Informatics 225
Computer Science 221

Information Retrieval

Lecture 5

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

A bit further on Web users:

How users evaluate search engines

Information Retrieval

How do users evaluate search engines?

- Major standard: Quality of pages

How do users evaluate search engines?

- Major standard: Quality of pages
  - Classic IR relevance: R(Q,D)
    - How you really define relevance is among your engine secrets R(Q,D,…).

How do users evaluate search engines?

- Major standard: Quality of pages
  - Classic IR relevance: R(Q,D)
    - How you really define relevance is among your engine secrets R(Q,D,…).
  - Also important:
    - Trust
    - Duplicate elimination (*with “show similar” option)
    - Readability
    - Fast access
    - No pop-ups

How do users evaluate search engines?

- **First :** what is **precision** and **recall**?

How do users evaluate search engines?

- **Precision**: *How relevant are the first few hits (how well can one reject non-relevant documents)?*

- **Recall**: *How many relevant hits are presented (Find all relevant documents)?*

Which one is more important for a web search engine?

How do users evaluate search engines?

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Reference: SEIRP, Bruce Croft, Donald Metzler, Trevor Strohman

- A = Set of relevant results for the query
- B = Set of retrieved results for the query
- ~A = Set of non-relevant results for the query
- ~B = Set of non-retrieved results for the query

How do users evaluate search engines?

- Precision: *How relevant are the first few hits (how well can one reject non-relevant documents)?*
- Recall: *How many relevant hits are presented (Find all relevant documents)?*
- **For web search: Precision is more important than recall**

How do users evaluate search engines?

- **For web search: Precision is more important than recall**

Eyetracking by Nielsen Norman Group nngroup.com NN/g

F-Shape reading patterns

How do users evaluate search engines?

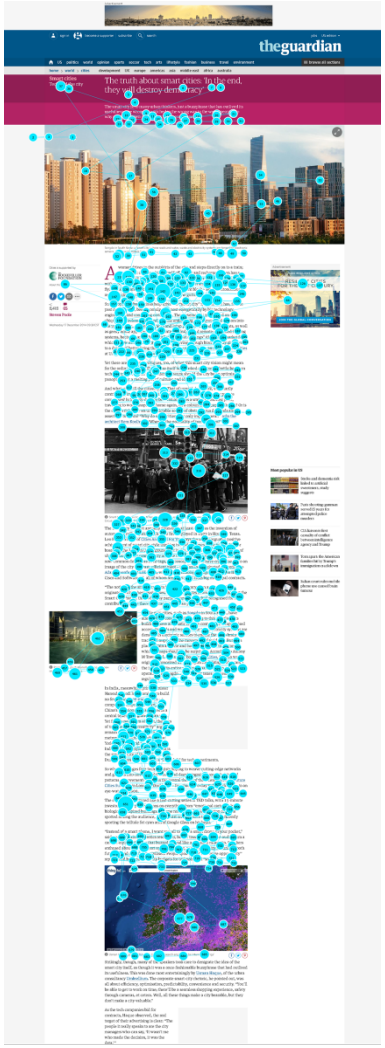
- Precision: *How relevant are the first few hits?*
- Recall: *How many relevant hits are presented?*
- For the web: Precision is more important than recall
- **When is recall important?**

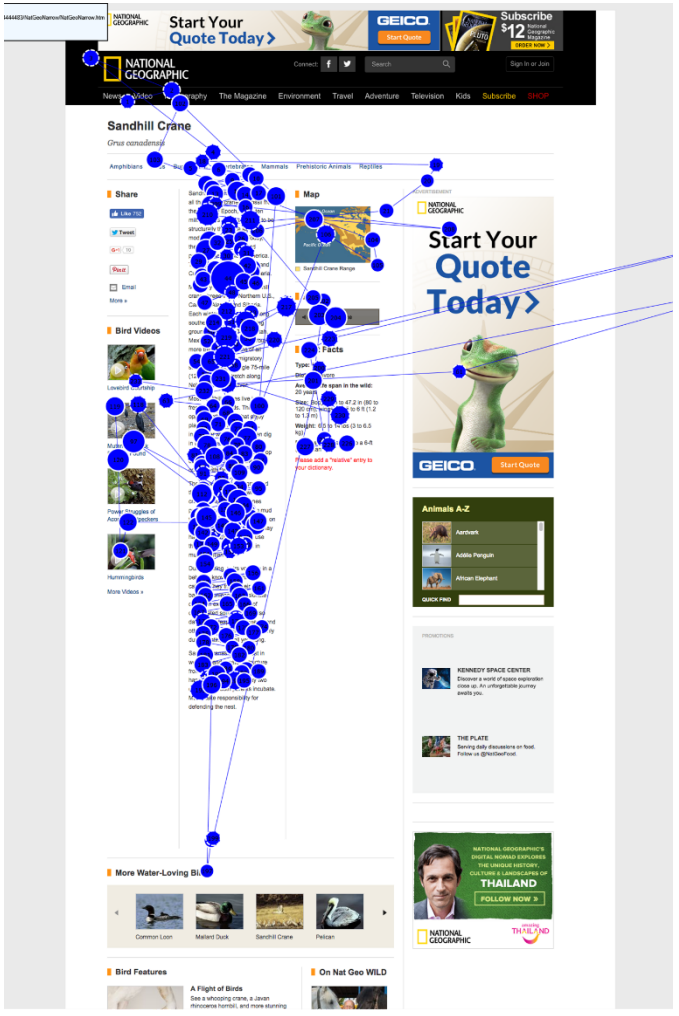
How do users evaluate search engines?

- **When is recall important?** When completeness is required

Commitment reading pattern : sources trusted *a priori, motivated users.*

Ex. Search engines for Law, Sciences.



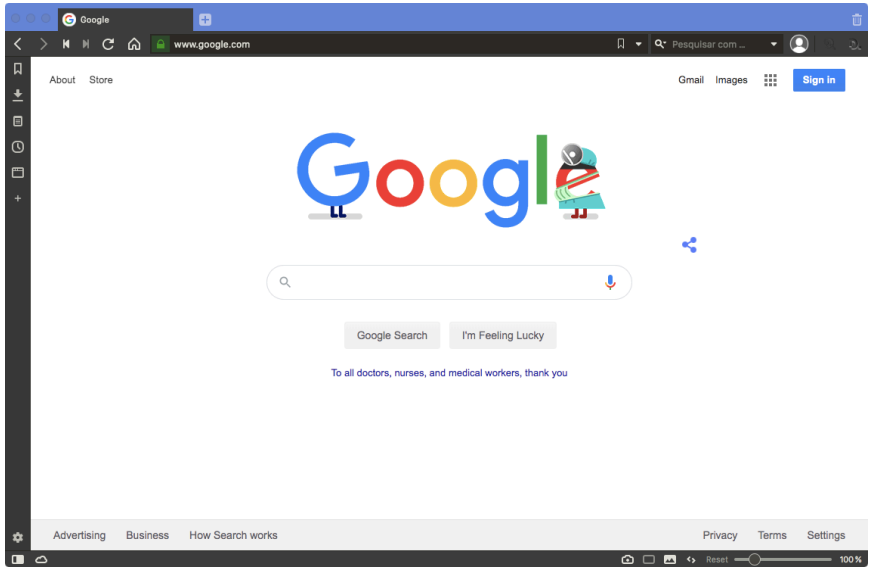


How do users evaluate search engines?

- Good UI: simple, no clutter

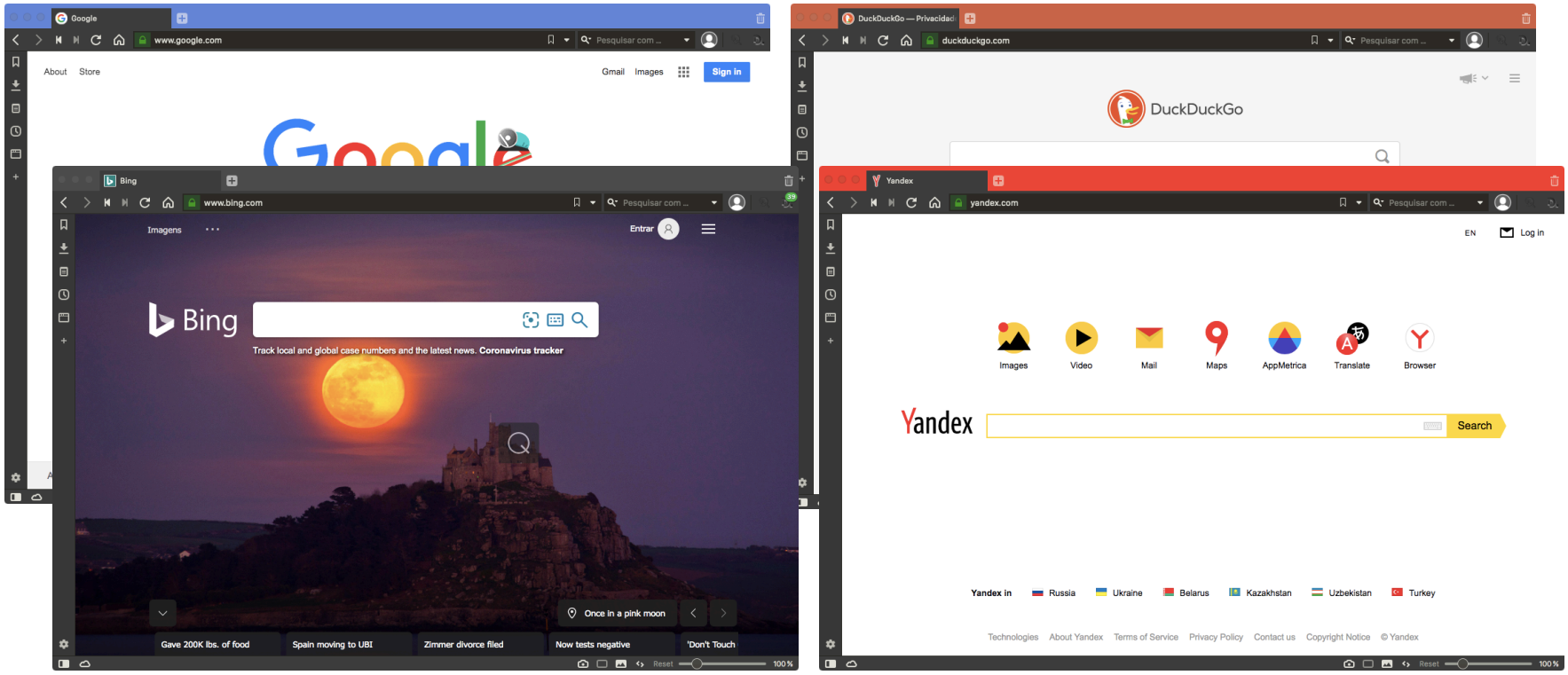
How do users evaluate search engines?

- Good UI: simple, no clutter



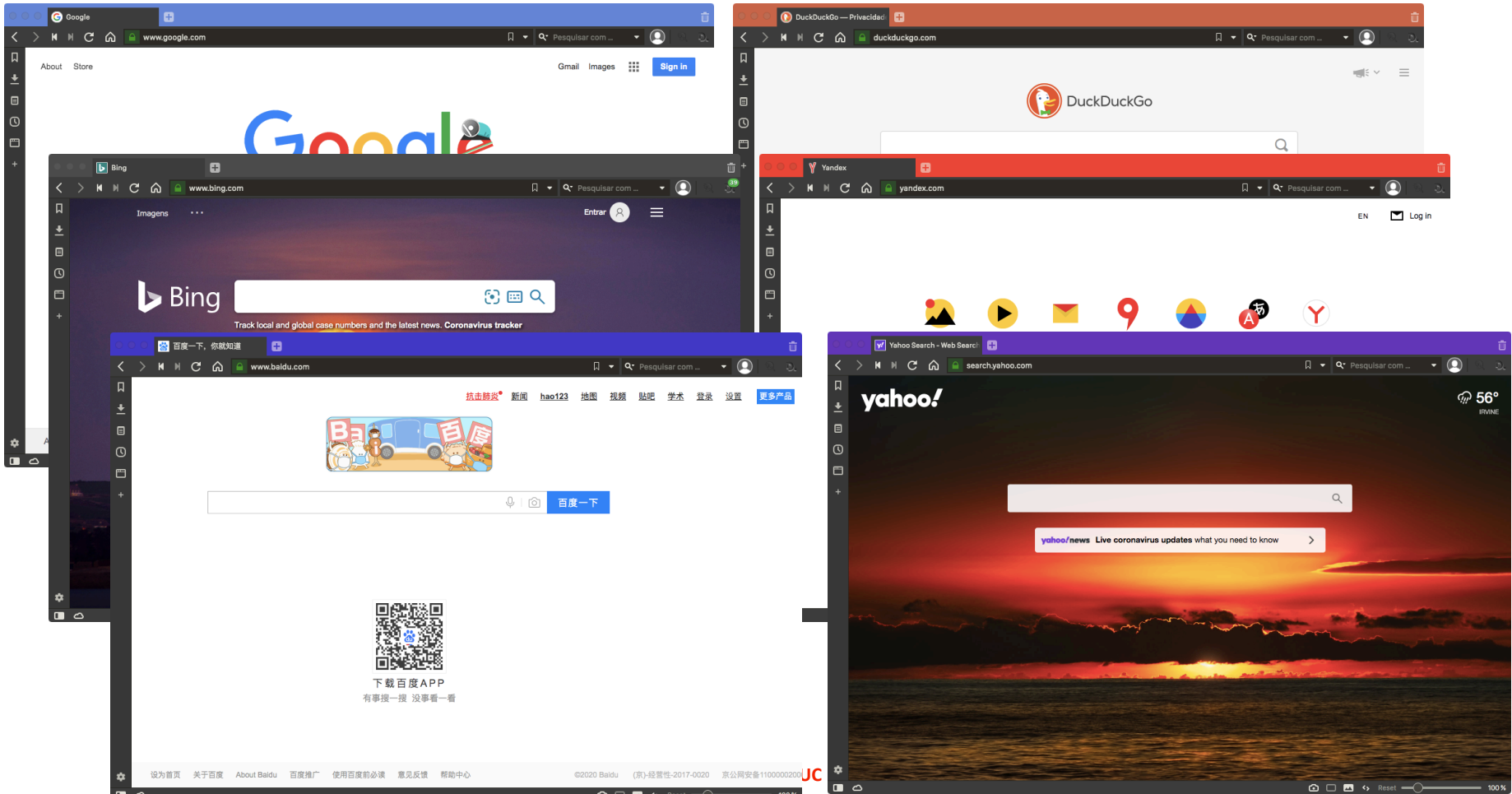
How do users evaluate search engines?

- Good UI: simple, no clutter



How do users evaluate search engines?

- Good UI: simple, no clutter



How do users evaluate search engines?

- Good UI: simple, no clutter
- Pre and post processing tools
  - Spell check / auto correct
  - Suggested alternative searches
  - Links to resources (maps, images, etc)
  - Able to deal with syntactic cues
    - ex: URL typed in search box, Math equation, etc.

Web Search business model : Advertising

Information Retrieval

Ads on Search

- Early synergy between search and ads
- First: keyword-based engines
  - Altavista (1995-1997), Excite, InfoSeek, Inktomi
- Paid placement ranking
  - Ads based on auction for keyword
  - Goto.com => Overture.com => Yahoo! => Google

Pagerank - a simple Definition

$$PR(i) = \sum_{j \in S_i} \frac{PR(j)}{L(j)}$$

- PR(i): Page rank of a given page I
- PR(j): Page rank of page j
- L(j): Number of forward link from page j

Ads on Search

- Link-based ranking pioneered by Google (1998+)
  - Links added the idea of authoritativeness to relevance
  - Blew away keyword-only engines
  - Great user experience, **looking for business model**

  - **Meanwhile, Goto/Overture's annual revenues near $1B**

Ads on Search

- Google added paid placement ads on site
  - Differentiated from search results
- Yahoo! build a similar engine
  - Bought Inktomi for search tech (a few hundred million USD!)
  - Bought Overture for paid placement tech (1.6 billion USD in 2003!)
    - Sponsored links placed alongside objective results.
- Google licensed paid placement tech from Yahoo!
  - Google's ad placement engine for many years (perhaps still today, nobody knows with certainty)

Ads on Search

- Google and most other search engines maintain that ads do not affect vendors' ranking in search
  - **But again: nobody knows for sure**…
- Ad placement:
  - Fully automated
  - Balance auction price and relevance
  - Targeted advertising : focus on audience

How companies pay for Ads on Search?

- Cost Per Mil (CPM)
  - Cost for showing the ad with 10^3-page shows
  - Important for branding campaigns
- Cost Per Click (CPC)
  - Cost for users clicking on the ad
  - Important for sales campaigns
- There are also other concepts
  - e.g. CPA: Cost per action/acquisition; if that click triggered some action; e.g. DuckDuckGo + Amazon.

Warning: Click Fraud

- $7.2 billion USD were lost between 2016-2018 (Guardian)
- Google calls this “invalid clicks”
- Search engines have sophisticated tools to detect fraud

Information Retrieval

Search Engine Optimization

Paid Ad Placement

- It costs money
- So...

Search Engine Optimization (SEO)

- **Tuning of a web page/site/app to rank highly** in search results for certain queries
- Alternative to paying for ad placement
  - Bonus: **visibility through rank = greater trust**
- **It's marketing**: getting your content to your audience

SEO

- Motives
  - Commercial
  - Political
  - Religious
  - Lobbying

SEO

- Motives
  - Commercial
  - Political
  - Religious
  - Lobbying
- Who does this?
  - Internally: webmasters, writers
  - Commercially: companies, consultants
  - Hosting services
  - Plugins to popular CMSs

SEO

- Essentially a **marketing – technology interface** area

- How to do it:
  - https://support.google.com/webmasters/answer/40349?hl=en
  - http://freetools.webmasterworld.com/category/seo-tools
  - Many books. E.g. :

SEO

- Ethical and unethical ways of doing it

- Legitimate approach:
  - Indexed age of the pages (older is better)
  - Good incoming links
  - Good content, well written, well organized, up to date
  - Good use of web standards/practices
  - Fast servers (quick response)

SEO

- Unethical approaches (aka spam):
  - fake pages
  - fake sites that point to your site
  - fake comments/engagement (bots!)
  - in short: “alternative facts” aka lies

- Sometimes the line between legitimate and illegitimate practices is hard to find. There’s a large grey area

SEO

- Search results (also) depend on which data center receives the query:
  - google.com vs. google.fr vs google.pt will show different results for a single query: e.g. “Paris”
- Different SEO strategies can be considered per country