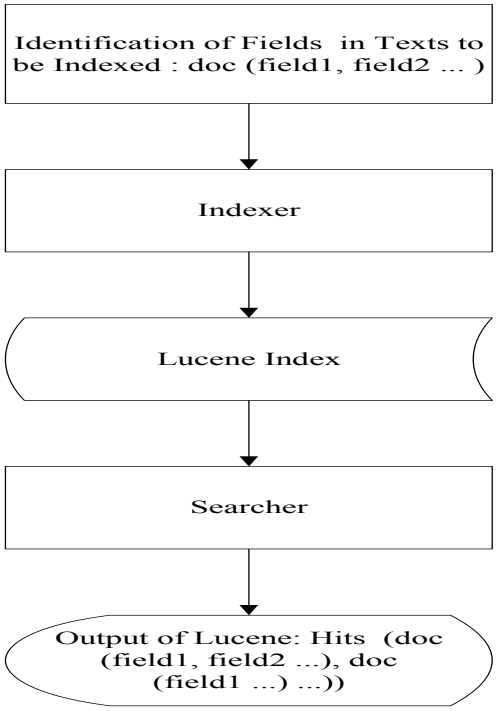
拾參、利用 Lucene 搜尋引擎檢索大量語料

Lucene 是一個以 Java 語言開發而成的全文搜尋引擎的套件，可以為各種檔製作倒置索引檔(Inverted File)，並透過簡單的 API 來檢索。Lucene 是免費開放程式碼(open source)。Lucene 可以將檔的每一個字建立索引，這樣讓搜尋就不需要逐字比對，檢索的效率可以大幅提高，Lucene 提供一組具有彈性且功能強大包括能夠解讀，過濾，分析檔，編排和使用索引的 API，讓使用者可以自訂功能。Lucene 雖然不是關連式資料庫，但可以透過類似關連式資料庫定義欄位元的方式達到關連式資料庫的功能。只要事先建立好索引檔，檢索的速度不會因為語料龐大的顯著降低檢索的速度，對於數十億詞的龐大語料庫而言，Lucene 搜尋引擎是一個不錯的選擇。下圖顯示 Lucene 欄位的建立與檢索的流程。



圖二十四 Lucene 欄位的建立與檢索的流程