Informatics 225
Computer Science 221

Information Retrieval

Lecture 2

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.

Announcements

- Assignment 1 has been posted.

Today

- Introduction
- Information Retrieval and Web Search

Our definition of information retrieval

The activity of searching and extracting information from a collection of information resources.

In this class : textual information from text documents

Web Search Basics

Introduction to Information Retrieval

Web Search Engines

- How often do you use web search engines?

Web Search Engines

- Without them, the Web wouldn't scale
  - No incentive in creating content unless it can be found
    - Taxonomies, bookmarks can't keep up
    - Or can they?

Web Search Engines

- Without them, the Web wouldn't scale
  - No incentive in creating content unless it can be found
    - Taxonomies, bookmarks can't keep up
    - Or can they?
- The Web is both a technology artifact and a social environment
  - For most people, Internet = Web,
    or even Internet = Google search

Web Search Engines

- Without them, the Web wouldn't scale
  - No incentive in creating content unless it can be found
    - Taxonomies, bookmarks can't keep up
    - Or can they?
- The Web is both a technology artifact and a social environment
  - For most people, Internet = Web,
    or even Internet = Google search
- Search engines make aggregation of interest possible:
  - Create incentives for niche players

Web Search Engines

- Search interaction makes “unlimited selection” stores possible
  - Amazon, Netflix, etc.

- Pew Research Center on *Search Engine Use 2012*

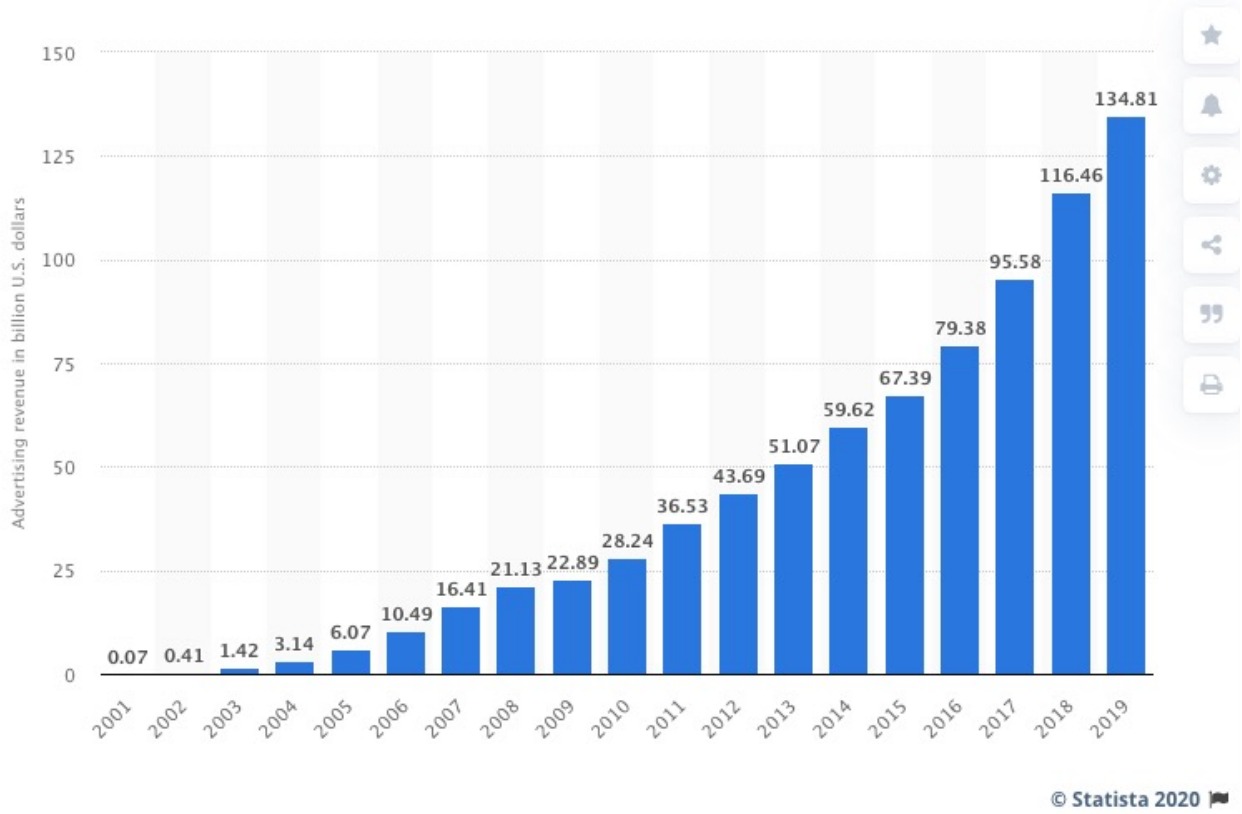
  *Search engines remain popular—and users are more satisfied than ever with the quality of search results—**but many are anxious about the collection of personal information by search engines and other websites.***

Web Search Engines

- Search turned out to be the [first] best mechanism for advertising on the Web:
  - web advertising revenue: Google (2019, global) **??????????**

Web Search Engines

- Search turned out to be the [first] best mechanism for advertising on the Web:
  - web advertising revenue: Google (2019, global) **$135B**



Web Search Engines

- Search turned out to be the [first] best mechanism for advertising on the Web:
  - web advertising revenue: Google (2019, global) **$135B**
  - Business model for ads changed over time

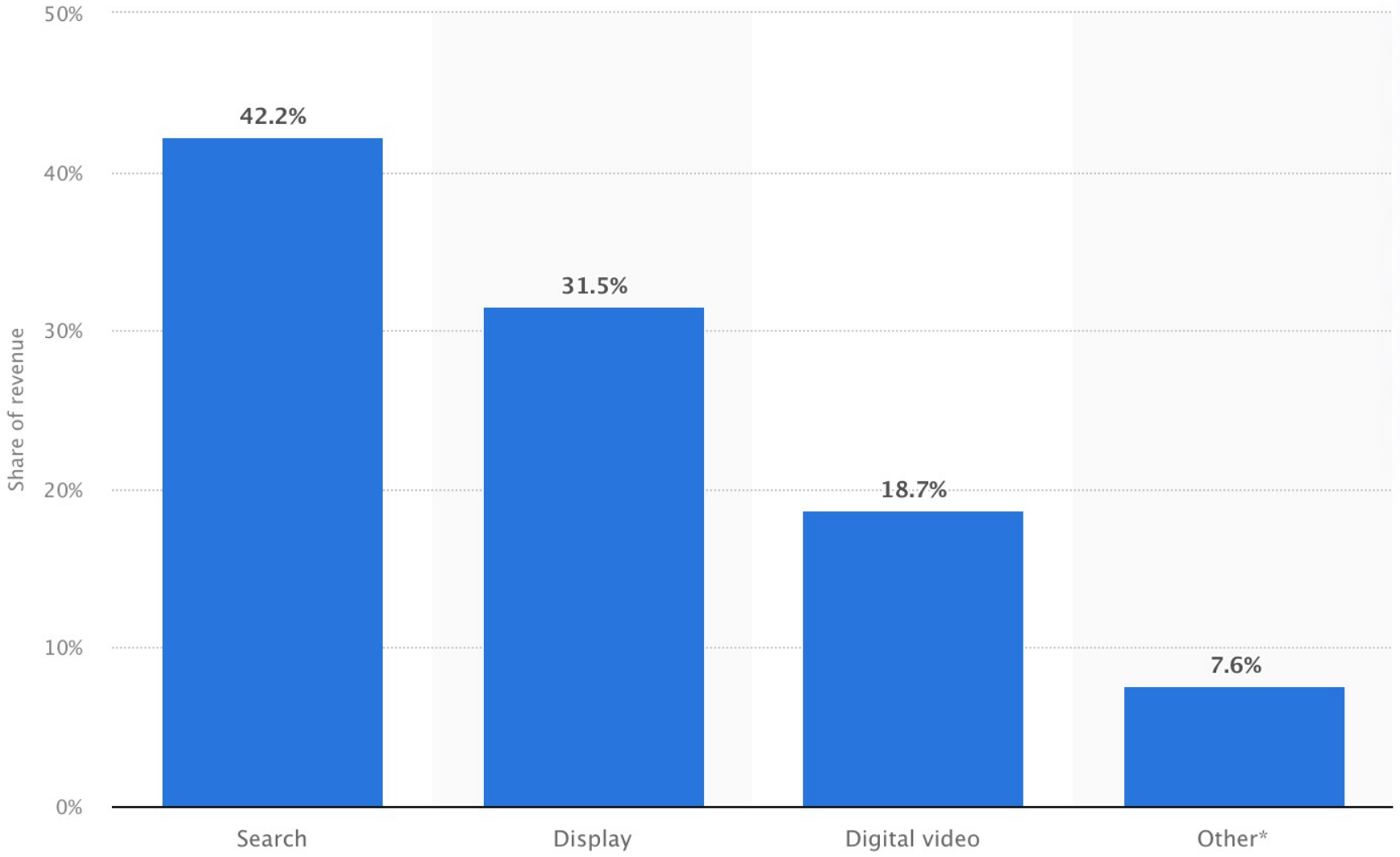
- Statistics and facts on the Digital Advertising Industry in the U.S. https://www.statista.com/topics/1176/online-advertising/ ::
  - *In 2021, digital ad. added up to more than **330 billion** U.S. dollars worldwide. According to the latest projections, digital Ad. revenue will amount to 460 billion U.S. dollars by 2024.*

Web Search Engines

- Search turned out to be the [first] best mechanism for advertising on the Web:
  - web advertising revenue: Google (2019, global) **$135B**
  - Business model for ads changed over time

- Statistics and facts on the Digital Advertising Industry in the U.S. https://www.statista.com/topics/1176/online-advertising/ :: *By 2021, digital ad. is projected to add up to more than **330 billion** U.S. dollars worldwide*
- Breakdown of online advertising revenue in the United States in 2020, by type https://www.statista.com/statistics/190458/categorie-breakdown-of-us-online-advertising-revenue-2010/

Web Search Engines



Search Engines

- Web search and Information retrieval
  - Essential to modern life
  - Central for your future professional life in Industry and Academia

Information Retrieval and Web Search

Introduction to Information Retrieval

Information Retrieval (IR)

- Information retrieval is a **field** concerned with the structure, analysis, organization, storage, searching, and retrieval of information (Salton, 1968)

- Primary focus: text
  - Scholarly papers, books, news, email, Web pages

- Other media:
  - Audio, images, video

Types of IR

- **Search (this course)**
- Filtering or tracking
  - Detecting documents of interest
- Classification
  - Labeling documents with pre-existing labels
- Question answering
  - Beyond searching for documents, searching for answers

Types of Search Engines

- **Web search (this course)**
- Vertical search
- Enterprise search
- Desktop search
- Peer-to-peer search

Database Engines vs. Search Engines

- What is the difference between a database search and a search engine?

Database Engines vs. Search Engines

- "Find accounts whose balance is greater than $100"
  - Requires data to be **well-structured** and **well-defined**
  - Typically involves not just **text** but **numbers**
  - Requires use of **formal language and logic**

- "Canvas UCI" or
  "Find how to merge two lists in Python"
  "Find the graph of y=log(x)"
  - **Vague** and varied information needs
  - **Informal** and **ad-hoc queries**
  - Requires engine to **infer meaning** of words and sentences

Core Issue: Relevance

- Textual similarity **[databases stop here !!]**
- Document context
  - Origin
  - Author
  - Popularity
  - …
- User context
  - Geographic location
  - Prior queries
  - Age
  - Preferred language
  - …
- Query context
  - Special symbols

Retrieval Models

- Conceptual designs of what to pay attention to when matching a query and a document
  - E.g. grammar vs. raw text, what parts of context, etc.
- Good retrieval model: finds documents relevant to the person who submitted the query
- Basis for *ranking algorithms*

Classic IR Assumptions

- **Corpus**: fixed document collection
- Goal: retrieve information content relevant to the information need

Classic IR Goal

- "Relevance"
  - For each **query Q**, and stored **document D**, **there exists** a relevance **score R(Q, D)**
  - **Maximize R(Q, D)**
    - Context is ignored
    - User is ignored
    - Corpus is static and centralized

Web IR

- The Web is huge (**cannot store** in any centralized memory!)
  - Corpus is not centralized!
- The Web changes all the time (**needs to update** constantly!)
  - Corpus is not static!
- There is information to avoid (adversarial IR! e.g. link bombing)
  - Context cannot be ignored!
- One interface for hugely divergent needs :
  - User cannot be ignored!