

工業工程專題

以文字雲技術支援互動搜尋

|指導教授|

侯建良 教授

|專題組員|

李政融 鍾岱穎

● 研究動機

現代搜尋引擎功能齊全，已是人們獲取資訊的主要途徑之一，但欲透過搜尋得到想要的資料，便必須輸入適當的關鍵字。若輸入錯誤的關鍵字，將可能導致搜尋方向與使用者之預期產生分歧，最後便無法得到想要的結果。因此，我們決定研發一套互動搜尋系統，藉由文字分析及文字雲的呈現，希望能夠增加使用者的搜尋效率。

● 摘要

本研究之互動搜尋系統有別於以往使用者的搜尋經驗，會對使用者的搜尋給予相對應的回饋，讓使用者能於搜尋過程中，依照系統的建議，及時調整搜尋方向，故並非只有使用者單方面對搜尋引擎進行搜尋。

本系統會對使用者搜尋到的網頁進行文字擷取，經過斷詞系統的分析後，再計算並排序所有網頁文字之TF-IDF值。此時，系統將提供兩種搜尋途徑給使用者。第一種是文字雲搜尋介面。文字雲Server端會依照字詞出現的頻率，利用TF-IDF值排除贅字後，藉由字體大小、顏色的變化，於Client端呈現出完整的文字雲。第二種方法是遞迴搜尋系統。遞迴搜尋系統透過對字詞頻率及TF-IDF值的統計，自動校正關鍵字並進行數次搜尋，遞迴結束後便會產生系統校正後的關鍵字。

本互動搜尋系統也參考現代瀏覽器之設計，以使用者為中心進行介面設計、模擬瀏覽器功能，方便使用者搜尋以及比較以上兩種搜尋途徑所得到的結果。

● 開發環境及相關技術

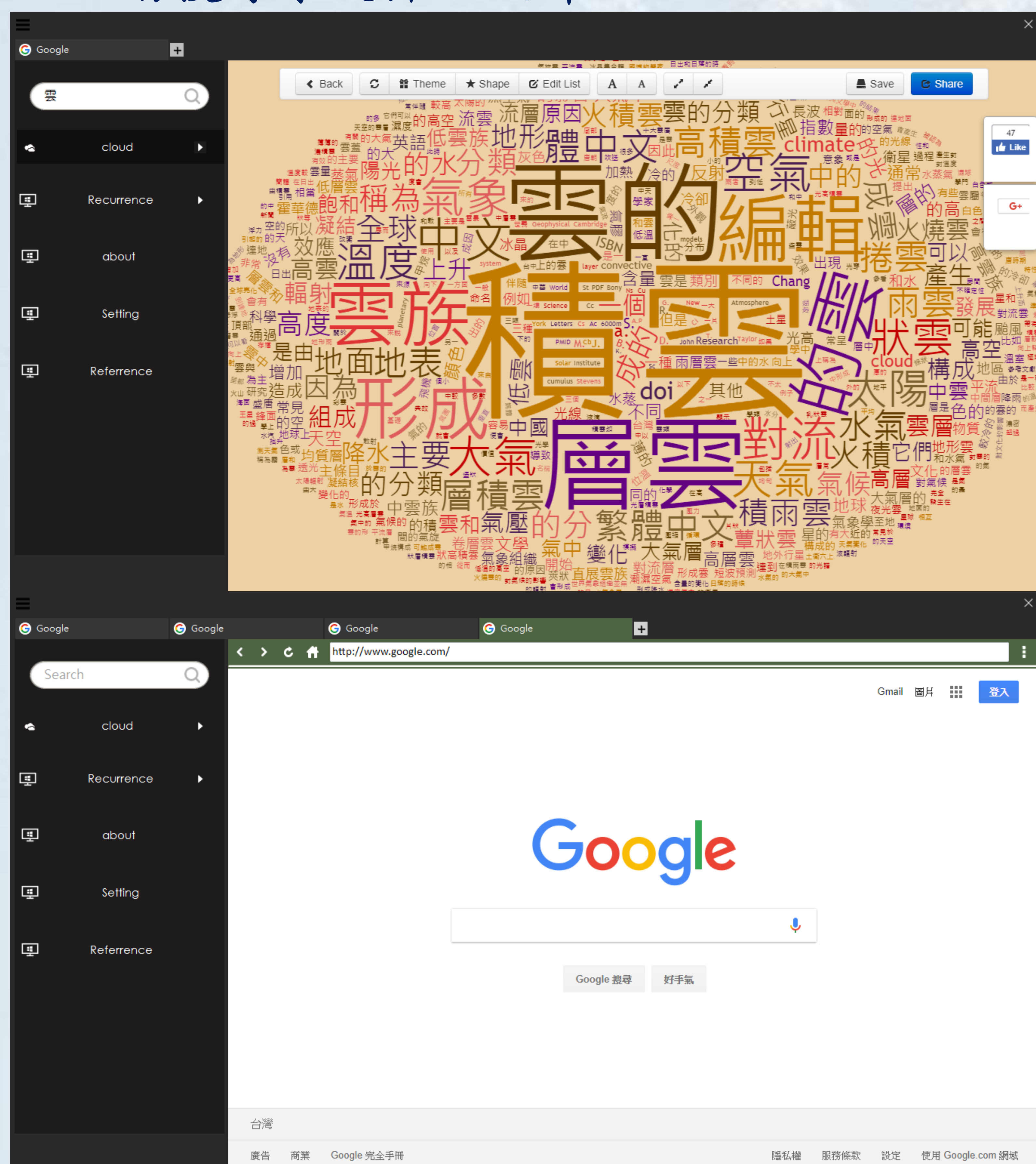
使用語言：

- Python (核心功能，如擷取文字、斷詞、TF-IDF、遞迴搜尋...等)
- C# (系統介面設計)
- JavaScript (文字雲)

開發環境：

- Microsoft Visual Studio 2017
- PyCharm Community Edition 2017.2.3
- WebStorm 2017.2.5

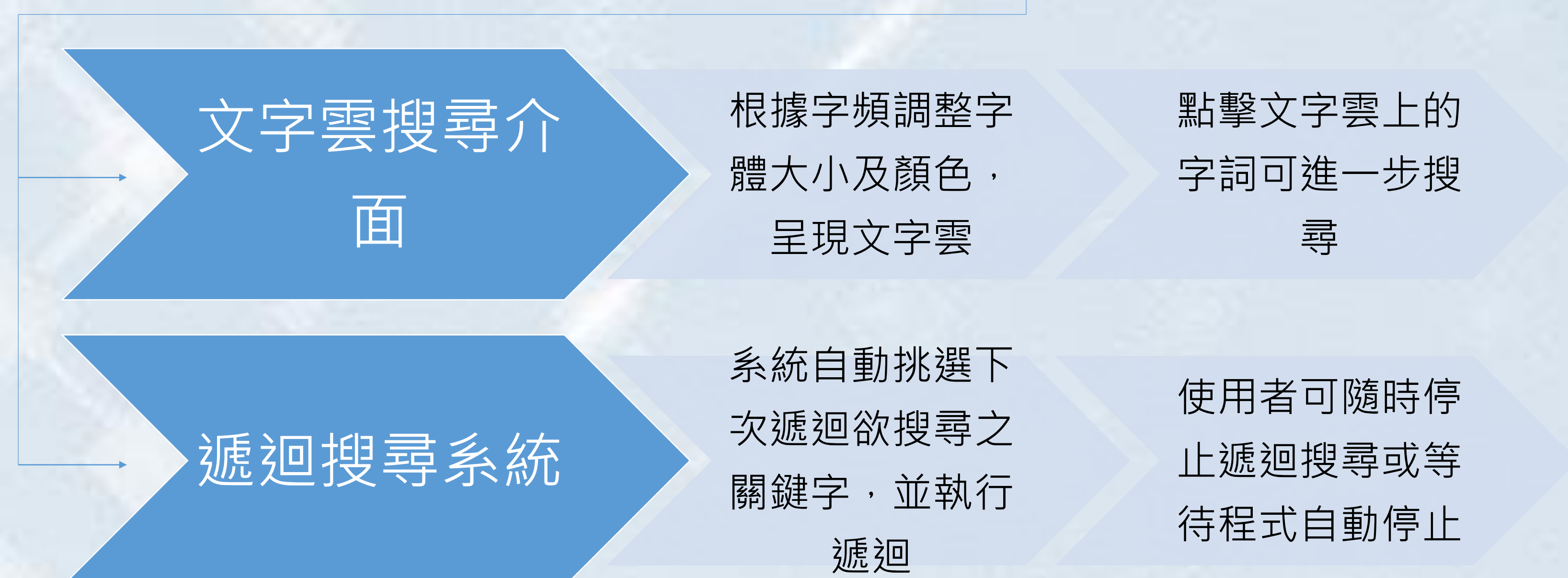
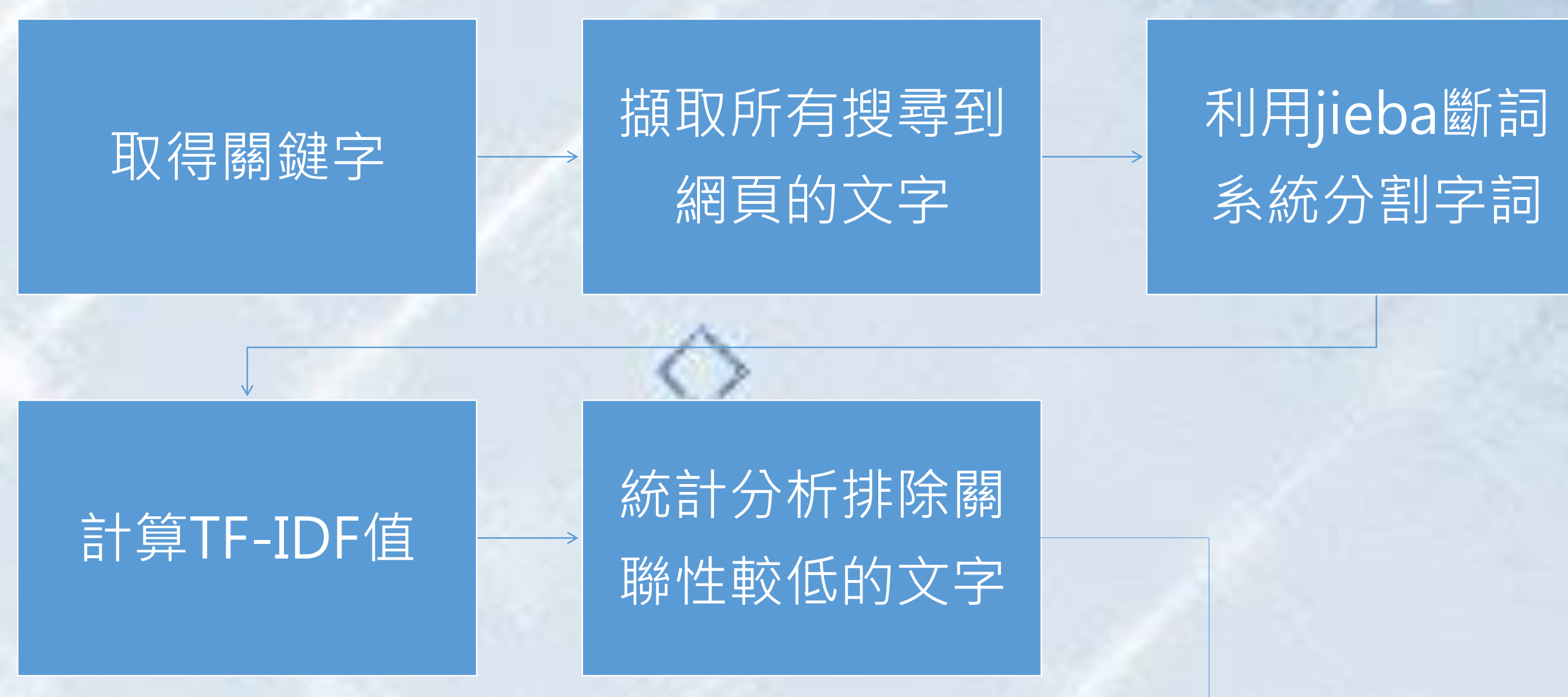
● 互動搜尋系統介面設計



● 系統架構



● 文字分析流程



● 系統實作

```
async def get_url(url):
```

根據搜尋引擎結果取得每列搜尋結果的網址

```
async def connect(url, headers=None, proxy=None, encoding='utf-8', raw=False):
```

取得網頁URL進行送出request請求，最後得到網頁的原始碼

```
def get_text(soup):
```

分析網頁原始碼，排除HTML Tag後抓取純文字

```
def split_into_word(article):
```

利用jieba斷詞系統進行斷詞

```
def get_tf_idf(seg_list, topK=10, withWeight=True, queue=None):
```

計算字詞的TF-IDF值，並依照大小排序

```
async def recursive_search(loop_count):
```

```
async def cloud_search():
```

執行遞迴搜尋或文字雲搜尋

106年度 工業工程專題
清華大學工業工程與工程管理學系

