

目錄

中文摘要	I
英文摘要	II
謝詞	III
目錄	IV
圖目錄	VI
表目錄	VIII
第壹章 緒論	1
1.1 研究背景	1
1.2 研究動機	2
1.3 研究目的	3
1.4 研究流程與步驟	3
1.5 研究範圍與研究限制	4
1.6 論文架構	5
第貳章 文獻探討	7
2.1 資料挖掘(Data Mining)	7
2.1.1 資料挖掘的分類	8
2.1.2 資料挖掘可找出規則種類	9
2.1.3 資料挖掘的演算法	11
2.2 網路上的規則挖掘(Web Mining)	14
2.2.1 網站使用狀況的挖掘(Web Usage Mining)	15
第參章 網頁-動作路徑的資料挖掘演算法	20
3.1 網頁-動作路徑	20
3.1.1 網頁-動作路徑的資料結構	20
3.1.2 網頁-動作路徑的 Path 與長度	23
3.2 網頁-動作路徑演算法(Page-Action Algorithm)	24
3.2.1 網頁-動作路徑的 Pattern(P-A Pattern)	25
3.2.2 網頁-動作路徑的 Pattern 掃描(P-A_ScanX)	26

3.2.3 網頁-動作路徑的門檻值(Large Threshold in P-A Algorithm) …	32
3.2.4 網頁-動作路徑的 Join(P-A_Join) ……………	37
3.3 網頁-動作路徑演算法的演算實作 ……………	38
3.3.1 不重覆計算的演算方式(PA_Scan1 Algorithm) ……………	38
3.3.2 重覆計算的演算方式(PAScan_N Algorithm) ……………	43
3.3.3 找出關聯規則(Find The Association Rules) ……………	49
3.4 一個可能真實例子的操作 ……………	50
3.5 本演算法的效率與效益 ……………	57
第肆章 雛形系統實作與數據分析 ……………	60
4.1 雛形系統的功能介紹 ……………	60
4.1.1 記錄處理子系統(P-A Log System)…………	60
4.1.2 網頁-動作資料挖掘子系統(P-A Mining System)…………	65
4.2 本雛形系統的執行效能 ……………	70
4.2.1 資料量不同對於網頁-動作路徑演算法的影響…………	73
4.2.2 Support 數不同對於網頁-動作路徑演算法的影響…………	74
4.2.3 資料型態不同對於網頁-動作路徑演算法的影響…………	76
第伍章 結論與未來研究方向 ……………	81
5.1 研究貢獻 ……………	81
5.1.1 學術上的貢獻 ……………	81
5.1.2 實務上的貢獻 ……………	81
5.2 後續研究建議 ……………	82
參考文獻 ……………	84

圖目錄

圖 1-1 本研究流程的示意圖	4
圖 2-1 Apriori 演算法的示意圖	12
圖 2-2 一個概念樹的例子	13
圖 2-3 Next Pass Large Threshold 與 Next Pass Large Sequence 演算法的概念	17
圖 4-1 雛型系統架構圖	61
圖 4-2 手動記錄器的執行頁面	62
圖 4-3 自動 Log 產生器的設定畫面	63
圖 4-4 序列轉換功能的執行畫面	64
圖 4-5 處理重覆計算的 support 總數功能的執行畫面	64
圖 4-6 PAScan1 的設定畫面	66
圖 4-7 PAScan1 的初始 Scan CS0 (手動模式)	66
圖 4-8 PAScan1 的 Scan CS1 (手動模式)	67
圖 4-9 PAScanN 的設定畫面(一)	67
圖 4-10 PAScanN 的設定畫面(二)	68
圖 4-11 找尋規則功能的設定畫面	69
圖 4-13 規則功能的呈現畫面	69
圖 4-14 PAScan1 在不同資料量所需的執行時間	73
圖 4-15 PAScanN 二種 Next Pass 參數在不同資料量所需的執行時間	74
圖 4-16 PAScan1 對同一資料庫在不同 MinS 所需的執行時間	75
圖 4-17 PAScanN 對同一資料庫在不同 MinS 所需的執行時間	75
圖 4-18 PAScan1 對不同平均網頁路徑長度所呈現的效能	77
圖 4-19 PAScanN 對不同平均網頁路徑長度所呈現的效能	77
圖 4-20 PAScan1 對不同平均動作路徑長度所呈現的效能	78

圖 4-21 PAScanN 對不同平均動作路徑長度所呈現的效能	78
圖 4-22 PAScan1 對不同網頁路徑的離散程度的效能呈現	79
圖 4-23 PAScanN 對不同網頁路徑的離散程度的效能呈現	79
圖 4-24 PAScan1 對不同動作路徑的離散程度的效能呈現	80
圖 4-25 PAScanN 對不同動作路徑的離散程度的效能呈現	80

表目錄

表 1-1 使用者在全球資訊網中可能的網頁與動作路徑.....	2
表 3-1 使用者瀏覽網頁路徑的 Log.....	20
表 3-2 使用者瀏覽動作的 Log.....	21
表 3-3 瀏覽頁面及動作路徑序列的資料庫.....	21
表 3-4 四種可能的掃描方式.....	27
表 3-5 部份 MainSequence 資料庫的資料.....	29
表 3-6 MainSequence 資料庫的資料.....	38
表 3-7 Candidate Sequence 0 及其 support 值：CS0.....	39
表 3-8 Large Sequence 0 及其 support 值：LS0.....	39
表 3-9 Candidate Sequence 1 及其 support 值：CS1.....	40
表 3-10 Large Sequence 1 及其 support 值：LS1.....	41
表 3-11 Candidate Sequence 2 及其 support 值：CS2.....	41
表 3-12 Large Sequence 2 及其 support 值：LS2.....	42
表 3-13 其它所有的 CS 與 LS 表.....	42
表 3-14 序列庫的 TPS 表.....	44
表 3-15 可重覆的 Candidate Sequence 0 及其 support 值：CS0.....	45
表 3-16 可重覆的 Candidate Sequence 1 及其 support 值：CS1.....	45
表 3-17 可重覆演算法其它的 CS·LS support 值.....	47
表 3-18 LS1 與 LS2 可產生出的規則.....	50
表 3-19 一個可能真實的 PageCode 與其代表	51
表 3-20 一個可能真實的 ActionCode 與其代表	51
表 3-21 一段可能真實的 Mainsequence 資料	52
表 3-22 本範例的預先掃描 CS ₀	53
表 3-23 本範例的 Large Sequence 1 及其 support 值：LS ₁	53

表 3-24 本範例的 Large Sequence 2 及其 support 值:LS ₂	54
表 3-25 本範例的 Large Sequence 1 及其 support 值:LS ₁	55
表 3-26 此例發生機率超過門檻值(80%)的規則.....	56
表 3-26 此例發生機率超過門檻值(80%)規則的解釋.....	56
表 4-1 測試資料有關參數.....	70
表 4-2 測試資料列表.....	72