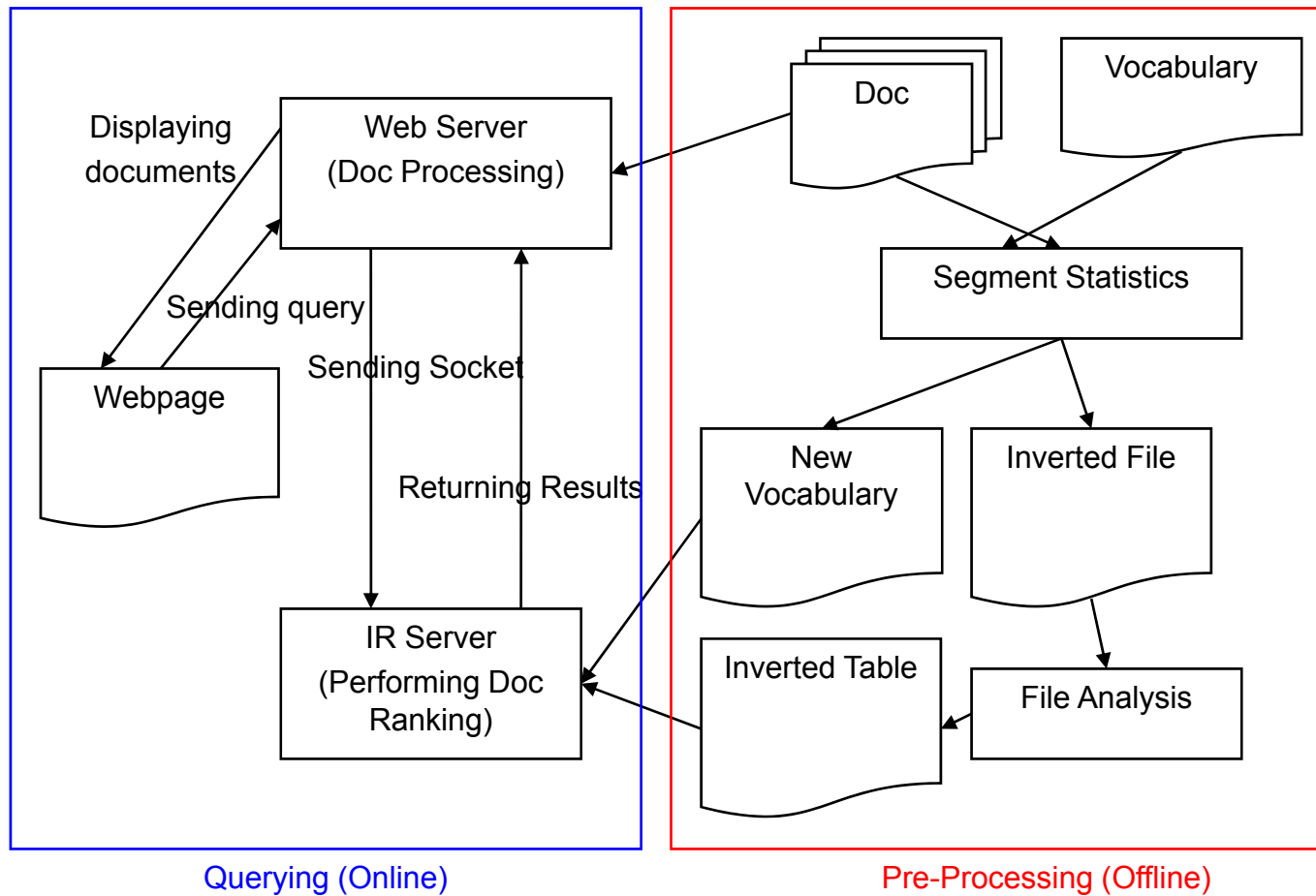


# **Information Retrieval 2009 - Description of Final Project**

# 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)

The screenshot shows a Microsoft Internet Explorer browser window displaying a search results page. The address bar shows the URL: `http://140.122.184.157/hw4/WebPage/search.php?query=%E3%AF%44%F4%AB%F3`. The search results are for the query "2002年8月~10月新聞搜尋" with the search term "陳水扁".

Annotations on the screenshot include:

- Query:** A box pointing to the search input field containing "陳水扁".
- Search time:** A box pointing to the text "約有607項符合陳水扁的查詢結果，這是第1至10項，共需0.0477380752563秒".
- Document ranking:** A box pointing to the list of search results on the left side of the page.
- Original document:** A box pointing to the full text of the first search result on the right side of the page.

The search results list includes:

- [陳水扁親自主持民進黨北高選戰指揮中心會議](#)
- [陳水扁說因吳淑珍不肯他尚不如願抱抱外孫](#)
- [陳水扁二十六日上街頭將作重大政策宣示](#)
- [國民黨要陳水扁說明一邊一國是指那個國家](#)
- [陳水扁一馬當先率領戰鬥團隊跑四百公尺](#)
- [面對唐榮被資遣員工抗議陳水扁以揮手回應](#)
- [民進黨中央以合成海報祝賀陳水扁當外公](#)
- [陳水扁對江澤民一中原則的制式回應不意外](#)
- [泰媒體大幅報導陳水扁一邊一國宣示](#)
- [土耳其媒體稱台灣取消軍演化解台海緊張情勢](#)

The first result's full text is displayed on the right, starting with "民主進步黨主席陳水扁今天首次主持北高選戰指揮中心會議".

# An Example System (Query: 師大)

http://140.122.184.157/hw4/WebPage/search.php?query=@vKj - Microsoft Internet Explorer

2002年8月~10月新聞搜尋

師大 [搜尋] [清除]

所有網頁 約有110項符合師大的查詢結果，這是第1至10項，共費0.0116889476776秒

<a href="#">台灣師大與金門衛星數位視訊教學試播成功</a>	台灣師大與金門衛星數位視訊教學試播成功
<a href="#">台灣師大與泰來拉隆功大學簽署教育合作</a>	國立台灣師範大學進修部今天透過衛星數位視訊系統自台北與金門進行視訊遠距同步教學試播獲得成功
<a href="#">台師大台科大舉行宣佈合併後首次會議</a>	將積極規劃把進修課程突破時空限制傳送到離島偏遠地區甚至世界各地
<a href="#">水上嘉年華會明天在台灣師大游泳館舉行</a>	
<a href="#">教育部說台科大台師大併校後不影響技職招生</a>	台灣師大是接受金門縣政府教育局委託開辦中小學及幼稚園現職教師特殊教育學分進修班
<a href="#">陳總統明令褒揚師大前校長孫亢會</a>	共有五十三名學員上課為突破時空因素限制
<a href="#">教部期盼台科大台師大合併激勵其他大學整併</a>	台灣師大進修部與年代電通公司合作今天進行數位視訊傳送教學試播
<a href="#">台灣師大前校長孫亢會以一</a>	上午進行的試播台灣師大潘裕豐教授於台北年代公司攝影棚主講特殊教育導論透過衛星同步傳送到金門收視點的金寧中小學師生並進行互動
<a href="#">陳舜田</a>	視訊畫質清晰穩定度高

頁數檢索: 1 2 3 4 5 6 7 8 9 10 下一頁