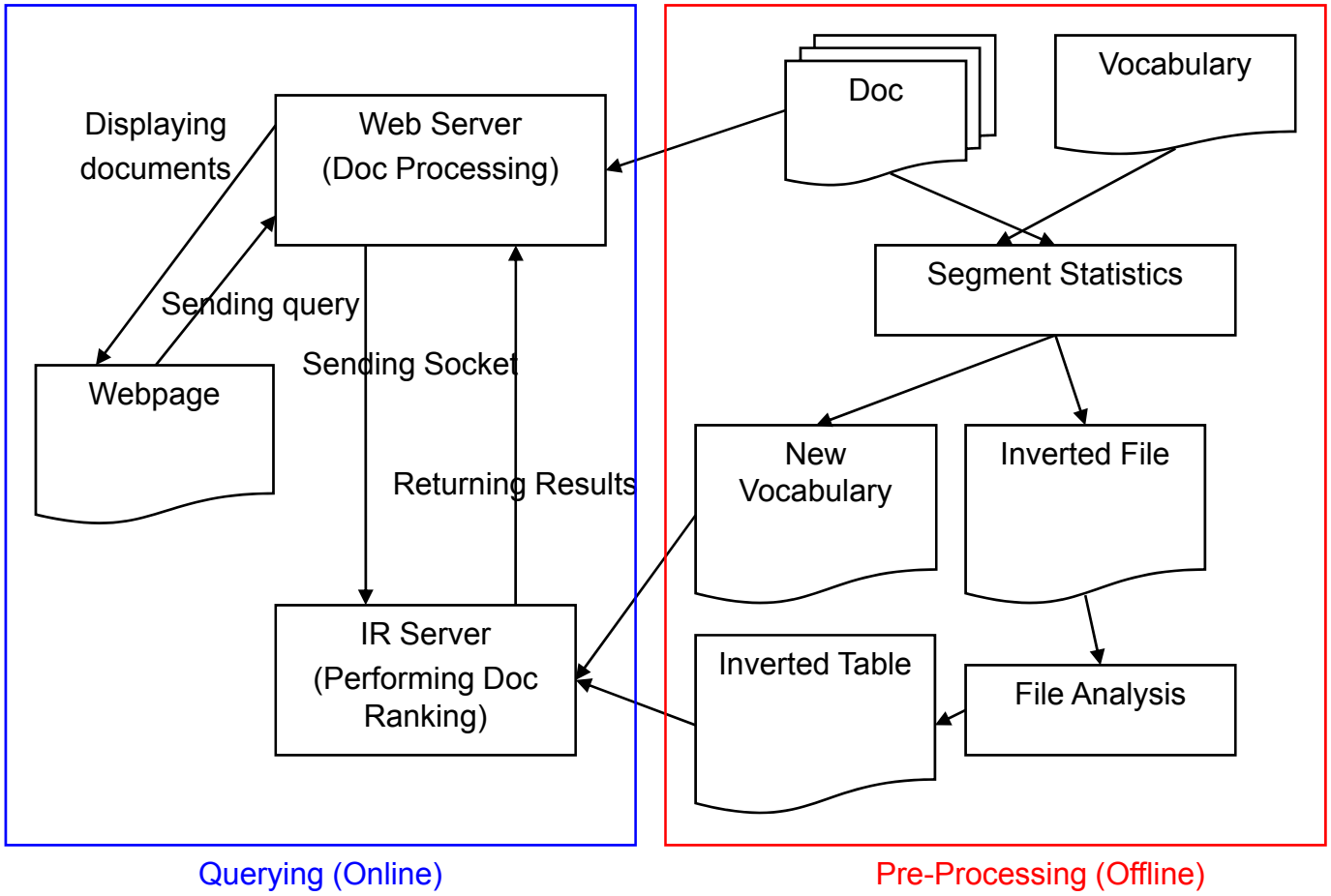
Information Retrieval 2009 - Description of Final Project

Task: Building a Web Retrieval System

- You have to build a client-server-based web retrieval system
- You can use any indexing features or retrieval models that have learned from this course
- You have to submit your results no later than January 22, 2010
- You should consult Dr. Berlin Chen or TA (Ms. Yu-Mei Chang) whenever you encounter any problem. (Please feel free to do that!)

Schematic Description of the Possible Implementation Procedure



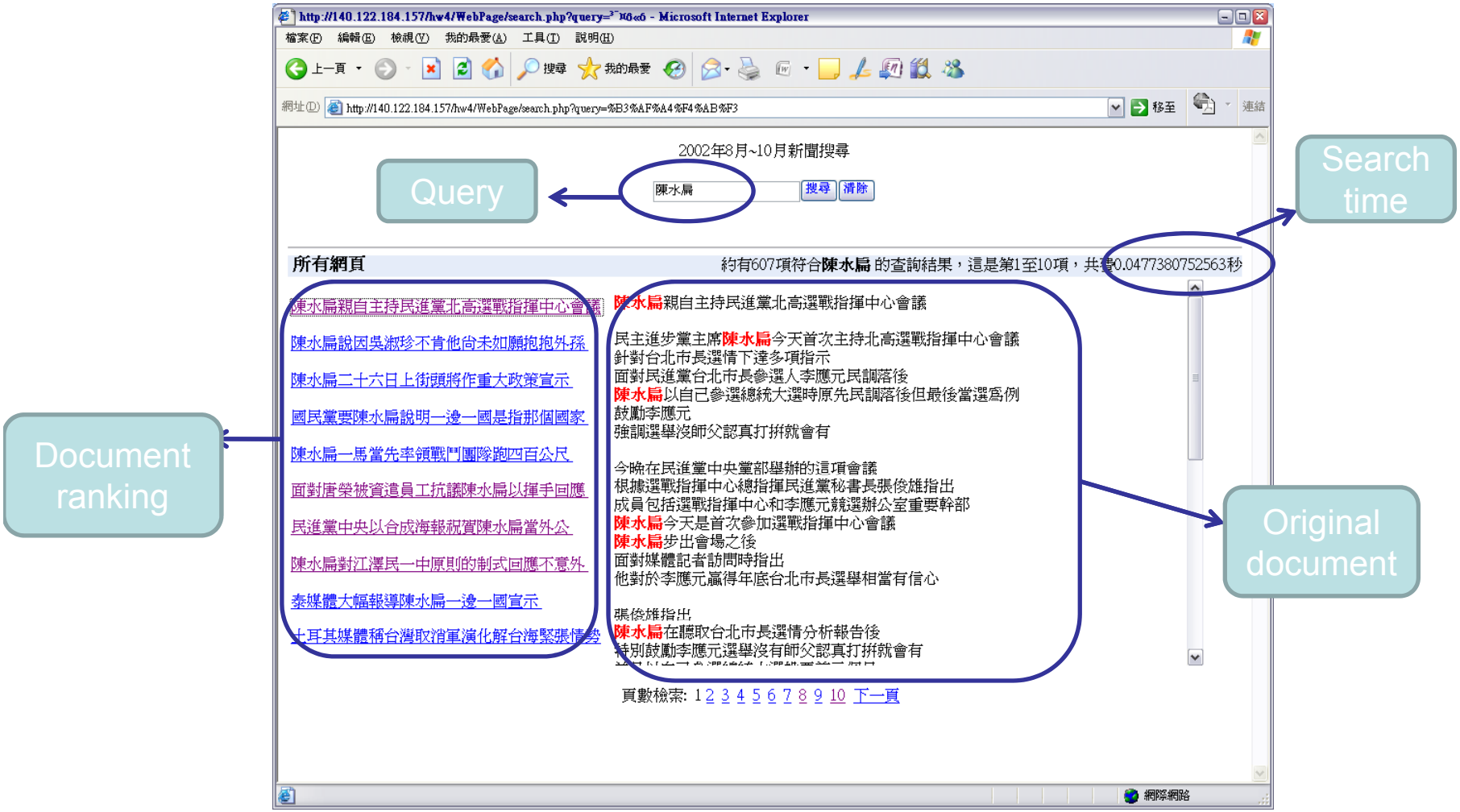
Preparing Data

- IR-WebSearchHW.rar
  - “Lexicon2003-72k.txt”: a vocabulary for Chinese word tokenization (segmentation)
  - “newvoc.txt”: containing a list of all distinct words that appear in the document collection
  - WebPage : this subdirectory contains information like index.htm, search.php and show.php
  - Web service code
  - Download: http://slp.csie.ntnu.edu.tw/IR-HW3/IR-WebSearchHW.rar
- WebIRCorpus.rar
  - Original documents: (story_pureText)-to show on the webpage
  - Documents resulting from word tokenization: (new_story_pureText)-for search
  - Download: http://slp.csie.ntnu.edu.tw/IR-HW3/WebIRCorpus.rar

More on the Programming Issues

- Building the inverted file
  - Output: the inverted file
  - Total: 39,027 documents
  - Goal: to speedup the search time
- IR-WebSearchHW.rar
  - Insert the “functionality” that you might need
    - E.g., void doquery( );void clearquery( ); segment( );…
- Web server package

An Example System (Query: 陳水扁)



An Example System (Query: 師大)