**資料**

**檢索**

**『IR』**

**組員：**

**資三A\_01156103\_柯皓元**

**資三A\_01156109\_陸建綱**

**資三A\_01156125\_林炳宏**

**指導老師：**

**柯淑津 教授**

* **Outline**
* **封面．．．．．．．．．．．．．．．．．．．．．p.1**
* **目錄．．．．．．．．．．．．．．．．．．．．．p.2**
* **目標、關鍵字．．．．．．．．．．．．．．．．．p.3**
* **Standard Boolean Model．．．．．．．．．．．．．p.3~6**
* **Extended Boolean Model．．．．．．．．．．．．．p.6~9**
* **Vector Space Model．．．．．．．．．．．．．．．p.9~13**
* **分析比較．．．．．．．．．．．．．．．．．．．p.13**
* **目標**

比較Standard Boolean Model，Extended Boolean Model，Vector Space Model這三個Model在查詢關鍵字上面的差別和優缺。

* **關鍵字**

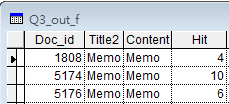
羽球AND ( 全民OR運動OR比賽 ) AND推廣

* **Standard Boolean Model**

以文章為單位 (沒排名)。

* **以下是查詢的結果：**

在第1808、5174、5176篇文章中都有中，分別中了4、10、6句。



* **優點：**

1. 結果可預測，容易解釋。
2. 包含很多特徵。
3. 處理速度快，因為不用事先運算，可以直接使用bigram。

* **缺點：**

1. 效果好壞決定在使用者上。
2. 簡單的Query沒辦法處理得很好，會找到太多不相關的。
3. 複雜的Query對使用者來說又太困難。

* **第1808篇 (合庫寒假育樂營 開始報名)：**



* **第5174篇 (金屬中心助木竹藤、羽球拍產業升級)：**



* **第5176篇 (吳宜倫承衣缽散播羽球種子)：**

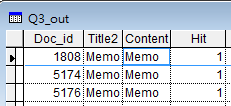


* **Extended Boolean Model**

以句子為單位 (有排名)。

* **以下是查詢的結果：**

在第1808、5174、5176篇文章中都有中。

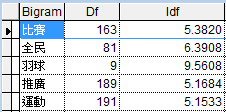


* **優缺點：**

1. 基本上和Standard Boolean Model很像。
2. 以句子為單位，要求更加嚴苛。
3. 可能找出更好的符合Query，因為以句子為單位，但也有可能沒有找到我們所希望的資料，因為要求過於嚴苛。

* **第1808篇 (合庫寒假育樂營 開始報名)：**
* **第5174篇 (金屬中心助木竹藤、羽球拍產業升級)：**
* **第5176篇 (吳宜倫承衣缽散播羽球種子)：**
* **Vector Space Model**
* **以下是查詢的結果：**

先算出它們的Df (出現次數)和Idf (重要度or鑑別率)。



* **優點：**

1. 數學為基礎的方法。
2. 會考慮document內部(tf) 和整體(idf) 出現頻率。
3. 允許部分match和排名。
4. 資料量大還是很有效的計算。

* **缺點：**

1. 缺少語意、句法的訊息 (完全以字詞判斷)。
2. 缺少Boolean Model的控制 (沒有查詢邏輯)。

* **結果發現一共有573篇中：**

**包含Standard Boolean Model和Extended Boolean Model找到的三篇。**

* **第5176篇 (吳宜倫承衣缽散播羽球種子)：**
* **第1329篇 (羽球兩大盛事 圓滿閉幕)：**

**用Standard Boolean Model和Extended Boolean Model所沒有找到的。**

* **第4216篇 (長春國小 羽球國手搖籃)：**

**用Standard Boolean Model和Extended Boolean Model所沒有找到的。**

* **第5174篇 (金屬中心助木竹藤、羽球拍產業升級)：**
* **第1808篇 (合庫寒假育樂營 開始報名)：**
* **分析比較**

1. Standard Boolean Model可以找出完全符合自己下的查詢，但是查尋得要下得很精準。
2. Extended Boolean Model跟Standard Boolean Model一樣，但是比Standard Boolean Model更精準，因為是以句子為單位。
3. 相較於Extended Boolean Model跟Standard Boolean Model，Vector Space Model比這兩種方法更加直覺，而且找出來的文章更符合我們希望的。
4. 但是Vector Space Model的分數愈小愈不符合，查出來的可能就不是我們所希望的，不過前幾名的都符合期待，甚至超越我們所想要的。