Topic 0: Introduction

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Administration

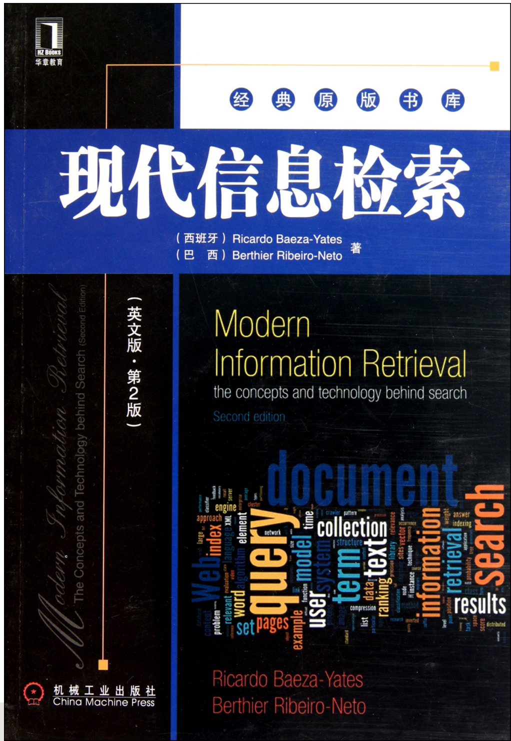
- **Lecturer:** David Lillis (david.lillis@ucd.ie)
- **Lectures:**
  - **Tuesdays** @ 1525 (Room 220, Teaching Building 4)
- **Labs:**
  - **Thursdays** @ 1330 (Room 512, Teaching Building 4). Starting Week 3 (15th March 2018)
- **Office Hours:**
  - By appointment - please email if you need to arrange a time to meet.

Virtual Learning Environment (VLE)

- We will use **Moodle** as the VLE for this module.
- I will post lecture slides and materials there throughout the semester.
- Located at: https://csmoodle.ucd.ie/moodle/course/view.php?id=653 (or scan the QR code)
- Enrolment key:
  - BDIC-IR-2018

Resources

- **Modern Information Retrieval** (2nd Edition) by Ricardo Baeza-Yates and Berthier Ribeiro-Neto (ISBN: 978-7-111-33174-2): in the library.
- **An Introduction to Information Retrieval** by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze: will be posted on Moodle.
- Academic Papers from conference proceedings (SIGIR, ECIR, TREC etc.): will be posted on Moodle.


经典原版书库
现代信息检索
（西班牙）Ricardo Baeza-Yates
（巴　西）Berthier Ribeiro-Neto 著
（英文版・第2版）
Modern Information Retrieval
the concepts and technology behind search
Second edition
Ricardo Baeza-Yates
Berthier Ribeiro-Neto
机械工业出版社
China Machine Press

However…

- A lecture is **not** just PowerPoint slides. It also consists of:
  - Everything I say during the lecture.
  - Any work that I do on the blackboard.
- Every student should bring a notebook to class:
  - Take notes on things that are not in the lecture slides.
  - Write down your own understanding of key points.
  - You remember more if you write things down.

What is Information Retrieval?

- For now, think of an internet search engine.
- As a user, there is some information that you want to know.
- The search engine has access to billions of web pages of information.
  - Its job is to let you know which pages contain the information that you are looking for.
- You describe what you want using a **query**.
- We will look at a more formal definition later.

Module Outline
(subject to change)

1. Introduction
2. A crash-course in the Python programming language.
3. The Information Retrieval (IR) Process
4. Data Structures for IR
5. IR Techniques
6. Evaluation of IR
7. Advanced Topic: Collection and Data Fusion
8. Advanced Topic: Web Search
9. Review

Assessment

- Final Exam (80% of final grade)
- Programming Assignment (20% of final grade)
  - Implement a simple search engine.
  - Before the programming assignment, we will use the labs to prepare you for this task.

What is Information Retrieval?

What is Information Retrieval (IR)?

- As computer and internet users, we use IR systems regularly:
  - Web Search Engines (e.g. Baidu, Bing, Google, etc.)
  - Desktop Search
  - Mobile Search
  - Library Catalogue Searches
  - Searching Individual Web Sites (e.g. newspaper archives, company document repositories, etc.)

What is Information Retrieval?

Information Retrieval (IR) deals with the **representation**, **storage**, **organization of**, and **access to** information items.

The representation and organization of the information items should provide the user with easy access to **the information in which he is interested**.

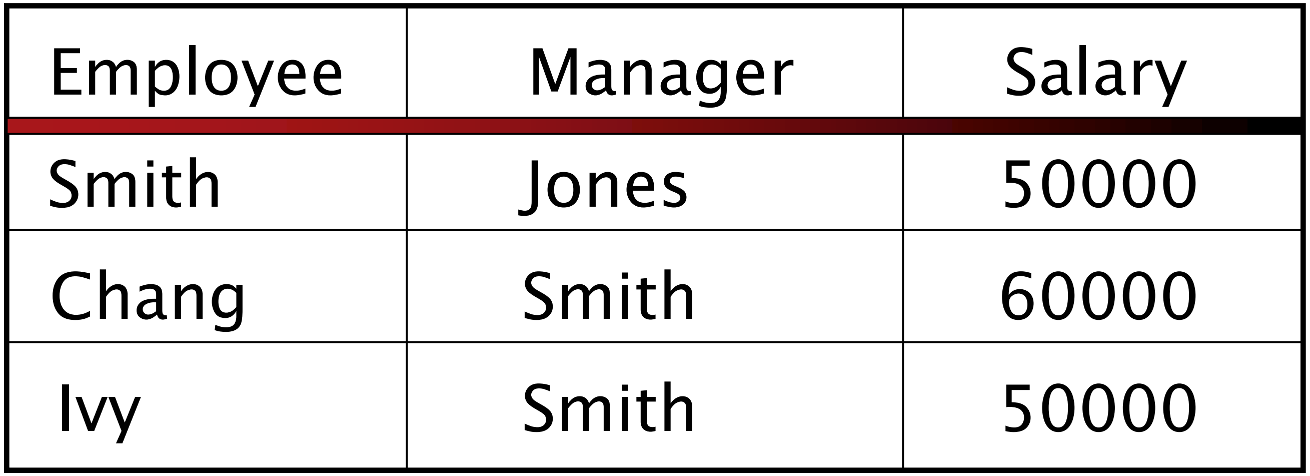
- Baeza-Yates and Ribeiro-Neto

What is Information Retrieval? Information Items

- This definition provides us with a number of issues that need to be addressed when considering what constitutes an IR system:
    - **Information Item:** Historically, “information items” have been books, documents or other written material that contain information in an **unstructured form**.
        - In IR, we frequently use the word “document” to refer to these items. However, in more modern times, IR systems have developed to cover other forms of information, and may be asked to include such things as video, images, etc.

What is Information Retrieval? Structured vs Unstructured Data

- Structured data tends to refer to information in “tables” (e.g. database, spreadsheet).



Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,
Salary < 60000 AND Manager = 'Smith'.

What is Information Retrieval? Structured vs Unstructured Data

- Unstructured data:
  - Typically refers to **free-form text** written in **natural language**.
    - "Natural" languages are those used by humans to communicate, e.g. English, Mandarin Chinese, French, etc.
  - Allows for:
    - **Keyword queries**, including operators (AND, OR, NOT, etc.)
    - More sophisticated **"concept" queries**, e.g.,
      - "find all web pages dealing with *drug abuse*".
  - In reality, most documents are **semi-structured** (e.g. we can identify the headings, titles and body text).
    - Also, it often fits the structure of the language it is written in.

What is Information Retrieval?: Representation

- **Representation:** In order to allow users to perform searches, documents must be **represented** in some way.
  - Searching through the raw text of millions of documents is a **very slow** process, so some **mathematical representation** of the information is typically used.
  - **Why?**
    - Computers are much quicker at doing mathematical calculations than they are at text processing.
    - We try to represent textual documents in some mathematical form so retrieval is faster (e.g. vectors, sets, bit-strings).

What is Information Retrieval? Storage, Organisation and Access

- **Storage and Organisation:** There are important considerations in terms of storing documents also. They must be organised so as to be **quickly accessible**. This is particularly important with large document collections.

- **Access:** The key attribute of an IR system is to allow users to access the information in question. This should happen in a **timely** and **efficient** manner.

What is Information Retrieval? Information Need

The representation and organization of the information items should provide the user with easy access to **the information in which he is interested**.

- Users only use IR systems when there is some information that they are interested in reading.
- We call this the **"information need"** of a user.
- This phrase was coined by Robert Taylor in 1962, and is defined as having four stages.

Information Need: 4 Stages

1. the actual, but unexpressed need for information **(visceral need)**: “a vague sort of dissatisfaction ... probably inexpressible in linguistic terms”.

2. the conscious within-brain description of the need **(conscious need)**: “an ambiguous and rambling statement”. i.e. “I know what I’m looking for but I don’t know how to say it”.

3. the formal statement of the question **(formalised need)**

4. the question as presented to the information system **(compromised need)**

Expressing an Information Need

- The third stage, the *formalised need*, is typically easy to express in a natural language:
  - What is the capital city of Cyprus?
  - Who are the 10 best golfers in the world?
  - What common arguments are made in favour of, or against, the use of nuclear power?
  - What were the effects of Barack Obama’s policies as President of the USA?

Presenting an Information Need to an IR System: Queries

- A **query** is the expression of an information need that is provided to an IR system to explain what information is required by the user. There are a number of different methods of expressing queries:
  - **Keyword-Based Querying:** A keyword (or list of keywords) is supplied to the IR system. This is by far the most common form of querying in web search. The average query consists of 2-3 keywords.
  - **Context Queries:** These specify sequences of words that should appear close together in documents that are retrieved. Physical proximity has semantic value.

Presenting an Information Need to an IR System: Queries

- **Types of Queries (continued)**
  - **Boolean Queries:** The oldest method of combining keywords allows a user to specify keywords that should or should not be contained in the documents that are desired, using the Boolean operators AND, OR and NOT.
    - e.g. “information AND (retrieval OR extraction) NOT data”.

Presenting an Information Need to an IR System: Queries

- **Types of Queries (continued)**
  - **Natural Language Queries:** Queries are provided in their natural form. To date, this form of querying has not been very successful, as Natural Language Processing is a difficult task.
    - For a long time, http://www.askjeeves.com was the most famous natural language web search engine, but it reverted to keyword-based searching in 2006.
    - Microsoft spent $100m to acquire http://www.powerset.com in 2008 in a new attempt at the concept.
    - Now, many search engines will try to automatically detect natural language queries, but support keyword search also.
      - Most voice-activated interfaces support natural language queries (e.g. Apple's Siri, Amazon's Alexa), although their understanding is often quite basic.

Example: the Boolean style query

- William Shakespeare was a famous English playwright. Suppose we want to search for characters in his plays.
- Which plays of Shakespeare contain the words ***Brutus*** *AND* ***Caesar*** but *NOT* ***Calpurnia***?
- As a first attempt, we could read the plays line-by-line to find those that contain ***Brutus*** and ***Caesar*** and then remove those that contain ***Calpurnia***.

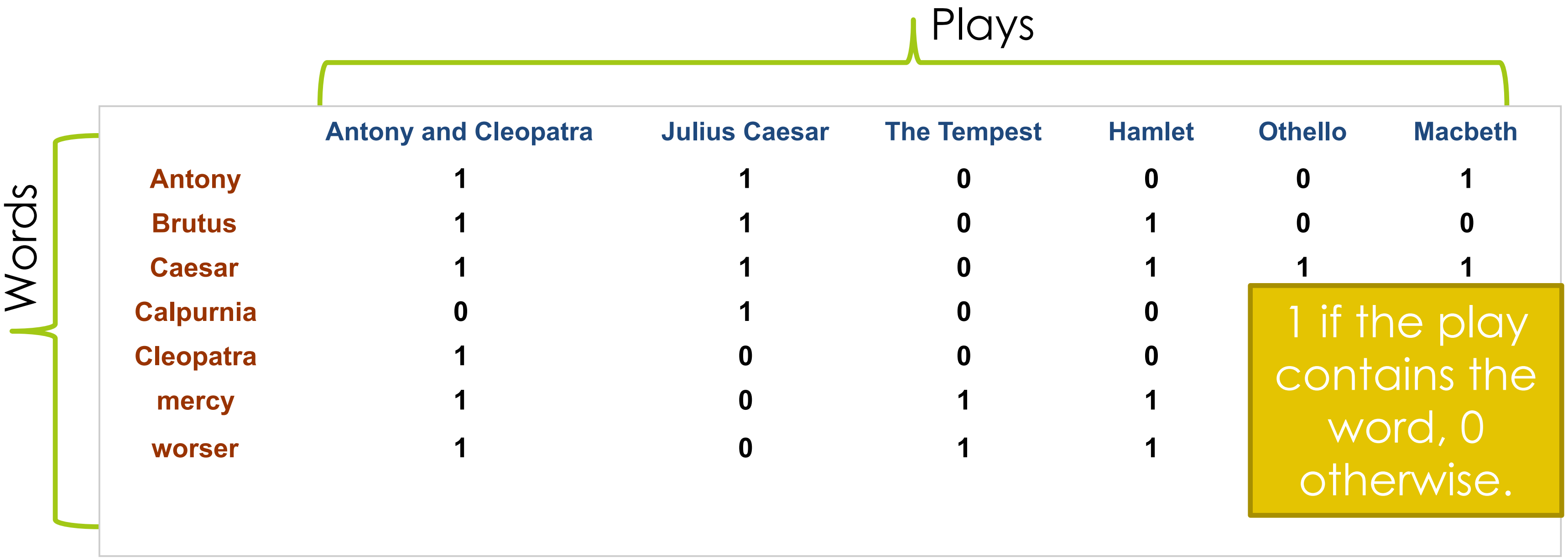
Example: the Boolean style query

- *"… we could read the plays line-by-line to find those that contain **Brutus** and **Caesar** and then remove those that contain **Calpurnia**."*
- Why is that not the answer?
  - Slow (for large corpora)
  - Other operations (e.g., find the word ***Romans*** near ***countrymen***) not feasible.
  - Ranked retrieval not possible (best documents shown first)
    - Later lectures!

A **corpus** is a collection of documents. The plural is **corpora.**

Example: The Boolean style query Term-document incidence matrix

- Instead, we store the information we need in some sort of data structure. Here is a **term-document incidence matrix** showing some of the words contained in some of Shakespeare's plays.



Plays

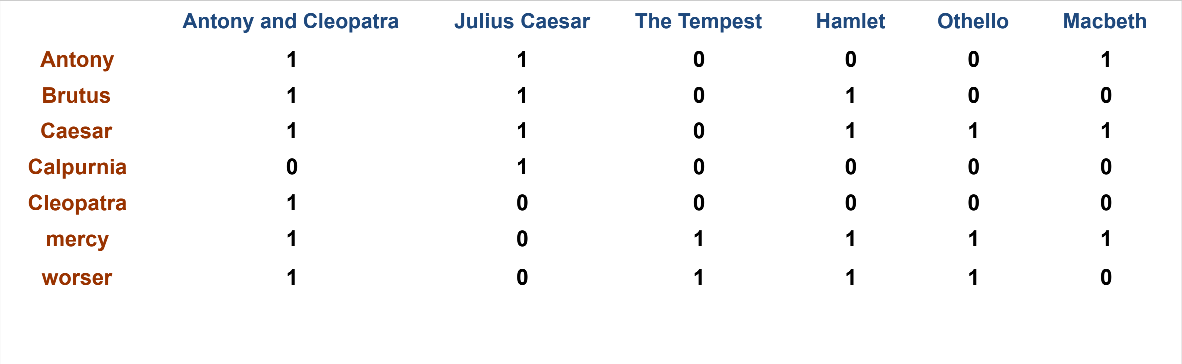
Words

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0		
Cleopatra	1	0	0	0		
mercy	1	0	1	1		
worser	1	0	1	1		

1 if the play contains the word, 0 otherwise.

Example: The Boolean style query Incidence vectors

- So we have an **incidence vector** for each word.
  - It consists of 1s (for the plays it appears in) and 0s (for those it does not appear in).
  - e.g.
    - ***Brutus:*** *110100*
    - ***Caesar****: 110111*
    - ***Calpurnia:*** 010000



	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Example: The Boolean style query Operators

- Our query is: ***Brutus*** *AND* ***Caesar*** *NOT* ***Calpurnia***
- To get "*NOT* ***Calpurnia***", we get the **complement** of the incidence vector for ***Calpurnia*** (using the bitwise NOT operator, which changes all 1s to 0s and all 0s to 1s).
  - 001000 → 110111
- Now we can use the bitwise AND operator to combine our three vectors:
  - 110100 AND 110111 AND 101111 = 100100
- Referring back to the term-document incidence matrix, we see that the answer is: **Antony and Cleopatra, Hamlet**

Bigger collections

- This approach can be effective, but how well does it **scale** to larger collections?
  - Consider *N* = 1,000,000 documents, each with about 1,000 words.
  - Average 6 bytes per word including spaces and punctuation.
  - 6GB of data in the documents.
  - Say there are *M* = 500,000 *distinct* terms among these.

Can't build the matrix

- 500,000 x 1,000,000 matrix has **half-a-trillion** 0s and 1s.
- But it has no more than one **billion** 1's.
  - matrix is extremely **sparse**.

Why?

- What's a better representation?
  - We only record the 1 positions, and not the 0s.
  - We will look at this in a later lecture.

Another problem: Ambiguous Queries

- Sometimes it is difficult to figure out what the information need was if we can only see the query: some queries are **ambiguous**.

- For example, if a user searches for “jaguar”, documents that discuss luxury cars may appear to be relevant, but will be of no use to a user who is researching big cats.

- Similarly, a search for “bank” could be:
  A river bank.
  A financial institution.
  A manoeuvre made by an aeroplane.

The Role of an IR System

The Role of an IR System

> An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquery. It merely informs on the **existence** (or non-existence) and **whereabouts** of documents relating to his request.

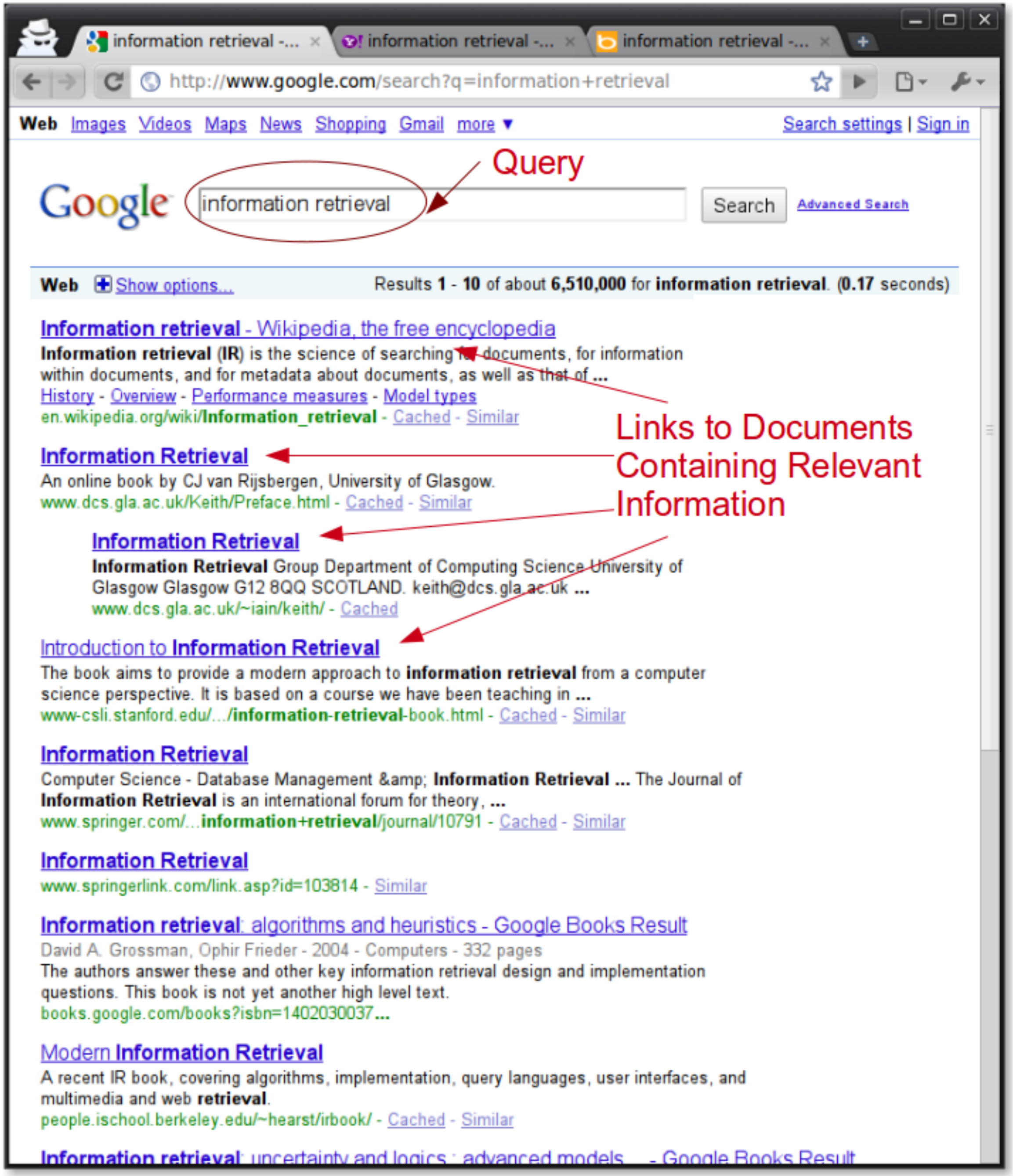
– C. J. van Rijsbergen

- **Existence:** Tells the user that a document exists that helps to satisfy the information need.
- **Whereabouts:** Tells the user where that document is.

- The user learns nothing unless they then read the document.

The Role of an IR System

- This helps us to define the role of an IR system:
  - The principal function of an IR system is to give users a **list of documents** that are **relevant** to the user's information need.
  - It makes no attempt to take a subset of the information contained in a document and present that to the user: it is the task of the searcher to do this.
  - With most systems, this list is ranked, with the document the system believes to be of most relevance at the top of the list.


information retrieval -... ×
information retrieval -... ×
information retrieval -... ×
http://www.google.com/search?q=information+retrieval
Web Images Videos Maps News Shopping Gmail more ▼
Search settings | Sign in
Google
information retrieval
Search
Advanced Search

Query

Links to Documents Containing Relevant Information

Web ⊞ Show options... Results **1** - **10** of about **6,510,000** for **information retrieval**. (**0.17** seconds)

Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (**IR**) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
History - Overview - Performance measures - Model types
en.wikipedia.org/wiki/**Information_retrieval** - Cached - Similar

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - Cached - Similar

Information Retrieval
Information Retrieval Group Department of Computing Science University of Glasgow Glasgow G12 8QQ SCOTLAND. keith@dcs.gla.ac.uk ...
www.dcs.gla.ac.uk/~iain/keith/ - Cached

Introduction to **Information Retrieval**
The book aims to provide a modern approach to **information retrieval** from a computer science perspective. It is based on a course we have been teaching in ...
www-csli.stanford.edu/.../**information-retrieval**-book.html - Cached - Similar

Information Retrieval
Computer Science - Database Management &amp; **Information Retrieval** ... The Journal of **Information Retrieval** is an international forum for theory, ...
www.springer.com/...**information+retrieval**/journal/10791 - Cached - Similar

Information Retrieval
www.springerlink.com/link.asp?id=103814 - Similar

Information retrieval: algorithms and heuristics - Google Books Result
David A. Grossman, Ophir Frieder - 2004 - Computers - 332 pages
The authors answer these and other key information retrieval design and implementation questions. This book is not yet another high level text.
books.google.com/books?isbn=1402030037...

Modern **Information Retrieval**
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web **retrieval**.
people.ischool.berkeley.edu/~hearst/irbook/ - Cached - Similar

Information retrieval: uncertainty and logics : advanced models ... - Google Books Result

Why is Information Retrieval Important?

- The term “Information Overload” coined by Alvin Toffler as far back as 1970.
- This describes the situation where there is so much information being made available to us that it is impossible for people to absorb all of it.
- This problem has become much worse as a result of the Internet, with millions of people publishing material every day.
- IR systems are vital in providing users with access to this information
- In March 2013, Baidu was reported to be handling 5 billion searches every day across all its services.

Google Searches over Time

Year	Number of Searches (approx)
19998	1,000,000,000 (1 billion)
2000	14,000,000,000 (14 billion)
2001-2003	Over 55,000,000,000 (55 billion)
2004-2008	73,000,000,000 (73 billion)
2009	Over 365,000,000,000 (365 billion)
2012-2015	1,200,000,000,000 (1.2 trillion)
2016	Over 2,000,000,000,000 (2 trillion)

Relevance

- This concept of "**relevance**" is very important in IR:
  - A document is ”relevant” if its content satisfies (or helps to satisfy) the user's information need.
  - The system alerts the user to the existence and location of a document: it's up to the user to read the document.
  - It is important to note that this is, in theory, a **subjective** concept, that is entirely up to the user to judge.
  - Remember that a query is just an expression of an information need.
    - For example, if a user searches for "jaguar", documents that discuss luxury cars may appear to be relevant, but will be of no use to a user who is researching big cats.

How do we know if the system is good?

- An IR system has been successful if the documents it returns satisfy the information need.
- The only person qualified to judge this is the user, which is a problem when trying to evaluate IR systems.
- Laboratory experiments use "test collections" for which **standard queries** have been created.
- Documents in these collections have been **judged** by human judges as to whether or not they are relevant to each query.
- This method becomes more difficult as document collections grow.

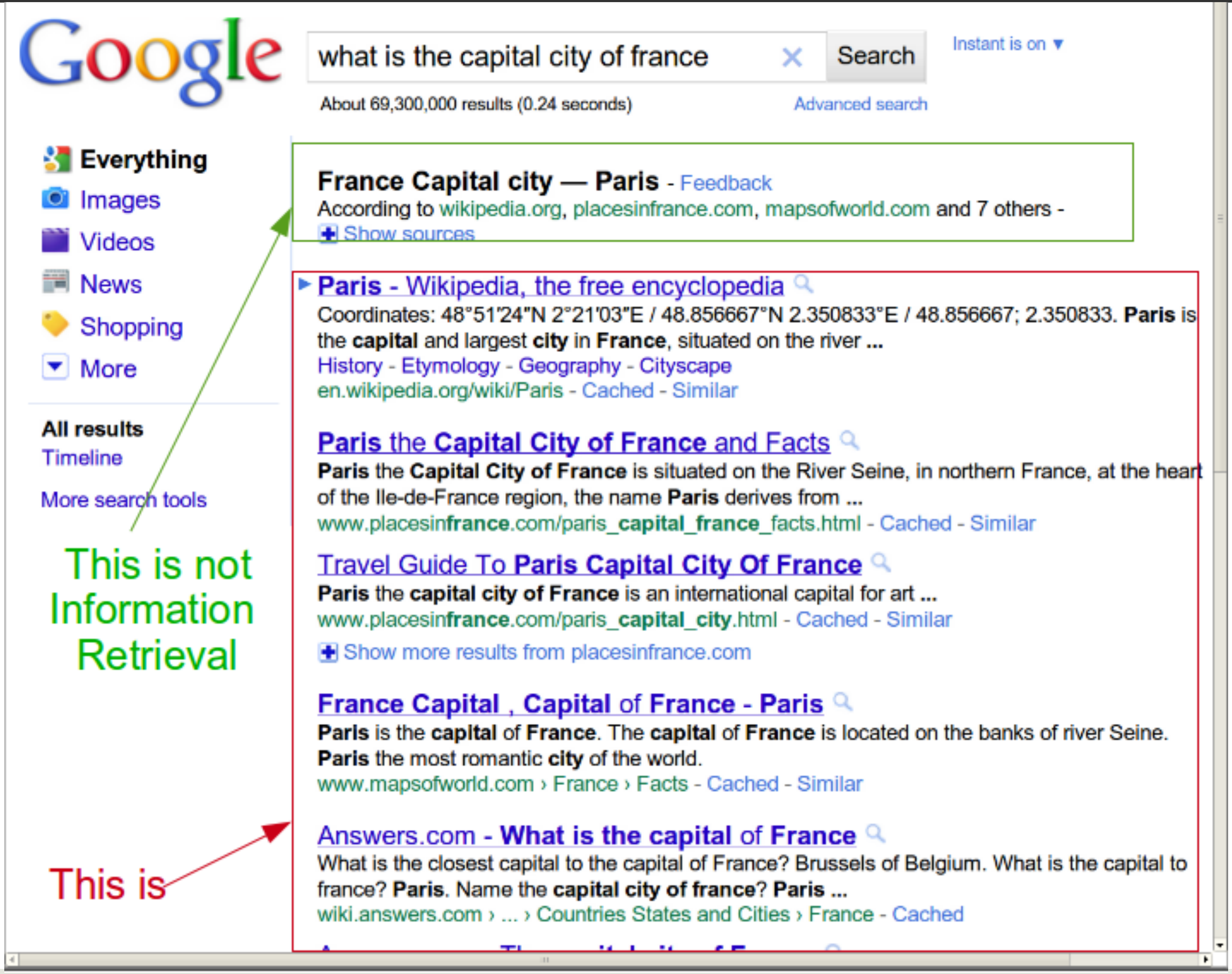
Related Research

There are a number of other research areas that are not traditionally considered to be “Information Retrieval”, but are closely related.

Question Answering

- **Question Answering:** Here, the system is designed to provide a full answer to a question, e.g.:
  - Question: What is the capital city of France?
  - Answer: Paris is the capital city of France.
- Be careful! Many mainstream search engines include some element of question answering as well as their Information Retrieval system.

Question Answering



About 69,300,000 results (0.24 seconds)

France Capital city — Paris - Feedback
According to wikipedia.org, placesinfrance.com, mapsofworld.com and 7 others -
Show sources

Paris - Wikipedia, the free encyclopedia
Coordinates: 48°51′24″N 2°21′03″E / 48.856667°N 2.350833°E / 48.856667; 2.350833. **Paris** is the **capital** and largest **city** in **France**, situated on the river ...
History - Etymology - Geography - Cityscape
en.wikipedia.org/wiki/Paris - Cached - Similar

Paris the **Capital City of France** and Facts
Paris the **Capital City of France** is situated on the River Seine, in northern France, at the heart of the Ile-de-France region, the name **Paris** derives from ...
www.placesin**france**.com/paris_**capital**_**france**_facts.html - Cached - Similar

Travel Guide To **Paris Capital City Of France**
Paris the **capital city of France** is an international capital for art ...
www.placesin**france**.com/paris_**capital**_**city**.html - Cached - Similar
Show more results from placesinfrance.com

France Capital , **Capital** of **France** - **Paris**
Paris is the **capital** of **France**. The **capital** of **France** is located on the banks of river Seine. **Paris** the most romantic **city** of the world.
www.mapsofworld.com › France › Facts - Cached - Similar

Answers.com - **What is the capital** of **France**
What is the closest capital to the capital of France? Brussels of Belgium. What is the capital to france? **Paris**. Name the **capital city of france**? **Paris** ...
wiki.answers.com › ... › Countries States and Cities › France - Cached

This is not Information Retrieval

This is

Information Extraction

- An **Information Extraction** system is designed to take unstructured text and try to create structured data from it.
- **Input:**
  - "Kate Hudson was born in Los Angeles, California, the daughter of Academy Award-winning actress Goldie Hawn and Bill Hudson, an actor, comedian, and musician. Her parents divorced eighteen months after her birth; she and her brother, actor Oliver Hudson, were raised in Colorado by her mother and her mother's long-time boyfriend, actor Kurt Russell."
- **Output (examples):** mother(Kate Hudson, Goldie Hawn), brother(Kate Hudson, Oliver Hudson), gender(Kate Hudson, female)

Overview

- What is Information Retrieval?
    - Unstructured data.
    - Representation, storage, organisation and access.
    - Information Need vs. Queries
        - Boolean query example.
        - Data structures are important!
    - The role of an IR system.
    - Why is IR important?
    - Related research.