find content?

- **Informational (~25%)** – want to learn about something
cancer
- **Navigational (~40%)** – want to go to that page
United Airlines
- **Transactional (~35%)** – want to do something (web-mediated)
 - Access a service
 - Downloads
Santa barbara weather
 - Shop
Mars surface images
- **Gray areas**
 - Find a good hub
 - Exploratory search “see what’s there”
Nikon D-SLR
Car rental Finland

Search Intent Analysis

- **Problem with keywords**
 - May not retrieve relevant documents that include synonymous terms.
 - “car” vs. “automobile” “UCSB” vs. “UC Santa Barbara”
 - May retrieve irrelevant documents that include ambiguous terms.
 - “bat” (baseball vs. mammal) “Apple” (company vs. fruit)
 - “bit” (unit of data vs. act of eating)
- **Taking into account the *meaning* of the words used.**
- **Taking into account the *order* of words in the query.**
 - Paris Hilton vs Hilton Paris
- **Adapting to the user based on direct or indirect feedback.**
- **Taking into account the *authority* of the source.**

Table of Content

- **Information Retrieval/Search Engine Architecture and Process**
- **Web Content and Size. Users Behavior in Search**
- **Advertisement & Impact to Business** 
 - Search Engine Optimization
- **Related Fields**

What is percent of users who can differentiate sponsored search links and algorithmic search results?

cannon camera - Yahoo! Search Results - Mozilla Firefox

File Edit View Go Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search?fr=ytf1-msgff&p=cannon%20camera&ei=UTF-8

Getting Started Latest Headlines Seeq — Search the W...

Y! cannon camera Search Web Mail My Yahoo! Basketball Games Music Answers

Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]

YAHOO! SEARCH cannon camera Search Advanced Search

Search Results 1 - 10 of about 4,070,000 for cannon camera - 0.20 sec. (About this page)

Did you mean: [canon camera](#)

Canon Camera at Circuit City
[www.CircuitCity.com](#) - Circuit City - Official Site. Free Shipping on Orders \$24 and Up.

Cannon Camera
[RitzCamera.com](#) - Huge Selection of Canon Cameras. Free Shipping & No Tax. Buy Today.

1. **Canon** (NYSE: [CAJ](#))
Global manufacturer of copy machines, fax machines, cameras, computer peripherals, and optical products.
[www.canon.com](#) - 23k - [Cached](#) - [More from this site](#)

2. **Canon Camera Museum**
Showcasing camera history, technology, and design.
[www.canon.com/camera-museum](#) - 22k - [Cached](#) - [More from this site](#)

3. **Canon Digital Cameras**
Official Canon site for its line of PowerShot and EOS digital cameras, photo printers, and film scanners.
[www.powershot.com](#) - 104k - [Cached](#) - [More from this site](#)

4. **Canon USA**
Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, cameras and lenses, and more.

SPONSOR RESULTS

Authorized Canon Cameras Pro Dealer
Buy Canon Cameras here.
Imageologists: Professional photographic...
[www.imageologists.com](#)

Canon Cameras
We Offer 3,500+ Digital Cameras.
Discover canon cameras.
[www.BizRate.com/canon](#)

Camera Cases and Bags
To know Bogen Imaging Inc, just take a look at the premium brands...
[www.bogenimaging.us](#)

Cannon Camera Battery Accessory
Spring Sale. 80% off. Valid till Apr-30. Free Ship coupon over \$30.
[www.cellphoneshop.net](#)

Higher slots get more clicks

How it works

Advertiser



I want to bid \$5 on
canon camera

I want to bid \$2 on
cannon camera

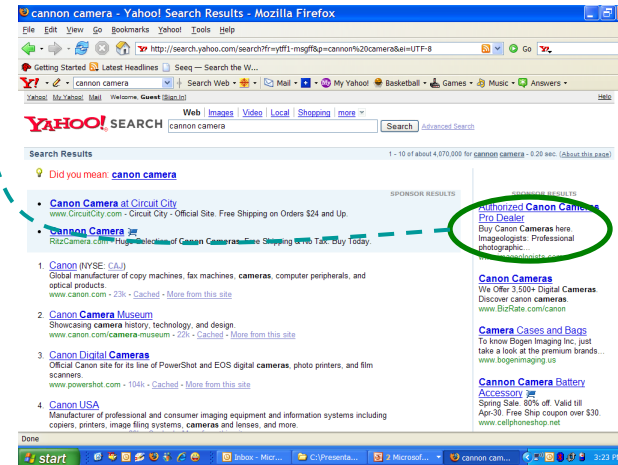


Ad Index

Sponsored
search engine

Engine decides when/where to show this ad.

Landing page

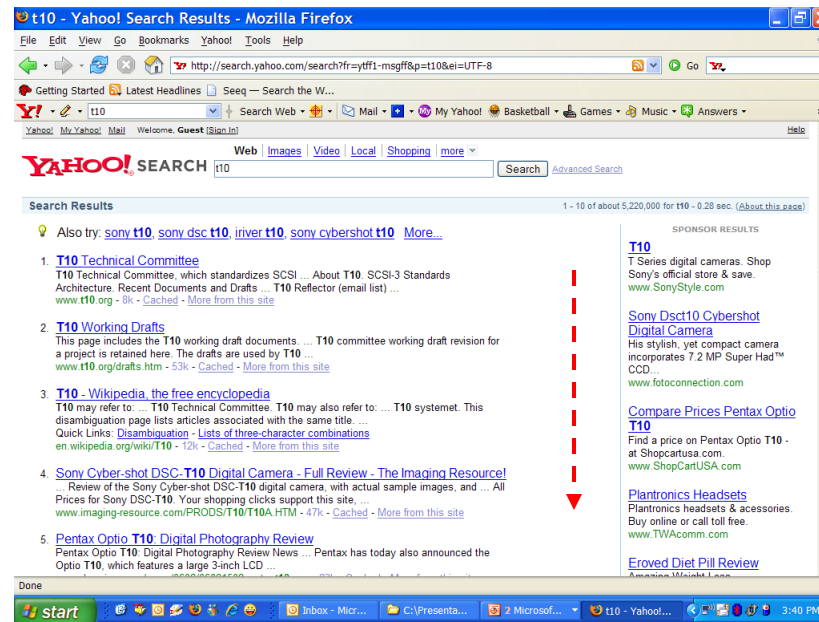


Engine decides how much to charge advertiser on a click.

Three sub-problems

1. Match ads to query/context
2. Order the ads
3. Pricing on a click-through

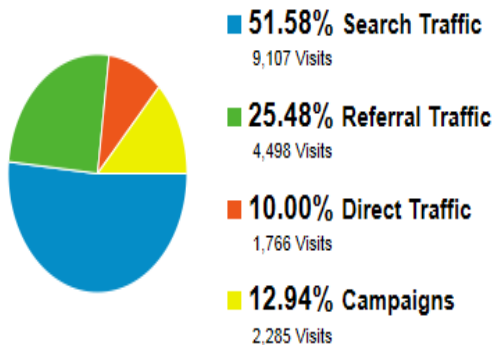
IR
Econ



Search Traffic is Important for Business:

Example of Site Traffic Analysis

17,656 people visited this site



Search Traffic

Keyword

Matched Search Query

Source

Referral Traffic

Source

Direct Traffic

Landing Page

Source	Visits	% Visits
google	8,795	96.57%
bing	106	1.16%
yahoo	96	1.05%
search	38	0.42%
ask	28	0.31%
aol	14	0.15%
avg	9	0.10%
images.google	9	0.10%
search-results	5	0.05%
babylon	3	0.03%

[view full report](#)

Paid placement vs Search Engine Optimization

- Paid placement costs money. What's the alternative?
- **Search Engine Optimization:**
 - “Tuning” your web page to rank highly in the search results for select keywords
 - Alternative to paying for placement
 - Thus, intrinsically a marketing function
 - Also known as **Search Engine Marketing**

Search engine optimization Strategies

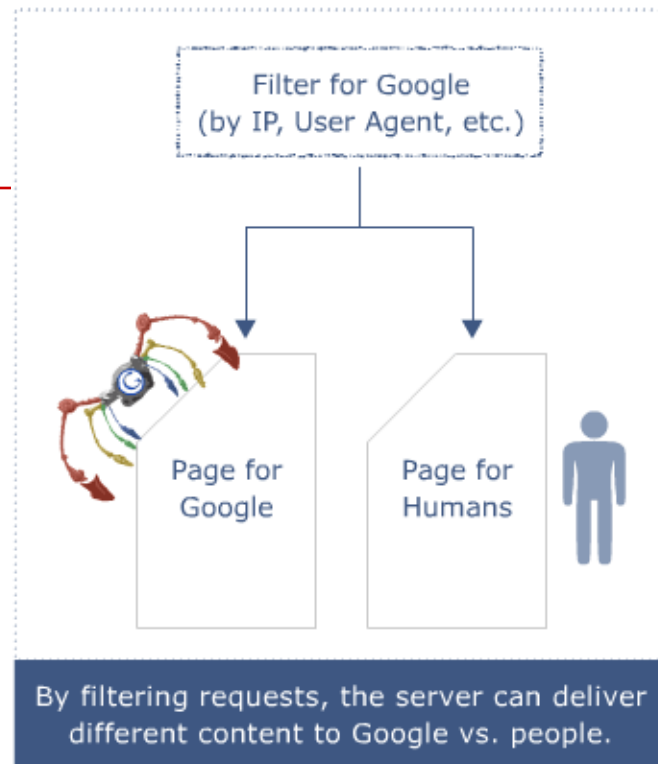
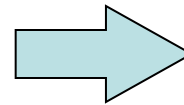
- **Early engines relied on the density of terms**
 - The top-ranked pages for the query *maui resort* were the ones containing the most *maui's* and *resort's*
- **SEOs responded with dense repetitions of chosen terms**
 - e.g., *maui resort maui resort maui resort*
 - Often, the repetitions would be in the same color as the background of the web page
 - Repeated terms got indexed by crawlers
 - But not visible to humans on browsers

Can't trust the words on a web page, for ranking.

SEO Strategy: Cloaking

Normal behavior: Web server delivers same content to people vs search engine crawlers

Cloaking: Web server delivers different content to people vs search engine crawlers



Web Images Maps News Shopping Gmail more ▼

Google matt cutts transcripts Search Blogs Search the Web Advanced Blog Search Preferences

Blog results

Published [View all web results for matt cutts transcripts](#)

[Last hour](#)
[Last 12 hours](#)
[Last day](#)
[Past week](#)
[Past month](#)
→Anytime
[Choose Dates](#)

Subscribe:
✉ [Blogs Alerts](#)
[Atom](#) | [RSS](#)

Matt Cutts Discusses the Importance of alt Tags - Mattcutts Video...
14 hours ago by power
If you look here, "Matt Cat, Emmy, Cutts, with some yarn> you can see this image tag, image source and an ALT tag which stand for alternative text, and if so somebody is using a screen reader, or they can't load the image for a reason, ...
[SEO BLOG - http://www.searchenginegenie.com/seo-blog/seo-blog.html](#)
[More results from SEO BLOG]

Matt Cutts 2007 SEO Wordpress talk at Wordcamp - Transcript
3 Apr 2008
Matt Cutts 2007 SEO Wordpress talk at Wordcamp - Transcript. Matt Cutts Wo...
[Cheap say viagra wordpress - http://cheap-say-viagra-wordpress.kbsbbs.com/](#)

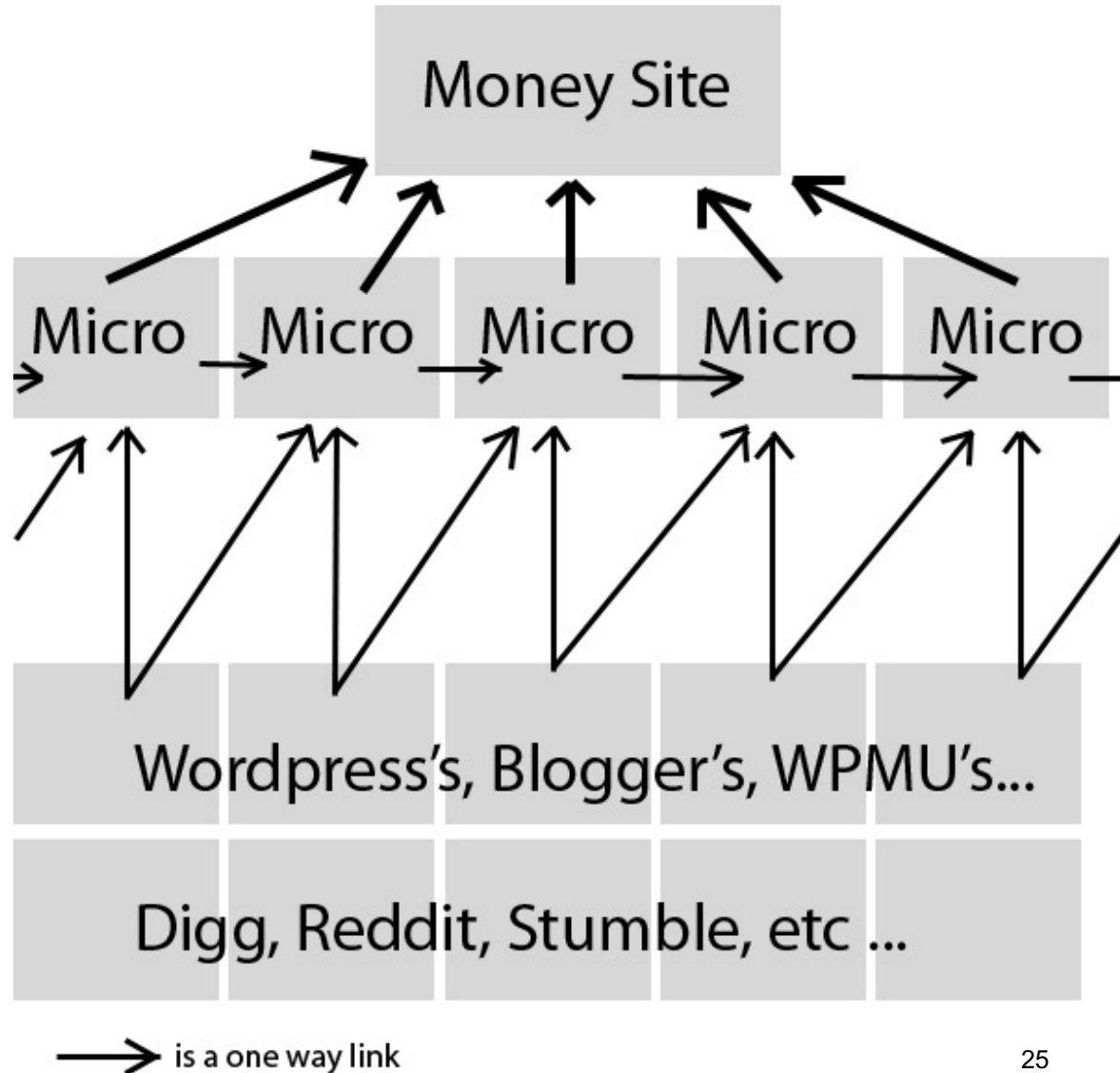
Session: Organic Listings Forum
17 Mar 2008 by Lee Odden
Greg: It's interesting that Matt Cutts coined the phrase PageRank sculpting. If you look at the toolbar PR, don't put a ton of credibility into it. But it can give you some indication if the site is trusted. ...
[Online Marketing Blog - http://www.tearankblog.com](#) References

Cloaking Site

Spamming with Link Farms

Page link support is important ranking feature.

SEO strategy:
Boost pagerank of a website with many artificial links



From Information Retrieval to Web Search

- **Challenging due to Large-scale and noisy data.**
 - retrieving relevant documents to a query.
 - retrieving from large sets of documents efficiently.
- **Relevance is a subjective judgment and may include:**
 - Simplest notion of relevance is that the query string appears verbatim in the document.
 - More:
 - Being on the proper subject.
 - Being timely (recent information).
 - Being authoritative (from a trusted source).
 - Satisfying the goals of the user and his/her intended use of the information (*information need*).

Related Areas

- **Information Management and Data Mining**
 - Information Science
 - Machine Learning and data mining
 - Natural Language Processing
 - Recommendation
 - Using statistics about the past actions of a group to give advice to an individual
- **Large-scale systems**
 - Database/data stores
 - Operating systems/networking support
 - Web language analysis
 - Compression/fast algorithms.
 - Fault tolerance/parallel+distributed systems

Course Topics and Workload

- **Information Retrieval & Web Search**
 - Indexing, compression, and online search
 - Document retrieval and ranking. Neural models.
 - Text mining including duplicate analysis.
- **Systems Support**
 - Online servers and offline computation.

No textbook. Weekly slides with references are posted.

Workload:

Group project (2 persons) using state-of-the-art techniques

Paper reviewing and presentation

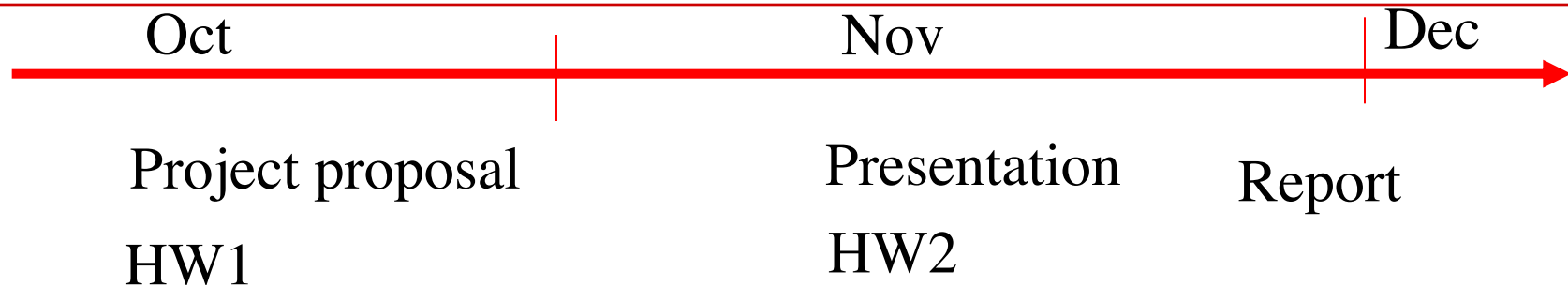
Implementation/evaluation.

Report.

HW exercises (25-30%); Class participation (3%)

Grading will be curved.

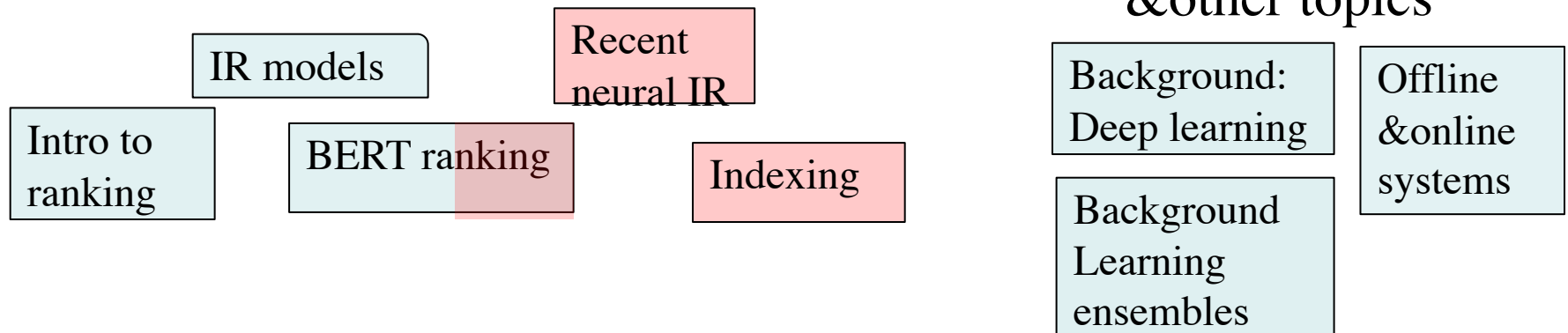
Timelines



Fast intro

Recent papers

Background
& other topics



Assume you have taken a machine learning course that covers basic learning/neural computing concepts

If not, review background slides (e.g. deep learning) earlier

Timeline & Additional Course Info

- **Oct:** HW1
- **Oct 19:** Short project proposal
- **Nov:** HW2
- **Nov (2nd week):** Paper presentation
- **Dec:** Project demo/interview. Final project slides/report

Additional course information

- http://www.cs.ucsb.edu/~tyang_class/293s22f
- Class discussion at Piazza (invite you based on the class roster)
- GradeScope: (Class code will be posted)
- We are in process of acquiring some GPU resource allocation for this class.