uc3m | Universidad Carlos III de Madrid

Lesson 2: RETRIEVAL AND ORGANIZATION OF INFORMATION

s

Library and Information Science Department

Contents

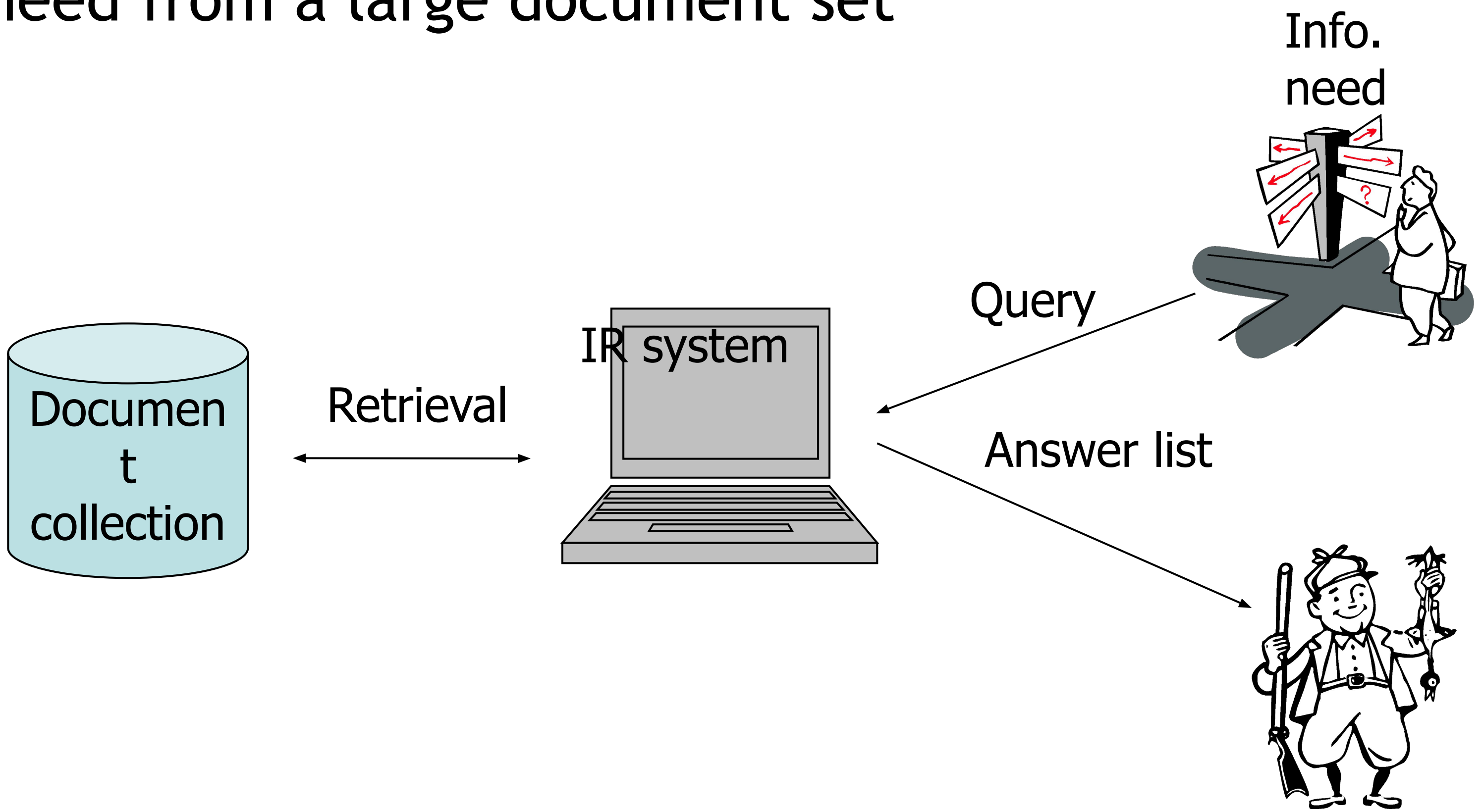
- □ Basic concepts of Information retrieval (pertinence, relevance, reliability, precision noise, silence, bias, etc. Deep Internet)
- □ Seeking and discovering digital information:
  - ▪ How to use a search engine: tools, utilities and recommendations
  - ▪ Search strategies for search engines
- □ Multidisciplinary databases.
- □ Internet search tools.

Outline

- □ What everybody knows about online searching...
- □ Search what?
- □ How to search?
- □ Where?
  - ▪ Selected resources versus Google etc.
  - ▪ Types of resources
- □ Search skills:
  - ▪ Preparation
  - ▪ Strategy
  - ▪ Query Formulation
  - ▪ Refining

Outline: the problem of IR

- □ **Goal** = find documents *relevant* to an information need from a large document set



Info. need

Query

IR system

Retrieval

Document collection

Answer list

Basic concepts of Information retrieval

Main problems in IR

- Document and query indexing
  - How to best represent their contents?
- Query evaluation (or retrieval process)
  - To what extent does a document correspond to a query?
- System evaluation
  - How good is a system?
  - Are the retrieved documents relevant? (precision)
  - Are all the relevant documents retrieved? (recall)

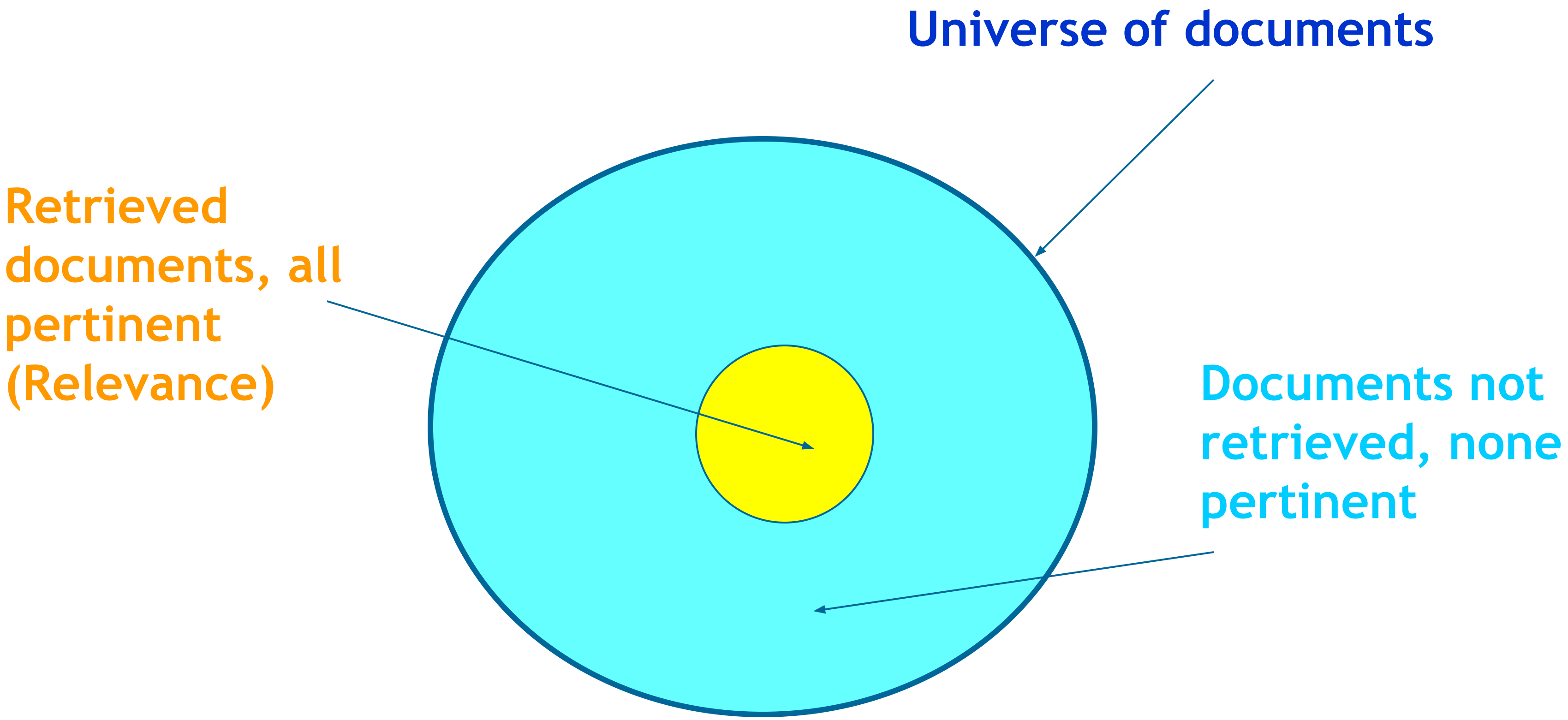
Relevance / Pertinence (Korfhage 1997)

- **Relevance:** Effective retrieved documents bearing the searched word (objective relevance)

- **Pertinence:** A retrieved document is useful for a particular information need (subjective relevance)

Strategy design: Success

What we dream of: the perfect strategy.



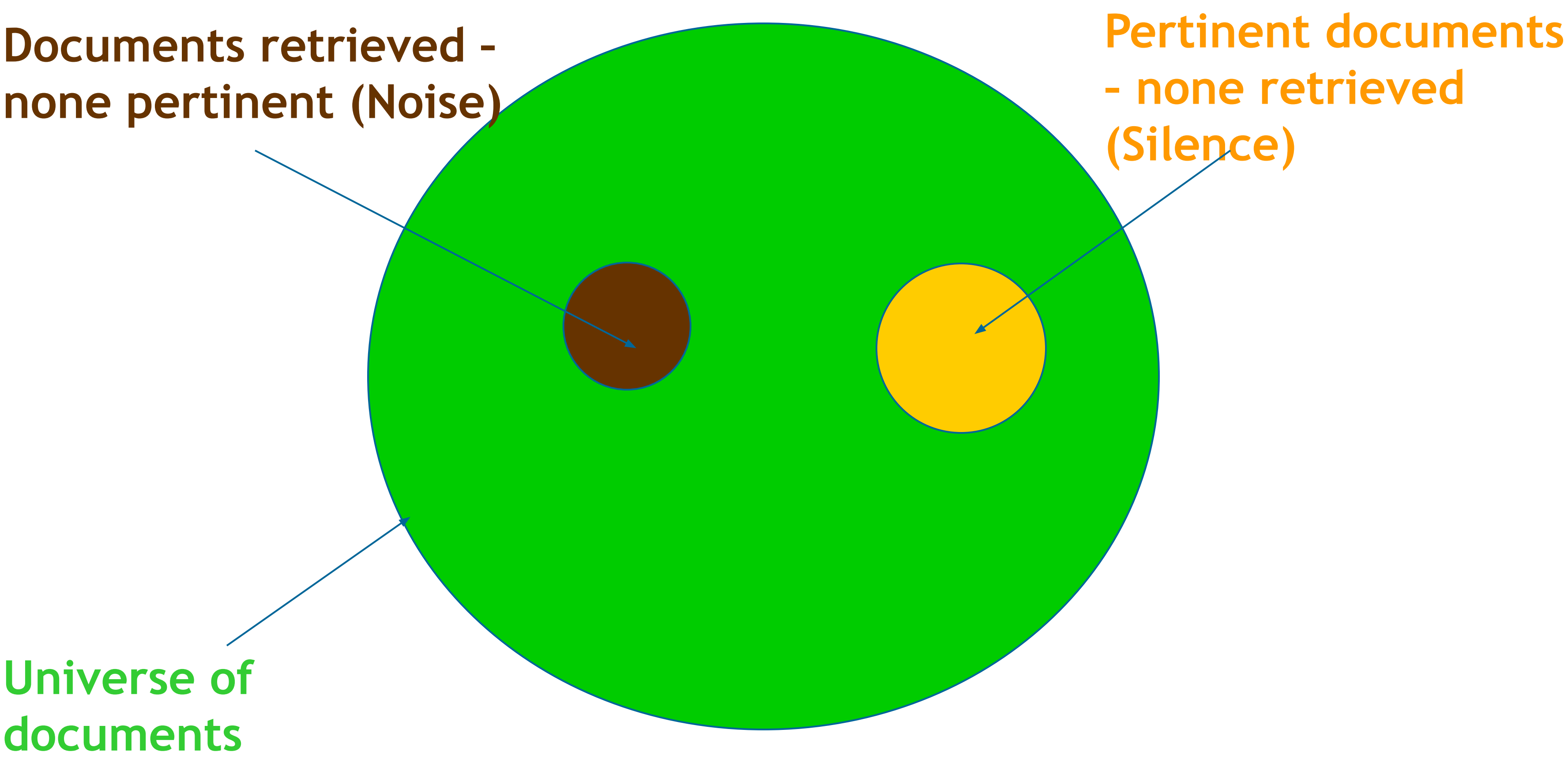
Universe of documents

Retrieved documents, all pertinent (Relevance)

Documents not retrieved, none pertinent

Strategy design: Failure

What you obtain sometimes: the worst possible case



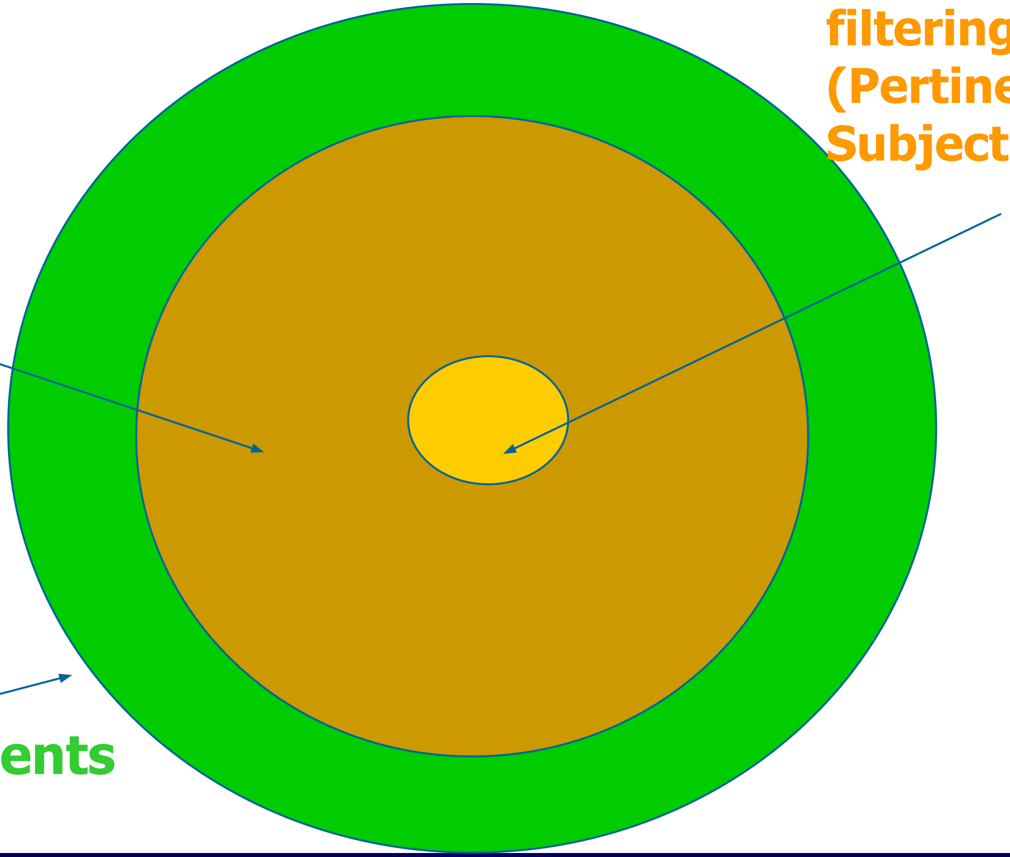
Documents retrieved - none pertinent (Noise)

Pertinent documents - none retrieved (Silence)

Universe of documents

Strategy design: Frequent case

A labor – intensive search:



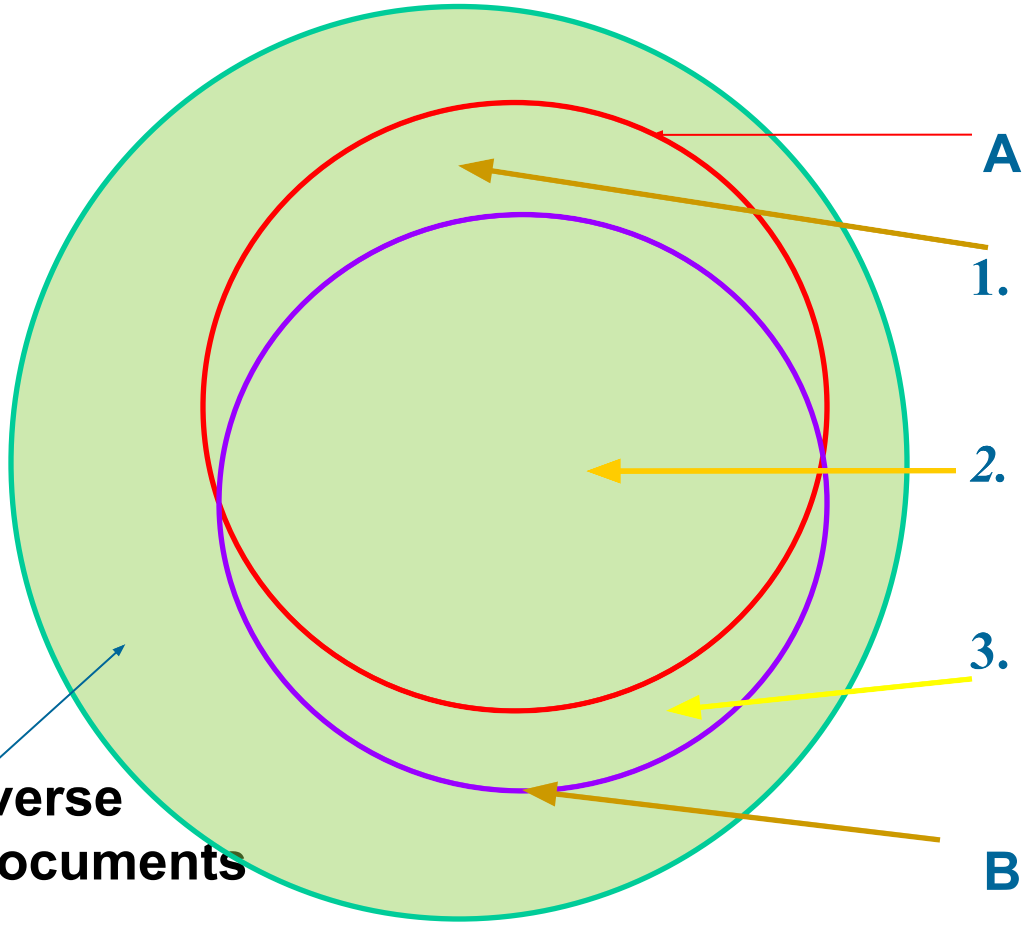
Pertinent documents retrieved after filtering results (Pertinence: Subjective relevance

All retrieved documents (Objective relevance)

Universe of documents

Strategy design: Our goals

Maximize 2
Minimize 1 & 3



A. Retrieved documents

1. **Retrieved documents – not pertinent (noise)**
2. ***Retrieved documents of interest (pertinence)***
3. **Pertinent documents – not retrieved (silence)**

B. Pertinent documents

IR concepts

- Relevance: results fulfill your query,
- Pertinence: results fulfill your information need
- Reliability: you can trust the quality of what you find,
- Recall: you retrieved a good % of what exists,
- Precision: you get only what you want, not much is irrelevant
- Noise: you get a lot of irrelevant hits
- Silence: You don’t get anything, you miss relevant hits
- Bias: you get only partial aspects of what’s available.

Recall

- In information retrieval, a measure of the effectiveness of a search
- Expressed as the ratio of the number of relevant records or documents retrieved in response to the query to the total number of relevant records or documents in the database
- In a database containing 100 records relevant to the topic "accounting" a search retrieving 50 records, 25 of which are relevant to the topic, would have 25 percent recall (25/100).

$$\text{Recall} = \frac{\text{Relevant retrieved documents}}{\text{Relevant documents in the system}}$$

Precision

- In information retrieval, a measure of search effectiveness, expressed as the ratio of relevant records or documents retrieved from a database to the total number retrieved in response to the query
- Ex. in a database containing 100 records relevant to the topic “accounting," a search retrieving 50 records, 25 of which are relevant to the topic, would have 50 percent precision (25/50).

$$\text{Precision} = \frac{\text{Relevant retrieved documents}}{\text{Total of retrieved documents}}$$

Noise

- No-relevant documents retrieved / the total of retrieved documents
- It is the inverse concept of precision
- To avoid noise:
    - Use specific terms
    - Use operators (AND & NOT)
    - Use search by phrase
    - Avoid confusing words (polysemy)
    - Make a good querying strategy

Silence

- □ Amount of relevant documents not retrieved / total of existing relevant documents
- □ It is the inverse of the recall.
- □ To avoid silence, we must:
  - Use operator OR
  - Use different varieties of a word (different languages)
  - Use query expansion (synonyms, etc.)

Search and discovering digital information: search engines

What is a Search engine?

- Several names: spiders, robots, bots, search engines, agents, web wanderers, wanders, web crawlers, engines, web ants, indexes, directories, etc.
- The most common/accepted name at international level is **Search engine**.
- A search engine is a software or set of software used for locating documents and information through the WWW.
- It does an automatic indexing of the web and record the web pages in a data base to retrieve them later.

Some examples

- □ Google: http://www.google.com - http://www.google.es
- □ Yahoo! Search: http://search.yahoo.com/ http://es.search.yahoo.com
- □ Bing: http://www.bing.com
- □ Altavista (what happen with the old one?): http://www.altavista.com - http://es.altavista.com
- □ Ask.com: http://www.ask.com - http://es.ask.com
- □ Gigablast: http://www.gigablast.com

Some examples

- Everyday new search engines appear…
- Everyday search engines disappears (Ex: Wisenut case, MSDewey case)
- Everyday some search engines are transformed
- Best resources to know what is going on about search engines world and search business are:
  - Search Engine Watch: http://searchenginewatch.com/
  - Alexa: http://www.alexa.com/
- Directories of search engines:
  - http://www.searchenginecolossus.com/
  - http://www.searchenginesdir.com

Search engines features

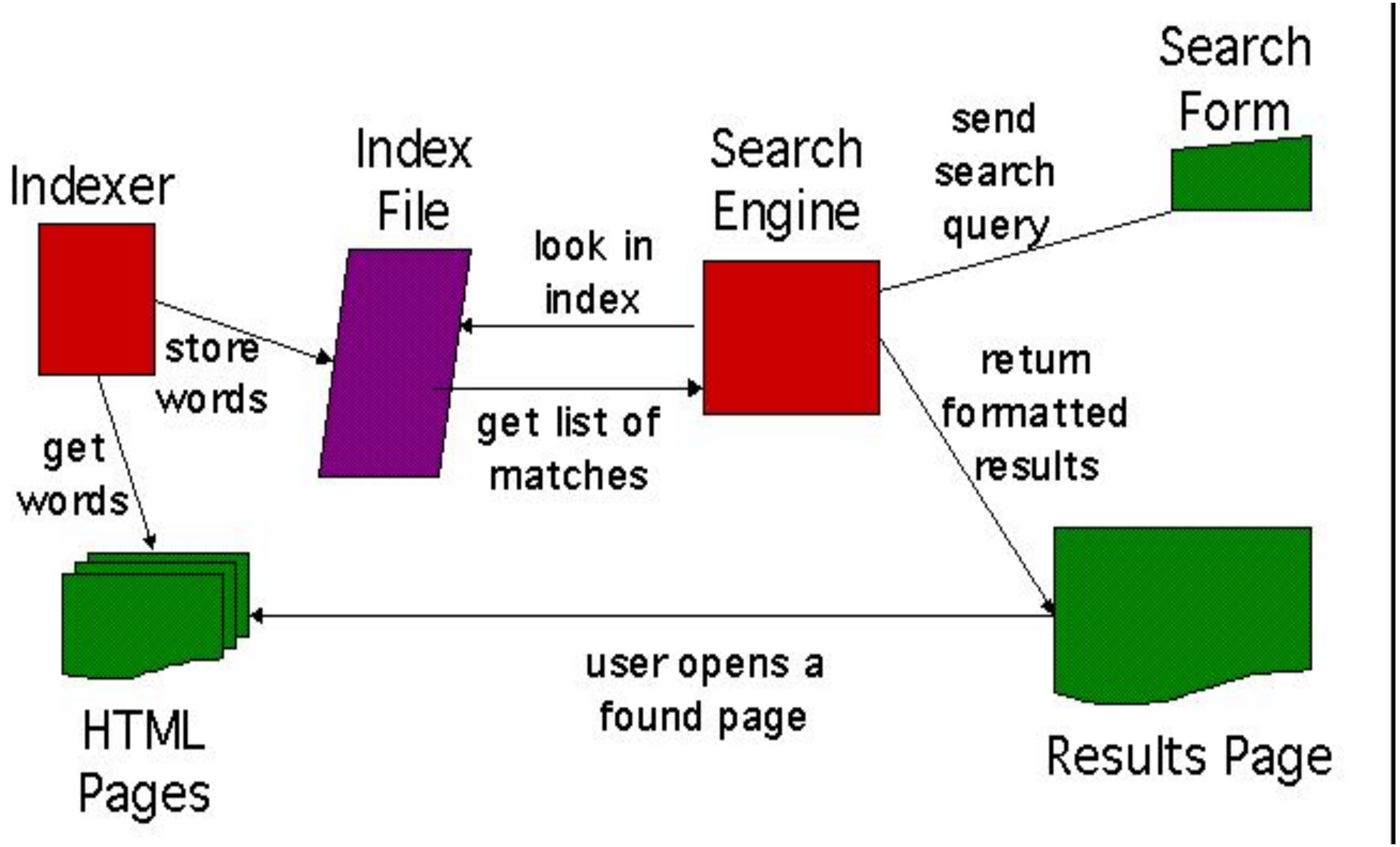
- Search systems based upon a software or robot that automatically indexes the Web.
- A Web search engine is a tool designed to search for information on the World Wide Web.
- Search results are usually presented in a list and are commonly called hits.
- The information may consist of web pages, images, information and other types of files.
- Some search engines also mine data available in news, books, databases, or open directories.
- Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

Search engines: components and access

- Search engines’ components:
  - a robot
  - automatic systems of analysis and indexing
  - a data base
  - a query system and query language
  - a Web interface

- Access:
  - Search by keywords introduced by a search interface
  - Sometimes we can search also by some fields

Search engines: components and access



IR through search engines

☺ Advantages:

☺ Exhaustivity

☺ We can find very specific resources

☹ Disadvantages:

☹ Variability of quality, no evaluation

☹ A lot of noise: too much results and sometimes duplicates

Search engines: scope and quality

□ Scope:

- Indexing all kind of web pages (text), also some other kind of internet resources (images, audio, video, news, yellow pages, blogs, rss, etc.)
- They can index the full text or parts of a document

□ Quality:

- Variable from a search engine to another
- Hits ranking based on:
  - Frequency/weight
  - HTML tags (<meta>) *(1999- spamming)*
  - Citations *(page-rank* de Google) – “link voting”

Search engines: how they work?

Do you know it?

https://www.google.com/search/howsearchworks/

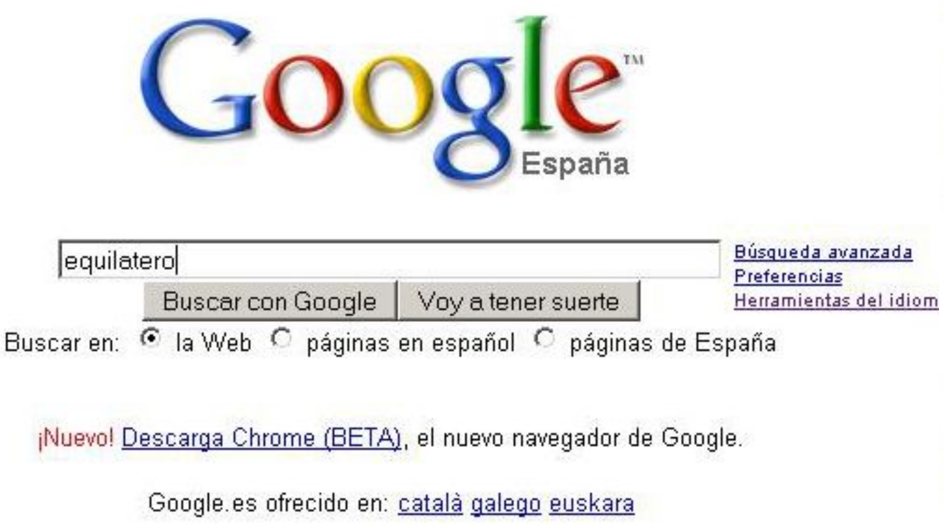
Every IR system (search engine, catalog, data base) has a HELP FILE to read and figure out how they work

□ All the search engines have a help where it is defined how they work, syntaxes, and some clues or advices to search:

- Google: http://support.google.com/websearch/?hl=en
- Yahoo! http://help.yahoo.com/l/us/yahoo/helpcentral/
- UC3M's OPAC: http://biblioteca.uc3m.es/iBistro_helps/English/power_search.html

Search engines: how they work?

- We enter a word through the search interface and we get a list of results ranked by RELEVANCE.
- Retrieval algorithms / ranking algorithms

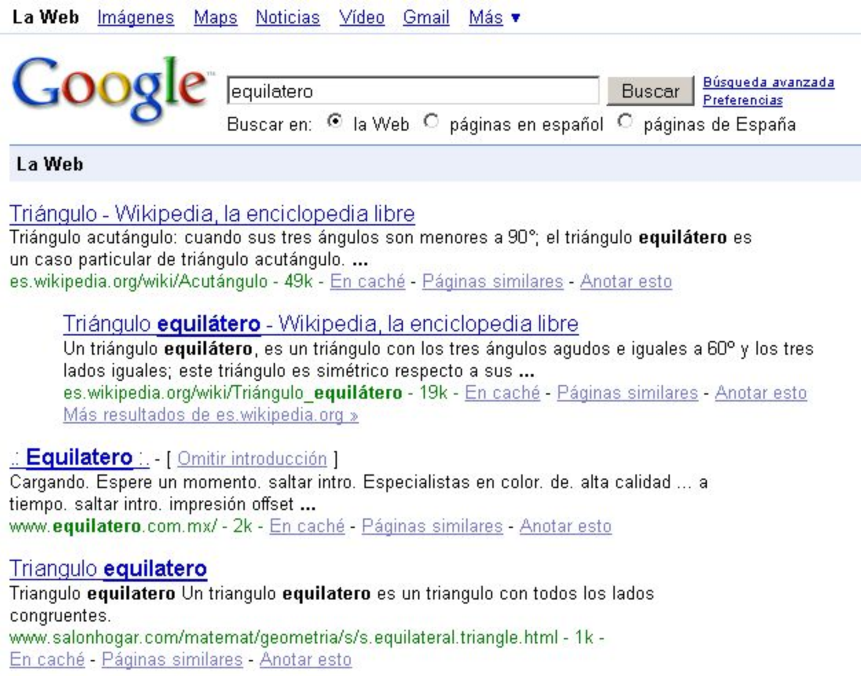


equilatero

Buscar en: ● la Web ○ páginas en español ○ páginas de España

¡Nuevo! Descarga Chrome (BETA), el nuevo navegador de Google.

Google.es ofrecido en: català galego euskara



equilatero

Buscar en: ● la Web ○ páginas en español ○ páginas de España

Triángulo - Wikipedia, la enciclopedia libre
Triángulo acutángulo: cuando sus tres ángulos son menores a 90°; el triángulo **equilátero** es un caso particular de triángulo acutángulo. **...**
es.wikipedia.org/wiki/Acutángulo - 49k - En caché - Páginas similares - Anotar esto

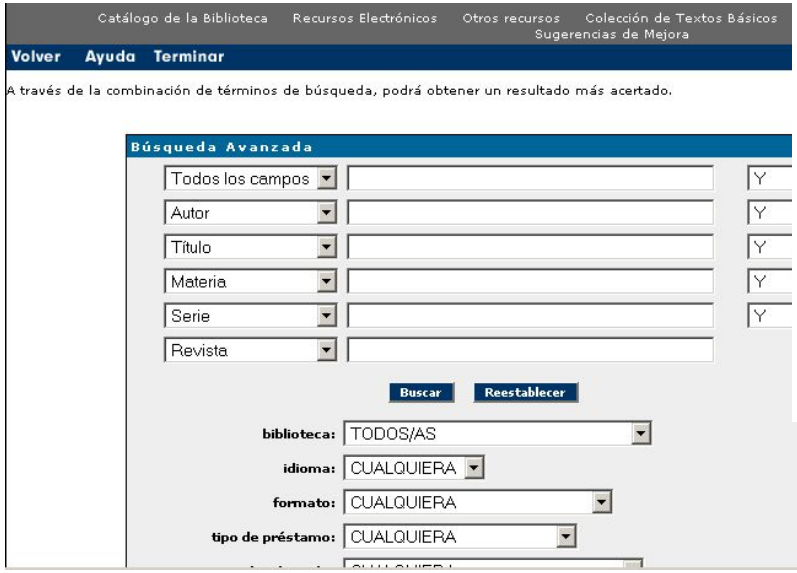
Triángulo **equilátero** - Wikipedia, la enciclopedia libre
Un triángulo **equilátero**, es un triángulo con los tres ángulos agudos e iguales a 60° y los tres lados iguales; este triángulo es simétrico respecto a sus **...**
es.wikipedia.org/wiki/Triángulo_**equilátero** - 19k - En caché - Páginas similares - Anotar esto
Más resultados de es.wikipedia.org »

.: **Equilatero** :. - [Omitir introducción]
Cargando. Espere un momento. saltar intro. Especialistas en color. de. alta calidad ... a tiempo. saltar intro. impresión offset **...**
www.**equilatero**.com.mx/ - 2k - En caché - Páginas similares - Anotar esto

Triangulo **equilatero**
Triangulo **equilatero** Un triangulo **equilatero** es un triangulo con todos los lados congruentes.
www.salonhogar.com/matemat/geometria/s/s.equilateral.triangle.html - 1k -
En caché - Páginas similares - Anotar esto

Search engines: how they work?

- The interface, the search syntax and how the search engine works is always similar (Internet, databases, Intranets, etc.)
- Common elements and particular elements.



A través de la combinación de términos de búsqueda, podrá obtener un resultado más acertado.

Búsqueda Avanzada

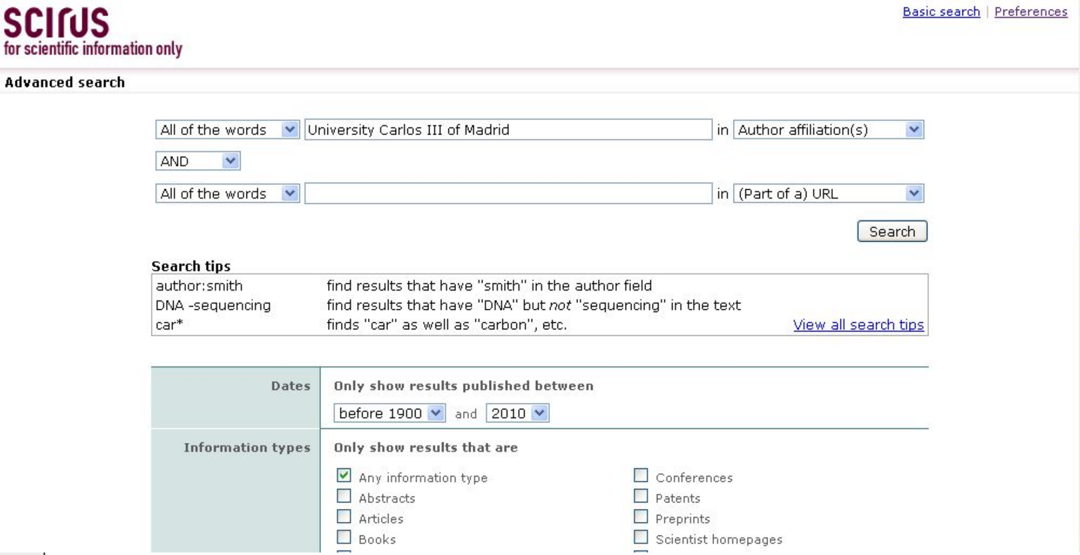
Todos los campos		Y
Autor		Y
Título		Y
Materia		Y
Serie		Y
Revista		

biblioteca: TODOS/AS

idioma: CUALQUIERA

formato: CUALQUIERA

tipo de préstamo: CUALQUIERA



Advanced search

All of the words: University Carlos III of Madrid in Author affiliation(s)

AND

All of the words: in (Part of a) URL

Search tips

author:smith	find results that have "smith" in the author field
DNA -sequencing	find results that have "DNA" but *not* "sequencing" in the text
car*	finds "car" as well as "carbon", etc.

View all search tips

Dates: Only show results published between before 1900 and 2010

Information types: Only show results that are

☑ Any information type
□ Abstracts
□ Articles
□ Books
□ Conferences
□ Patents
□ Preprints
□ Scientist homepages

Criteria to chose a Search engine (or why we use Google)

We should chose a search engine regarding…

□ Speed

□ Quality of results

□ Size of the data base (exhaustivity)

□ Data base updating

□ Easiness

□ Advanced search

□ Additional options

Barker, 2003. What Makes a Search Engine Good?
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SrchEngCriteria.pdf

Type of search engines

- **Web resources**
  - Web 1.0
  - Web 2.0
- Software & files
- People and institutions
- Listservs
- News

Internet/Deep Web

- Directories/ Index
  - general
  - specialized
- Search Engines
- Portals
- Metasearchengines
- Agents
- Web-Rings
- etc.

When to use a Search Engine?

□ When we know enough about we want to retrieve (familiar subjects)

□ When we want to know an exhaustive knowledge about a keyword of concept

□ When we have to do very specific search (boolean operators, parts of a document)

□ When we need specific data.

How to search and searching strategies (clues, operators and filters)

We must remember…

- In Internet… the information :-(
  - Overload of information
  - Information not very well structured
  - Change of location and time
  - Credibility/reliability?
  - Information overload
  - User level

General search tips before you start...

- Read help screens, instructions, advice (tips, hints), tutorials, descriptions, of each database or search engine.
    - Underlying principles are the same, but applied differently in each of the resources.
- Experiment with all buttons, links, menus, etc...
- Read the periphery of the screen and scroll a lot.
- Write search terms in the language/s of the documents you search for in the database!!
- Use the advanced search menu!!
    - It is more effective than basic searching (...and easier...it has guided functions).
- Try different terms, use those seen in documents already retrieved

Search strategies: Preparing the search

- Objective: match the query with records of stored materials.
- First, self-diagnose information needs, focus on and specify the problem, the “unknown”.
- Identify & verbalize the question in several ways.
- Analyze the question, select clues to be used to formulate the strategy.
- Translate those clues into a language and strategy compatible with the system (machine or human, or other).
- Formalize language and strategy in a mode compatible with the device or agent.

Selection of clues and expression of the query

- Predict:
  - how authors have written
  - how indexers have analyzed what authors have written
  - how analytics (clues) were recorded.
- Use variations of expression.
- If you don’t know well the subject coverage of the database, begin with general terms.
- Specify more than one aspect or point of view of the subject.

How is information processed and stored in a Database?

- DB have a structure (fields) & language
- Uniform criteria for selecting, processing and recording
- Formal analysis & Content analysis
  - Tries to infer at the same time the intentions of the author and of the searcher
  - Multidimensional
- Selection of resulting clues
- Translation into the system's language
  - Words, phrases, codes, numbers, etc.
  - Control of the vocabulary and the subjects expressed
  - Rules, syntaxes, indexing systems, classification schemes
- May include, in addition, full text / raw data

Translating search clues

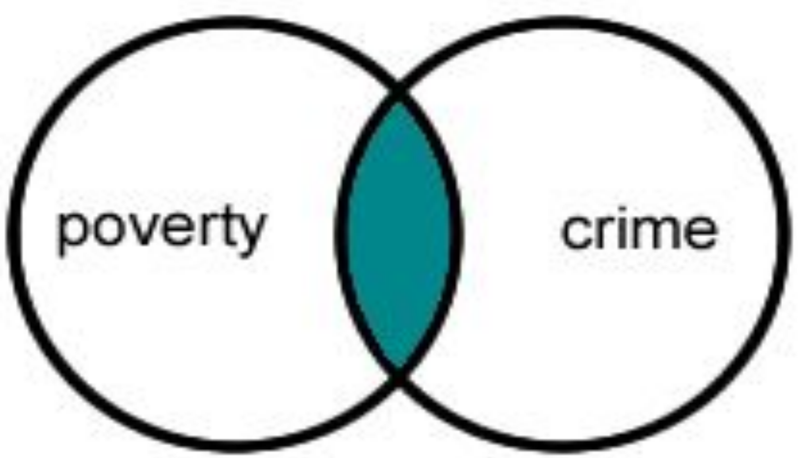
- Clues can be words, terms, expressions, formulas, phrases, dates, numbers, codes, etc. and the relationships between them.
- Translation is done in different ways depending on system characteristics:
  - search equations / queries
  - fill-in forms or query menus
  - indexes or automated thesauri
  - use of codes and classification schemes or taxonomies, etc.
  - folksonomies
- In “friendly” systems: auxiliary functions (interface guides the translation).
- Command languages: more powerful, efficient and precise, but need training.

Boolean operators

- Logical operations applied to different search terms in a searching system
- When using these operators we will get the documents according with that conditions
- Boolean logic consists of three logical operators:
  - OR
  - AND
  - NOT

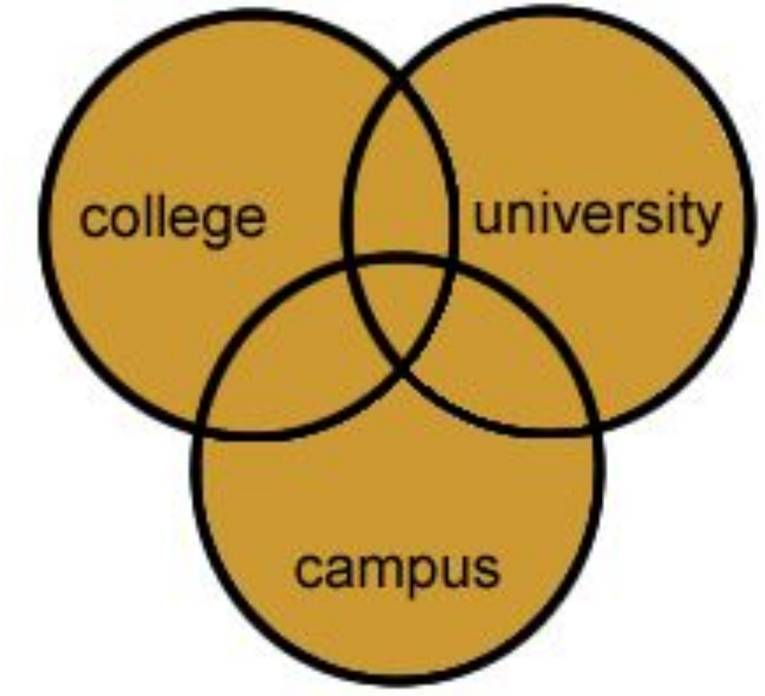
Boolean operators: AND

- Default one in a lot of Search engines (Google)
- We will get all the documents that have the first AND the second keywords.



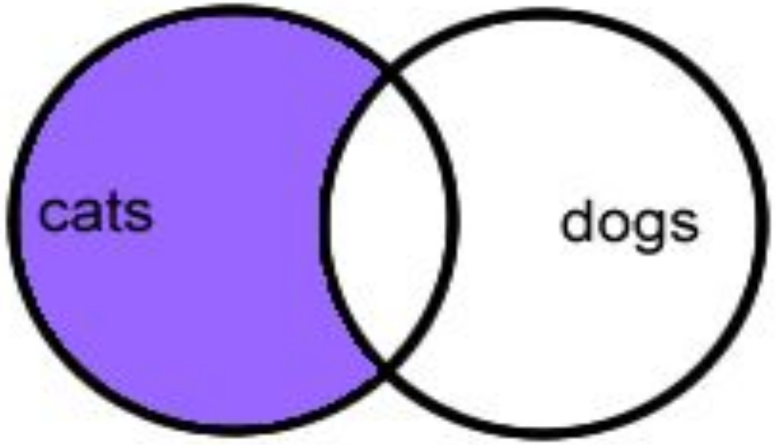
Boolean operators: OR

- We will get all the documents having the first keyword OR the second one □ documents having either one.



Boolean operators: NOT (-)

- We will get the documents that do NOT have the term
- We use this operator to filter documents from a previous search. Ex.:



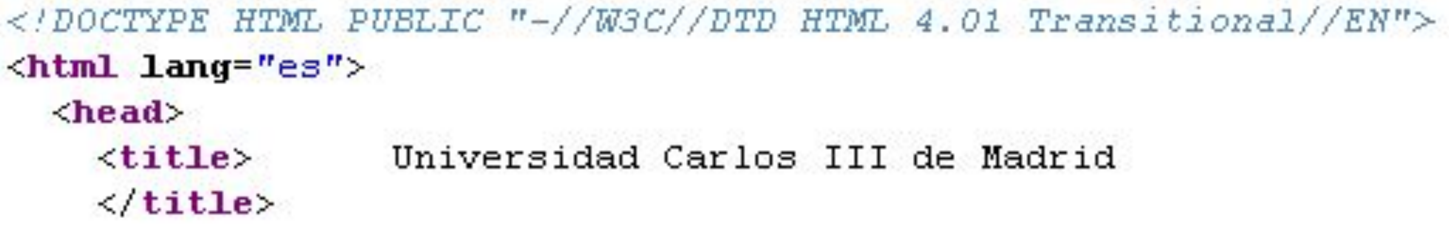
Other Search Operators

- There are other operators to improve search results or make our searches more precise. Such us:
  - **Exact phrase**. Usually in “…” Ex. “Information society” □ all the words in that order
  - **SAME**. Ex. cooking SAME carrots □ documents having those words in the same paragraph.
  - **WITH**. Ex. Economy WITH inflation □ documents having those terms in the same sentence/statement
  - **NEAR**. Ex. Money NEAR crisis □ Documents having those terms following one to another
  - **Adjacency.** Ex. Information ADJ Market □ documents having the first term just before the second one

Other operators

- **Shorten.** *, $, ?
  - Ex. Eco* □ it will retrieve documents having Economy, Economics, Ecosystem… etc.
  - Ex. *conduc* □ it will retrieve, for example, semiconducting…
  - Ej. ho?e □ home, hole, hose, etc.

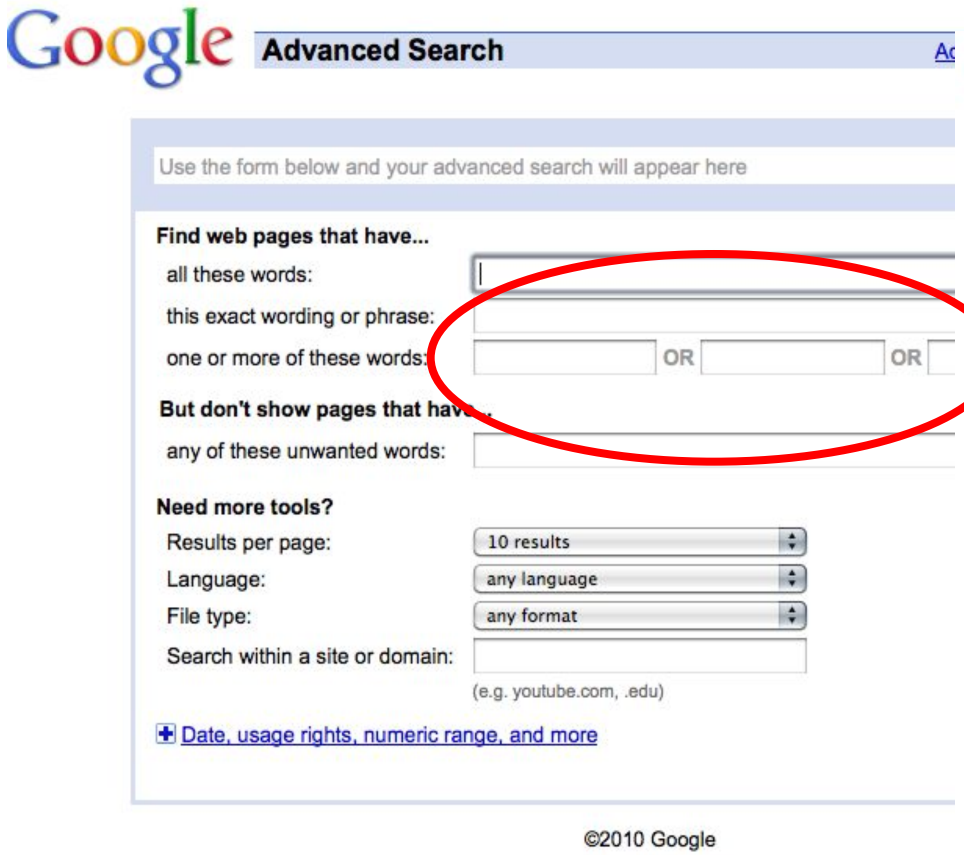
- **Search by fields.** intitle, inurl, link, site, etc. Ex. intitle:universidad □ documents having “universidad” in the title element of HTML



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html lang="es">
  <head>
    <title>      Universidad Carlos III de Madrid
    </title>
```

How to use the operators?

- Advanced search…



Google Advanced Search Ac

Use the form below and your advanced search will appear here

Find web pages that have...

all these words:

this exact wording or phrase:

one or more of these words: OR OR

But don't show pages that have...

any of these unwanted words:

Need more tools?

Results per page: 10 results

Language: any language

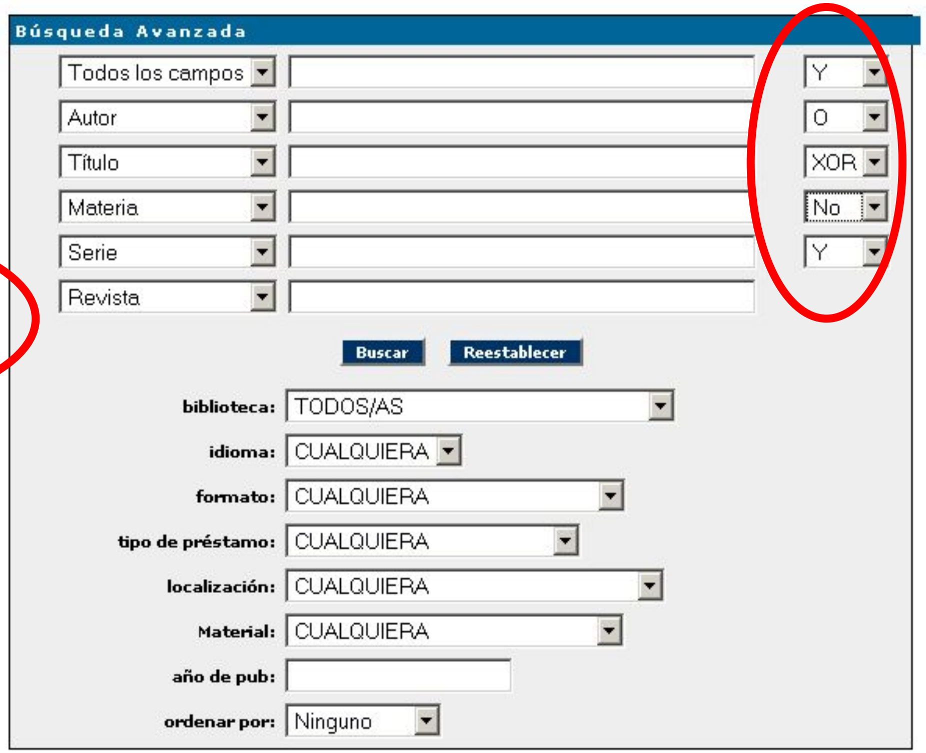
File type: any format

Search within a site or domain:

(e.g. youtube.com, .edu)

Date, usage rights, numeric range, and more

©2010 Google



Búsqueda Avanzada

Todos los campos	Y
Autor	O
Título	XOR
Materia	No
Serie	Y
Revista	

biblioteca: TODOS/AS

idioma: CUALQUIERA

formato: CUALQUIERA

tipo de préstamo: CUALQUIERA

localización: CUALQUIERA

Material: CUALQUIERA

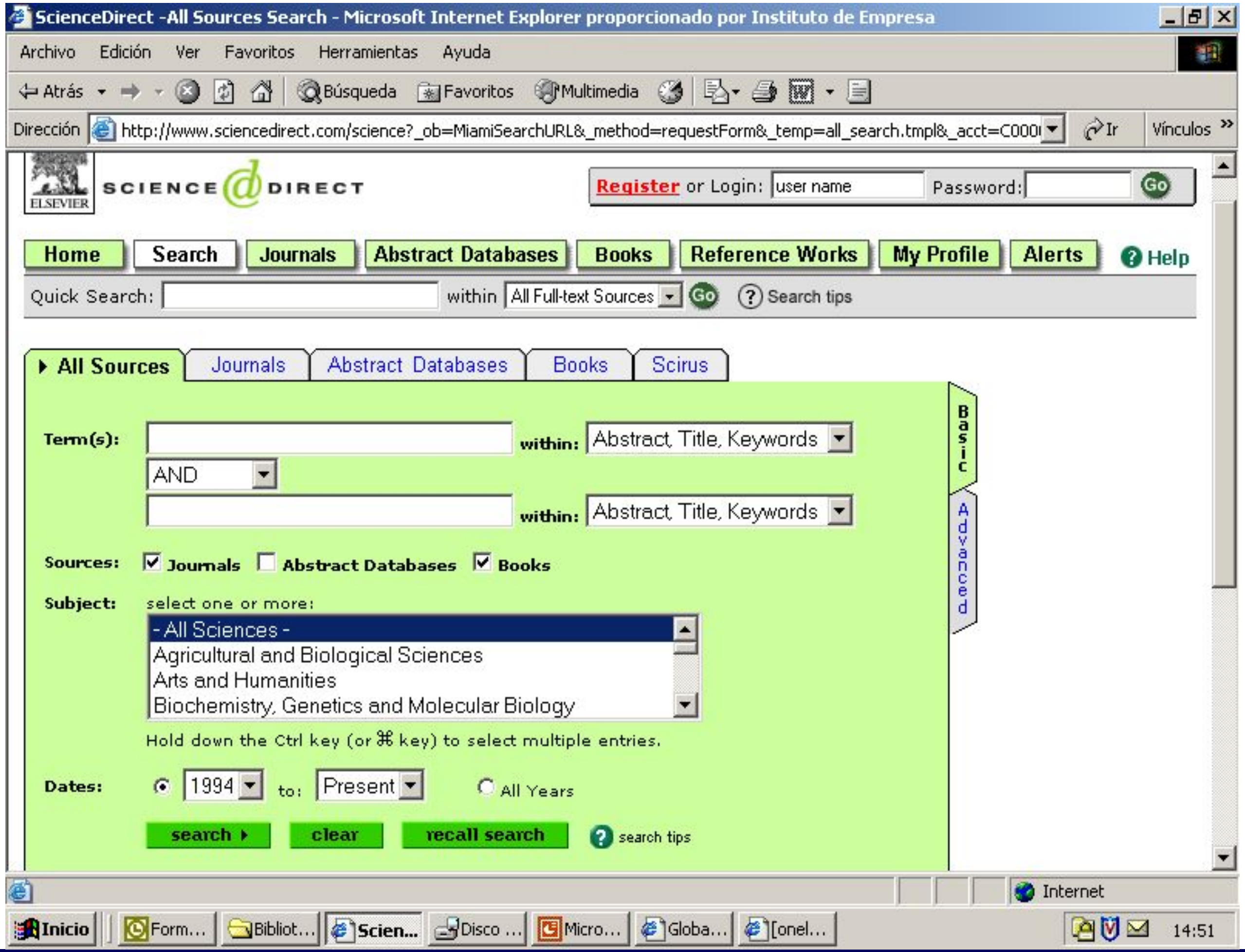
año de pub:

ordenar por: Ninguno

How to search. Query.

- Manipulate a search engine.
- System does most of the filtering.
- Sometimes complex or not user-friendly
- Different languages, syntaxes.
- If a meta –search engine is available, for several databases (federated search), more friendly but less precise.
- Advanced searching requires some training.
- Effective both for known references and for new research.
- Faster than browsing.
- Less information “escapes” (silence)
- Serendipity is also possible.

Example of query menu in a Data Base



ScienceDirect -All Sources Search - Microsoft Internet Explorer proporcionado por Instituto de Empresa

Dirección: http://www.sciencedirect.com/science?_ob=MiamiSearchURL&_method=requestForm&_temp=all_search.tmpl&_acct=C000

Register or Login: user name Password:

Quick Search: within All Full-text Sources

Term(s): within: Abstract, Title, Keywords

AND

within: Abstract, Title, Keywords

Sources: ☑ Journals □ Abstract Databases ☑ Books

Subject: select one or more:
- All Sciences -
Agricultural and Biological Sciences
Arts and Humanities
Biochemistry, Genetics and Molecular Biology

Hold down the Ctrl key (or ⌘ key) to select multiple entries.

Dates: ● 1994 to: Present ○ All Years

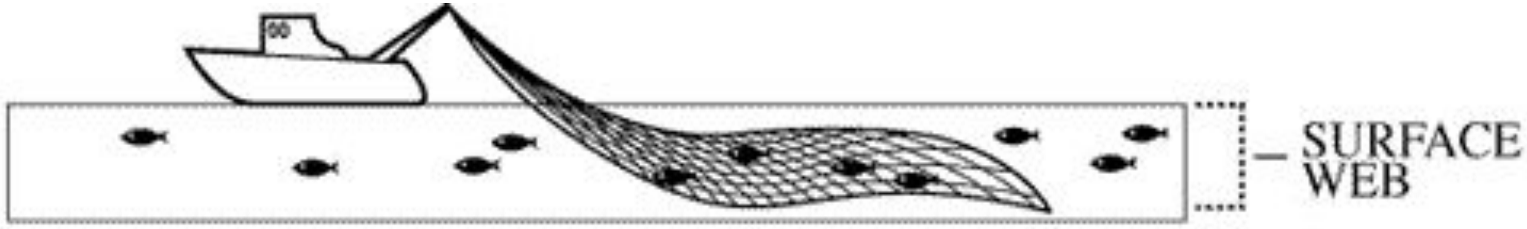
Think about searching…

- What will retrieve a search engine or Information Retrieval System in the following queries:
    - “Automation and Electronics” -Universidad
    - Informatic* AND “civil code”
    - “Digital Equipment Corporation” OR DEC
    - Prize NEAR Nobel
    - Brussels AND NOT “Brussels sprouts”
    - What about using other operators/filters?

uc3m | Universidad Carlos III de Madrid

Basic concepts of Information Retrieval on the Web: Deep Internet/ Invisible Web

What Search Tools Index



- From a White Paper produced by BrightPlanet.com LLC, July 2000. Available at www.completeplanet.com.

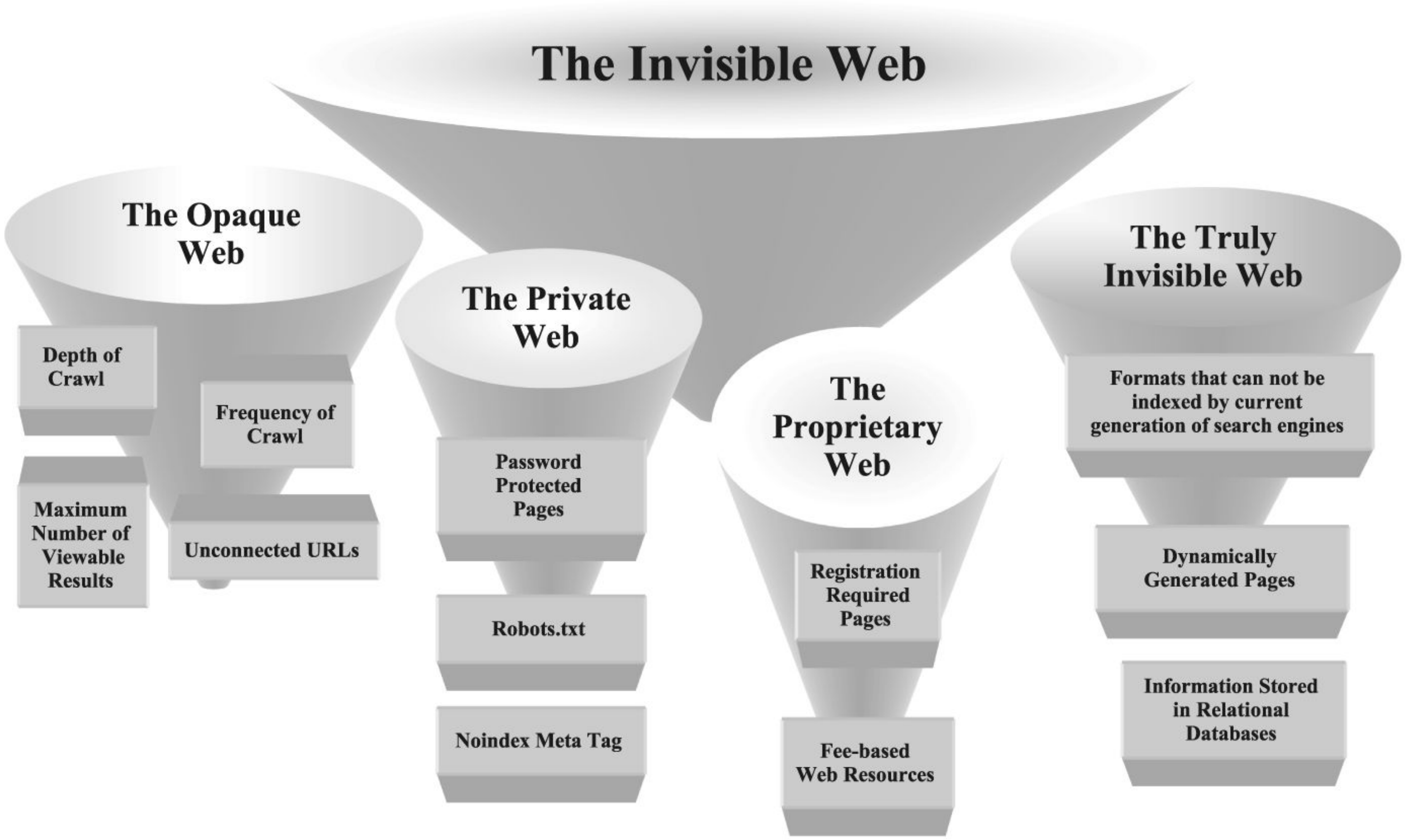
Invisible Web / Deep web

- 60/40□ 60 bigger web sites of the Invisible Web contain 40 times more information than all the visible Web (BrightPlanet).
- Search engines improvements (ej. pdf, doc, ppt)
- In the future, the invisible web could be smaller, but it will not disappear.

Invisible Web / Deep web

- Today, the invisible web means:
  - Data bases
  - Library catalogs and other bibliographic data bases
  - Data bases of electronic journals
  - Documents in formats/web technologies not good for indexing (ASP or PHP)
  - Interactive tools newsgroups or listservs
  - Material not linked or hidden in the servers
  - Statistical resources in different knowledge bases
  - Etc.

Different kinds of Invisible Web


The Invisible Web
The Opaque Web
Depth of Crawl
Frequency of Crawl
Maximum Number of Viewable Results
Unconnected URLs
The Private Web
Password Protected Pages
Robots.txt
Noindex Meta Tag
The Proprietary Web
Registration Required Pages
Fee-based Web Resources
The Truly Invisible Web
Formats that can not be indexed by current generation of search engines
Dynamically Generated Pages
Information Stored in Relational Databases