# 應用資料採礦及演化式計算概念於數位學習路徑分析

何俊輝* 蔡崇煒** 楊竹星* 莊嘉育**
國立中山大學資訊工程學系
886-7-5254333
* {jhho, csyang}@cse.nsysu.edu.tw
** {d9134804, m9134657}@student.nsysu.edu.tw

## 摘要
隨著資訊科技的迅速進展，人類的學習方式也隨之而改變。學習方式不再侷限於課堂上的學習，進而增加了網路數位學習模式。在許多的資料採礦 (Data-Mining) 或知識管理(Knowledge Management)的研究中，皆在致力於發展出更好的學習方式、更佳的知識累積及傳承。在這些研究中，找出較佳學習的路徑是一項重要的研究議題。這類相關的問題在資料採礦中，深受研究學者喜愛。最具代表性的是Agrawal 所提的方法，之後的學者在這類研究上所做的研究大多與之相關。由於問題定義的限制，大部分求解的方法皆需要花費許多的計算時間或記憶體空間。本研究中，我們將重新定義問題使之適合於網路學習平台上，並提出一快速計算的方法-EMS，應用在網路教學 (e-Learning)上發掘學生的學習路徑。這項方法除了非常具有效率，並且可以簡易地運算、分析不同時期長時間累積的資訊。

## 關鍵字
資料挖礦、知識管理、網路教學

## 1. 簡介
近年來，由於科技進步使得人類經常面臨大量資料，而從這之中擷取對於人類有用資訊，已發展許多熱門研究方法：分群(Clustering)[2]、分類(Classification)、關聯法則(Association rule)[12]及循序特徵(Sequence Pattern)[13]等等。其中，關聯法則及循序特徵這兩個研究議題非常近似，最大的差別就是循序特徵除了找出項目(item-set)的關連性，並需考慮項目的順序性。在這些研究中，發展一有效方法從大量的 log 記錄中找出項目的關連性及順序性，一直以來都是 Data Mining 的研究學者所致力研究的目標。

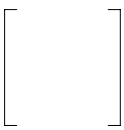
從 90 年代開始，許多學者逐漸發現這類研究議題的重要性。最初是由 Agrawal 及 Srikant[12-13] 兩位學者提出關於關聯法則及循序特徵問題，並且發展出如何求解的方法。而後來陸續有許多學者亦投入這個循序特徵研究領域進行研究，例如 Masseglia 的 PSP[5]演算法、Han 的 FreeSapn[6]及 PrefixSpan[8]演算法、Zaki[10]的 SPADE 演算法。除了上述提升效率的方法。更有許多學者亦利用這些資料採礦的方法應用 e-Learning 的學習路徑探勘[3]，或是將資料採礦的技術應用於其他 e-Learning 的相關研究[11,15]，其共同的目的皆是希望這些技術或方法可以促進建立一更有效的學習平台。

在本研究中我們除了重新定義這項問題以符合數位學習的分析，並提出一個快速方法-EMS。第 2 章將說明問題定義，第 3 章中介紹本研究所提出的方法，3.1 節將介紹系統的架構，3.2 節介紹本研究所提的演算法之基本概念，並詳細說明演算法。第 4 章為系統成果及效能分析。第 5 章將本研究之成果加以探討，並說明結論以及系統未來之研究發展。

## 2. 問題定義
在許多分析數位學習過程中關於學員學習行為模式的研究[1]，其中有些研究是將學員的學習過程記錄在系統的 Log 檔，經由萃取之

後，再利用 Agrawal 或其他學者在 Sequence Pattern 研究中所提的演算法進行分析。而這類的方法通常需要花費許多系統的計算時間或大量的記憶體空間。在本研究中，針對 e-Learning 系統實際應用所需顧及的層面重新考量問題之定義，其中，包含系統回應時間、建議的學習路徑、不同群體之間學習歷程的比較與分析等等。因此，本研究定義的問題和傳統求解 Sequence Pattern 的研究會有些許不同

在問題定義中，令 $I = \{i_1, i_2 ...., i_n\}$，其中 I 表所有的學習歷程紀錄，$i_1, i_2 ...., i_n$ 表示不同的學習行為集合，n 表示 I 所含的子集合數量。其中，$i_1 = \{i_{11}, i_{12}, .....i_{1m1}\}$，表示學習歷程的順序組合。m1 表示 i1 所含的子集合數量。而 $i_1 = \{i_{11}, i_{12}, ....., i_{1m1}\}$ 可以視為 log 檔中瀏覽 e-Learning 課程教材的紀錄，若其中 $i_{11}$ 為教材 A，$i_{12}$ 為教材 B，$i_{13}$ 為教材 D，$i_{14}$ 為教材 C，$i_{15}$ 為教材 E，則表示其瀏覽課程的行為模式為教材 A→教材 B→教材 D→教材 C→教材 E 所組成。從大量的記錄中找出較為重要的行為模式是本研究所的重點。而如何判斷教材間瀏覽的順序關係重要程度可以從教材 X→教材 Y 這項關係在所有紀錄中出現的頻率而訂定。把所有的歷程紀錄 I 轉換成二維矩陣可以表示教材間的關係。下圖一及圖二表示課程教材之間被瀏覽的順序以及次數，圖一的二維矩陣可對應至圖二的教材間之關係。從下圖中我們可以發現瀏覽教材 A 之後有可能瀏覽教材 B、教材 C、教材 D，瀏覽教材 B 之後只會瀏覽教材 C，瀏覽教材 E 之後只會瀏覽教材 B，而瀏覽其他的教材 C 及教材 D 並不會繼續瀏覽其他的教材。



**圖一、二維矩陣**

**圖二、教材間瀏覽順序及次數**

有了上述的矩陣，可以把問題轉成圖論中搜尋最短路徑的問題，並進行求解。所求得之解可以挖掘到較為重要的學習歷程組合。上例中 A→B→C 及 E→B→C 是較最重要的組合之一。因此，問題的定義如下：

Given: 給一個資料集合 I，且 $I = \{i_1, i_2 ...., i_n\}$，$i_1, i_2 ...., i_n$ 表示不同的學習行為集合。其中，$i_1 = \{i_{11}, i_{12}, ....., i_{1m1}\}$，表示學習歷程的順序組合。由這些資料集合可以產生出一個 $n \times n$ 的二維矩陣(如圖一)。

Objective:找出數組由許多節點所組成的集合 $R$，$R = \{r_1, r_2 ...., r_n\}$，$r_1, r_2 ...., r_n$ 表示各種的學習行為歷程集合，$r_1 = \{r_{11}, r_{12}, ....., r_{1m1}\}$，表示學習歷程的順序組合。所求得之解 $R$ 為資料集合 I，較為重要的學習歷程組合。

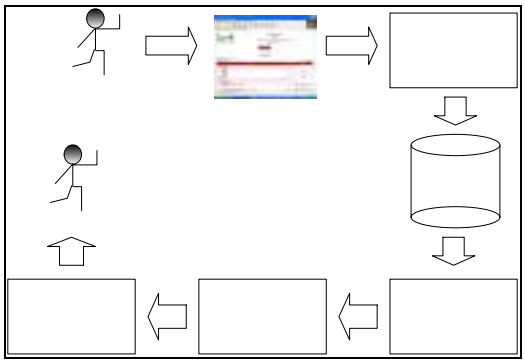
上述的問題定義與 Agrawal 所提的問題定義最大的差異是，不考慮列出最長且高於 support 值的序列集合，而是找出所有瀏覽紀錄中相對重要(較常發生)，但不見得是最長的序列組合。換言之，在 Agrawal 的研究中所產生的最終結果若為{A, B, C}，其 support 值=2，為{A, B}及{B, C}所組成，support 值為 3 及 4。在上述問題定義之下，最終所找出結果則為{B, C}。

## 3. 經驗矩陣系統(EMS)

本研究除了發展一個新的方法之外，並在 GLMS(called Global Learning Management System)平台上實作此項系統。

### 3.1 系統架構

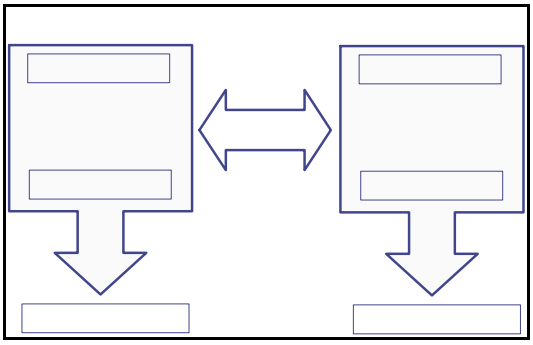
本研究所設計之 EMS 系統，其發展目的是希望可以將使用者使用教材的行為有系統的予以紀錄，藉由整合分析這些紀錄，達到不同學年度(團體)學習者的行為模式分析比較，使得學習知識得以被累積。圖三描述學習者與系統之間的關係，Generation n 指的是某一個班級。步驟 1.當這些學員在 GLMS 數位學習系統進行瀏覽教材時，其行為會被一 Daemon 行程監視追蹤，依據其使用者 ID 及 session ID 等參數紀錄至資料庫，不是藉由傳統的 system Log 方式紀錄。在步驟 2.中，我們從資料庫中取出這些紀錄，並且將其轉變成有用的資訊，以供步驟 3 中進行使用者行為分析。當步驟 3 分析完成之後，步驟 4 將這些結果以圖形及文字的方式呈現，使之後的學習者(generation n+1)以及授課教師瞭解先前使用者的學習行為。



**圖三、系統紀錄學員使用教材之過程以及建議學習路徑**

圖四中，我們可以看到藉由系統分析出的有用知識，可以再重複利用並且累積。舉例來說，1998 年，班級 A 的學習模式經由分析後成為知識，在 1999 年班級 A 的學員可以參照上一年度的學習途徑，而這些知識可以被累積，至 2004 年時的班級 A 已經可以吸取先前歷年來所有的知識，避免不需要的錯誤嘗試學習，直接以最好的途徑進行學習。除了以時間為導向的知識累積，系統並可以將同一年度不同班級的知識進行整合、分析及比較。教師可以藉由不同的班級所呈現出的學習行為找出較佳的教材編排方式，或較佳的學習路徑。最終，系統可以依圖四中知識整合方式，以較小空間存放大量的學習歷程樣本紀錄。因此在推薦學習路徑時，較具時效性。在所有的 Data Mining 過程中，其中有一項令人感興趣的是如何將這些大量紀錄保留，並進行分析。

一般的作法是將系統紀錄進行萃取，減少紀錄的欄位。在本系統中，系統除紀錄所有的使用教材紀錄，並且將其轉換成一 2 維的矩陣存放，這項 2 維的矩陣並不會隨者使用者登錄次數或使用者人數的增加變大，其維度是由教材的數量所構成的。舉例來說，若教材數量為 10，那矩陣就是 10×10，不會改變。只要分析過一次，其資訊就已成為知識。當第一次分析 10 人的行為模式後，其結果就存放於矩陣，若要再增加 10 人的行為模式，只需要將後 10 人的行為累加至矩陣之後，分析矩陣即可。不需要重複到資料庫去查詢先前 10 人的行為紀錄。



**圖四、知識在不同的時間及空間累積**

### 3.2 演算法

在此，將介紹上節中圖三內的步驟 3 如何進行行為模式分析的演算法。

Step1: Transform Log Data to Information

Step2: Create the Experience Matrix (EM)

Step3: Transform EM to EM'

Setp4: Compute every edge on EM' that Satisfied Condition W

Step5:Build the Set of Learning Path by ACO Algorithm (Evolutionary Algorithm)

在 Step 1 中，我們首先將資料庫所紀錄的 Log Data 轉換成資訊，以便於系統中在進行計算。例如，所有的這些學員所瀏覽的教材為 Course 13-1-1、Course 13-1-2、Course 13-1-3，則可以轉換成 1、2、3。若一筆學員的學習歷程為{ Course 13-1-1, Course 13-1-3, Course 13-1-2}，可轉換成{1, 3, 2}。若另一筆學員的學習歷程為{ Course 13-1-1, Course 13-1-2}，則可轉換成{1, 2}。接下來在 Step 2 中，建至一個 3X3 的二維矩陣，並將上述的已經經過轉換過的學員學習紀錄累加至這項二維矩陣中(圖五)。並將教材被瀏覽的紀錄進行累加(圖六)。

**圖五、Experience Matrix-1**

**圖六、教材瀏覽紀錄累加**

在 Step3 中，則是將 EM 矩陣進行些許的轉換，進而轉換成可以圖論中求解最短路徑的問題。首先，我們先把每一筆學習歷程的紀錄有條件的加入矩陣之中。舉例來說，若{1, 2, 3, 4}為一筆學習歷程，則除了 1→2，2→3，3→4 的路徑需要加一(已經在 Step 2 中計算)，並且需要考慮 1→3，1→4 及 2→4 這些可能情況，因為在 1 之後也有除了 2 之外尚有 3 及 4 可能發生。在圖五的例子中，1→2 即是在 Step3 所需增加累計的路徑。因此圖五的矩陣會變成圖七的矩陣。除了上述計算，並把兩點之間的值 reverse(由大變小，小變大)。舉例來說，圖七的矩陣元素有{0, 1, 2, 3}，maximum 的值為 3(M)，所有非 0 的元素皆套用 M - EMij +1 公式計算，2-2+1-1, 2-1+1=2。因此經由轉換後的矩陣會由圖七的矩陣成為圖八的矩陣。

**圖七、Experience Matrix-2**

**圖八、Experience Matrix-3**

在 Step 4 中，我們會根據使用者給定的路徑重要性權重 IM 來分析哪些路徑是需要被計算，而哪些路徑則不需要。序列 E={2, 2, 1}為圖八矩陣中所有非 0 值的集合，且經過排序(由大至小)。假設 IM 給定為 20%，則 W 為 1，IM 給定為 50%則 W 為 2。換言之，就是利用 IM 依照等級去索引序列 E={2, 2, 1}(%33.3, %66.6, %99.9)，透過這種方式，我們可以去分析不同重要程度的學習歷程。因此，若選擇的 IM 為 50%，W 為 2 索引到的值為 2，故其值大於 2 以上的 edge 則轉換成 0，如圖九。

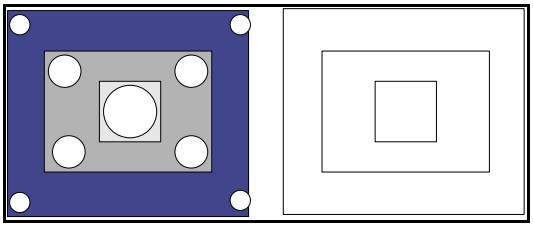
**圖九、Experience Matrix-4**

在經由 Step1 至 Step4 的矩陣計算之後，我們可以得到一個具有意義的矩陣 EM'。這個矩陣內的路徑關係也就是所有的使用者瀏覽教材的關係以及教材之間被瀏覽順序關係。考量系統的回應時間，因此利用啟發式演算法來求解

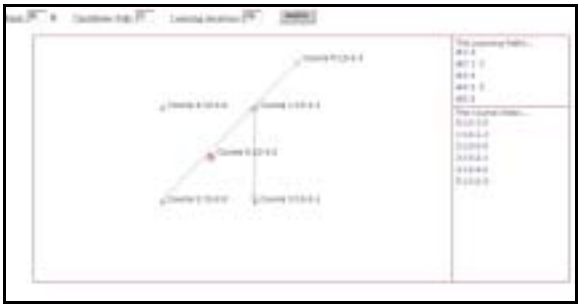
在本系統中是較為合理方法。當然啟發式演算法在近年應用於求解最短路徑的研究，已有非常多的成果。例如 Neural Network[14]、Genetic Algorithm[7]、Ant Colony System[9]、Tabu Search[4]等等。在此研究中，這些啟發式方法所擔任的角色是找出解的工具，故在 Step 5 中，可以應用上述的任何一種方法。而本研究目前是以改良過的 Ant Colony System 進行求解。我們修改 ACS 的更新函式及搜尋方式，更新費洛蒙的方式跟傳統 ACS 在解 TSP 問題類似，Ant 不需要繞行回出發點，因此省去這段費洛蒙的加總。此外，在 Global Update 時，最佳的路徑的值-Distance 的大小則是以 DP 及 SL 兩種權重計算，Distance = 0.5×DP + 0.5×SL。其中 DP 表示此路徑是否重要(越多人瀏覽過則越重要)，SL 表示此路徑的單元數(越多單元的路徑越好)。這兩項權重是可以改變，目前研究中，是將這兩項因素視為同等重要。在未來的研究中，將會比較不同的演算法在本系統中的效能。

## 4. 效能分析

在前文中我們說明系統如何設計以及其中的演算法如何運算，本章將介紹本研究的成果。圖十表示教材的顯示規則，越重要的教材放置位置越靠近中央，以 Level、Level2、Level3 的方式逐漸向外發散。圖十一中可以看到我們將課程之間的關係如何以圖形顯示，並且將學習路徑顯示在課程之間。圖十一上方的參數設定介面則是提供使用者在使用本系統時，可以調整、設定的參數值。其中 Rank 為前文中所提到的路徑重要性權重 IM，系統會根據使用者所設定的權重顯示不同的路徑畫面以及規劃 Learning Path，至於 Candidates Path 及 Learning iterations 是給使用者設定啟發式演算法的一些參數。本研究中，Candidates Path 代表使用多少的組解進行分析比較，而 Learning iterations 則是使用啟發式演算法所需執行的 iterations。圖十一右邊則是列出系統所建議的學習路徑及課程名稱與索引數字的對應。



**圖十、教材顯示規則**

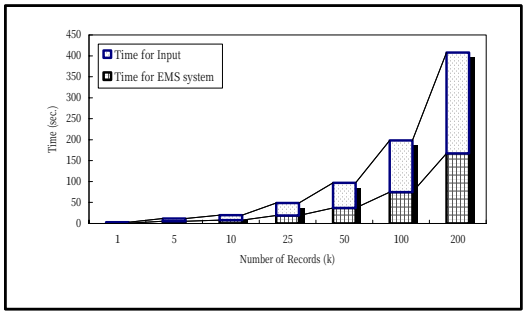


**圖十一、Experience Matrix-4**

在本研究中，我們模擬產生資料，測試系統在不同的資料數量下的計算時間。

**表一、每筆交易紀錄中平均瀏覽 5 筆教材(時間單位為秒)**

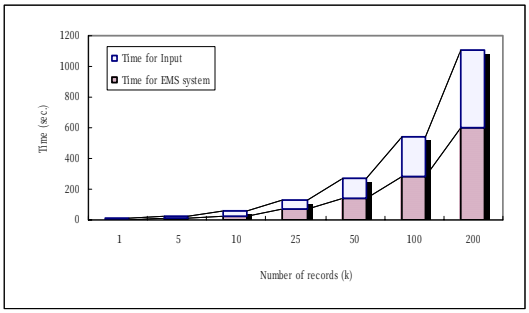
| # of Data (k) | 1 | 5 | 10 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|
| Execute Time 1 | 2 | 7 | 12 | 30 | 60 | 123 | 241 |
| Execute Time 2 | 1 | 5 | 8 | 19 | 37 | 75 | 167 |
| Execute Time 3 | 3 | 12 | 20 | 49 | 97 | 198 | 408 |




**圖十二、每筆交易紀錄中平均瀏覽 5 筆教材**

**表二、每筆交易紀錄中平均瀏覽 10 筆教材(時間單位為秒)**

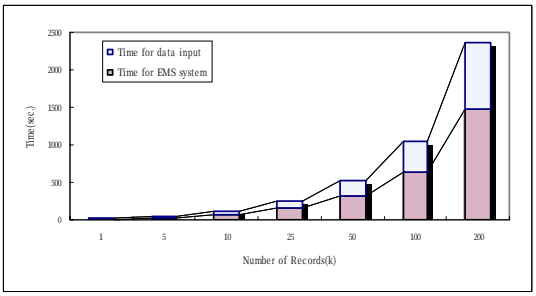
| # of Data (k) | 1 | 5 | 10 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|
| Execute Time 1 | 10 | 13 | 25 | 62 | 124 | 249 | 502 |
| Execute Time 2 | 4 | 15 | 29 | 70 | 141 | 287 | 599 |
| Execute Time 3 | 14 | 28 | 54 | 132 | 265 | 536 | 1101 |




**圖十三、每筆交易紀錄中平均瀏覽 10 筆教材**

**表三、每筆交易紀錄中平均瀏覽 15 筆教材(時間單位為秒)**

| # of Data (k) | 1 | 5 | 10 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|
| Execute Time 1 | 4 | 22 | 41 | 102 | 206 | 419 | 875 |
| Execute Time 2 | 9 | 33 | 69 | 158 | 313 | 628 | 1485 |
| Execute Time 3 | 13 | 55 | 106 | 260 | 628 | 1047 | 2360 |




**圖十四、每筆交易紀錄中平均瀏覽 15 筆教材**

**表四、總和比較**

| # of Data (k) | 1 | 5 | 10 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|
| Avg. 5 items | 3 | 12 | 20 | 49 | 97 | 198 | 408 |
| Avg. 10 items | 14 | 28 | 54 | 132 | 265 | 536 | 1101 |
| Avg. 15 items | 13 | 55 | 106 | 260 | 628 | 1047 | 2360 |

## 5. 結論

EMS 是我們重新定義 Data Mining 中 Sequence Pattern 這項問題後，所發展出的系統及演算法。本研究中，運用我們現有的 e-Learning 平台建置此一系統，並且實作出研究中所提出的演算法。這項方法的特色是在極為巨量的資料中，快速地求得合理的解，以建議授課教師調整教材的順序，以及提示修課的學員一些較佳的學習路徑。此外，對於知識的累積以及班級間學習方式的分析比較提供一項較佳的方法。

這項研究除了有上述優點，並與啟發式演算法進行合併(可使用不同的方法)，在目前我們是以 ACS 當作找尋最佳路徑的方法，未來的研究，我們將會引進不同的方法如 GA、Tabu Search、SA 等等著名的演算法，並比較其效能，以找出最佳的求解策略。

## 6. 參考文獻


[1] 李建億、吳孟淞、吳政道、李育強,"在全球資訊網學習環境中學習歷程樣式發掘法之研究",台灣區網際網路研討會(TANET'2000),臺灣, 2000 年

[2] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data Englewood Cliffs, N.J.:Prentice Hall, 1988.

[3] Catherine Stones and Stephen Sobol, "DMASC: A Tool For Visualizing User Paths Through A Web Site", Database and Expert Systems Applications, pp.389-393, 2002

[4] F.Glover, "Tabu search-part I", ORSA Journal of computing, vol. 1, no. 3, pp. 190-206, 1989

[5] F. Masseglia, F. Cathala and P. Poncelet, "The PSP Approach for Mining Sequential Patterns." Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, Nantes, France, Vol 1510, pp. 176-184, 1998

[6] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M-C. Hsu, "Freespan: Frequent pattern-projected sequential pattern mining," Proceedings of the International Conference of Knowledge Discovery and Data mining, pp. 355-359 ,2000

[7] J.H.Holland, Adaption in Natural and Artificial System, Boston,MA: MIT Press, 1992.

[8] J. Pei, J.Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U.Dayal and M-C. Hsu, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix Prejected Pattern Growth,” Proceeding of the International Conference of Data Engineering, H eidelberg, Germany, pp. 215-224, 2001

[9] Marco Dorigo and Luca Maria Gambardella, “Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem”, IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, April 1997

[10] M. J. Zaki, “SPADE: An Efficient Algorithm for Mining Frequent Sequences,” Proceeding of Machine Learning Journal, special issue on Unsupervised Learning, Vol. 42 Nos. 1/2, pp. 31-60 , 2001

[11] Osmar R. Za¨ıane, “Building a Recommender Agent for e-Learning Systems”, International Conference on Computers in Education, vol.1, pp.55-59, 2002

[12] R. Agrawal and R.Srikant, “Fast Algorithm for Mining Association Rules in Large Databases,” Proceedings of The 20th International Conference on Very Large DataBases, pp. 487-499 , 1994

[13] R. Agrawal and R.Srikant, “Mining sequential patterns,” Proceedings of The International Conference on Data Engineering, Taipei, Taiwan, pp. 3-14, 1995

[14] T.Kohonen,"Self-organized formation of topologically correct feature maps", Biol.Cybern., vol.43, pp.59-69,1982

[15] Weinan Wang and Osmar R. Za¨ıane, “Clustering Web Sessions by Sequence Alignment”, Database and Expert Systems Applications, pp.394-398, 2002