中華大學資訊工程學系專題系統開發

成果報告

PDF 知識萃取系統

PDF Knowledge Extractor System

指導老師：曾秋蓉 教授

專題組員：謝承諭、龔聖華、張浩澤

專題編號：PRJ-2011-CSIE-10018

執行期間：100 年 03 月 至 101 年 08 月

目錄

壹、簡介

1.1 研究背景

生活在現代的資訊環境中，電腦及網際網路提供了人們快速獲得資訊的管道，而隨著資訊量的暴增，對資訊的管理也變得更加重要。由於近來知識管理活動盛行，各種企業在資料的處理歸納上，也不再局限於書面上的資料，其中，PDF 文件為最常用的儲存管理方式，所以在如何管理這些眾多的 PDF 文件並能做到有效的查詢，便是我們首要的課題。

1.2 研究動機

有鑑於現今知識管理活動的盛行，很多公司都積極推動知識管理活動，但傳統的知識管理大多需要花費大量時間在尋找資料，或等待有經驗的同仁協助。有鑑於目前各公司內都儲存許多過去所撰寫的知識文件，若能充份利用這些知識文件將有助於同仁快速解決問題。

PDF文件的特色在於檔案小、在網路上的通用性高，若是能將PDF文件切割分類成各種片段，並儲存在資料庫上進行有效的管理與查詢，便能促使知識管理活動成功達成目標。透過有效的知識管理使公司內同仁能夠在短時間內取得所想要的資料以解決多數遭遇問題。

1.3 研究目的

為了解決前述問題，本專題所要進行的是 PDF 文件知識萃取的程式研發，主要目的在萃取 PDF 文件中的文字段落、圖片、表格以及建立其資料的相互關連性。

本系統將萃取出的資料建立成知識素材，當使用者有需求時可透過檢索系統找到正確以及有用的素材解決他的需求，也使公司內儲存的大量資料得以有效的重複使用。這也是本專題研發 PDF 知識萃取系統的主要目的。

貳、相關知識

2.1 PDF 簡介

PDF (Portable Document Format 的簡稱，意思是「便攜式檔案格式」)由 Adobe Streams 在 1993 年用於檔案交換所發展出的檔案格式。

PDF 長期以來一直被用作交換和瀏覽商業檔案的格式。不過 Adobe 一直保留了該格式的版權，直到 2007 年 2 月因行業壓力被迫向 ISO 送出了標準化申請。

2009 年 9 月 1 日，作為電子文件長期保存格式的 PDF/Archive (PDF/A)經中國國家標準化管理委員會批准已成為正式的中國國家標準。

開放標準 — PDF 現已成為正式的開放標準 ISO 32000。ISO 32000 由國際標準化組織負責維持，今後也將以確保 PDF 的完整性與長久性為目標不斷進行研發，為現今超過十億個的 PDF 檔提供開放標準。

多平台 — 可在各種平台上檢視 PDF 檔案，包括 Windows®、Mac OS 和多種行動平台（例如 Android™）。

可擴充性 - 全球有超過 2,000 家廠商提供以 PDF 為主的解決方案，包括建立、增效模組、諮詢、訓練和支援工具。

受信任與可靠性 — 現今有超過 1 億 5 千萬份 PDF 文件在網路上公開流傳，而全球各地的政府機關與企業中所使用的 PDF 檔案也難以數計，證明有許多企業機關都仰賴 PDF 擷取資訊。

完整豐富的檔案內容 — PDF 檔案看起來和原始文件無異，並保留了原始檔案的資訊 — 文字、繪圖、多媒體、視訊、3D、地圖、彩色圖片、相片甚至商業邏輯 — 毋需考慮建立這些檔案使用的應用程式，甚至可從多種格式編譯為單一 PDF 文件夾。

提高安全性 — 您可在使用 Acrobat 或 Adobe LiveCycle® ES2 軟體建立的 PDF 文件上加上數位簽名或密碼保護。

可搜尋 — 文件和中繼資料的文字搜尋功能讓您可輕鬆在 PDF 文件中進行搜尋。

無障礙環境支援 — PDF 文件可與無障礙環境支援技術一起使用，讓殘障人士也能存取資訊。

2.2 iTextSharp 簡介

iTextSharp 是源自於著名開放原始碼網站 Sourceforge 的一個軟體，是一個用來處理 PDF 檔案的 JAVA 函式庫，透過 iTextSharp 不僅可以產生出 PDF 文件，也可以將 XML 或 HTML 直接轉換成 PDF 檔案。

iTextSharp 包含了三大功能，分別是產生 PDF 文件、閱讀 PDF 文件及修改 PDF 文件，主要在文件加密、資料萃取、標記及簽章的部份有很大的貢獻，如圖 2-1。

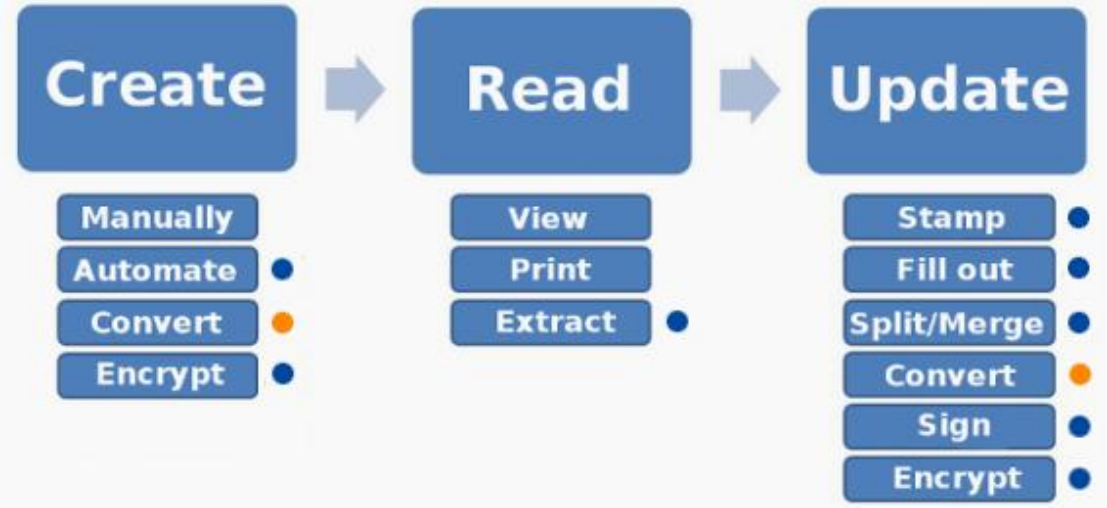


圖 2-1 iTextSharp 函式圖

參、專題進行方式

3.1 初步概念

為了讓使用者透過檢索系統找到正確以及有用的知識來解決他的問題，我們將開發一套 PDF 文件知識萃取系統，預期該系統可將 PDF 文件的文字段落、圖片及表格等知識素材萃取出來存放到資料庫，並藉由資料關聯性等方法，分類存取，達到更有效率的搜尋。

3.2 工作分配

組員	工作與職責
謝承諭	知識萃取模組開發 檢索系統開發
龔聖華	網頁設計 投影片製作
張浩澤	PDF 文件蒐集 海報製作

3.3 開發環境與平台

開發平台

Windows XP & Windows 7

開發工具

MicroSoft Visual Studio 2010
MicroSoft SQL Server 2008
Internet Information Services (IIS)
Adobe Acrobat 9.3

程式語言

Visual Basic
ASP.net 4.0

參考函式

iTextSharp.dll
Acrobat.dll

3.4 研究方法

PDF 知識萃取系統由四個模組所組成的，分別是圖片萃取模、表格萃取模、文字段落擷取模組及圖片表格與文字段落關聯模組。由管理者來上傳 PDF 文件經由 PDF 知識萃取系統萃取出知識素材後，透過知識素材特徵化模組將分解出的素材建成有用的知識庫，可以讓使用者透過使用者平台與主動式知識檢索模組到知識庫找尋他所需要的知識素材，如圖 3-1 系統架構圖。

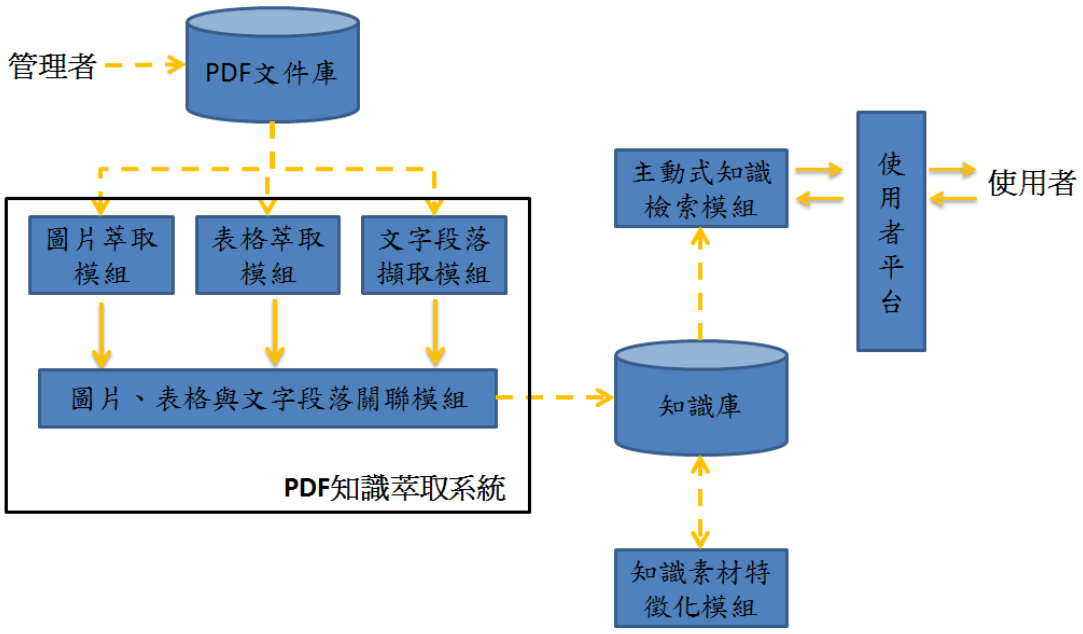
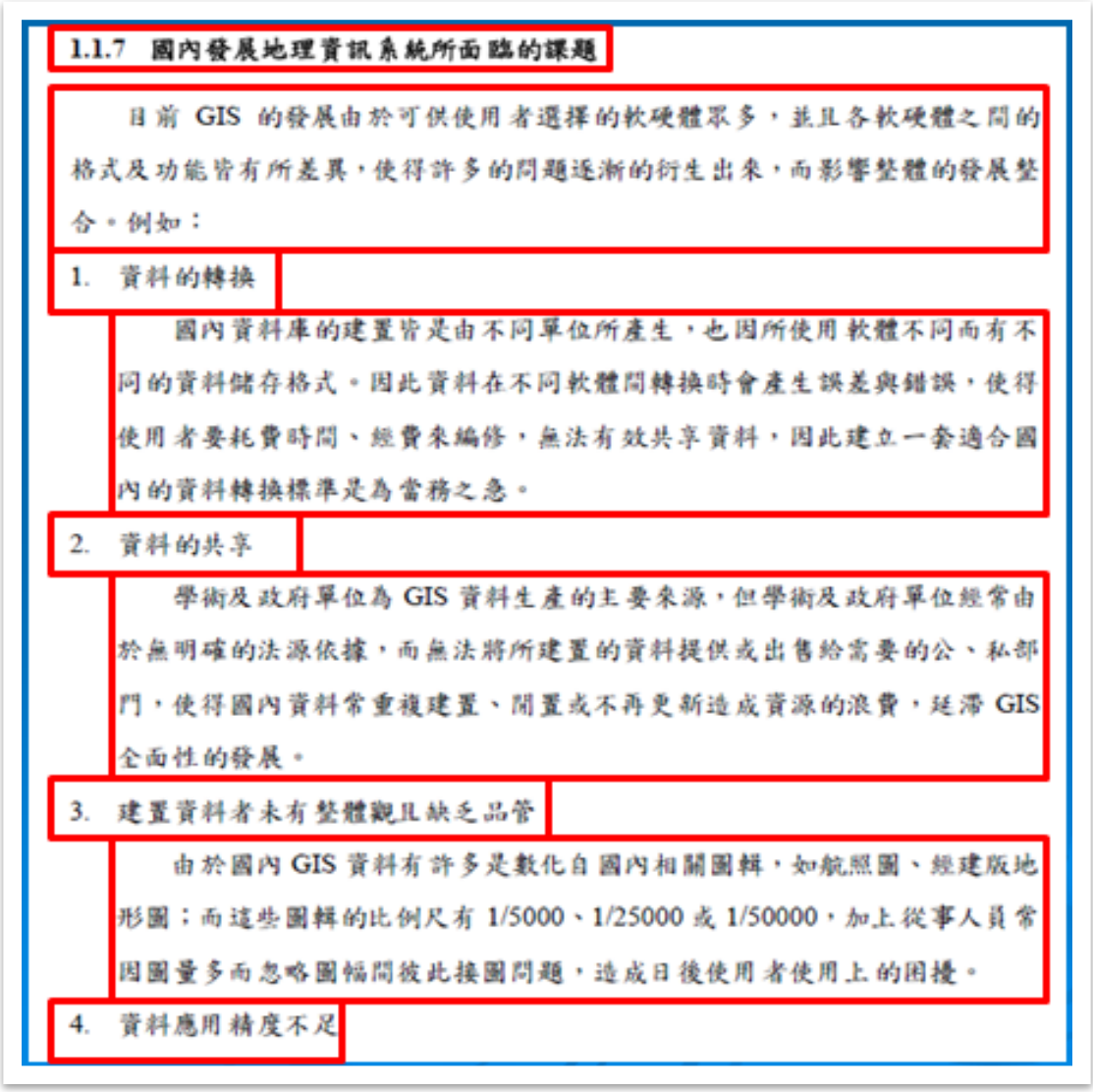


圖 3-1 系統架構圖

3.4.1 文字段落擷取模組

利用 iTextSharp 提供之功能函數裡的 PdfTextExtractor.getTextFromPage 逐頁掃描出 PDF 檔內所有的文字存入 StrText 中，並以 Split 將所有的文字以空白加換行做段落的分割後存入 StrTextSplit()中，最後以先後順序來記錄文字段落及其所在的位置，如圖 3-2。


1.1.7 國內發展地理資訊系統所面臨的課題

目前 GIS 的發展由於可供使用者選擇的軟硬體眾多，並且各軟硬體之間的格式及功能皆有所差異，使得許多的問題逐漸的衍生出來，而影響整體的發展整合。例如：

1. 資料的轉換

國內資料庫的建置皆是由不同單位所產生，也因所使用軟體不同而有不同的資料儲存格式。因此資料在不同軟體間轉換時會產生誤差與錯誤，使得使用者要耗費時間、經費來編修，無法有效共享資料，因此建立一套適合國內的資料轉換標準是為當務之急。

2. 資料的共享

學術及政府單位為 GIS 資料生產的主要來源，但學術及政府單位經常由於無明確的法源依據，而無法將所建置的資料提供或出售給需要的公、私部門，使得國內資料常重複建置、閒置或不再更新造成資源的浪費，延滯 GIS 全面性的發展。

3. 建置資料者未有整體觀且缺乏品管

由於國內 GIS 資料有許多是數化自國內相關圖輯，如航照圖、經建版地形圖；而這些圖輯的比例尺有 1/5000、1/25000 或 1/50000，加上從事人員常因圖量多而忽略圖幅間彼此接圖問題，造成日後使用者使用上的困擾。

4. 資料應用精度不足

圖 3-2 PDF 段落分割

3.4.2 圖片萃取模組

利用 iTextSharp 提供之功能函數裡的 PdfTextExtractor.getTextFromPage 逐頁掃描出 PDF 檔內所有的文字存入 StrText 中，並以 split 將所有的文字以空白加換行做段落的分割後存入 StrTextSplit()中，再透過正規表示式來篩選看是否為圖片說明，如圖 3-3。

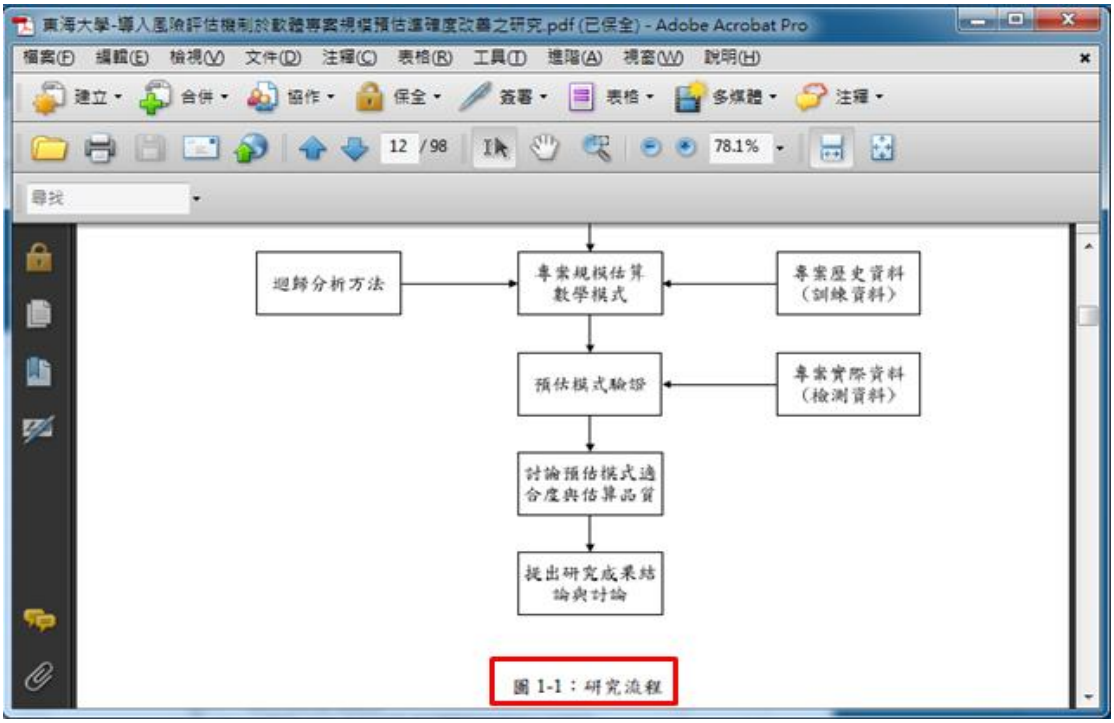


圖 3-3 圖片說明範例

3.4.3 表格萃取模組

利用 iTextSharp 提供之功能函數裡的 PdfTextExtractor.getTextFromPage 頁掃描出 PDF 檔內所有的文字存入 StrText 中，並以 split 將所有的文字以空白加換行做段落的分割後存入 StrTextSplit()中，再透過正規表示式來篩選看是否為表格說明，如圖 3-4。

圖 3-4 表格說明範例

3.4.4 圖片、表格與文字段落關聯模組

用圖片編號(例如圖 1.1)及表格編號(例如表 2-1)與文字段落做比對，將有包含圖片編號或表格編號的文字段落與該圖片/表格建立關聯，如圖 3-5、圖 3-6。

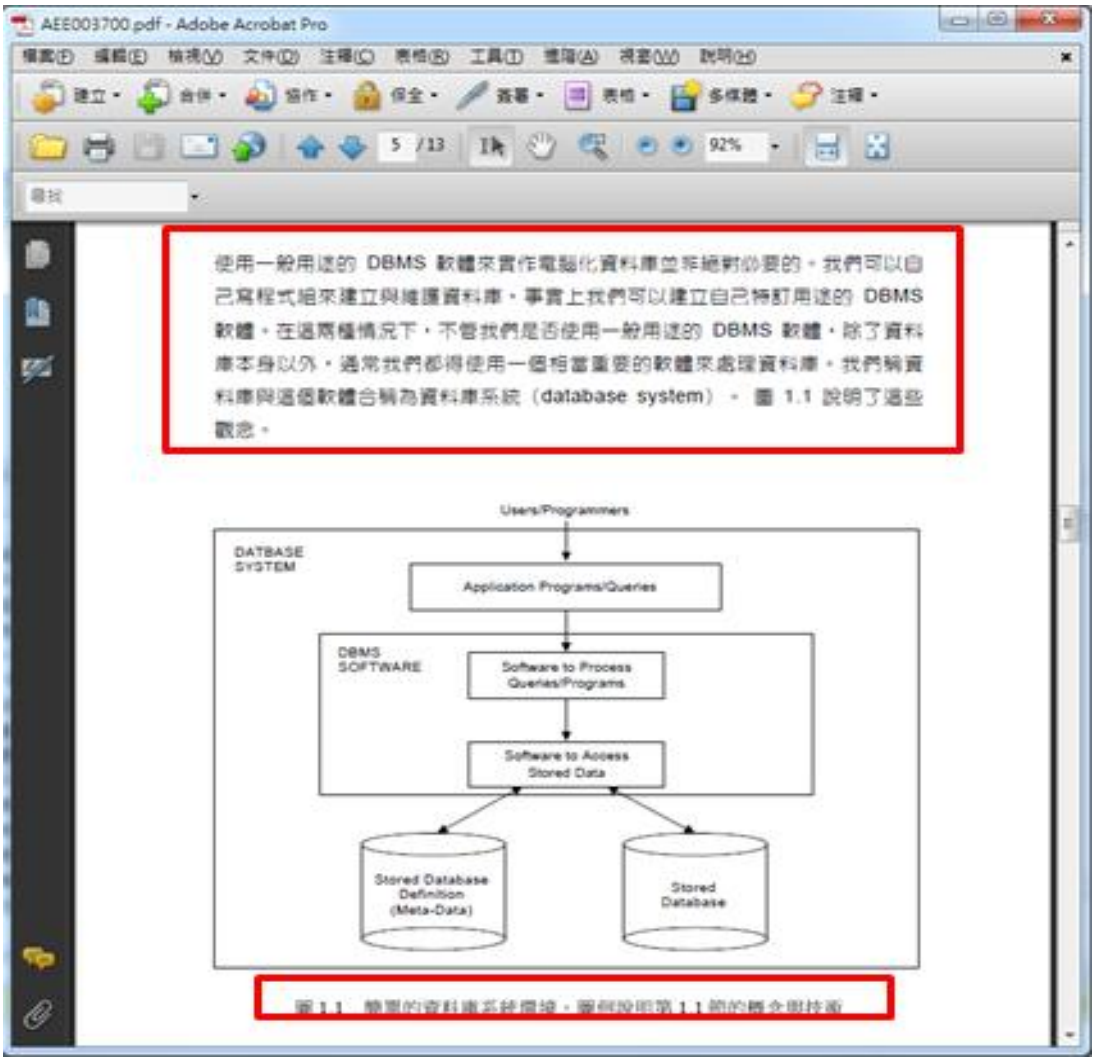


圖 3-5 圖片關聯範例

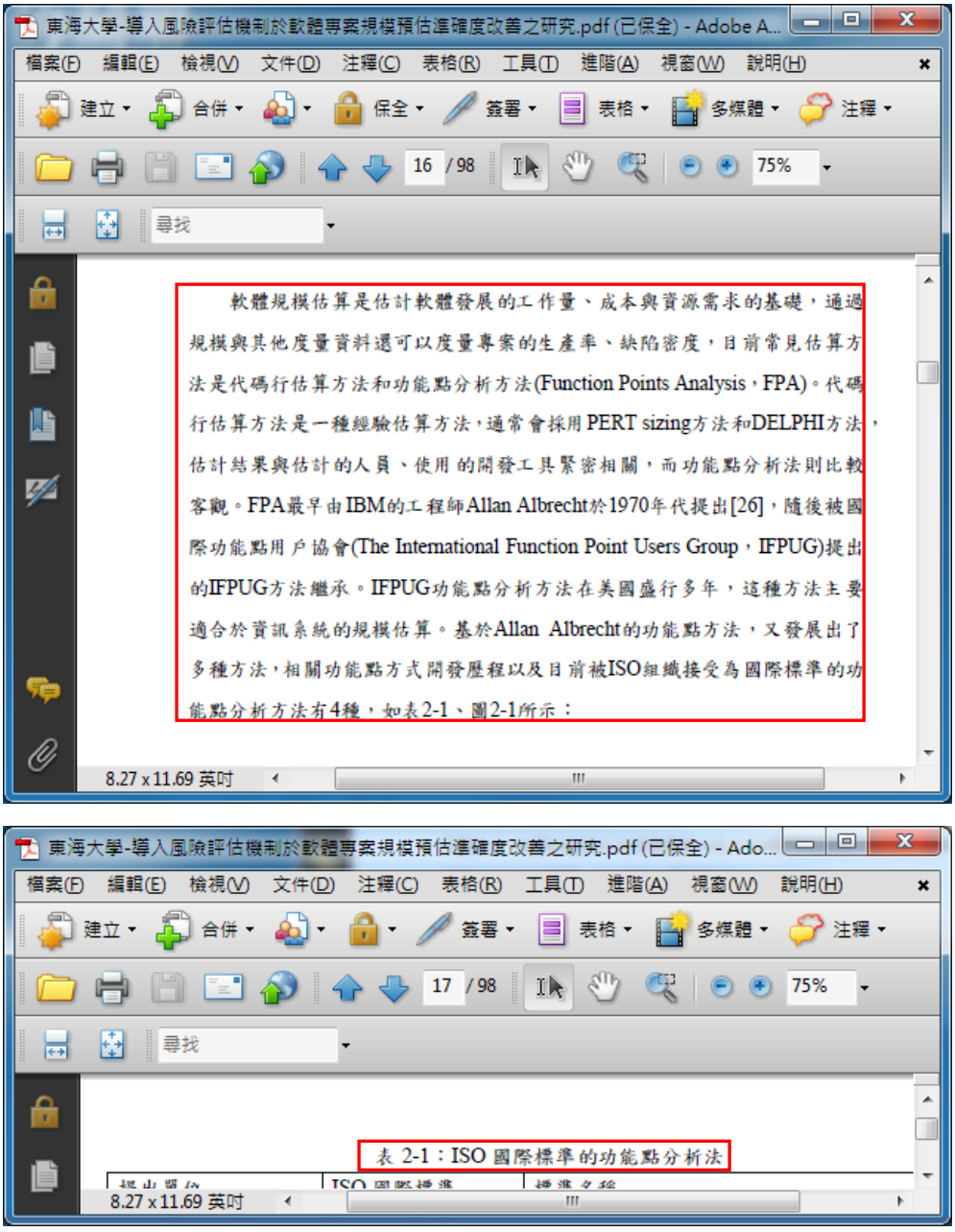

東海大學-導入風險評估機制於軟體專案規模預估準確度改善之研究.pdf (已保全) - Adobe A...
檔案(F) 編輯(E) 檢視(V) 文件(D) 注釋(C) 表格(R) 工具(T) 進階(A) 視窗(W) 說明(H)
建立 合併 保全 簽署 表格 多媒體 注釋
16 / 98
75%
尋找
軟體規模估算是估計軟體發展的工作量、成本與資源需求的基礎，通過規模與其他度量資料還可以度量專案的生產率、缺陷密度，目前常見估算方法是代碼行估算方法和功能點分析方法(Function Points Analysis，FPA)。代碼行估算方法是一種經驗估算方法，通常會採用PERT sizing方法和DELPHI方法，估計結果與估計的人員、使用的開發工具緊密相關，而功能點分析法則比較客觀。FPA最早由IBM的工程師Allan Albrecht於1970年代提出[26]，隨後被國際功能點用戶協會(The International Function Point Users Group，IFPUG)提出的IFPUG方法繼承。IFPUG功能點分析方法在美國盛行多年，這種方法主要適合於資訊系統的規模估算。基於Allan Albrecht的功能點方法，又發展出了多種方法，相關功能點方式開發歷程以及目前被ISO組織接受為國際標準的功能點分析方法有4種，如表2-1、圖2-1所示：
8.27 x 11.69 英吋
東海大學-導入風險評估機制於軟體專案規模預估準確度改善之研究.pdf (已保全) - Ado...
檔案(F) 編輯(E) 檢視(V) 文件(D) 注釋(C) 表格(R) 工具(T) 進階(A) 視窗(W) 說明(H)
建立 合併 簽署 表格 多媒體 注釋
17 / 98
75%
尋找
表 2-1：ISO 國際標準的功能點分析法
8.27 x 11.69 英吋

圖 3-6 表格關聯範例

3.4.5 主動式知識檢索模組

使用者的問題首先會被斷詞，以向量表示，向量中的每一個元素代表每一個詞的重要程度。接著使用者問題的向量會與知識素材的向量進行內積的計算並做相似度比對。將知識素材由高到低排序後取前面幾名相似度較高的知識素材提供給使用者參考，如圖 3-7。

圖 3-7 使用者搜尋結果

當使用者選擇瀏覽某一個相關的知識素材時，系統會將與知識素材有關的參數記錄在網址上，並傳遞至呈現 PDF 的頁面，透過分析網址來得知使用者所挑選的知識素材，並呈現相對應的 PDF 文件及該知識素材所在的頁面給使用者，如圖 3-8、圖 3-9。

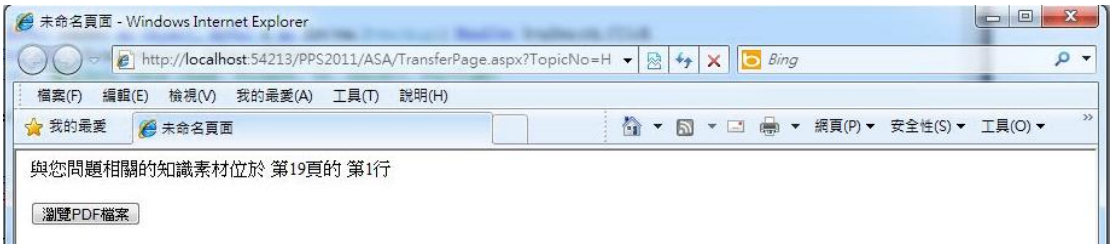


圖 3-8 搜尋結果的相關資訊

圖 3-9 知識素材所在頁面

3.5 遭遇問題與解決方案

在本專題執行過程中，碰到了兩個比較困難的問題，分述如下：

問題一：無法直接從 PDF 內部萃取出內容

在 PDF 格式中內容串流將會被隱藏在某個 obj 裡面如下

16 0 obj
<</Contents 18 0 R/CropBox[0.0 0.0 595.32 841.92]/MediaBox[0.0 0.0 595.32 841.92]/Parent 12 0 R/Resources<</ColorSpace<</CS0 23 0 R>>/Font<</TT0 25 0 R>>>>/Rotate 0/StructParents 0/Type/Page>>
endobj

首先由 16 0 obj 內標記部份得知 PDF 的內容被存放在 18 0 obj 裡。

18 0 obj
<</Filter/FlateDecode/Length 157>>stream

H?　D ⑳??爾 M????-j"選　n2ofv+!{DQ?骾 0 妦????惋?a8??^?[F 憊弗 y　i???

---

6 欹?犴??嵽?/r g?　栿蹕 C 燕　x^?M 蕨 X??

}

)? endstream endobj

接著可從 18 0 obj 裡面的 stream 部份取出 PDF 的部分內容。但由於編碼問題過於複雜加上有版本上的問題，目前我們還沒辦法解讀這串亂碼，故沒辦法透過此方法把內容還原。

解決方案：利用 iTextSharp 提供之功能函數來處理

利用 iTextSharp 提供 PdfTextExtractor.GetTextFromPage(ReaderFile, iReadPage)的方法將指定頁數中的文字逐一掃描，並以空格加換行來進行段落的切割，藉以取出 PDF 文件中的內容。

問題二：無法完整萃取圖片與表格

原先專題的計畫是將 PDF 轉成 XML 再從中進行圖片的萃取，但因由於 PDF 文件轉換成 XML 後圖片所在的位置與原始文件的內容不符合，故無法得知圖片正確的所在位置，且有些圖片也會在轉換後變成許多零碎的圖片，難以還原成原貌，故沒辦法將圖片完整萃取出來。

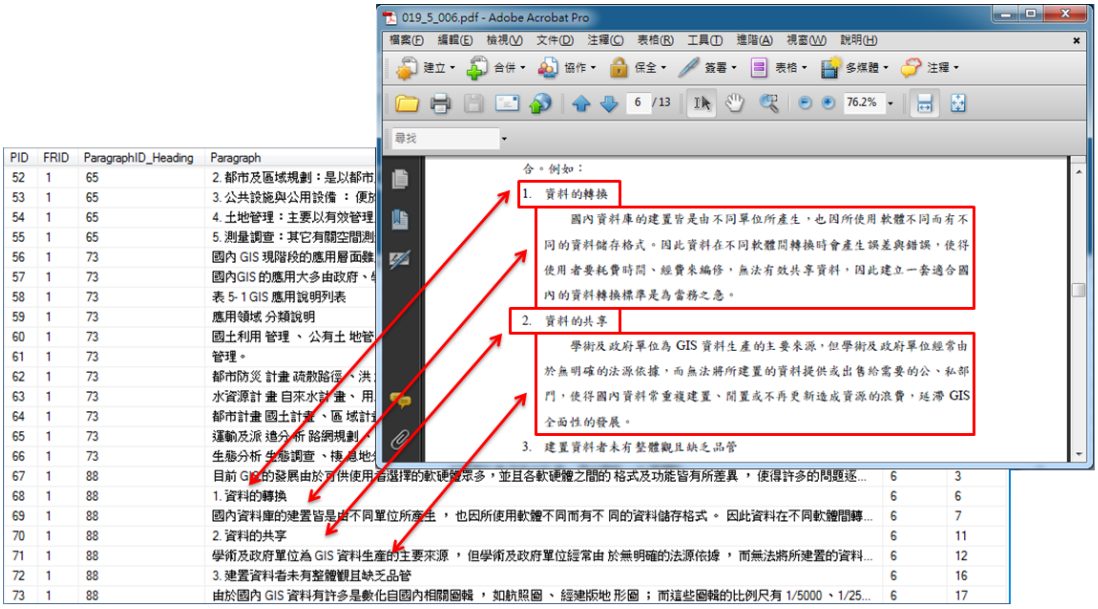
解決方案：取出圖表說明並開啟原始的 PDF 檔供使用者瀏覽

利用資料庫語法來可得知圖表所在的位置，因此可使用 adobe acrobat 函式庫中的 Response.ContentType = "application/pdf"及 Response.WriteFile(getpath)搭配網址中的參數指定開啟 PDF 所在的頁數。這樣一來使用者將可以透過圖說明來找到他所需要的圖片。

肆、研究成果

4.1 文字段落萃取

將文字段落從 PDF 中萃取出來存入資料庫中，並記錄它的內容及其所在的頁數與行數，如圖 4-1。

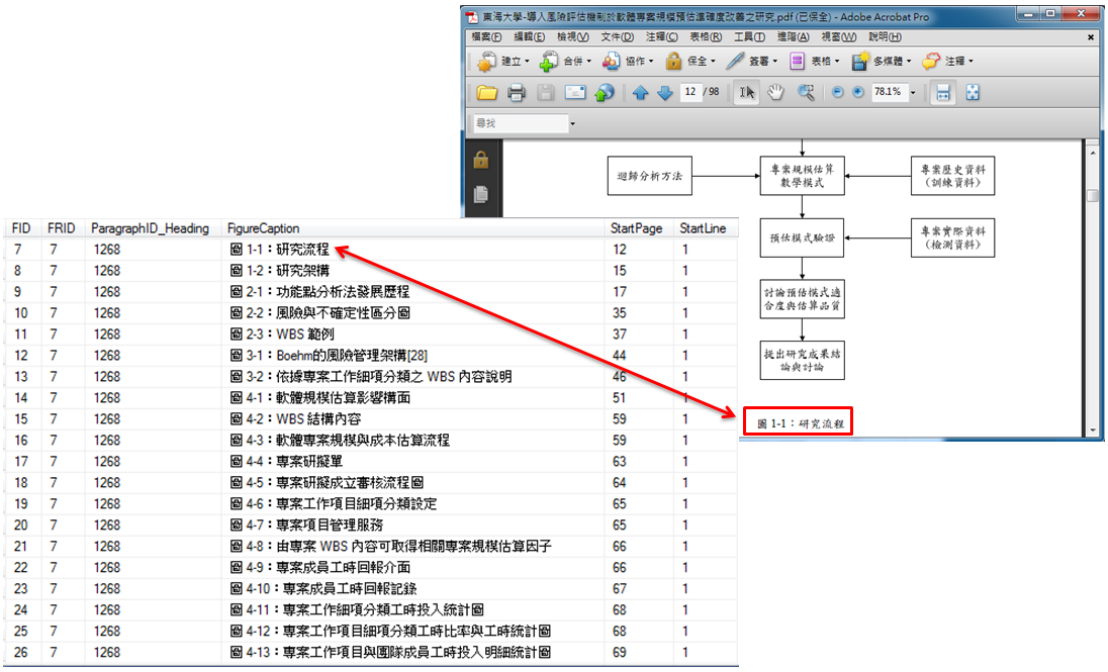


PID	FRID	ParagraphID_Heading	Paragraph		
52	1	65	2. 都市及區域規劃：是以都市		
53	1	65	3. 公共設施與公用設備 ：便		
54	1	65	4. 土地管理：主要以有效管理		
55	1	65	5. 測量調查：其它有關空間測		
56	1	73	國內 GIS 現階段的應用層面雖		
57	1	73	國內GIS 的應用大多由政府、		
58	1	73	表 5- 1 GIS 應用說明列表		
59	1	73	應用領域 分類說明		
60	1	73	國土利用 管理 、 公有土 地管		
61	1	73	管理。		
62	1	73	都市防災 計畫 疏散路徑、洪		
63	1	73	水資源計 畫 自來水計畫、 用		
64	1	73	都市計畫 國土計畫、區 域計		
65	1	73	運輸及派 遣分析 路網規劃、		
66	1	73	生態分析 生態調查 、棲 息地		
67	1	88	目前 GIS 的發展由於可供使用者選擇的軟硬體眾多，並且各軟硬體之間的 格式及功能皆有所差異 ， 便得許多的問題逐...	6	3
68	1	88	1. 資料的轉換	6	6
69	1	88	國內資料庫的建置皆是由不同單位所產生 ，也因所使用軟體不同而有不 同的資料儲存格式 。 因此資料在不同軟體間轉...	6	7
70	1	88	2. 資料的共享	6	11
71	1	88	學術及政府單位為 GIS 資料生產的主要來源 ， 但學術及政府單位經常由 於無明確的法源依據 ， 而無法將所建置的資料...	6	12
72	1	88	3. 建置資料者未有整體觀且缺乏品管	6	16
73	1	88	由於國內 GIS 資料有許多是數化自國內相關圖幅 ， 如航照圖 、 經建版地 形圖 ； 而這些圖幅的比例尺有 1/5000 、1/25...	6	17

圖 4-1 文字段落萃取範例

4.2 圖片知識萃取

將圖片說明從 PDF 中萃取出來存入資料庫中，並記錄它的內容及其所在的頁數與行數，如圖 4-2。

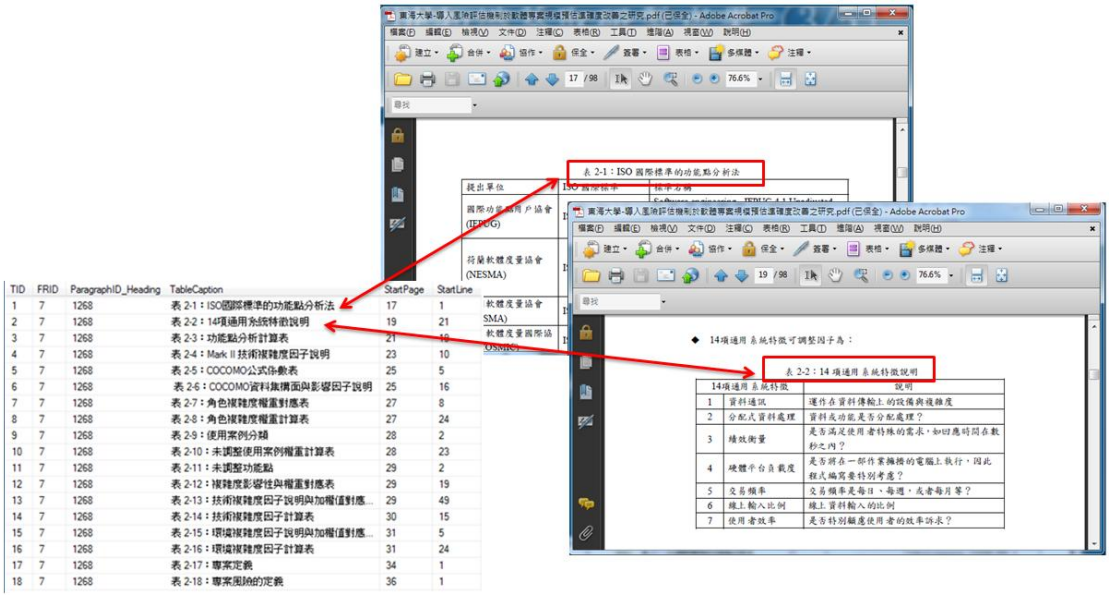


FID	FRID	ParagraphID_Heading	FigureCaption	StartPage	StartLine
7	7	1268	圖 1-1：研究流程	12	1
8	7	1268	圖 1-2：研究架構	15	1
9	7	1268	圖 2-1：功能點分析法發展歷程	17	1
10	7	1268	圖 2-2：風險與不確定性區分圖	35	1
11	7	1268	圖 2-3：WBS 範例	37	1
12	7	1268	圖 3-1：Boehm的風險管理架構[28]	44	1
13	7	1268	圖 3-2：依據專案工作細項分類之 WBS 內容說明	46	1
14	7	1268	圖 4-1：軟體規模估算影響構面	51	1
15	7	1268	圖 4-2：WBS 結構內容	59	1
16	7	1268	圖 4-3：軟體專案規模與成本估算流程	59	1
17	7	1268	圖 4-4：專案研擬單	63	1
18	7	1268	圖 4-5：專案研擬成立審核流程圖	64	1
19	7	1268	圖 4-6：專案工作項目細項分類設定	65	1
20	7	1268	圖 4-7：專案項目管理服務	65	1
21	7	1268	圖 4-8：由專案 WBS 內容可取得相關專案規模估算因子	66	1
22	7	1268	圖 4-9：專案成員工時回報介面	66	1
23	7	1268	圖 4-10：專案成員工時回報記錄	67	1
24	7	1268	圖 4-11：專案工作細項分類工時投入統計圖	68	1
25	7	1268	圖 4-12：專案工作項目細項分類工時比率與工時統計圖	68	1
26	7	1268	圖 4-13：專案工作項目與團隊成員工時投入明細統計圖	69	1

圖 4-2 圖片說明萃取範例

4.3 表格知識萃取

將表格說明從 PDF 中萃取出來存入資料庫中，並記錄它的內容及其所在的頁數與行數，如圖 4-3。



TID	FRID	ParagraphID_Heading	TableCaption	StartPage	StartLine
1	7	1268	表 2-1：ISO國際標準的功能點分析法	17	1
2	7	1268	表 2-2：14項通用系統特徵說明	19	21
3	7	1268	表 2-3：功能點分析計算表	21	19
4	7	1268	表 2-4：Mark II 技術複雜度因子說明	23	10
5	7	1268	表 2-5：COCOMO公式係數表	25	5
6	7	1268	表 2-6：COCOMO資料庫構面與影響因子說明	25	16
7	7	1268	表 2-7：角色複雜度權重對應表	27	8
8	7	1268	表 2-8：角色複雜度權重計算表	27	24
9	7	1268	表 2-9：使用案例分類	28	2
10	7	1268	表 2-10：未調整使用案例權重計算表	28	23
11	7	1268	表 2-11：未調整功能點	29	2
12	7	1268	表 2-12：複雜度影響性與權重對應表	29	19
13	7	1268	表 2-13：技術複雜度因子說明與加權值對應...	29	49
14	7	1268	表 2-14：技術複雜度因子計算表	30	15
15	7	1268	表 2-15：環境複雜度因子說明與加權值對應...	31	5
16	7	1268	表 2-16：環境複雜度因子計算表	31	24
17	7	1268	表 2-17：專案定義	34	1
18	7	1268	表 2-18：專案開始的定義	36	1

圖 4-3 表格說明萃取範例

4.4 圖片與文字段落關聯

將圖片說明與文字段落做比對，如果有相同片段即取出文字段落編號及圖片說明編號並將建立關聯，如圖 4-4。

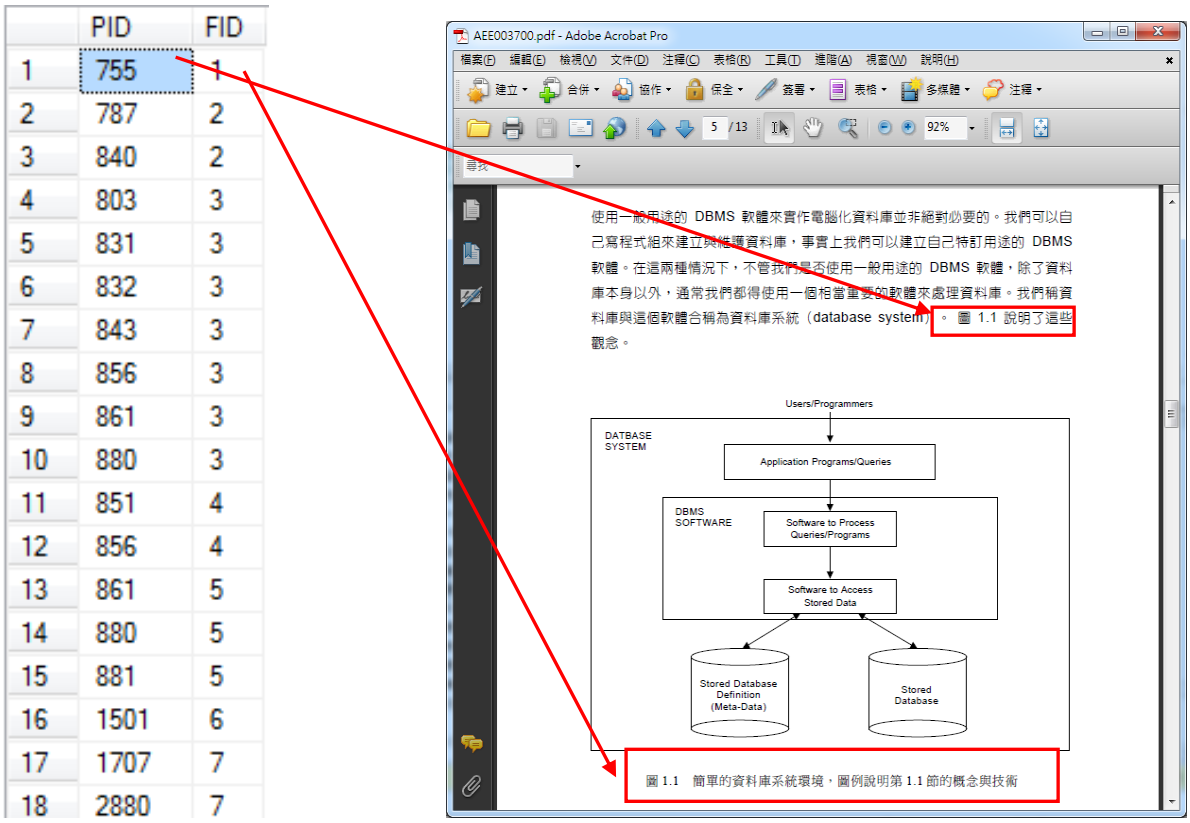


圖 4-4 圖片與文字關聯範例

4.5 表格與文字段落關聯

將表格說明與文字段落做比對，如果有相同片段即取出文字段落編號及表格說明編號並將建立關聯，如圖 4-5。

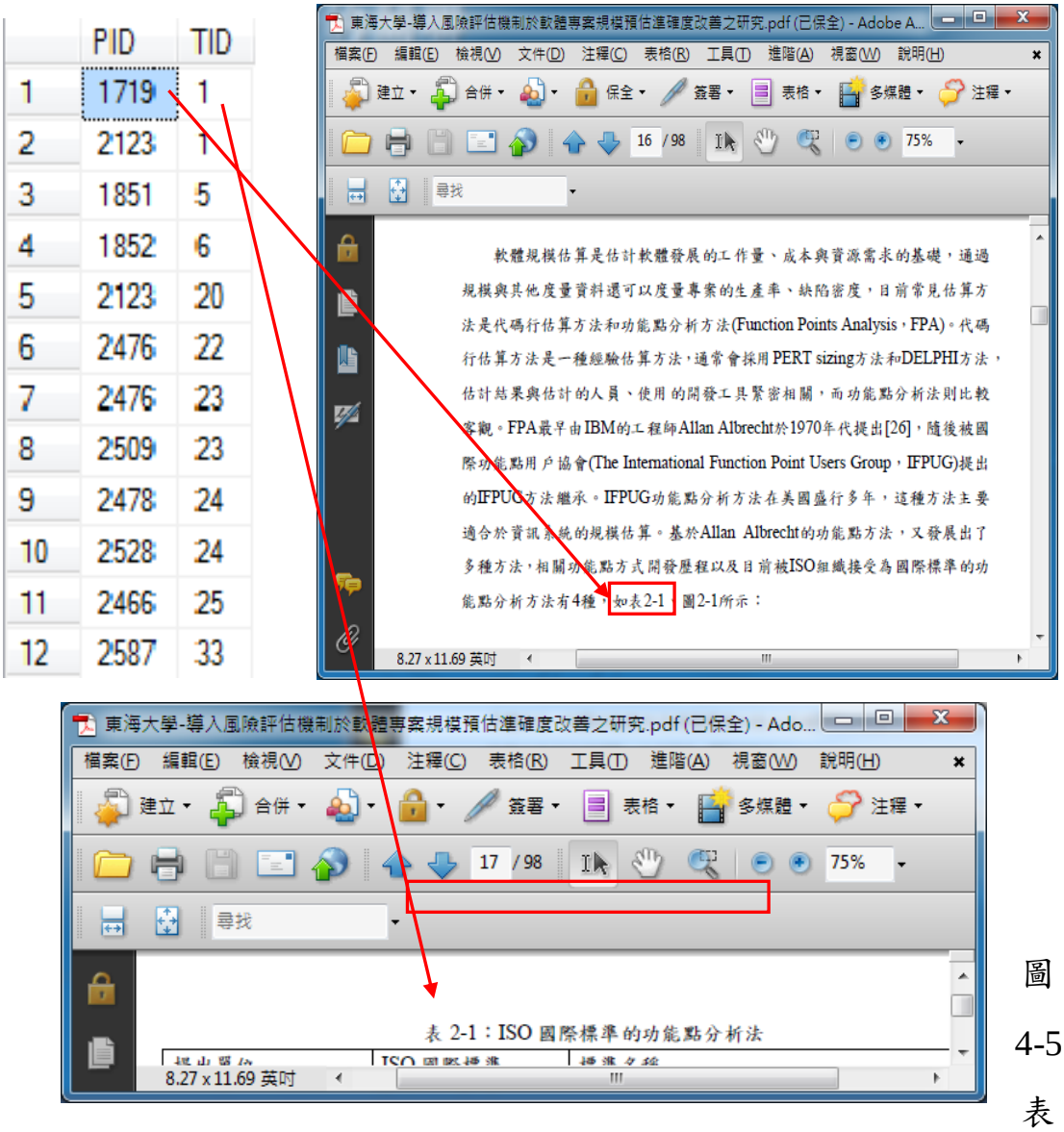


圖 4-5 表格與文字關聯範例

伍、評估與展望

經過這次的專題，我們學到了很多課程中學不到的技術，提升最多的當然是撰寫程式的能力，不過也獲得了一些解決問題的能力及面對事情應有的態度。實作方面花最多時間的多半是在解決一些未知的問題，不過在學長的帶領及循序漸進的找尋之下，最終問題都一一解決，最後系統順利的完成。系統目前還有一些可以進一步改進及優化之處，因此往後我們會朝優化系統的方向繼續努力。

未來可以加強改善以下功能：

(1) 將檢索系統搜尋的關鍵字以螢光筆作為標記，更容易發現資料所在位置。

(2) 讓使用者自行上傳文件並建立相關知識庫。

陸、銘謝

感謝曾秋蓉教授與學長的幫助和指導，在製作專題過程，我們到困難時，給予幫助及建議。

柒、參考資料

[1] 施威銘 “Visual Basic 2010 程式設計” 旗標

[2] 陳會安 “ASP.NET4.0” 旗標

[3] Adobe Acrobat “PDF Reference” Adobe Acrobat
[4] Adobe Acrobat http://www.adobe.com/
[5] iText http://itextpdf.com/