資訊檢索第一步

國家圖書館　知識服務組

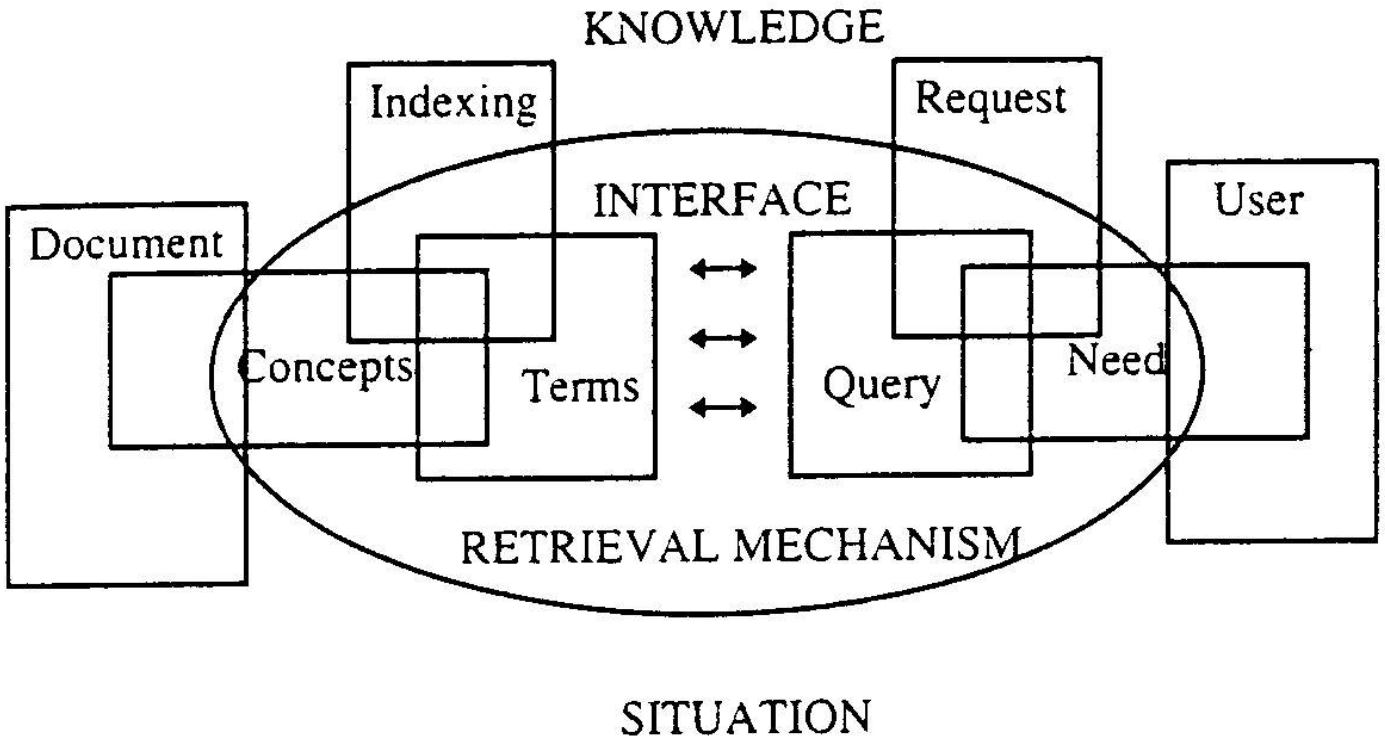
綱要

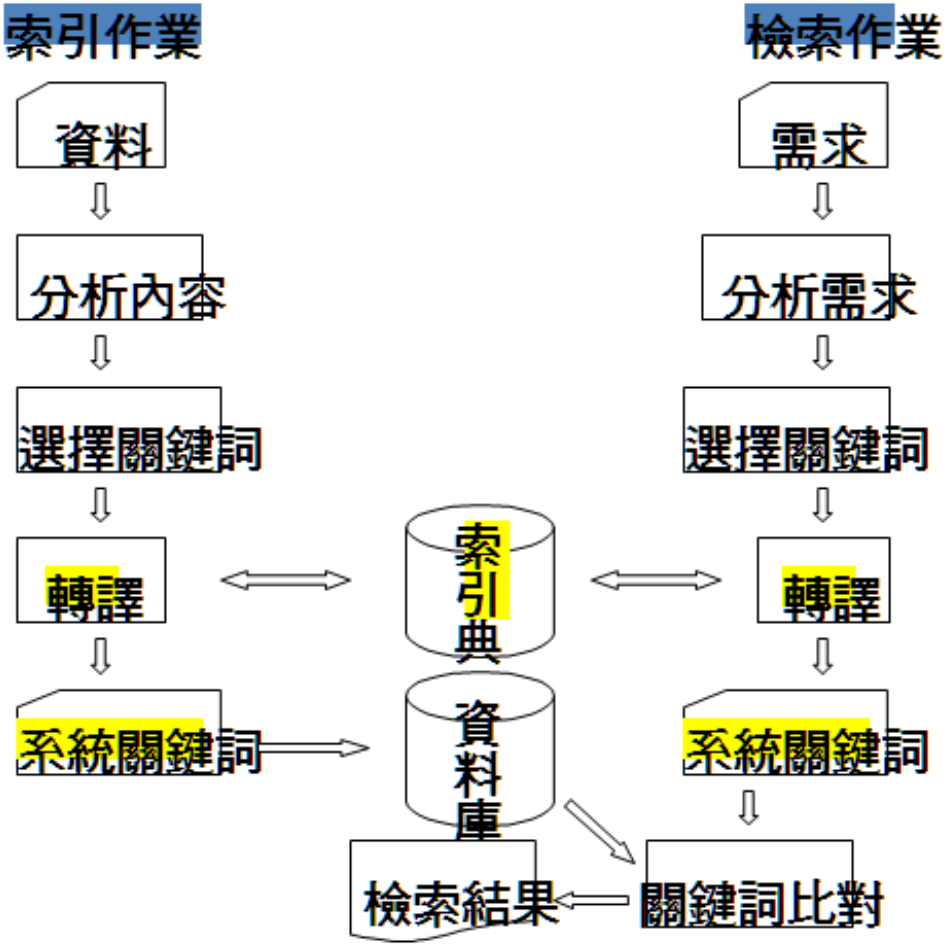
- 資訊檢索之基本概念
- 檢索策略
- 檢索技巧

資訊檢索之基本概念

- IS & R
- 自然語言 vs. 控制字彙
- precision vs. recall
- 布林邏輯運算元
- 切截
- 相近運算元
- Known item search vs. subject search

資訊檢索與儲存 (IS & R Model)




索引作業
資料
分析內容
選擇關鍵詞
轉譯
系統關鍵詞
檢索作業
需求
分析需求
選擇關鍵詞
轉譯
系統關鍵詞
索引典
資料庫
檢索結果
關鍵詞比對

自然語言

自然語言是相對於人工語言的一種人類語言，也是最合乎人類教談行為的溝通方式，它依循著人類自然進化而發展，成為人和人之間溝通的最基本工具，如中文、英文、日文等都是自然語言。

控制字彙

- Maintenance
  UF Preventive maintenance
  Upkeep

- Preventive maintenance use **Maintenance**
  Upkeep use **Maintenance**

回收率 & 精準率

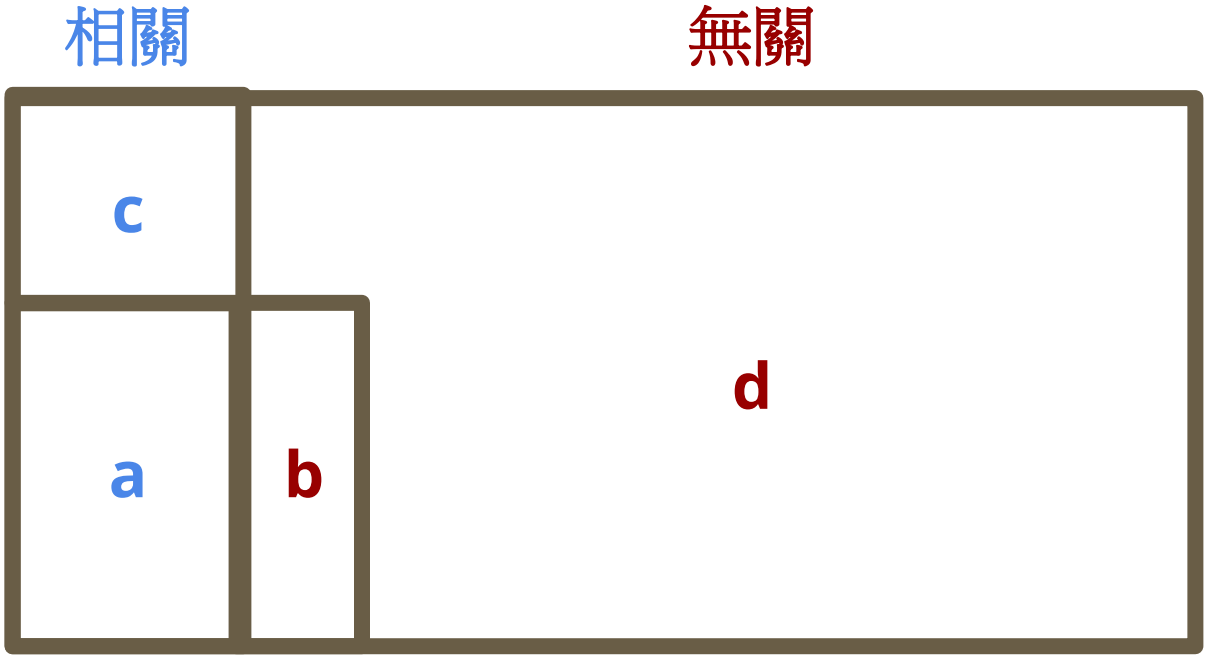
Recall(回收率/查全率) & Precision(精確率/查準率)

	相關	不相關
檢索到	a	b
未檢索到	c	d
總數	a+c	b+d

$$\text{回收率} = \frac{a}{a+c} = \frac{\text{檢索所得之相關文章篇數}}{\text{資料庫中所有相關文章篇數}}$$

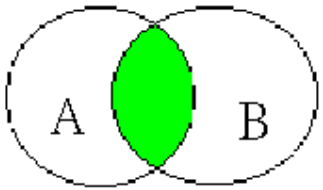
$$\text{精確率} = \frac{a}{a+b} = \frac{\text{檢索所得之相關文章篇數}}{\text{檢索所得之所有書目篇數}}$$

査全率=a/(a+c)　　査準率=a/(a+b)

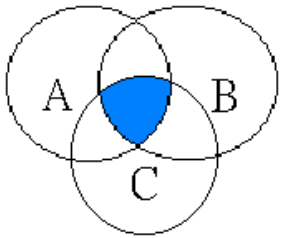


布林邏輯(AND交集、OR聯集、NOT剔除)

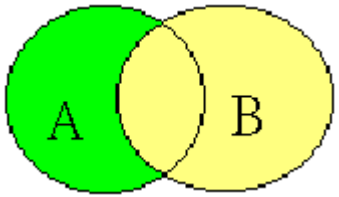
A **AND** B



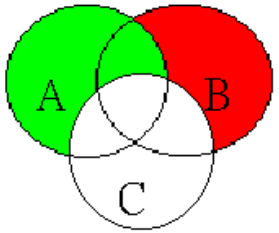
A AND B AND C



A **OR** B



A OR B **NOT** C



切截(truncation)

右切截 單複數、詞性不同	**詞間切截** 單複數	美式 / 英式拼音
Library, libraries, librarian, librarians, librarianship --> librar*	Woman, Women --> Wom#n	Color, Colour --> Colo#r

相近運算元 (adjacent/near)

ANALOG*	ADJ1	DIGITAL*	482
ANALOG*	NEAR1	DIGITAL*	506

Known item search

已知書目之檢索，即精確檢索

- 用已知的書目資料來檢索，包括：作者、題名、期刊名、出版商、出版年...

Subject search

主題檢索

- 想檢索一下到底有那些關於某主題的文獻存在

檢索策略

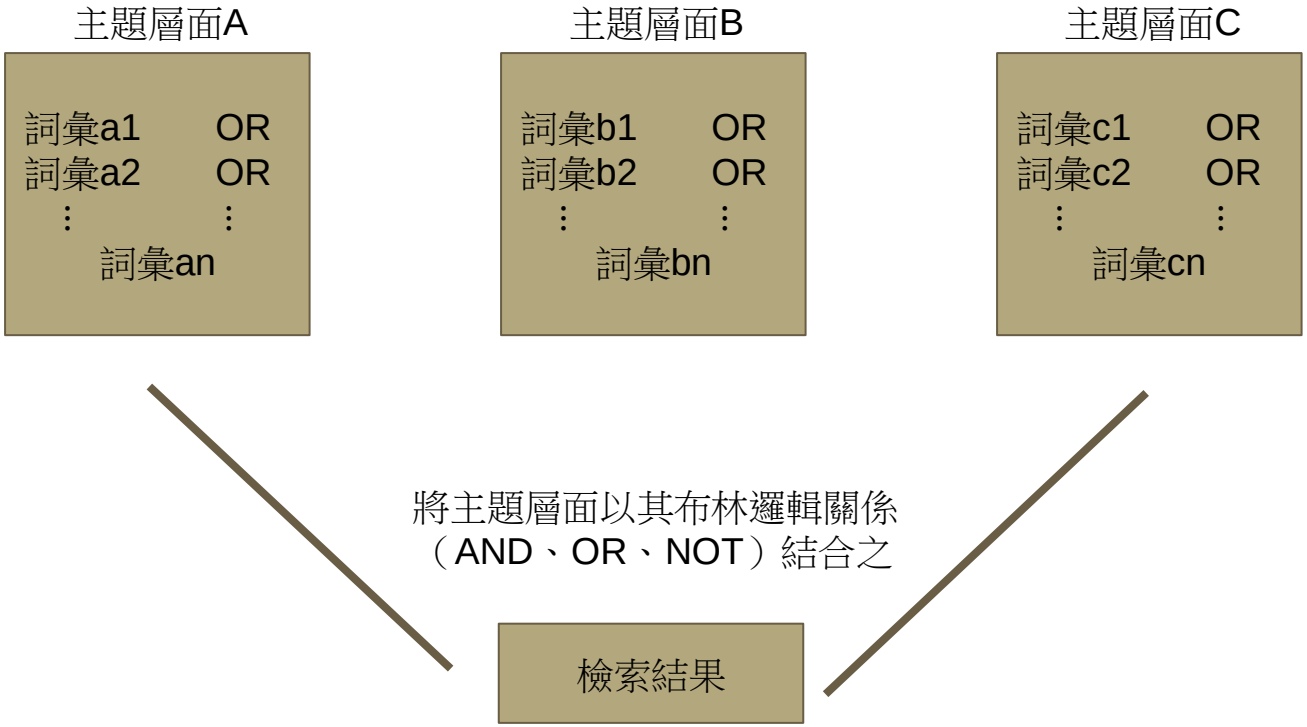
針對一檢索問題之通盤考量或全面性規劃

- 分區組合檢索法 (Block Building)
- 引用文獻滾雪球法 (Citation Pearl Growing)
- 簡易檢索 (Brief search)
- 主題層面連續檢索 (successive facet strategies)
- 主題層面配對檢索 (pairwise facets strategies)

分區組合檢索法

1. 選擇資料庫
2. 確定問題之主要概念及其布林邏輯關係
3. 依序找出代表每個概念之所有詞彙
4. 將各概念下所有詞彙以“OR”連結
5. 將步驟4所得結果以步驟2所決定之布林邏輯關係進行結合
6. 依步驟1至步驟5 規劃檢索敘述
7. 輸入檢索敘述
8. 評估檢索成果

分區組合檢索法示意圖


主題層面A
詞彙a1 OR
詞彙a2 OR
詞彙an
主題層面B
詞彙b1 OR
詞彙b2 OR
詞彙bn
主題層面C
詞彙c1 OR
詞彙c2 OR
詞彙cn
將主題層面以其布林邏輯關係
（AND、OR、NOT）結合之
檢索結果

引用文獻滾雪球法

- 事先掌握一篇或數篇相關文章，利用這些相關文章找尋更多相關的文章，如此相關文章就像雪球一樣越滾越大
- 在資訊檢索上的應用：以相關文章的關鍵字或敘述語繼續檢索
- 是由precision反向追求recall的方法
- 通常必須進行多次檢索，才能找到足夠的相關文章

簡易檢索

最常見的檢索

通常用簡單的幾個關鍵字，加上布林邏輯的組合

快速，同時檢索到的文章不多，recall低

適用情形：

- 檢索者只想閱讀“幾篇”相關文章
- 執行已知書目檢索時
- 檢索概念相當專指 (specific) 時

主題層面連續檢索

- 在決定檢索問題的主題層面之後，必須確認各主題層面的優先順序
- 在最專指概念或是可能產生最少資料的概念輸入系統後，如果產生太多資料，再輸入其他次要概念與之結合
- 直到檢索者認為檢索筆數可以接受為止

主題層面連續檢索 1/2

適用情形：

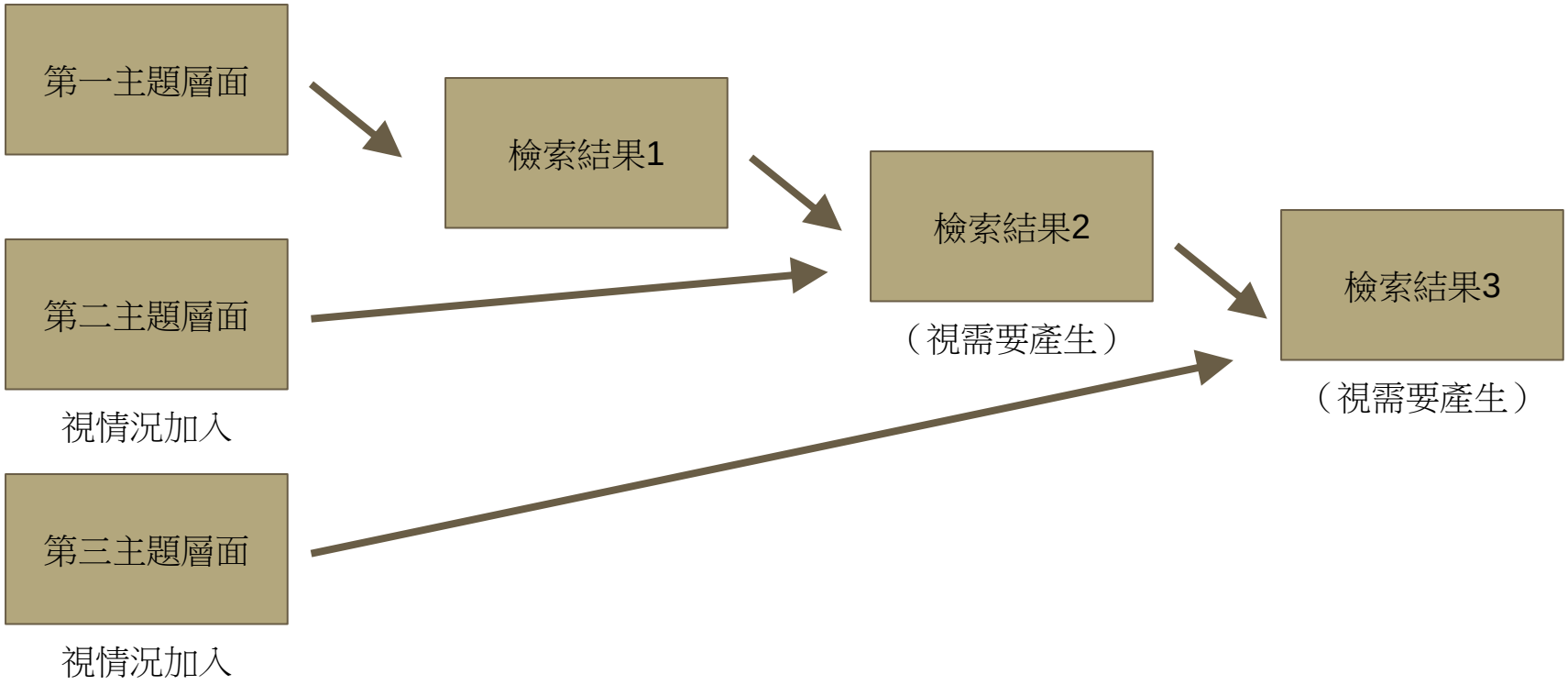
- 當所有主題層面以布林運算元結合，可能產生零筆資料時
- 當檢索問題中有一至二個主題層面涵義相當模糊時
- 當檢索問題具備其他非主題之檢索條件時（如：資料類型、語文、出版年代），可將此非主題檢索條件視為第一個檢索概念

主題層面連續檢索 2/2

適用情形：

- 當檢索者寧願忍受誤引，而不願失去相關文章時
- 當加入其他主題層面所花費的時間和金錢，可能會超過直接列出檢索結果，每筆一一審視時
- 當相關文獻過少，檢索者願意檢視一些相關度較低的文章時

主題層面連續檢索示意圖



主題層面配對檢索

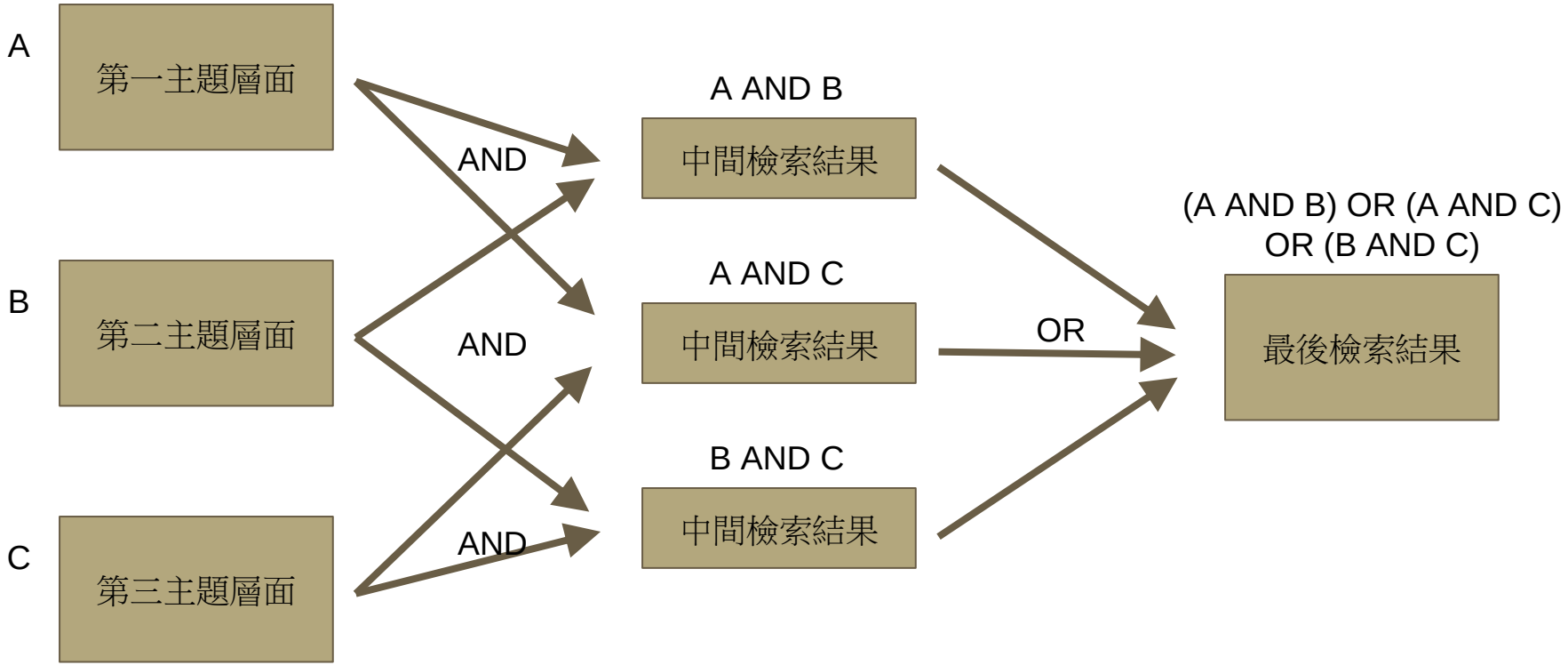
是先將主題層面兩兩配對，並取其交集

也就是取任意二主題層面的交集而後聯集之

適用情形：

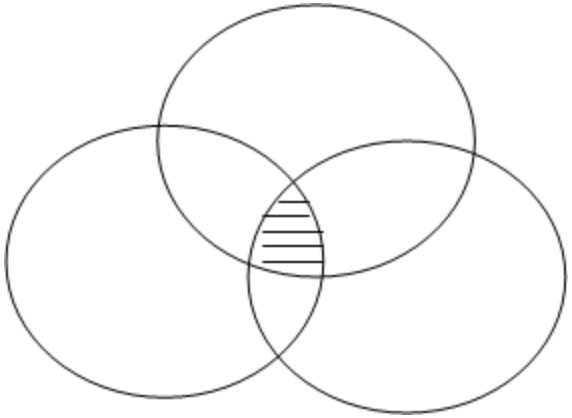
- 當所有主題層面都同樣重要時
- 當主題層面之專指性或模糊性相差不大時
- 當將所有主題層面結合可能導致零筆資料時

主題層面配對檢索示意圖



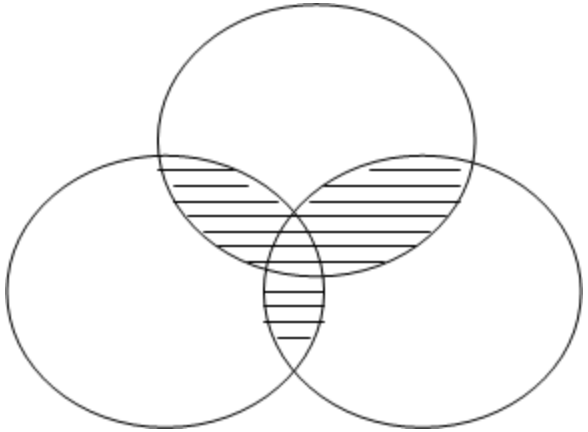
比較圖

分區組合檢索



A AND B AND C

主題層面配對檢索



(A AND B) OR (B AND C) OR (A AND C)

檢索技巧

為完成特性目的所採取的行動

- 當檢索所得資料筆數過多時（通常指誤引太多）
- 當檢索所得資料筆數過少時（包括零筆資料）
- 當檢索者想提高 recall 時
- 當檢索者想提高 precision 時

當檢索所得資料筆數過多時

- 是否過份簡化問題？
- 是否需要重新釐清檢索概念？
- 是否使用了正確的布林邏輯運算元？
- 是否使用過份含混或一般性之名詞？
- 是否應考慮使用控制字彙？
- 是否相近運算元限制過鬆？
- 是否切截應用過鬆？

當檢索所得資料筆數過少時 1/2

- 是否將問題過份複雜化？
- 是否真有文獻探討該主題？
- 是否每個概念都使用足夠的檢索詞彙來表達？
- 是否相近運算元限制過緊？

當檢索所得資料筆數過少時 2/2

- 是否使用了正確的布林邏輯運算元？
- 是否有語法或拼字上的錯誤？
- 是否該改用自然語言進行檢索？
- 是否考慮使用切截？

當檢索者想提高 recall 時 1/2

- 增加同義詞和類同義詞的數目
- 使用較廣義的檢索詞彙
- 以自然語言檢索代替控制字彙檢索
- 檢索其他主題欄位
- 刪除布林邏輯運算元“AND”及“NOT”

當檢索者想提高 recall 時 2/2

- 增加切截的範圍
- 使用較鬆的相近運算元
- 刪除一些非主題之檢索限制（如：年代、資料類型）
- 刪除一主題層面

當檢索者想提高 precision 時

- 刪除部份類同義詞或是詞意含糊的檢索詞彙
- 使用專指性較高的詞彙進行檢索
- 當有適當的控制字彙工具時，盡量使用其來代替自然語言
- 增加一主題層面
- 使用“ NOT” 除去不相關文章
- 減弱切截的範圍
- 加上非主題之檢索限制（如年代、資料類型）