An Introduction to Information Retrieval

Kuan-Yu Chen (陳冠宇)

2017/09/14 @ TR-509, NTUST

Information Retrieval

- Information Retrieval is a broad area of computer science
  - It mainly focuses on proving the **users** with easy access to **information** of **their interest**

- IR deals with the representation, storage, organization of, and access to information items
  - documents, Web pages, online catalogs, structured records, multimedia objects

The Goal of IR – 1

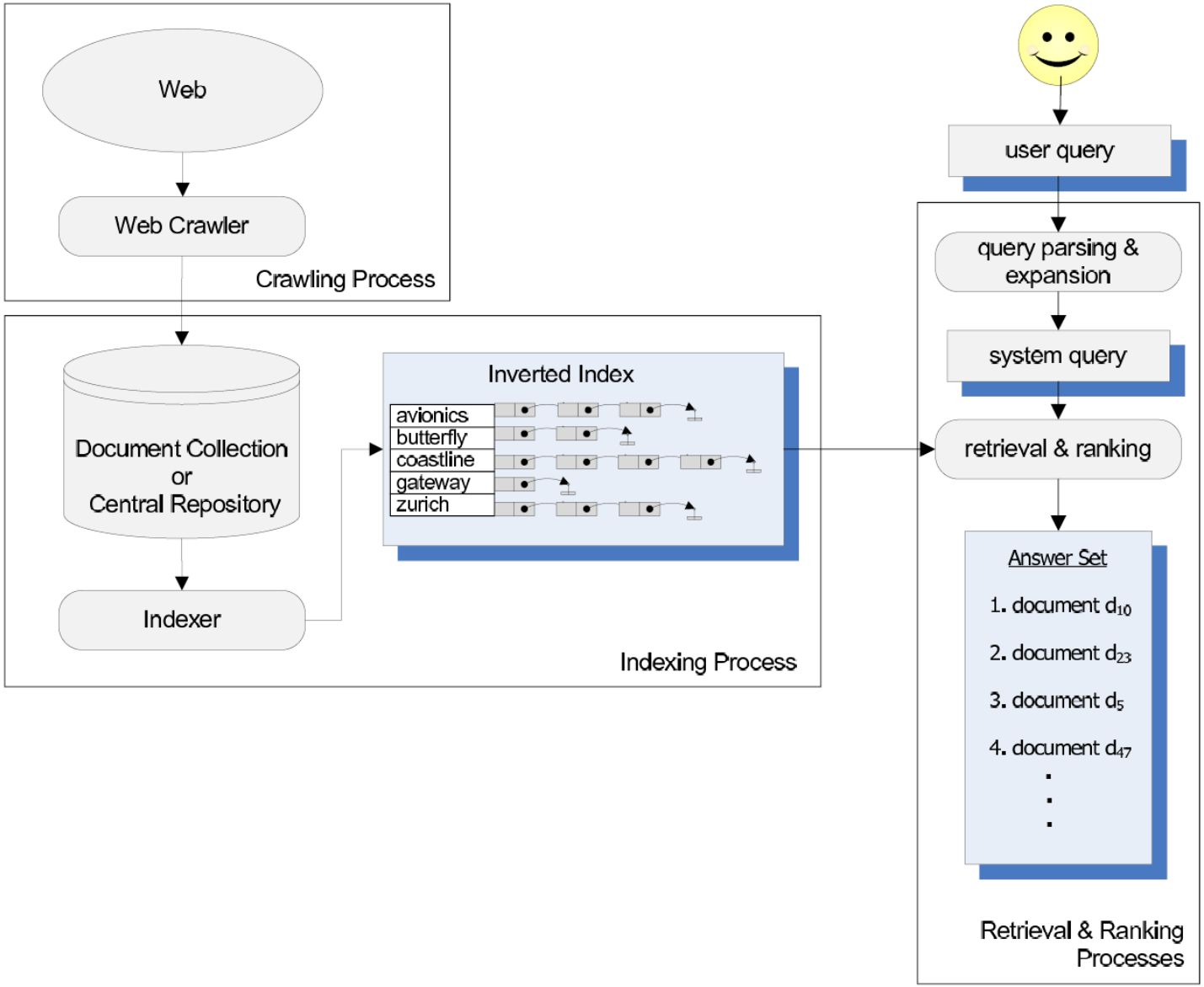
- Early goals of the IR area
  - Indexing text
  - Searching for useful documents in a collection

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

The Goal of IR – 2

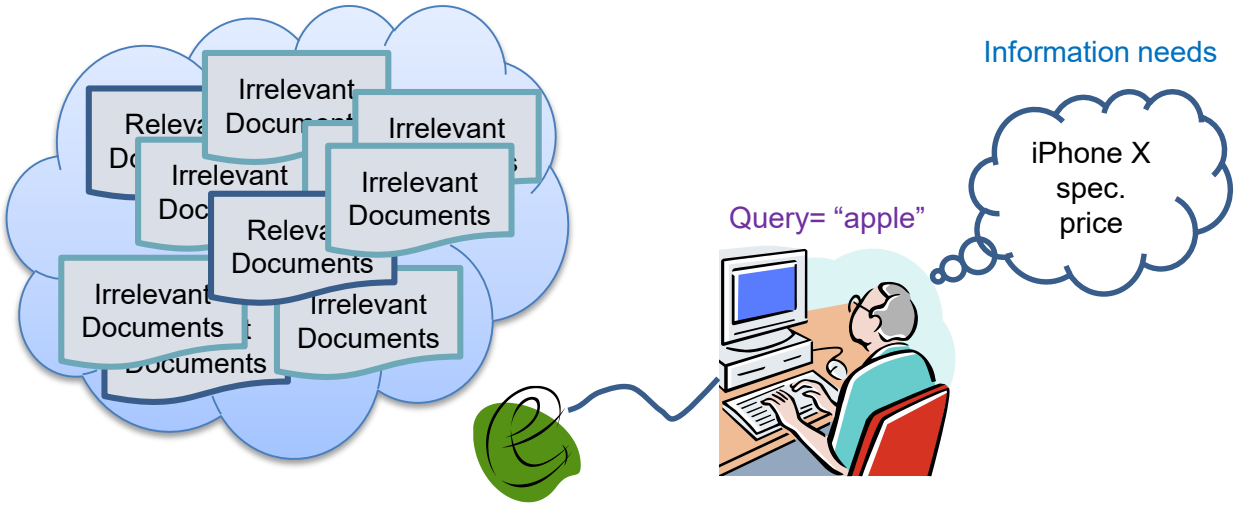
- In terms of research, the area may be classified into two distinct and complementary points
  - Computer-centered
    - Building efficient indexes/representations
    - Processing user queries with high performance
    - Developing ranking algorithms
  - Human-centered
    - Studying the behavior of the user
    - Understanding the information need

General Architecture of the IR System



IR Problems – 1

- Users of modern IR systems, such as search engine users, have **information needs** of varying complexity
  - A full description of the user information need is not a good query to be submitted to the IR system
  - Instead, the user translate this information need into a query
    - a set of ***keywords***, or ***index terms***, which summarize the user information need
  - The key goal of the IR system is to retrieve information that is useful or relevant to the user



IR Problems – 2

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

- The definition of 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 nonrelevant items as possible***

- The notion of **relevance** is of central importance in IR

About "Relevance"

- Relevance is a personal assessment that depends on the task being solved and its context

- Relevance can change with
  - Time
  - Location
  - device

Until now, no IR system can provide perfect answers to all users all the time!

Information Retrieval vs. Data Retrieval

- Data Retrieval
  - concentrates on determining which documents of a collection contain the keywords in the user query
  - can't satisfy the user information need
  - provides a solution to the user of a database system

- Information Retrieval
  - retrieves information about a subject
  - aims at satisfying the user information need
  - offers a way to the user of a information system

The Web – History 1

- Berners-Lee worked in Geneva at the CERN
  - Conseil Européen pour la Recherche Nucléaire

- In CERN, researchers who wanted to share documentation with others had to reformat their documents to make them compatible with an internal publishing system

- Berners-Lee reasoned that it would be nice if the solution of sharing documents were decentralized

The Web – History 2

- In 1990, Berners-Lee
  - Wrote the HTTP protocol
  - Defined the HTML language
  - Wrote the first browser, which he called World Wide Web
  - Wrote the first Web server

- In 1991, he made his browser and server software available in the Internet

- The Web was born!

The Web

- Since its inception, the Web became a huge success
  - Well over 20 billion pages are now available and accessible in the Web
  - More than one fourth of humanity now access the Web on a regular basis
  - The e-Publishing era

How the Web Change the Search

- The Web is composed of pages distributed over millions of sites and connected through hyperlinks
  - The size of the collection
  - A new phase in the IR process, introduced by the Web, is called "crawling"
- The volume of user queries submitted on a daily basis
  - Performance and scalability have become critical characteristics of the IR system
- In the context of Web, predicting relevance is much harder than before
  - New sources of evidence: hyperlinks and user clicks
  - Web Spam
- Privacy, Security, Copyright, Cross-language

Questions?

kychen@mail.ntust.edu.tw