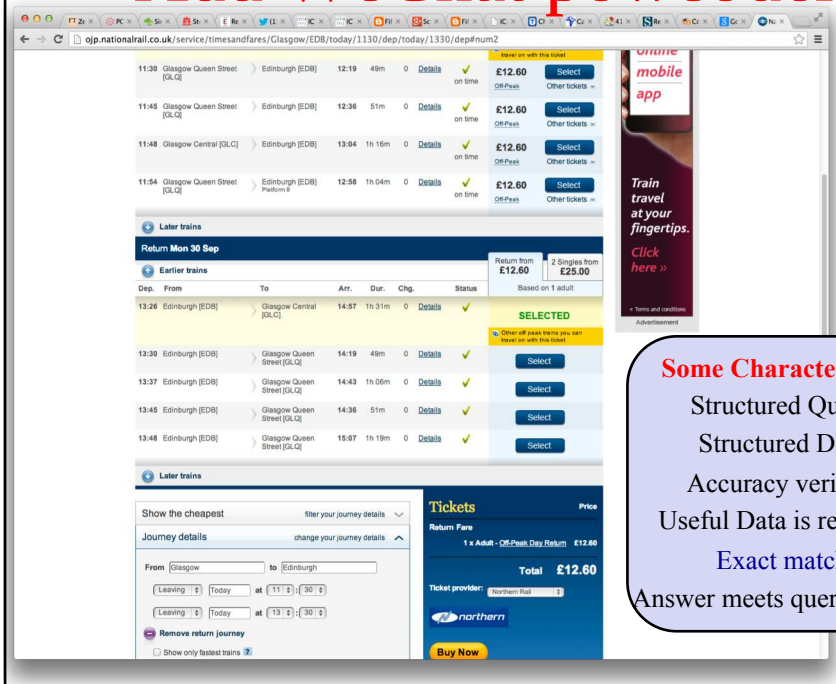
Introduction to Information Retrieval

Information Retrieval

Iadh Ounis

2022



Some Characteristics

Structured Query

Structured Data

Accuracy verified

Useful Data is returned

Exact match

Answer meets query criteria

2

Information Retrieval

Information Retrieval is the *science of search engines*

How best to address the information needs of users…

- *Effectively*: Get the right information to a user!
- *Efficiently*: Get it to users quickly!

Information Needs

4

Search Engines

- The most visible application of information retrieval technologies are search engines
  - Search engines have evolved since their initial conception 70 years ago

Search Engines

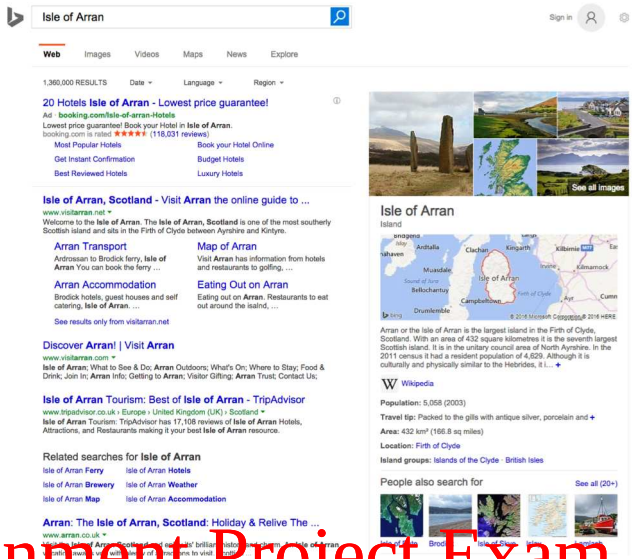
YAHOO! Google Baidu 百度 bing alltheweb find it all TEOMA Yandex — Why?

- Quite effective (at some things)
- Highly visible (some are very widely used)
- Commercially successful (some of them)
  - Google is one of the biggest corporations in the world
    - Underlying technology for searching …..
- What goes behind the scenes?
  - How do they work?

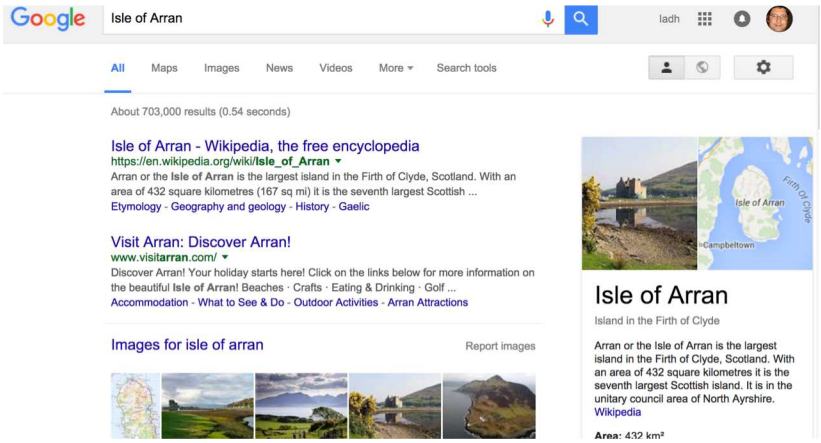
- Let us have a look at the commercial Search Systems!

6

On Bing

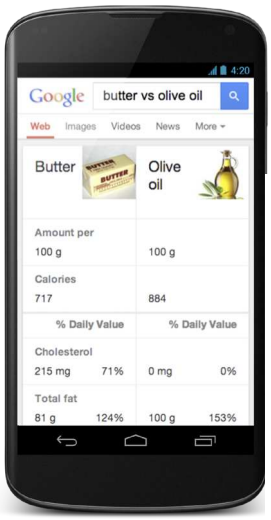


On Google



What features make Google & Bing different?

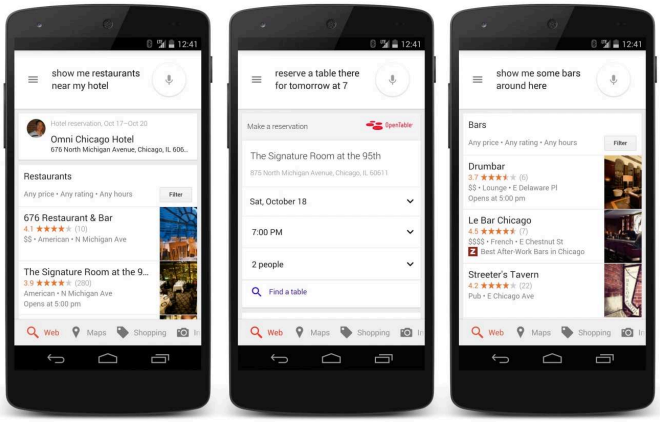
8



How do these systems work?

What are the commonalities & their functions?

Is there more to IR than Web Search?



In This Course, We Ask …

- What makes a system like Google or Bing tick?
  - How does it gather information?
  - What tricks does it use?
  - Expanding beyond the Web?
- How can those approaches be made better?
  - Natural language understanding?
  - Machine learning?
  - User interactions?
- What can we do to make things work quickly?
  - Fast computers? Caching?
  - Compression?
- How do we decide whether it works well?
  - For all queries? For special types of queries?
  - On every collection of information?
- What else can we do with the same approach?
  - Other media? Other tasks?

10

Definitions of Information Retrieval

- Salton, 1968
  - Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information
- Needham, 1977
  - The complexity arises from the impossibility of describing the content of a document, or the intent of request, precisely, or unambiguously
- General definition
  - Retrieval of **relevant information** from data sources which were not originally intended for access (e.g. unstructured data)
  - What does this mean?
    - Text (most often - e.g. searching newspaper articles or searching the Web)
    - Images… Video … Audio….

Documents vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes)
  - e.g. bank records with account numbers, balances, names, addresses, social security numbers, date of birth, etc.
- Easy to compare fields with well-defined semantics to a query in order to find matches

(Unstructured) text is more difficult

12

Imprecision in IR

- Most algorithms in Computer Science have a "right" answer. In contrast, a heuristic tries to guess something close to the right answer.
- Consider the three problems:
  - Sort the following ten integers
  - Find the highest integer
  - Find the beers made by X (i.e. SELECT … FROM …WHERE …)
- Now consider:
  - Find the documents most relevant to "hippos in the zoo".

IR techniques are essentially heuristics because we do not know the right answer

What Makes a Document Relevant?

- It contains the query terms?
- It contains ALL of the query terms?
- It contains the query terms many times? (…but what if it is a long document?)
- It contains the query terms many times in a short document?
- It contains the query terms close together?
- It is fresh/recent?
- It contains terms similar to the query terms?
- It is authoritative (has many links)?
- It doesn't contain too many ads?
- It doesn't contain too many different, unrelated words (e.g. spam)
- It has been clicked by many other people for the same query?

ALL of these are heuristics. They are not guaranteed to get a correct, relevant document for all users

14

Why to Retrieve?

Task type ?

Time available varies

- I need to find some information
  - Who is the head of college of science & engineering?
  - How to get to the LUX city centre from LUX airport?
  - *What is the upcoming topic in IR research?*
  - *What to do this weekend in Glasgow?*
- Where to search?
  - Newspaper articles, Web pages, Scholarly materials (ACM/IEEE Digital Library), Emails, Tweets, ..., Images (Flickr), …, Videos (YouTube)
  - … and many more? Including a combination of all in your own desktop, … or in your enterprise…. or external sources like the Web!

Exact Need ⟷ Vague need

15

What do We Mean by "Information"?

- How is it different from "Data"?
  - Information is data in context
  - Databases contain data and produce information
  - IR systems contain and provide information

- How is it different from "Knowledge"?
  - Knowledge is a basis for making decisions
  - Many "knowledge bases" contain decision rules

16

What Do We Mean by "Retrieval"?

- Find something that you are looking for:
  - **Ad hoc** search
    - Find documents "about this" topic-x
  - **Known item** search
    - Find the University of Glasgow home page
  - **Answer** seeking
    - What is the capital of Belgium?
  - Directed **exploration**
    - Who makes video conferencing systems?
  - **Decision** making
    - Best places to stay in Paris
  - **Expert** search
    - Who knows about Stable Marriage in my organisation?
  - Etc.

Role of Interaction

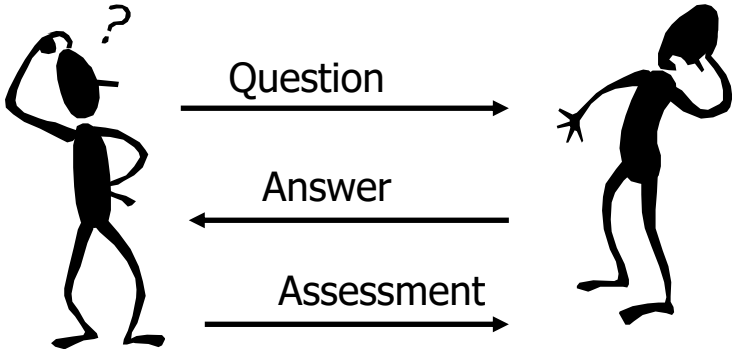
Scenarios & Applications

- Web search was a "killer app"
  - Developing advertising models on the web ....
- Today there are many retrieval applications:

- These types of content share common properties:
  - Text content + some metadata (e.g. title, author, date for papers; subject, sender, destination for email)

18

A Question-answer scenario



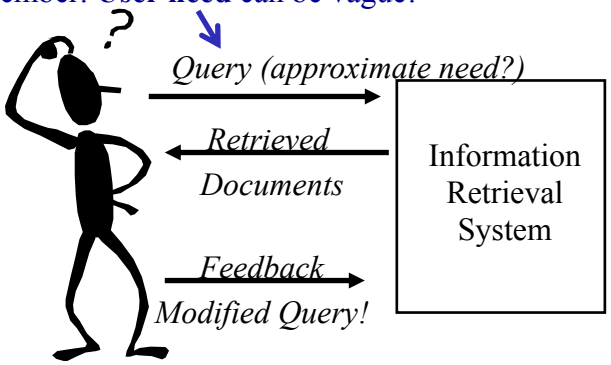
Note 1: Asking a good question can be as hard as answering it

Note 2: The objective of the **search engineer** is to automate the above process

Search/IR Engineers are increasingly sought in industry

Retrieval Process

Remember: **User need** can be vague!



Even if the information need is Exact- there is a problem!

IR System capabilities

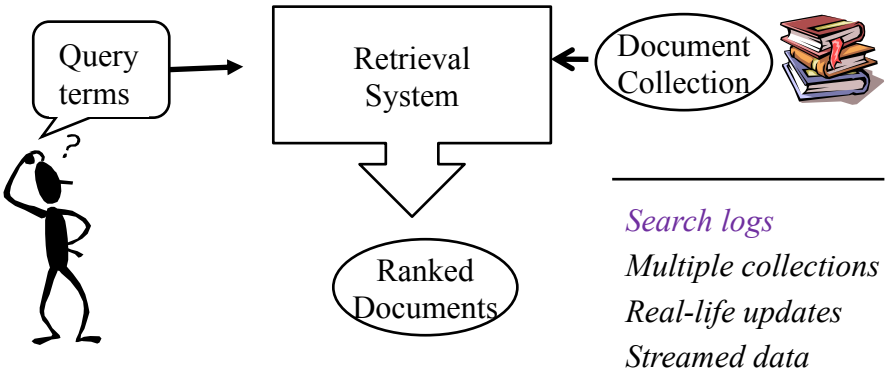
Given few query words

Infer what a user wants

Fetch relevant documents as fast as possible

Present to users in a way they understand

An IR System



Search logs
Multiple collections
Real-life updates
Streamed data
Social media

Relevance ... Effectiveness ... Efficiency

- Relevance
  - If the query and the document are **about** the same topic (known as the **Topical Relevance**) – Remember that a user's query can be vague!!
- Effectiveness
  - Looks into the **quality** of the retrieved set. Do they largely contain **relevant** documents? (role of **Retrieval Models**)
- Efficiency
  - Return results as fast as possible (role of search engine **architecture** - indexes) – architecture of the IR systems: distributing computation, updating indexes, etc.

22

IR H/M: General Information

- Structure
  - Zoom lectures: Fridays 10:00-12:00 & [14:00-15:00]
  - Labs: BO1028/Teams (2 groups alternating every lab) on Fridays 14:00-15:00 – Schedule of lab sessions will be posted on Moodle
  - Q/A forums on Teams
- Assessment
  - Assessed coursework (20%) & Final exams (80%)
    - Coursework: Ex1 (4%), Ex2 (8%), Ex3 (8%)
- Lecturer (s):

Iadh Ounis (Email: **iadh.ounis@glasgow.ac.uk**)
Craig Macdonald (Email: **craig.macdonald@glasgow.ac.uk)**
… & various tutors/guests


Yours truly


Craig

@iadh
craig_macdonald


Xiao


Sarawoot


Sasha


Sean

Objectives

- At the end of this course you will be able to…
  - Explain the process of retrieval
  - Build an IR system and deploy it for practical applications

    You may need a retrieval component in your own project work
  - Explain how IR systems are evaluated
  - Understand the web search engine architecture
  - Explain advanced IR technologies & Applications
    - Learning to rank; diversification; personalization, ...
- Empower you with a necessary set of skills
  - Develop and deploy IR systems or related technologies

24

Planned Svllabus

Fundamental Topics

WK 1 - *Starting on Monday 10th January* - Introduction and IR Systems Architecture

WK 2 - *Starting on Monday 17th January* - Term Weighting Models (Vector Space)

WK 3 - *Starting on Monday 24th January* - Evaluation of IR Systems

WK 4 - *Starting on Monday 31st January* - Interactive Retrieval & Relevance Feedback

WK 5 - *Starting on Monday 7th February* - *Advanced Retrieval Models (Probabilistic; Language Models; DFR, etc.)*

Advanced/Emerging Topics

WK 6 - *Starting on Monday 14th February* - Advanced IR Models; Machine Learning for IR & Learning to Rank

WK 7 - *Starting on Monday 21st February* - Learning to Rank; Web Search (introduction)

WK 8 - *Starting on Monday 28th February* - Web Search (Link Analysis, etc.); Web Search (Crawling, etc.)

WK 9 - *Starting on Monday 7th March* - *Web Search Evaluation;* Efficient IR (Compression, Pruning, DAAT/TAAT)

WK 10 - *Starting on Monday 14th March - Neural Networks for IR; Revision*

Teaching Resources

- Text Books – Recommended (available online!!)
  - Search Engines (Information Retrieval in Practice), Bruce Croft, Metzer, Strohman, Addison Wesley 2015
  - Introduction to Information Retrieval , Manning, Raghavan, Schutze (eds), Cambridge University Press 2008

- Course Web page is available on Moodle
  - Updated regularly as we go
  - News/material regarding the course will be posted on Moodle; Interactive quizzes

- Relevant (online) Seminars
  - Mondays, 3-4pm; All are welcome
  - Announced on the school web pages

26

First 3 Weeks: Look at the Basics of IR

- Architecture of the System
- Concepts of Relevance & Ranking
- Text Normalisation techniques
- IR Evaluation