Lecture 1:
Basic Information Retrieval and Web Search

01204453 Web Information Retrieval and Mining

Department of Computer Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand.

Department of Computer Engineering Kasetsart University

MIKE LABORATORY

Outline

- Introduction
- Brief History
- The IR Problem
- The IR System
- Web Search

Motivation

- Information Retrieval (IR)
  - How to store
  - How to extract/represent
  - How to query
  - How to access/search
  - How to present
- Serve the users' (complex) information need
- Why Information Retrieval (not Data Retrieval)?

Terminology

- Data Retrieval
  - Documents containing the keywords in the user query
  - Language with clearly defined conditions
  - Single error means total failure
  - Deal with a well defined structure and semantics data
- Information Retrieval
  - Documents concerned more information about a subject
  - Query translated from the user information need
  - Inaccurate and small errors are likely to go unnoticed
  - Deal with natural language and semantically ambiguous text
- Knowledge Discovery

Information Retrieval

- Early goals of the IR area
  - Indexing text and searching for useful documents in a collection

- Nowadays, research in IR
  - Modeling
  - Web search
  - Text classification
  - System architecture
  - User interfaces
  - Data visualization
  - Filtering
  - Languages
  - …

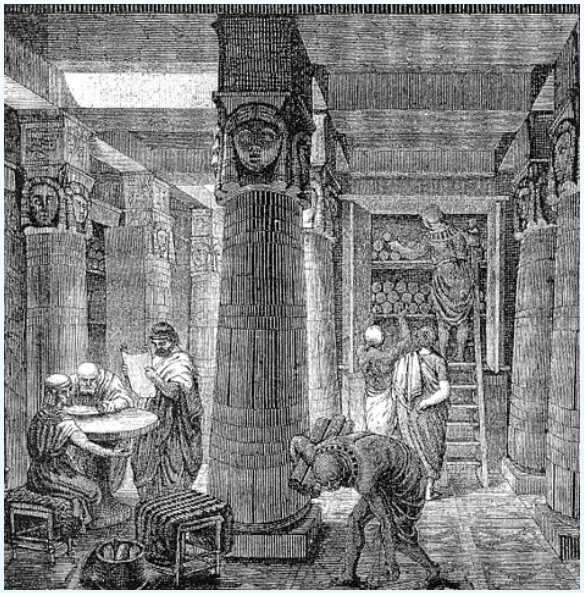
The Brief History of IR

Early Developments

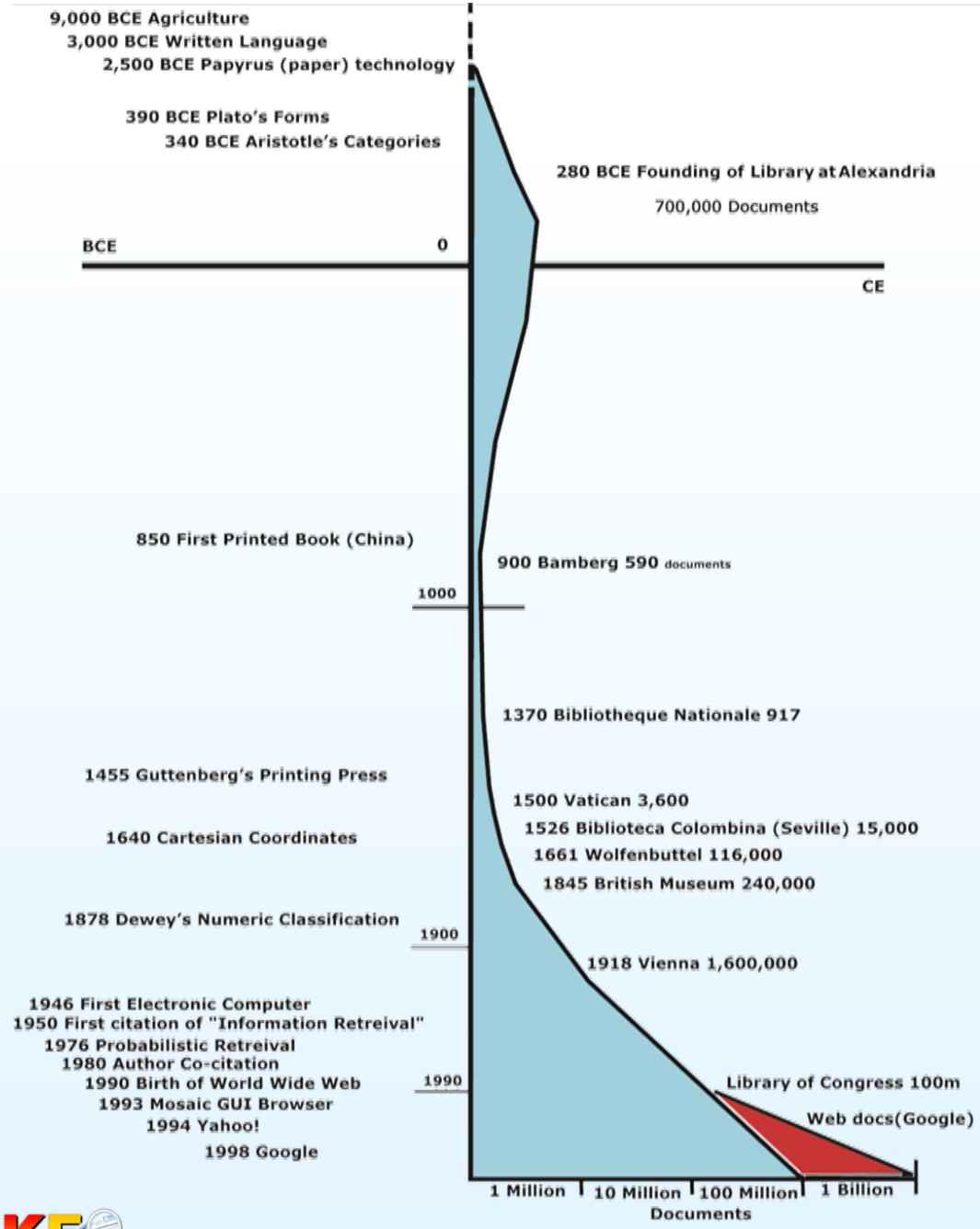
- For more than 5000 years, man has organized information for later retrieval and searching
  - This has been done by compiling, storing, organizing, and indexing
- For holding the various items, special purpose buildings, called libraries, are used
  - The oldest known library was created in Ebla, in the Fertile Crescent, between 3000 and 2500 BC

Early Developments

- By 700 BC, Assyrian king Ashurbanipal created the library of Nineveh, on the Tigris river
  - Contained more than 30,000 clay tablets
- By 300 BC, Ptolemy Soter, a Macedonian, created the Great Library in Alexandria



- Nowadays, libraries are everywhere


9,000 BCE Agriculture
3,000 BCE Written Language
2,500 BCE Papyrus (paper) technology
390 BCE Plato's Forms
340 BCE Aristotle's Categories
280 BCE Founding of Library at Alexandria
700,000 Documents
BCE
0
CE
850 First Printed Book (China)
900 Bamberg 590 documents
1000
1370 Bibliotheque Nationale 917
1455 Guttenberg's Printing Press
1500 Vatican 3,600
1526 Biblioteca Colombina (Seville) 15,000
1640 Cartesian Coordinates
1661 Wolfenbuttel 116,000
1845 British Museum 240,000
1878 Dewey's Numeric Classification
1900
1918 Vienna 1,600,000
1946 First Electronic Computer
1950 First citation of "Information Retreival"
1976 Probabilistic Retreival
1980 Author Co-citation
1990 Birth of World Wide Web
1990
Library of Congress 100m
1993 Mosaic GUI Browser
Web docs(Google)
1994 Yahoo!
1998 Google
1 Million
10 Million
100 Million
1 Billion
Documents

Early Developments

- Since the volume of information in libraries is always growing, it is necessary to build specialized data structures for fast search—the indexes
- For centuries indexes have been created manually as sets of categories, with labels associated with each category
- The advent of modern computers has allowed the construction of large indexes automatically

Early Developments in IR

- During the 50's, research efforts in IR were initiated by pioneers such as Hans Peter Luhn, Eugene Garfield, Philip Bagley, and Calvin Moores, who allegedly coined the term Information Retrieval
- In 1962, Cyril Cleverdon published the Cranfield studies on retrieval evaluation
- In 1963, Joseph Becker and Robert Hayes published the first book on IR
- In the late 60's, key research conducted by Karen Sparck Jones and Gerard Salton, among others, led to the definition of the TF-IDF term weighting scheme

Early Developments in IR

- In 1978, the first ACM SIGIR International Conference on Information Retrieval was held in Rochester
- In 1979, van Rijsbergen published a classic book entitled Information Retrieval, which focused on the Probabilistic Model
- In 1983, Salton and McGill published a classic book entitled Introduction to Modern Information Retrieval, which focused on the Vector Model

Libraries and Digital Libraries

- Libraries were among the first institutions to adopt IR systems for retrieving information

- Initially, such systems consisted of an automation of existing processes such as card catalogs searching
  - Restricted to author names, titles

Libraries and Digital Libraries

- In the next generation, Increased search functionality was added
  - E.g., subject headings, keywords, query operations

- Nowadays, the focus has been on improved graphical interfaces, electronic forms, hypertext features

IR at the Center of the Stage

- Previously, IR was an area of interest restricted mainly to librarians and information experts
- A single fact changed these perceptions—the introduction of the Web, which has become the largest repository of knowledge in human history
- The Web also introduced difficult task on finding useful information
- Searching on the Web is all about IR and its technologies

Thus, almost overnight, IR has gained a place with other technologies at the center of the stage

The IR Problem

The IR Problem

- Users of modern IR systems have information needs of varying complexity
  - An example of the simple information need:

    Look for the link to the homepage of a company or institution
  - An example of the complex information need:

    Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation

The IR Problem

- This full description of the user need is normally not good for querying the IR system
- Instead, the user would first translate into a query
  - A set of keywords summarized his/her information need
- Given the user query, the key goal of the IR system is to retrieve the useful or relevant information

The IR Problem

- Thus, the IR system must rank the information items according to a degree of relevance to the user query

- The IR problem:

  The key goal of an IR system is to retrieve all the items that are relevant to a user query, while retrieving as few non-relevant items as possible

- The notion of relevance is of central importance in IR

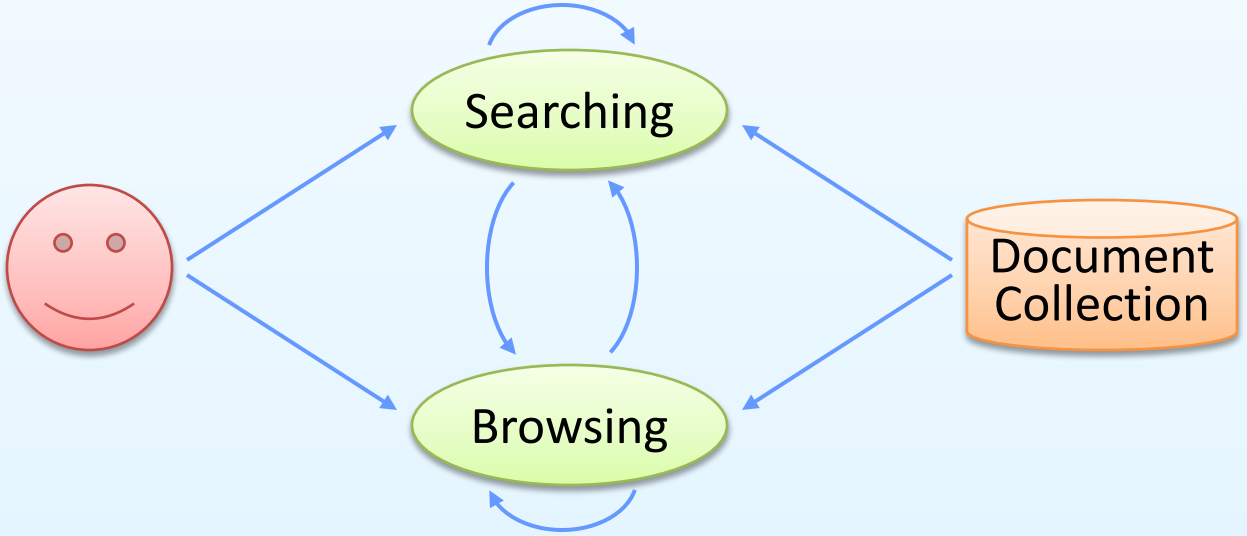
Basic Concepts

- The effective retrieval of relevant information is directly affected by 2 factors:
  - The user task
  - The logical view of the documents

The User Task

- The action when the user first creates a query and then submits to the system, called "searching" or "querying"

- For a poorly defined or inherently broad query, the user need to perform "browsing" or "navigating" the retrieved documents



Document

- A single unit of information
  - Typically, text in a digital form
    - e.g., book, article, passage, e-mail, web page, etc.

- Documents in a collection are frequently represented through a set of index terms or keywords (directly extracted or expertly generated) to provide *a logical view*

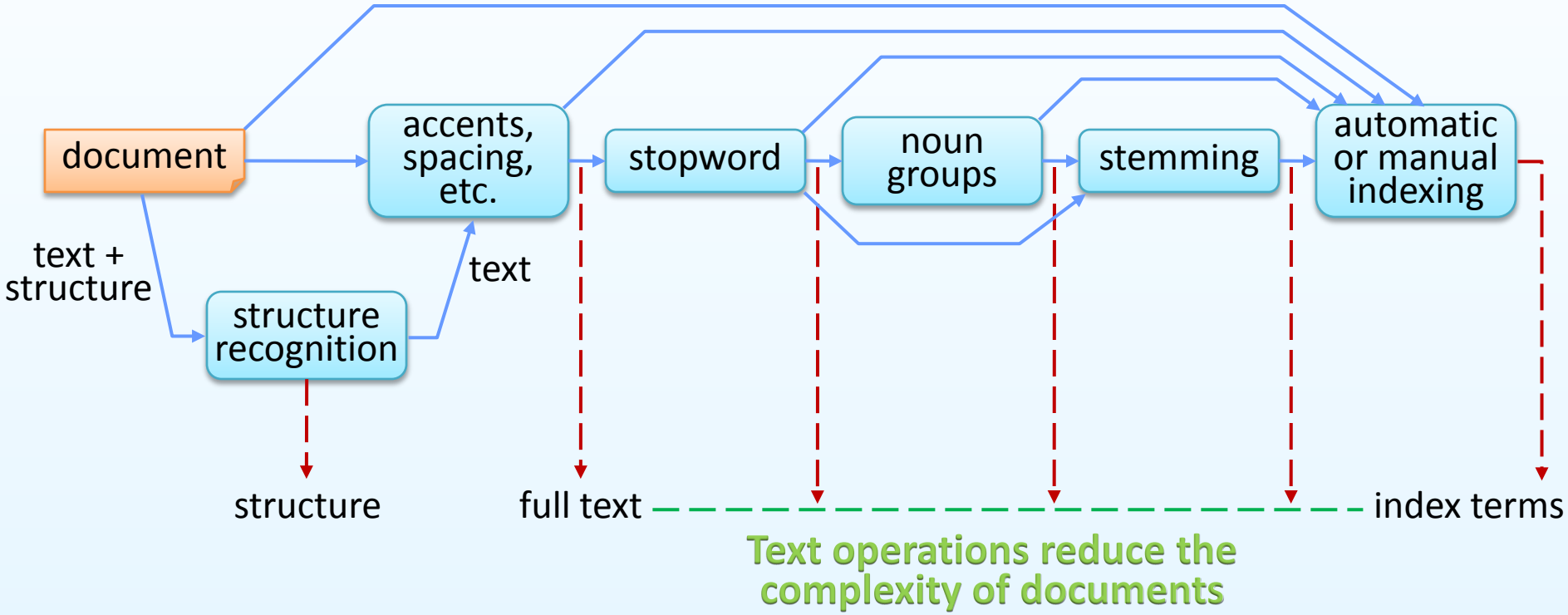
The Logical View of the Documents

- The use of full text
  - ⇒ Provide the most complete logical view
  - ⇒ Take higher computational costs

- The use of a small set of categories generated by human experts
  - ⇒ Provide the most concise logical view
  - ⇒ Might lead to retrieval of poor quality

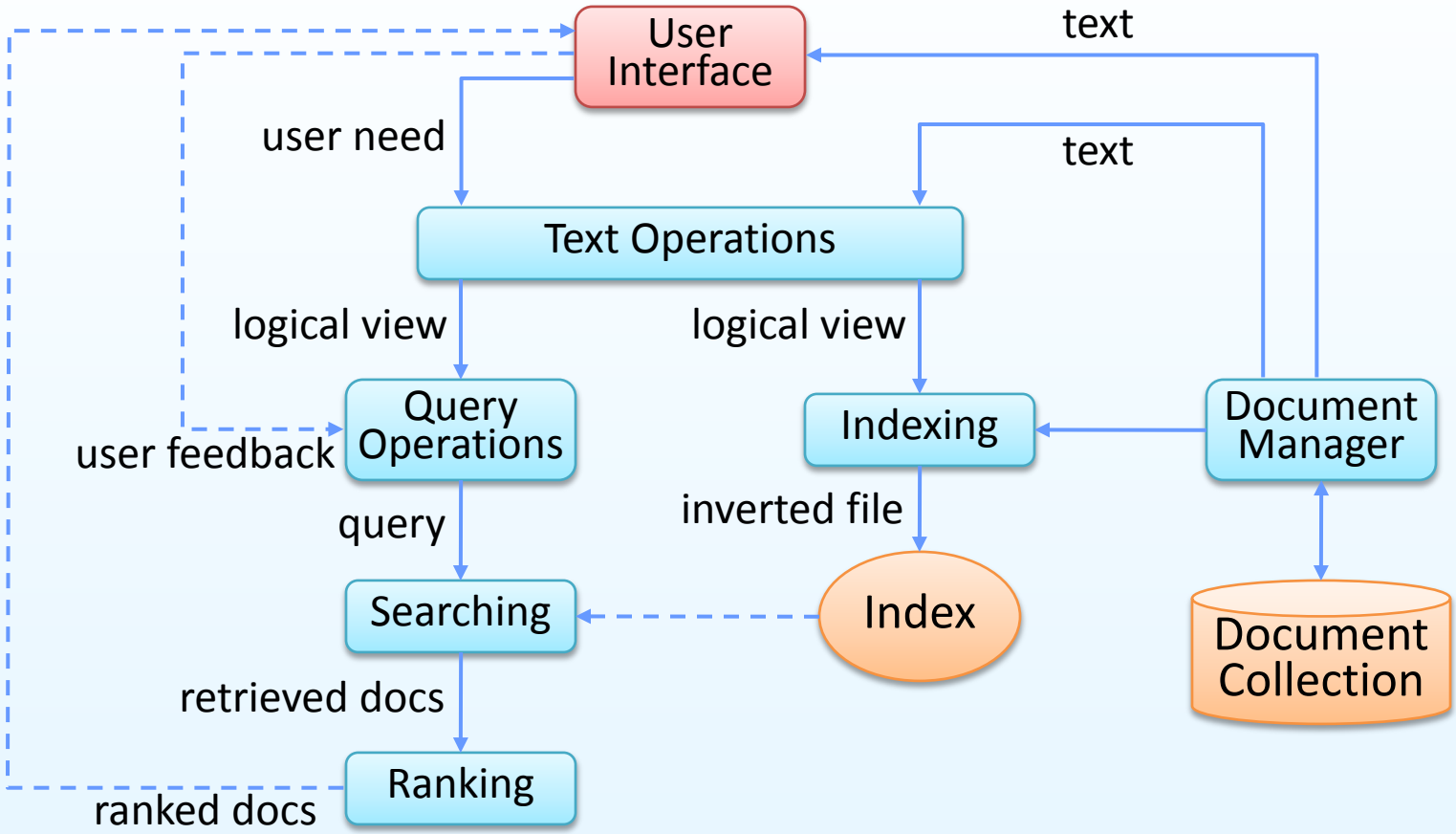
The Logical View of the Documents

- Several intermediate logical views might be adopted an IR system



The IR System

The Retrieval Process



Web Search

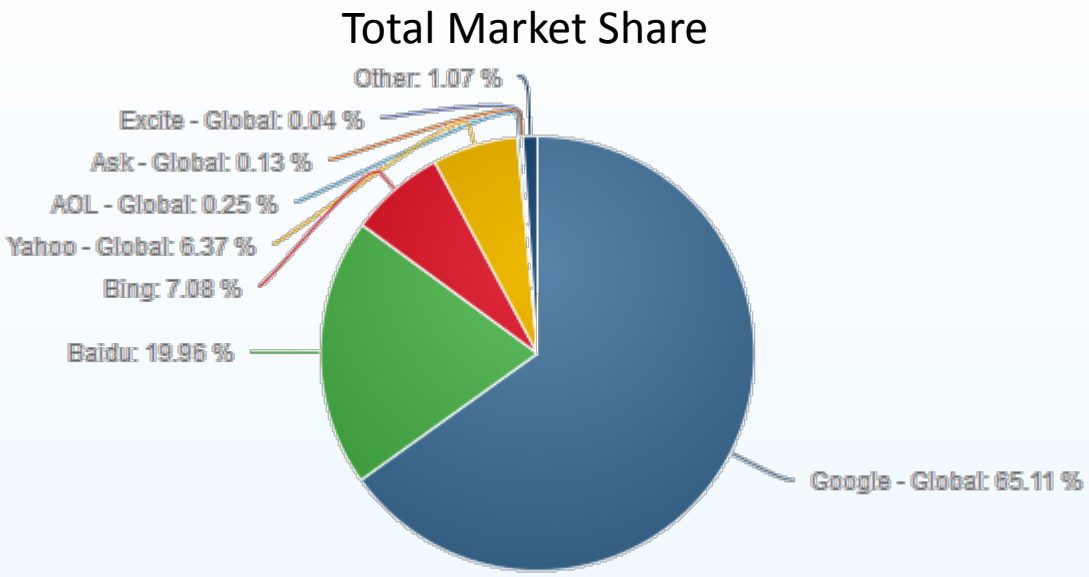
Web Search

- Internet is the largest source containing several kinds of data.
- Web search is one of the IR applications that concerns hypertext documents (web pages) in the Internet.
- Dealing with the web search has still been very challenging for years.

Characteristics of the WWW

- Very huge resource
  - Static and dynamic web pages
  - Not easy to measure the size of the WWW
- Proliferation and dynamic nature
- Decentralized content publishing
- Different quality of content
- Various natural languages
- Wide variation of presentations
  - Colors, fonts, structures, etc.

Search Engine Market Share



Search Engine	Total Market Share
Google - Global	65.11%
Baidu	19.96%
Bing	7.08%
Yahoo - Global	6.37%
AOL - Global	0.25%
Ask - Global	0.13%
Excite - Global	0.04%

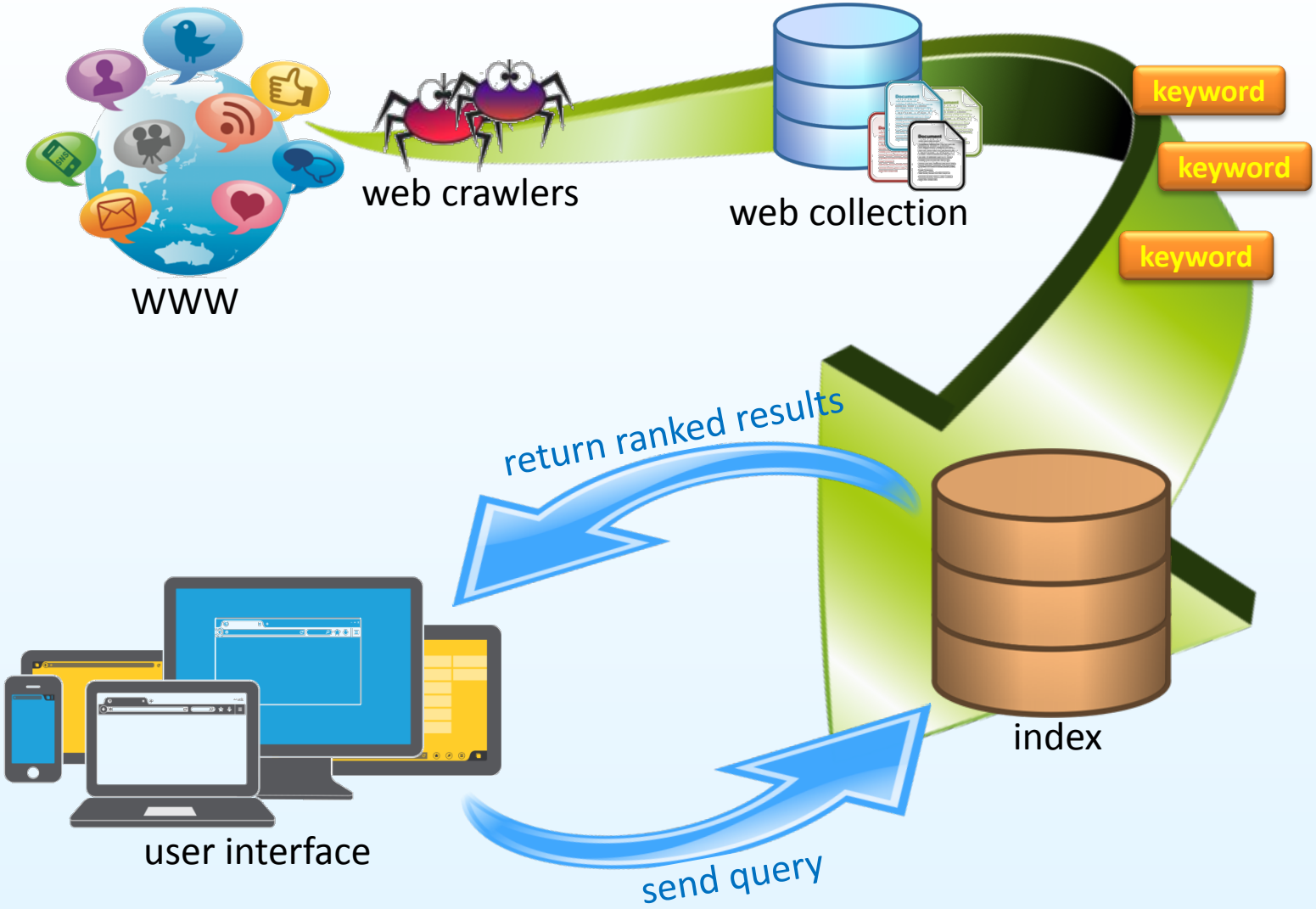
August 2016 → http://marketshare.hitslink.com

Search Engine Trend

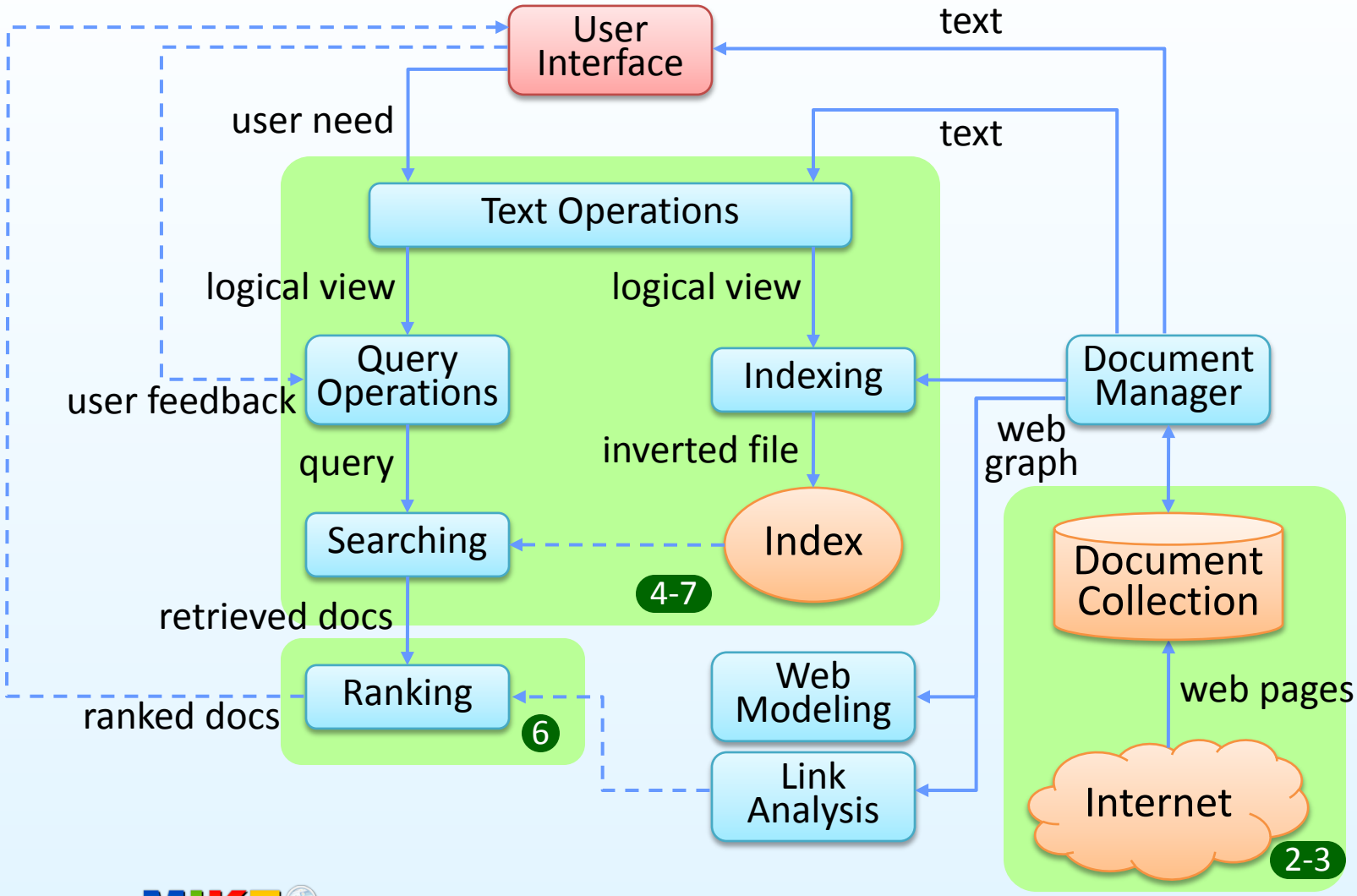
Year	Annual Number of Google Searches	Average Searches Per Day
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000 **Googles official first year*	9,800

June 2015 → http://www.statisticbrain.com/google-searches/

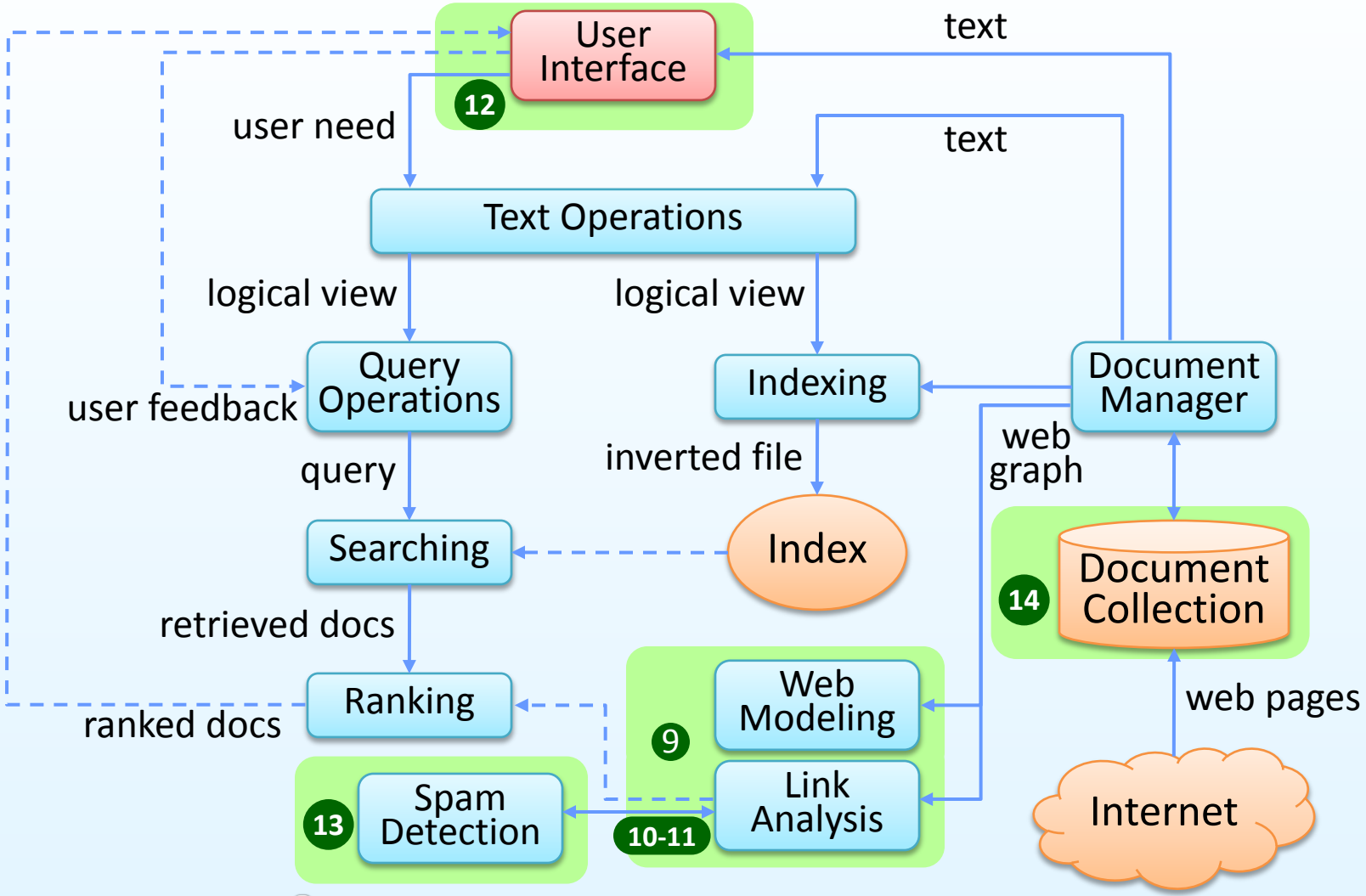
Web Search Engine Architecture



Course Organization (before Midterm)



Course Organization (after Midterm)



Any Question?