University of Padua, Italy
Department of Electronics and Computer Science
Information Management Systems Group

IMS Research Group

Information Retrieval on the Web

Maristella Agosti

Outline of First Part

- Background
- Traditional IR
- "IR on the Web": Terminology/Definitions and History
- Types of Tools for Performing IR on the Web
- Architecture and Components of IR Web Tools

ESSIR 2000

European Summer School in Information Retrieval
September 11-15, 2000 - Villa Monastero, Varenna, Italy

IR on the Web

Background

- Hypertext Information Retrieval (HIR) before the Web
- Automatic Construction of Hypertexts for Information Retrieval
- Combining Browsing (Navigation) and Searching

Traditional IR

Collection of **documents**

A collection of documents is a **set of documents** which is related to a specific **context of interest** (e.g. a specific subject or thematic area, a time span).

Selection from a given **collection** of **documents** of those documents that are of interest in relation to a specific **information need.**

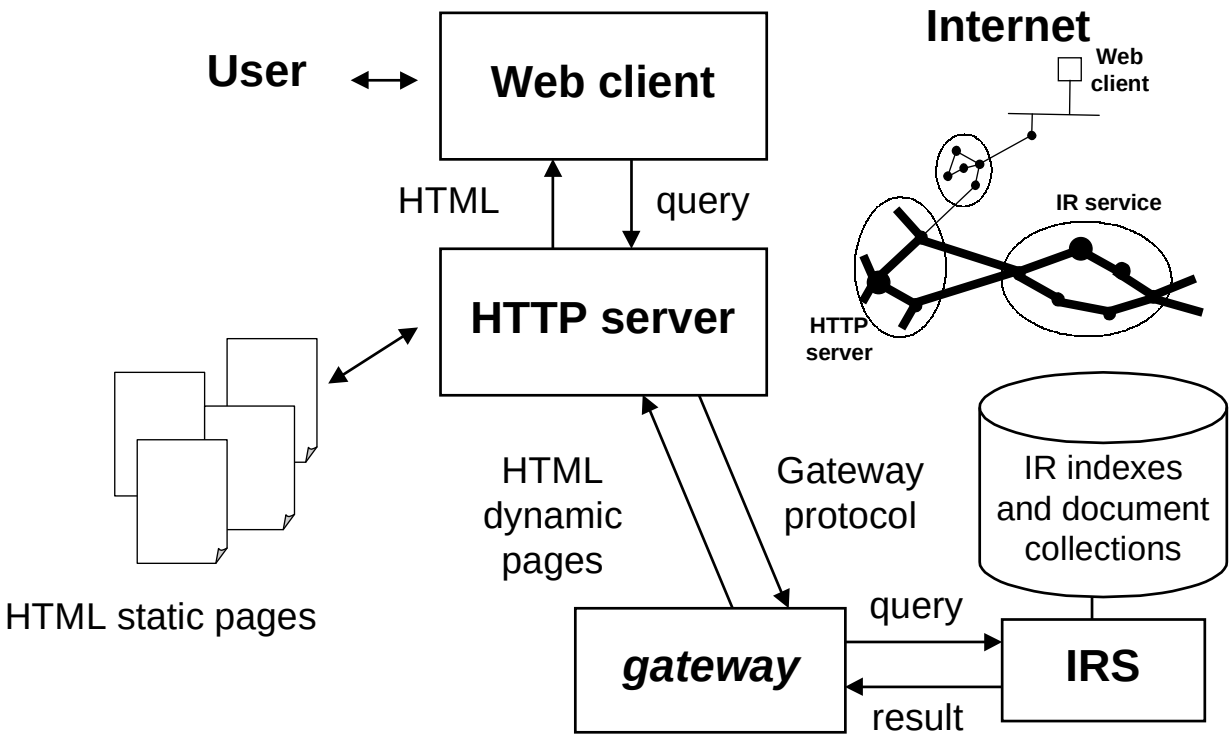
Traditional IR

Automatic IR system

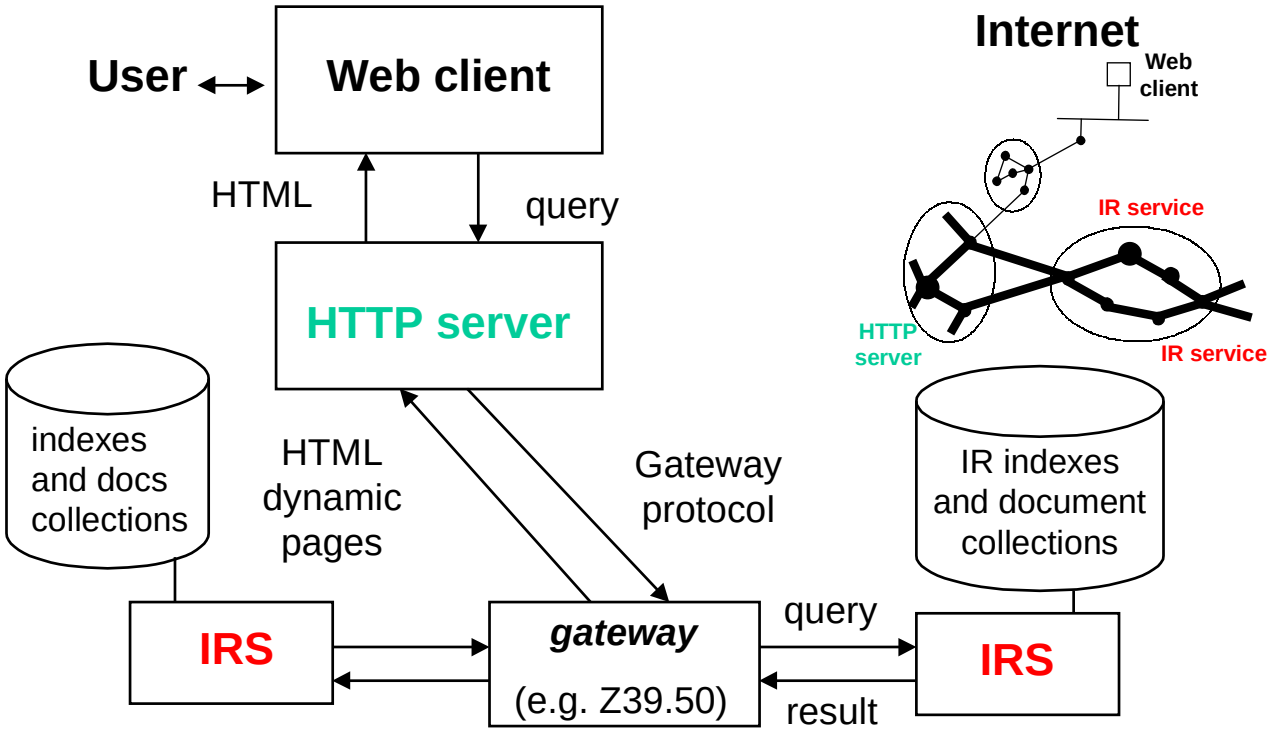
Choice of the probably relevant documents is made in an **automatic** way by the IR system which answers to **queries**.

Indexing process is applied to the **full text** of the documents.

Traditional IR through a browser Web



Traditional IR distributed over the Web



Terminology/Definitions

- A **Web page** corresponds to a **document** in traditional IR.
- Web pages are really **different in size** and in the **type of files** that can constitute them (text, graphics, sound, …).
- IR on the Web considers as **collection of documents** of interest the part of the **Web** which is **publicly indexable**, this excludes pages that cannot considered for indexing

  (e.g. pages with authorisation requirements, excluded using the robot exclusion standard, or pages dynamically generated).

Location of Web pages by navigation

Navigation

- direct request by knowing the correct URL
- indirect request of a page using the hypertext link presents in an available Web page
- availability of a Web page by a “narrowcast” service.

Location of Web pages by searching or “IR on the Web”

Searching

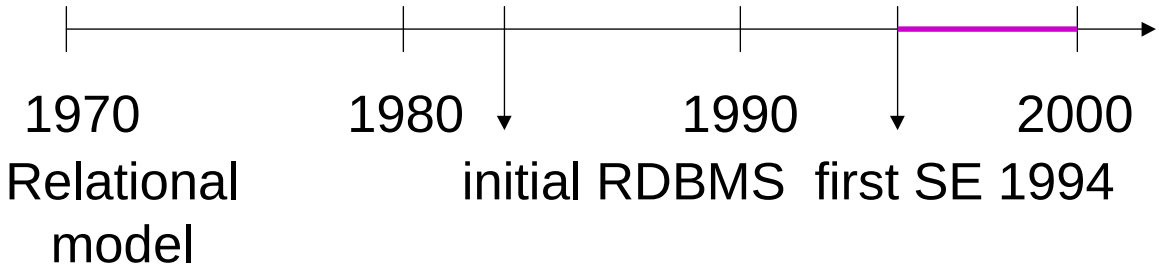
- **Web search service** which makes use of a **Web search engine**, which is a collection of complex software tools that implement IR functions using as **collection of documents** the indexable Web.
- Indexable Web is made up by pages that are related/associated by links.

History

- Web is available from less than 10 years.
- Web Search Engines that allow the user to search for information using the “**full text**” of **entire Web documents** (entire *textual part*) from 1994:
  - April 1994: **WebCrawler**, University of Washington
  - July 1994: **Lycos**, Carnegie Mellon University.

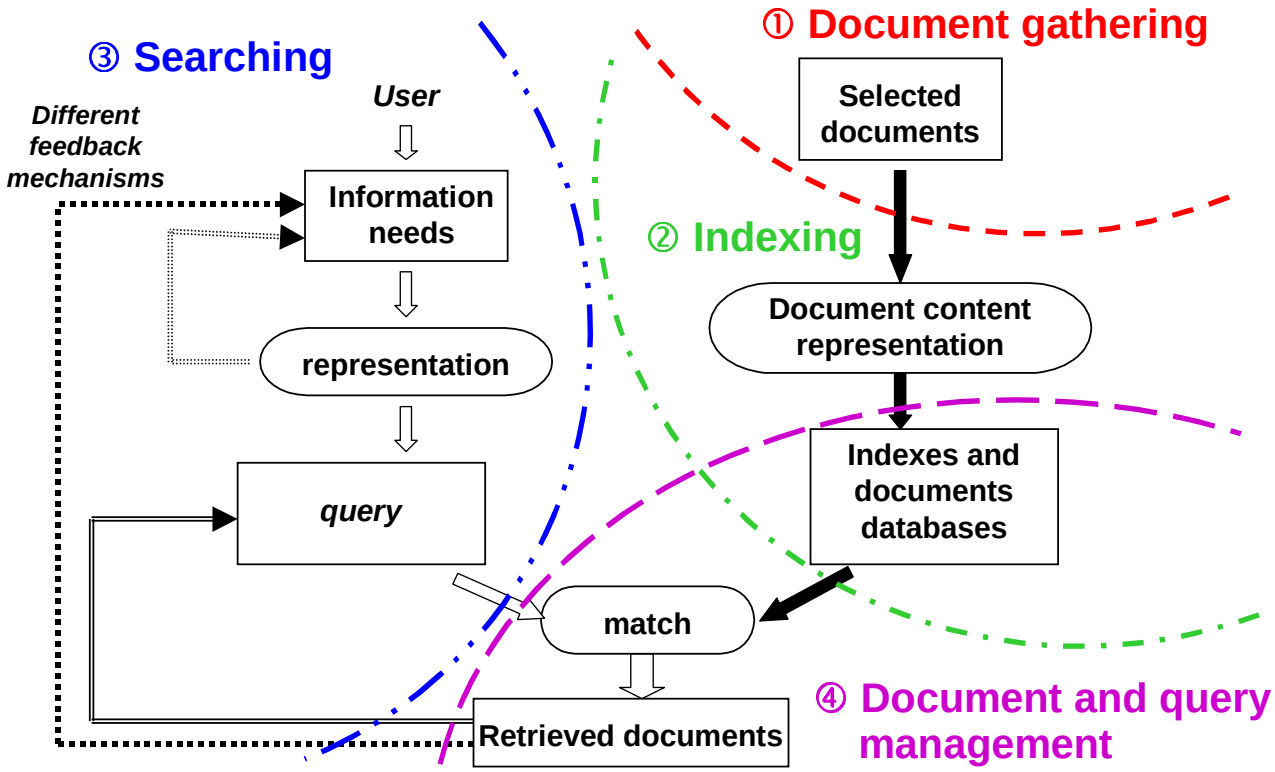
History

Web Search Engines are in their infancy.
Parallelism with DBMS development:

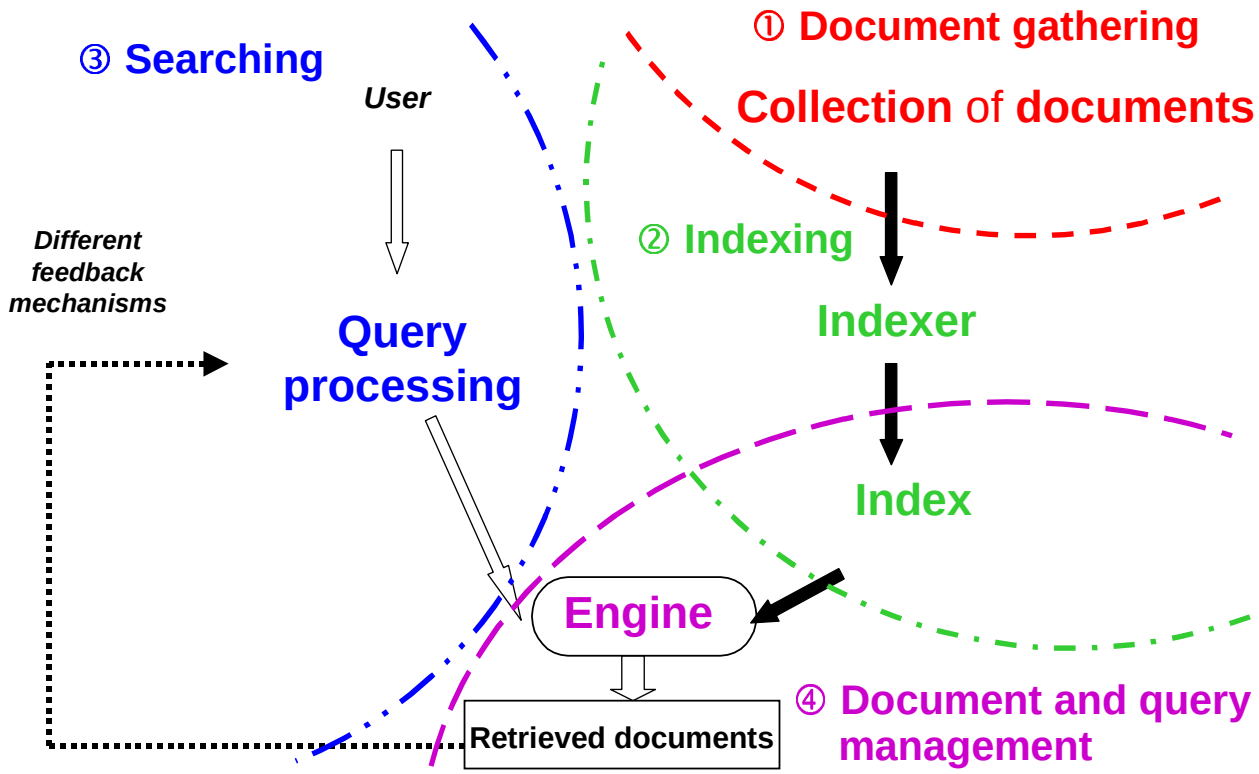


After 5-6 years there is a wide **range of tools**, so it is worthwhile to identify some aspects, that give a sort of **reference** for understanding the features that are relevant to IR.

Phases of the IR Process



Architecture of a traditional IRS



① Document gathering

IR on the Web

Target: construction of the **collection of the Web documents** that form the universe of interest that the software tools have to index and manage.

The user is going to search and retrieve "relevant" documents from this **subset** of the **real Web**.

Output of this phase:

- "**virtual collection**" (docs discarded after indexing)
- "**physical collection**" (docs maintained).

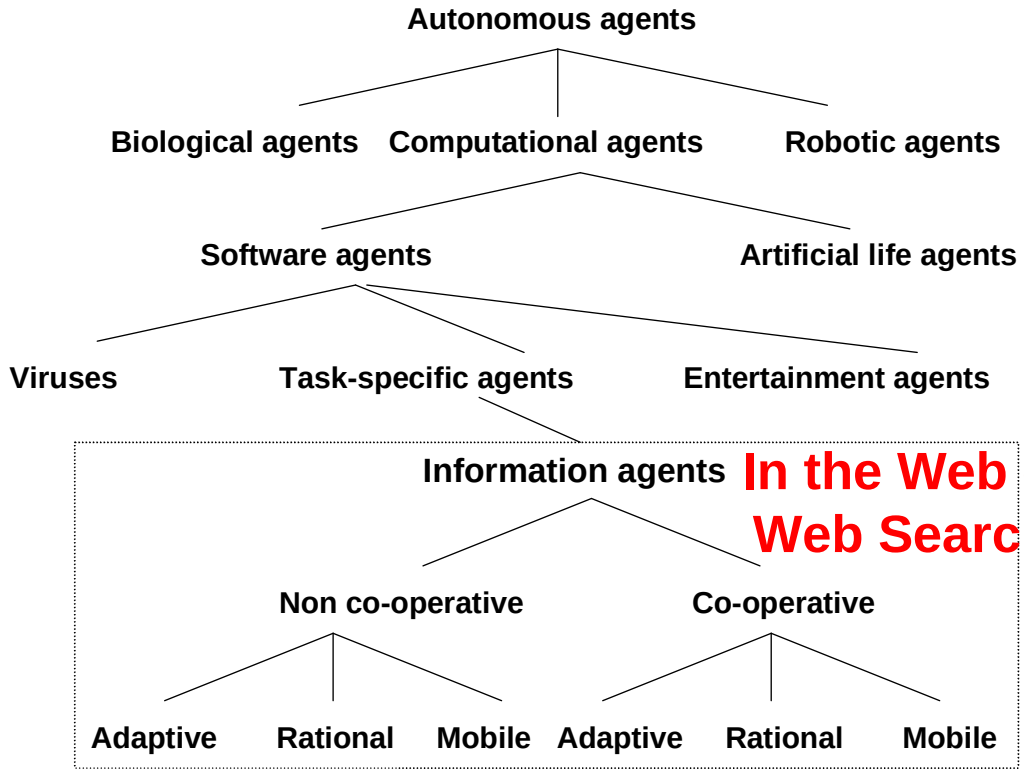
① Document gathering
IR on the Web - SW Tools

Software tools can gather the documents for building the collection mainly in two different ways:

1. Web pages submitted by users or companies to the search engine

2. Web pages collected by a **Web Search Agent (WSA)** of the search engine that is an information agent (see Klusch, 1999); typical names: spider, crawler, robot.

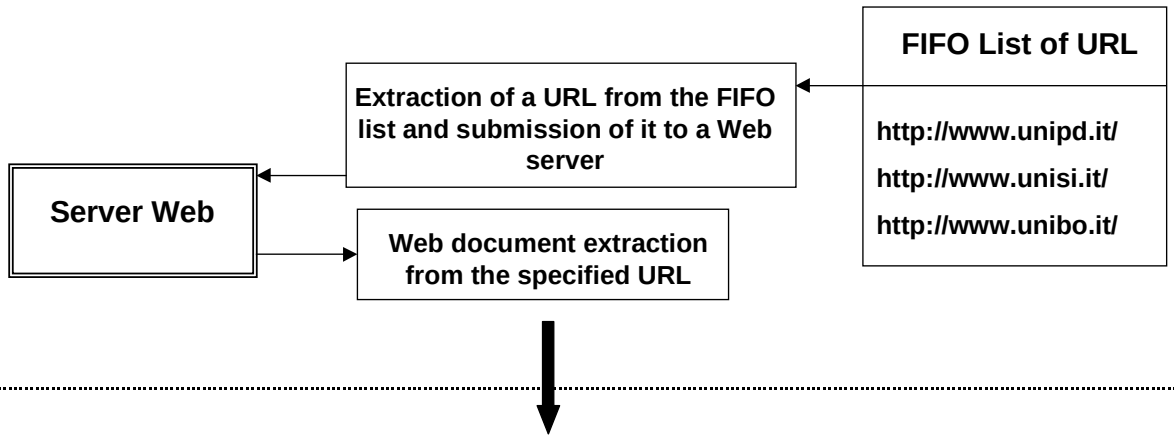
A Classification of Information Agents by Klusch, 1999



In the Web context: Web Search Agent

① Document gathering

IR on the Web - *Web Search Agent (WSA)*

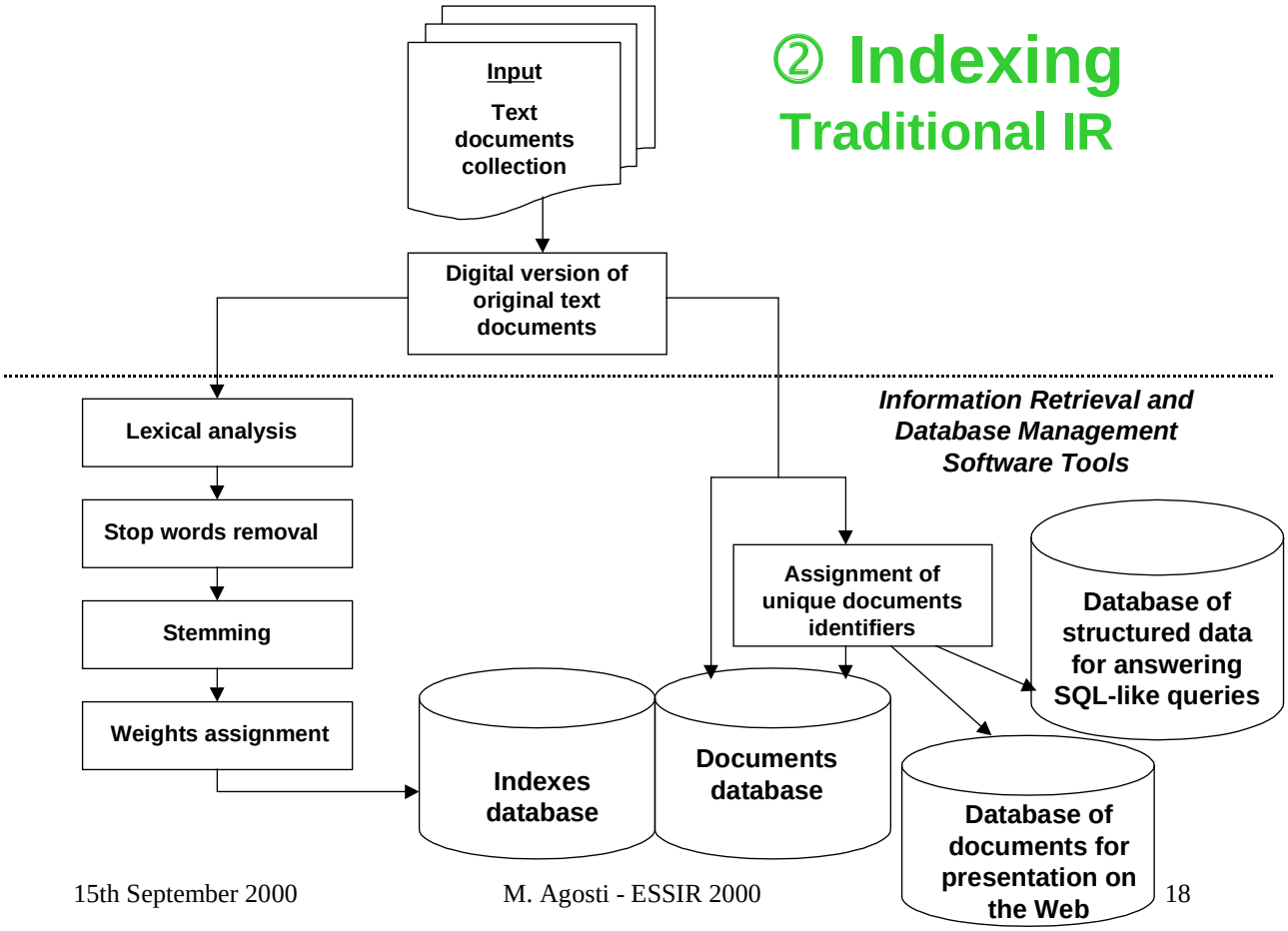


Indexer

Web documents indexing

② Indexing

Traditional IR



② Indexing
IR on the Web

Query-based Engines:
indexes are automatically built

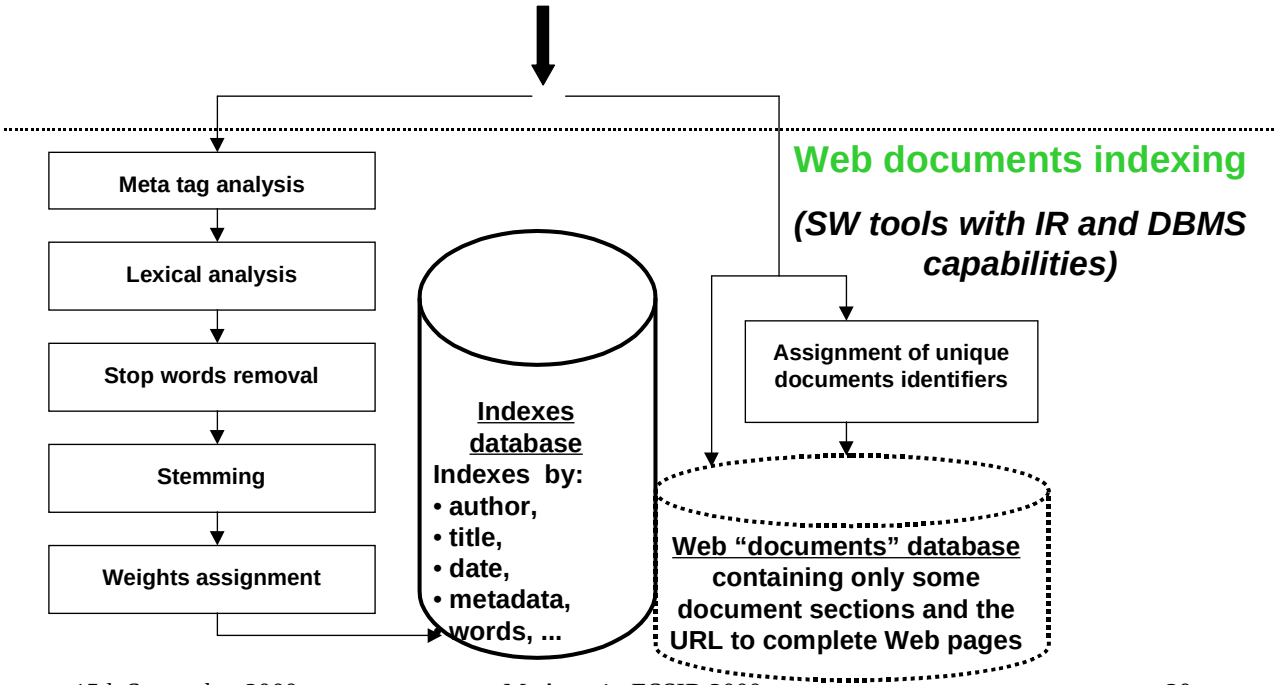
Classified Lists:
subject directory catalogues are manually built

automatic

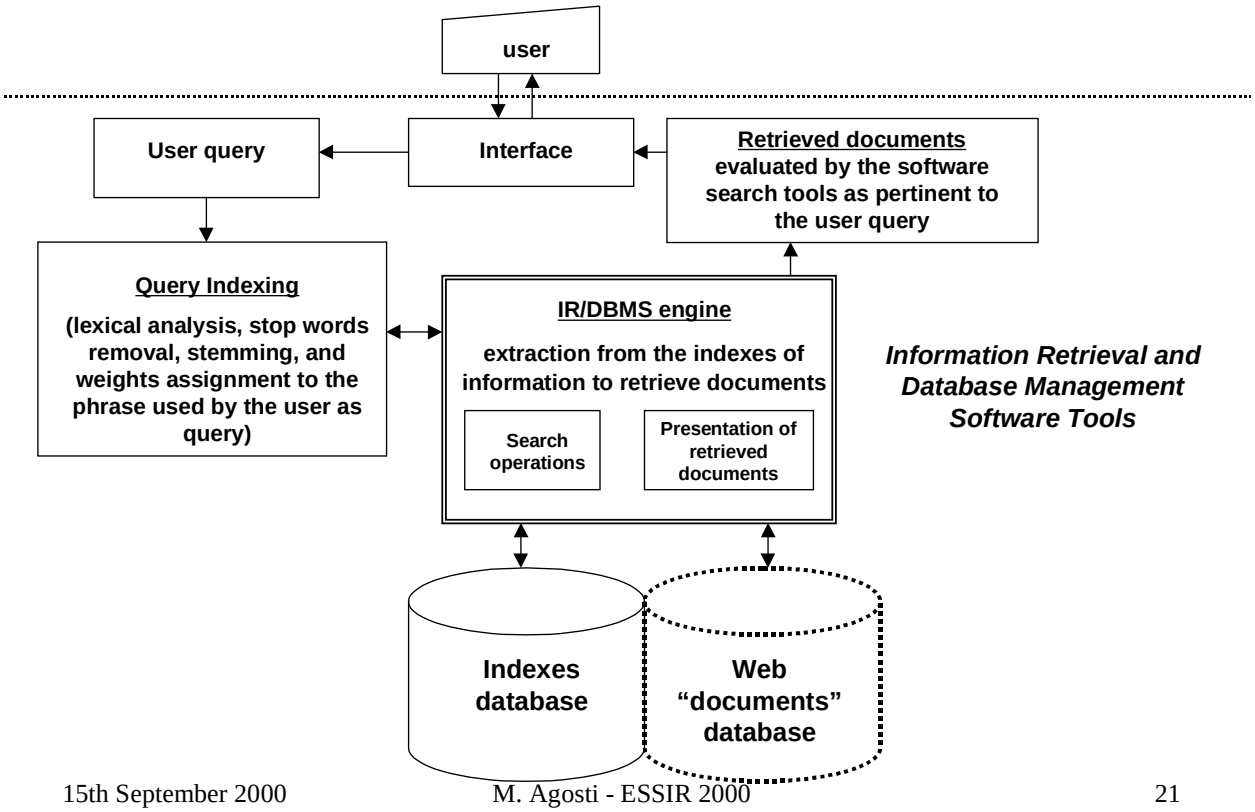
manual

② Indexing
IR on the Web

Web Search Agent (WSA)



③ Searching



④ Document and query management

"**virtual collection**" (docs discarded after indexing) **versus** "**physical collection**" (docs maintained)

The decision of maintaining the original version of the Web document which has been indexed can be made, because the document can change over time and the new one can be really different from the version which has been indexed, and it can constitute a surprise for the user when it sees it as present in the answer of its query.