

行政院國家科學委員會專題研究計畫 成果報告

探勘使用者移動行為模式機制之研發與實作

計畫類別：個別型計畫

計畫編號：NSC92-2211-E-009-068-

執行期間：92年09月01日至93年07月31日

執行單位：國立交通大學資訊工程學系

計畫主持人：彭文志

報告類型：精簡報告

處理方式：本計畫涉及專利或其他智慧財產權，1年後可公開查詢

中 華 民 國 93 年 11 月 1 日

行政院國家科學委員會專題研究計畫成果報告

探勘使用者移動行為模式機制之研發與實作

Design and Implementation of Mining User Moving Patterns in a Mobile Computing Environment

計畫編號：NSC 92-2211-E009-068

執行期限：92年09月01日至93年07月31日

主持人：彭文志教授

執行機構：國立交通大學資訊工程系

中英文摘要

中文摘要：

關鍵詞：資料探勘、使用者移動行為模式、行動資料處理

本研究計畫的目的在於利用通聯紀錄(call detail records)，以期能夠在不增加行動通訊系統額外之儲存空間，探勘使用者的移動行為模式。在此計畫中，我們提出一系列之演算法，用以探勘使用者移動行為模式。首先，通聯記錄需轉換為移動序列。我們提出演算法 LS 用以精確地決定相類似的移動序列。利用空間-時間區域性(亦即在相近的時間內使用者的位置不會相距太遠)的特性，演算法 TC 將時間上相臨近的通聯記錄予以分群，使得同一群之通聯記錄具有空間-時間區域之特性。利用迴歸分析的概念，我們研發演算法 MF 用以推導出使用者移動行為模式函數。為了驗證所提出探勘機制正確性，我們建構行動計算系統模擬環境，對此一機制進行效能分析。模擬實驗結果顯示所提出的機制能有效的探勘使用者的移動行為模式且利用現有之通聯紀錄便可逼近使用者真實的移動行為模式。

英文摘要：

Keywords : Data mining, user moving patterns, mobile data management

In this project, by exploiting the log of call detail records, we present a solution procedure of mining user moving patterns in a mobile computing system. Specifically, we propose algorithm LS to accurately determine similar moving sequences from the log of call detail records so as to obtain moving behaviors of users. By exploring the feature of spatial-temporal locality, which refers to the feature that if the time interval among consecutive calls of a mobile user is small, the mobile user is likely to move nearby, we develop algorithm TC to cluster those call detail records whose time intervals are very close. In light of the concept of regression, we devise algorithm MF to derive moving functions of moving behaviors. Performance of the proposed solution procedure is analyzed and sensitivity analysis on several design parameters is conducted. It is shown by our simulation results that user moving patterns obtained by our solution procedure are of very high quality and in fact very close to real user moving behaviors.

報告內容

計畫緣由與目的

隨著第三代行動通訊系統的研發，未來的無線通訊系統將能夠提供更多元化的服

務，如使用者可透過無線通訊界面到提供應用服務的伺服器進行資料擷取，或是透過伺服器收播有興趣之影像資料如新聞、電影等資料流。可預期的是隨著行動計算環境的成熟，使用者將可在任何時間任何地點擷取所需要的資料，且資料更具多樣性。在此行動計算環境中，使用者可漫遊其中，而探勘使用者移動行為模式 (user moving patterns) 將極具研究價值，其重要性可由兩方面來觀之：其一、從使用者的觀點：使用者移動行為模式有助於提供個人化的加值服務，如藉由使用者的移動路徑，可更有效率且更精準的提供其所需之位置依存資料 (location dependent data)，其二由電信廠商或是資訊內容提供者的觀點：使用者移動行為模式，可以精準地提供使用者所需之位置依存資料，更可以利用使用者移動行為模式對系統的效能做最佳化，如使用者資料配置 (data allocation) 或是頻道資源分配之最佳化。因此，研發在行動計算環境上探勘使用者移動行為模式機制是一極為重要之研究課題。

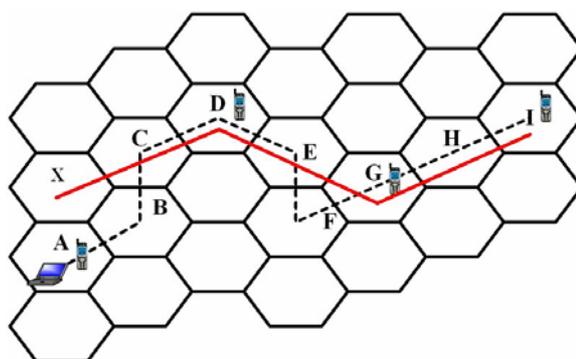
探勘使用者移動行為模式雖已引起學術界與業界之重視，然目前仍有許多技術上的挑戰。就使用者移動行為的資料記錄方面而言，如採用移動記錄 (moving log) 記錄所有人的移動行為時，資料量將會非常的龐大，且需要大量的儲存設備，在現有行動通訊系統中，其所花費的成本過高。因此，如何在增加現有的行動通訊的移動記錄儲存成本，研發在行動計算環境中，有效率的探勘使用者移動行為模式為本計畫之研究重點。明確的說，在本計畫中，我們將根據行動通訊系統中現有的通聯記錄 (call detail records)，研發資料探勘技術，探勘使用者的移動行為模式。

Uid	Date	Time	Cellid
1	01/03/2004	03:30:21	A
1	01/03/2004	09:12:45	D
1	01/03/2004	20:30:21	G
1	01/03/2004	21:50:32	I

表一、通聯記錄範例

表一所示為通聯記錄之範例，Uid 代表使用者的代號，Date 為通聯記錄之日期，Time 為接或收電話的時間，而 Cellid 則代表在接收電話時基地台之代碼。在通聯記錄中，Date 和 Time 屬於時間的資訊，而 Cellid 則是屬於空間的資訊。因此，通聯記錄隱含使用者的移動行為模式。

然通聯記錄包含了使用者片段的移動行為模式，在本計畫中，我們欲從通聯記錄中探勘使用者之移動行為，希冀所探勘之行為模式能精確地逼近真實之移動行為。圖一所示為利用通聯記錄所探勘之行為模式。虛線為使用者真實移動模式，實線為利用通聯記錄所建構的使用者移動模式。由圖一所示，透過通聯記錄所探勘之移動行為模式，即可大略地估算使用者真實移動行為。



圖一、利用通聯記錄探勘使用者移動行為

研究方法

假設使用者的移動是具有週期性，且使用者在鄰近的時間中的位置不會相隔太遠。moving section 為基本之時間單位。moving record 為記錄使用者在某個 moving section 中

有通聯記錄之基地台代碼以及次數。moving sequence 為由 ε 個 moving record 所形成之序列。第 i 個 moving sequence 表示為： $MS_i = \langle MR_i^1, \dots, MR_i^\varepsilon \rangle$ ，其中 MR_i^j 為第 i 個 moving sequence 中，第 j 個 moving record。本計劃中，給定一通聯記錄，我們須先將其轉換成數個 moving sequences，而後透過所研發之演算法探勘使用者行為模式。

時間投影序列 (Time Projection Sequence) $TP_{MS_i} = \langle a_1, \dots, a_n \rangle$ ，其中

$MR_i^{a_j} \neq \{\}$, $a_1 < \dots < a_n$ 。叢聚型時間投影序列 (Clustered Time Projection Sequence)

$CTP(TP_{MS_i}) = \langle CL_1, \dots, CL_x \rangle$ ，其中 CL_i 為第 i

個 cluster。以表一為例，假設 moving section 為四小時，一個 moving sequence 有 6 個 moving record，可得知其 moving sequence 為： $MS_1 = \langle \{A:1\}, \{\}, \{D:1\}, \{\}, \{\}, \{G:1, I:1\} \rangle$

而 $TP_{MS_1} = \langle 1, 3, 6 \rangle$ ，而 TP_{MS_1} 之叢聚型時間投影

序列為： $CTP(TP_{MS_1}) = \langle \{\}, \{3, 6\} \rangle$ 。

探勘使用者移動行為模式 (User Moving Pattern) 可分為以下三個步驟：

1. 資料收集步驟 (Data Collection Phase)

利用演算法 LS 自每 w 個移動序列 (moving sequences) 探勘出移動序列的規律性。

2. 時間叢聚步驟 (Time Clustering Phase)

利用演算法 TC 產生叢聚型時間投影序列 (Clustered Time Projection Sequence)

3. 迴歸步驟 (Regression Phase)

利用演算法 MF 自 Clustered Time Projection Sequence 導出使用者移動行

為模式函數。

當通聯記錄資料累積至一定量時，則可將通聯記錄之資料轉換成多個 moving sequences。通聯記錄雖然隱含了使用者移動的規律移動行為，但也可能包含某些不規律的移動行為。例如：使用者在平日是朝九晚五的上班族。在假日出遊時可能產生部份通聯記錄。然此部份通聯記錄所含之移動行為和其規律性移動行為模式有所不同，這些通聯記錄將被視為雜訊。因此，在 Data Collection Phase 中，將自 w 個 moving sequences 中，過濾掉此種雜訊而粹取出相類似的使用者移動行為。利用時間-空間區域性 (spatial-temporal locality) 之特性，亦即在相近的時間內使用者的位置不會相距太遠，演算法 TC 利用此特性將時間投影序列轉換成叢聚時間序列。演算法 MF 同時考慮時間和空間，利用回歸分析產生使用者移動路徑函數。

I. 資料收集步驟 (Data Collection Phase)

在此步驟，我們將從 w 個 moving sequences 中取出相近似的 moving sequence 並且將其合併為一個 aggregate moving sequence (簡稱 AMS)。其中，AMS 為由數個 Large Moving Record (簡稱 LMR^j) 所構成之序列。 LMR^j 記錄在不同的 moving sequences 的同一 moving section) 中，出現次數頻繁的基地台之代號以及其出現次數。首先，我們求出每個 moving section 之頻繁出現基地台代號後，而後利用此 AMS，將自 w 個 moving sequence 中粹取出相近似的 moving sequences，以便獲得頻繁之基地台代號出現次數。 MR_i^j 與 LMR^j 相似度函數為：

$$C(MR_i^j, LMR^j) = \frac{|\{x \in MR_i^j \cap LMR^j\}|}{|\{x \in MR_i^j \cup LMR^j\}|}$$

因此， MS_i 和 AMS 相似度函數定義為：

$$sim(MS_i, AMS) = \sum_{j=1}^{\varepsilon} |MR_i^j| * C(MR_i^j, LMR^j)$$

。給定一門檻值 $match_min_sup$ ，若 $sim(MS_i, AMS) \geq match_min_sup$ 則將 MS_i 視為相近似 moving sequence，表示其移動行為相近似具規律性。

Algorithm LS

input: w moving sequences with their length being ε ,
two threshold: $vertical_min_sup$ and
 $match_min_sup$

output: Aggregate moving sequence AMS

```

1 begin
2 for j = 1 to  $\varepsilon$ 
3 for i = 1 to  $w$ 
4  $LMR^j$  = large 1-itemset of  $MR_i^j$ ;
   (by  $vertical\_min\_sup$ )
5 for i = 1 to  $w$ 
6 begin
7  $match = 0$ ;
8 for j = 1 to  $\varepsilon$ 
9 begin
10  $C(MR_i^j, LMR^j) =$ 
    $|x \in MR_i^j \cap LMR^j| / |y \in MR_i^j \cup LMR^j|$ ;
11  $match = match + |MR_i^j| * C(MR_i^j, LMR^j)$ ;
12 end
13 if  $match \geq match\_min\_sup$  then
14 accumulate the occurring counts of
   items in the aggregate moving sequence;
15 end
16 end

```

在表二中，分別有五個 moving sequences MS_1 到 MS_5 。設 $vertical_min_sup$ 為 2，則可以得到：

$$LMR^1 = \{A, B\}, LMR^2 = \{A\}, \dots, LMR^5 = \{H\}$$

。設 $match_min_sup$ 為 2，可算出

$$sim(MS_1, AMS) = 1 * \frac{1}{2} + 1 * \frac{1}{1} + 0 + 1 * \frac{1}{2} + 1 * \frac{0}{1} = 2$$

$$sim(MS_2, AMS) = 3, \quad sim(MS_3, AMS) = 2,$$

$$sim(MS_4, AMS) = 3 \text{ 和 } sim(MS_5, AMS) = \frac{1}{2}。$$

	1	2	3	4	5
MS_1	A:14	A:2		F:1	I:2
MS_2			C:8	C:1, D:1, F:1	H:1, G:4
MS_3	A:1	C:1			H:1
MS_4	A:1, B:1	A:1	F:9		
MS_5	B:4	D:4	H:1		A:1, B:2
AMS	{A:16, B:1}	{A:3}	{}	{D:2, F:3}	{H:2}

表二、演算法 LS 範例

因此，可以 MS_1 到 MS_4 視為與 AMS 相近似的 moving sequences。透過計算相近似之 moving sequence 中基地台出現次數，可得到 $AMS = \langle \{A:16, B:1\}, \{A:3\}, \{\}, \{D:2, F:3\}, \{H:2\} \rangle$ 如表二所示。

II. 時間叢聚步驟(Time Clustering Phase)

產生 AMS 後，經由先前時間投影序列的定義，便可得到 TP_{AMS} 。根據時間-空間區域性的特性，演算法 TC 可將相鄰近的時間分成一群，如此便可以得知在每一群的時間內，使用者移動的距離不會太遠。藉由演算法 TC 將產生 $CTP(TP_{AMS})$ 。

在演算法 TC 中，對於某一個 cluster CL_i ，引入一測定其離散程度的測度

$$Sd(CL_i) = \frac{1}{m} \sum_{k=1}^m (n_k - \frac{1}{m} \sum_{j=1}^m n_j)^2。$$

其中 n_k 為 CL_i 中第 k 個元素。根據此測度，給與一門檻值 σ^2 ，使每一群中的離散程度不致於太高。如此一來便可達到在群內的時間中，使用者移動距離不至於太遠。

Algorithm TC

input: a Time Projection Sequence TP_{AMS} ,
threshold δ and σ^2
output: a Clustered Time Projection Sequence
 $CTP(TP_{AMS})$

```
1 begin
2 Group the numbers whose differences are within  $\delta$ ;
3 mark all clusters;
4 while there exist marked clusters and  $\delta \geq 1$ 
5 for each marked clusters  $CL_i$ ;
6 if  $Sd(CL_i) \leq \sigma^2$ 
7 unmark  $CL_i$ ;
8  $\delta = \delta - 1$ ;
9 for all marked clusters  $CL_i$ 
10 group the numbers whose differences are
    within  $\delta$  in  $CL_i$ ;
11 end while
12 if there exist marked clusters
13 for all marked cluster  $CL$ 
14  $k = 1$ ;
15 repeat
16  $k++$ ;
17 Divide evenly  $CL$  into  $k$  groups;
18 until the spread degree of each group  $\leq \sigma^2$ ;
19 end
```

演算法 TC 中， δ 是用來進行 distance-based clustering 的參數。第 4-11 行是檢驗一個 cluster 的 $Sd(CL_i)$ 是否較 σ^2 小，若不是，則代表該 cluster 內的時間距離太遠，因此進行細分，將 δ 進一步減小。在 12-18 行中，當 δ 值減少後仍無法符合條件時，進行更細分割的過程，此情況會發生在當一個 cluster 內的時間是連續的數字時。因此，令此 cluster 長 n ，可以證明當在將此 cluster 細分為 k 個子 cluster 時，若要其子 cluster 之 Sd 值合為最小，則要將原本長 n 之 cluster 等分成 k 個子 cluster。

Run	δ	σ^2	Clusters of a time projection sequence
0	3	1.6	$\langle\{1,2,3,4,5,9,10,14,17,18,20\}\rangle$
1	3	1.6	$\langle\{1,2,3,4,5\}^*,\{9,10\},\{14,17,18,20\}^*\rangle$
2	2	1.6	$\langle\{1,2,3,4,5\}^*,\{9,10\},\{14\},\{17,18,20\}\rangle$
3	1	1.6	$\langle\{1,2,3,4,5\}^*,\{9,10\},\{14\},\{17,18,20\}\rangle$
4	0	1.6	$\langle\{1,2,3,4,5\}^*,\{9,10\},\{14\},\{17,18,20\}\rangle$
5	0	1.6	$\langle\{1,2,3\},\{4,5\},\{9,10\},\{14\},\{17,18,20\}\rangle$

表三、演算法 TC 執行範例

在剛開始時，設定 δ 為 3， σ^2 為 1.6。 TP_{AMS}

為： $\langle\{1,2,3,4,5,9,10,14,17,18,20\}\rangle$ 。第一回合結束後，分成 $\{1,2,3,4,5\}$ ， $\{9,10\}$ ， $\{14,17,18,20\}$ 三個 cluster。但 $Sd(\{1,2,3,4,5\})$ 和 $Sd(\{14,17,18,20\})$ 均大於 1.6，因此將 δ 降為 2。此時 $\{14,17,18,20\}$ 在第二回合分成兩個 cluster $\{14\}$ 及 $\{17,18,20\}$ 。兩群的 Sd 值均符合條件。但是 $\{1,2,3,4,5\}$ 不符合條件，因此，降低 δ 值。然 $\delta=0$ 亦無法使 $Sd(\{1,2,3,4,5\}) < 1.6$ 。此時，將 $\{1,2,3,4,5\}$ 先嘗試平均 2 等份，也就是分成 $\{1,2,3\}$ $\{4,5\}$ 兩個 cluster。因為 $Sd(\{1,2,3\}) < 1.6$ ，符合門檻值 σ^2 ，演算法執行完畢，演算法 TC 停止。最後得到得結果是： $CTP(TP_{AMS}) = \langle\{1,2,3\}\{4,5\}\{9,10\}\{14\}\{17,18,20\}\rangle$ 。

III. 迴歸步驟(Regression Phase)

利用自 AMS 及 $CTP(TP_{AMS})$ ，在此步驟，利用演算法 MF 可以導出使用者的移動函數 (moving function)，而此移動函數可預測使用者移動模式 (User Moving Pattern)。

假設給定之 $AMS = \langle LMR^1, \dots, LMR^e \rangle$ 以及 $CTP(TP_{AMS}) = \{CL_1, \dots, CL_k\}$ ，其中 CL_i 為自演算法 TC 得到之第 i 個 cluster。對 $CTP(TP_{AMS})$ 中之每一個 cluster，可以導出以 $E_i(t) = (\hat{x}_i(t), \hat{y}_i(t), \text{valid_time_interval})$ 表示之 moving function。其中， $\hat{x}_i(t)$ 表示 x 軸分量的 moving function， $\hat{y}_i(t)$ 表示 y 軸分量的 moving function，valid_time_interval 為此 moving function 適用之時間。

假設 $CL_i = \{t_1, \dots, t_n\}$ ，其中 t_i 為在 CL_i 一個 moving section。首先，先自 AMS 中取出 moving section 自 t_l 到 t_n 的資料 (包括頻繁出現之基地台代號以及其次數)。為了要導出 moving function，必須先自電信業者處得到基地台的地理位置座標資訊。因此，可以將資料表示成以下數對： $(t_l, x_l, y_l, w_l), \dots, (t_n, x_n, y_n, w_n)$ ，其中 (x_i, y_i) 為在 AMS 中基地台之座標，

w_i 則為 AMS 中記錄在該基地台的通聯次數。因此，利用演算法 MF，可將使用者的路徑表示成一個三維空間(即 t-x-y)上的函數 F。使用者位置(即 x 座標和 y 座標)與時間有關，因此，此函數 F 應為時間的函數。然而，在位置座標中，x 座標和 y 座標彼此並無關係，須將三維空間(t-x-y)分成兩個二維的空間(t-x)及(t-y)來考慮，則一個資料點(t_i, x_i, y_i, w_i)可分別視為在空間(t-x)上的(t_i, x_i, w_i)以及空間(t-y)上的(t_i, y_i, w_i)。因此，分別對在兩個空間上的資料點分別求算其迴歸函數 $\hat{x}_i(t)$ 及 $\hat{y}_i(t)$ 。則對某一個時間點 $t = t_0$ ，便可以用 $(\hat{x}_i(t_0), \hat{y}_i(t_0))$ 估計使用者在此時間點的座標值。而此 cluster 的 valid_time_interval 則設定為 $[t_l, t_n]$ 。

對於 cluster 外的時間點，因為無法捕捉關於使用者足夠的資訊，因此，在 cluster 間的移動函數可利用 cluster 較邊緣的點進行內插法導出。如此一來便可以保證整個 moving function 在每一個時間點均可以預測使用者的位置。

Algorithm MF

input: AMS and clustered time projection sequence $CTP(TP_{AMS})$
output: A set of moving functions $F(t) = \{U_0(t), E_1(t), U_1(t), \dots, E_k(t), U_k(t)\}$

```

1 begin
2   initialize  $F(t)$ =empty;
3   for i= 1 to k-1
4     begin
5       Doing regression on  $CL_i$  to generate  $E_i(t)$ ;
6       Doing regression on  $CL_{i+1}$  to generate  $E_{i+1}(t)$ ;
7        $t_1$  =the last number in  $CL_i$ ;
8        $t_2$  =the first number in  $CL_{i+1}$ ;
9       Using inner interpolation to generate
           $U_i(t) = (\hat{x}_i(t), \hat{y}_i(t), (t_1, t_2))$ ;
10      Insert  $E_i(t), U_i(t)$  and  $E_{i+1}(t)$  in  $F(t)$ ;
11    end
12  if  $(1 \notin CL_1)$ 
13    generate  $U_0(t)$  and Insert  $U_0(t)$  into  $F(t)$ ;
14  if  $(k \in CL_k)$ 
15    generate  $U_k(t)$  and Insert  $U_k(t)$  into  $F(t)$ ;
16  return  $F(t)$ ;
17 end

```

第 5-6 行為對兩個 cluster 做出其 moving

function。第 9 行則是利用內插法算出在兩個 cluster 間的 moving function。第 12-13 行表示當第一個 cluster 若無第一個 moving section，則將第一個 cluster 所算出的 moving function 代入 t_1 ，令其為(a,b)，則 $U_0(t) = (a,b)$ ，其 valid_time_interval = $[1, t_1)$ 。第 14-15 所指的情形類似。

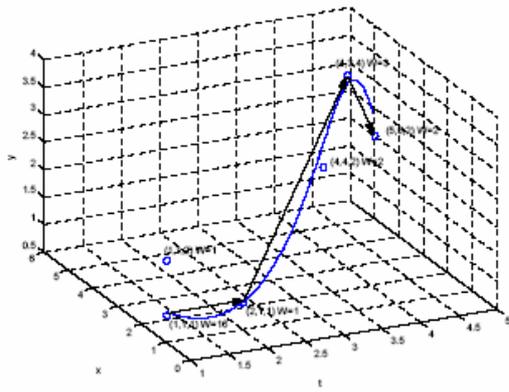
t_i	item	x_i	y_i	w_i
1	A	1	1	16
1	B	1	2	1
2	A	1	1	1
4	D	4	2	2
4	F	3	4	3
5	H	5	2	2

表四、通聯資料集合

假設某一 $CTP(TP_{AMS})$ 中的其中一個 cluster 為 $\{1,2,4,5\}$ ，則將 AMS 及基地台座標匯整成如表四。首先先對空間(t-x)上的資料點做加權迴歸。每一點的權重為 w_i 。因為在 (t,x) 中，有五個相異點，因此求算一四次方程式來逼近。其方程式

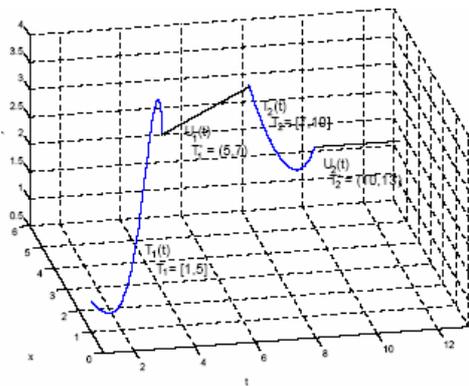
$$\hat{x}_i(t) = 3.191 - 3.807t + 1.918t^2 - 0.324t^3 + 0.022t^4$$

而對空間(t-y)上的資料點做加權迴歸時，每一點的權重亦為 w_i ，因為在(t,y)中，有六個相異點，因此求算一五次方程式來逼近。其方程式 $\hat{y}_i(t) = 2.144 - 1.497t + 0.368t^2 + 0.026t^3 + 0.022t^4 - 0.006t^5$ 。則可求算出 $E_i(t) = (\hat{x}_i(t), \hat{y}_i(t), [1,5])$ ，圖二為所求得之移動函數。

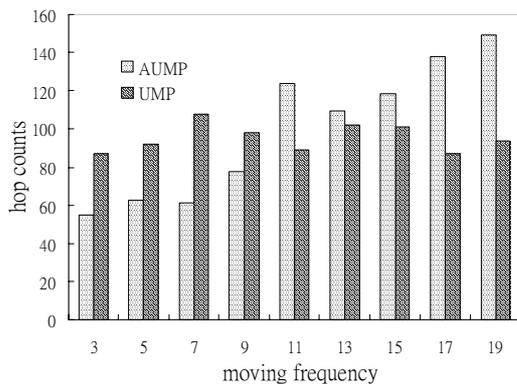


圖二、空間(t-x-y) moving function

對 CTP(TP_{AMS}) 每一個 cluster 均進行如上步驟後，將每一條 cluster 內的 moving function 以內插法相連即可。圖三為所示一近似移動行為模式範例。



圖三、moving function 範例



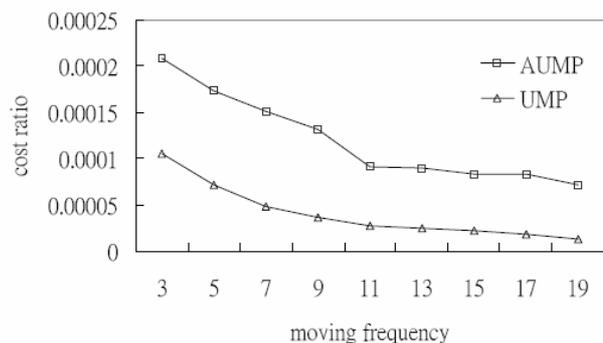
圖四、UMP 及 AUMP 之 hop count

實驗結果

為了模擬在行動環境中的基地台，我們

使用 8*8 的 mesh network，每一個節點代表一個基地台，共有 64 個基地台。使用者在一個 moving section 中的移動次數是一個介於 $mf-2$ 和 $mf+2$ 常態分布值。使用者在一個 moving section 中有 cf 次可決定是否通話。而在移動時，會將使用者移動的方向設定分別一個機率，並且為了保持週期性，使用者移動至使用者出發位置之方向的機率為 P_{back} ，而前往其他方向的機率為 $(1-P_{back})/(n-1)$ ，其中 n 為該使用者所移動之基地台數。在此計劃中，我們採用 hop count 做為測量精確度之一測度。

在[18]中探勘使用者移動行為模式之結果為 UMP，此研究中探勘出之結果以 AUMP 表示。首先，我們將研究 mf 對 AUMP 之影響。若 $w=10$ ， $cf=3$ ， $\epsilon=12$ ，由圖四中可得知當使用者移動越快速時，AUMP 的準確度不若 UMP。考慮一測度 $cost\ ratio = (1/hop\ counts) / data\ amount$ ，此一測度將資料量加入考量，結果如圖八所示，AUMP 相對於 UMP 表現較佳。亦即縱使 AUMP 精確度較不佳，但在考慮儲存資料量成本下，AUMP 能用較少之資料量達到較高之精確度。

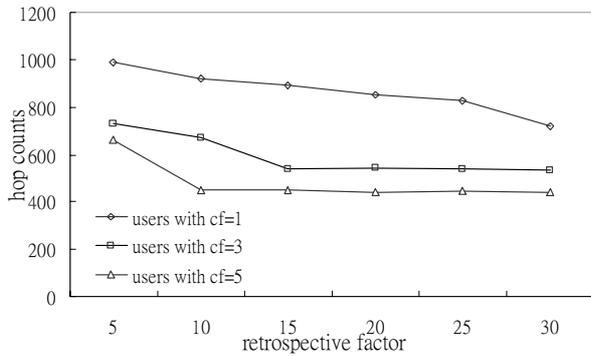


圖五、UMP 及 AUMP 之 cost ratio 比較

在 $\epsilon=12$ ， $mf=3$ ， $min_vertical_sup=20\%$ ， $min_match_sup=2.4$ ， $\sigma^2=0.25$ 的情形下，考慮 w 與 cf 變動時對精確度的影響，圖六為其結果。由此圖可得知，當使用者有較高的使用頻率則可收集較少之 moving sequence 即可

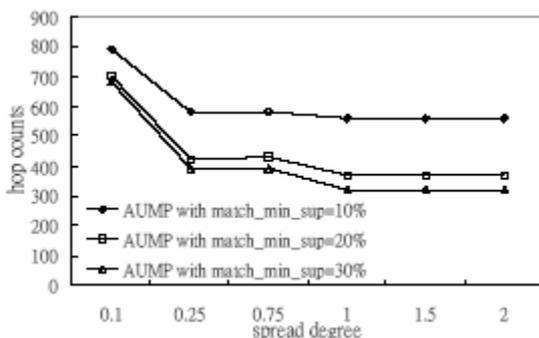
捕捉使用者的移動行為。若使用者之使用頻率較低，則需收集較多之 moving sequence 方可有效捕捉其移動行為。

另外，對離散程度之門檻值 σ^2 之影響進行討論。首先設定 $\epsilon=12$ ， $mf=3$ ， $vertical_min_sup=20\%$ 。Match_min_sup 設定



圖六、AUMP 在 w 變動時對 hop count 影響

成 1, 3, 5。圖七為其實驗結果。由此圖可得知當所給的 σ^2 高於 0.25 時可以得到較精確的結果，而在 σ^2 高於 0.75 其精確度並無太大改善。原因是因為在演算法 TC 中，若 σ^2 低於 0.25 時，則每一個 moving section 會自己成為一個 cluster，如此一來會使得無法利用空間-時間區域性之特性而造成 hop count 大增。若 σ^2 介於 0.25 與 0.75 間，因為可利用空間-時間區域性之特性，因此精確度增加。在 σ^2 大於 0.75 時，可看出若 match_min_sup 較高者，亦即選出較能代表使用者行為之 moving sequence 者具有增加精確度的功用。



圖七 門檻值 σ^2 對精確度的影響

結果與討論

在計畫中，我們提出利用通聯記錄來探勘使用者移動行為模式機制，包含一系列演算法，此機制並不用需要額外的儲存空間。探勘的方法分為以下的步驟：演算法 LS 粹取出類似的使用者移動行為之 moving sequence。經過演算法 LS，可獲取通聯記錄資料中隱含之行為模式。利用空間-時間區域性的特性，演算法 TC 將較近的時間聚集至同一群 (cluster)。透過演算法 TC 所分出來的時間區間以及演算法 LS 探勘出之類似使用者移動行為，演算法 MF 可導出使用者的移動函數。利用此移動函數，可以推測出使用者在某時間點的座標，並可以轉換成基地台的代碼。

在模擬實驗中，證實所提出的方法確實可以有效的探勘出高品質的使用者移動行為模式。

計畫成果自評

由上述的結果與討論可知利用我們所提出的架構確實可以有效的降低資料儲存量，並且可以達到有效利用通聯記錄達到探勘使用者移動行為模式之目的。本計劃研究成果已撰寫成論文，並已投至國際研討會，目前待審中。

參考文獻

- [1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceeding of 20th International Conference on Very Large Data Bases (VLDB)*, September 1994.
- [2] I. F. Akyildiz and Joesph S. On Location Management for Personal Communications

- Network. In *Proceeding of IEEE Communications Magazine*, pages 138-145, September 1996.
- [3] B. Bruegge and B. Bennington. Applications of Mobile Computing and Communication. *IEEE Personal Communication*, pages 64-71, February 1996.
- [4] S. K. Das and S. K. Sen. Adaptive Location Prediction Strategies Based on a Hierarchical Network Model in Cellular Mobile Environment. In *Proceeding of the 2nd International Mobile Computing Conference*, pages 131-140, March 1996.
- [5] N. Davies, G. S. Blair, K. Cheverst, and A. Friday. Supporting Collaborative Application in a Heterogeneous Mobile Environment. *Computer Communication Special Issues on Mobile Computing*, 1996.
- [6] M. H. Dunham. Mobile Computing and Databases. *Tutorial of International Conference on Data Engineering*, February 1998.
- [7] EIA/TIA. Cellular Radio Telecommunication Intersystem Operations. 1991.
- [8] A. Elmagarmid, J. Jain, and T. Furukawa. Wireless Client/Server Computing for Personal Information Services and Applications. *ACM SIGMOD RECORD*, 24(4):16-21, December 1995.
- [9] K.-Y. L. Guo-Hui Li and T.-W. Kuo. Location Update Generation in Cellular Mobile Computing Systems. In *Proceeding of International Workshop on Parallel and Distributed Real-time System*, April 2001.
- [10] J.-T. H. Hsiao-Kuang Wu, Ming-Hui Jin and C.-Y. Ke. Personal Paging Area Design Based On Mobile's Moving Behaviors. In *Proceeding of IEEE INFOCOM*, 2001.
- [11] R. V. Hogg and E. A. Tanis. *Probability and Statistical Inference*. Prentice-Hall International Inc., 1997.
- [12] T. Imielinski and B. R. Badrinath. Querying in Highly Mobile and Distributed Environment. In *Proceeding of the 18th International Conference on Very Large Data Bases*, pages 41-52, August 1992.
- [13] N. Krishnakumar and R. Jain. Escrow Techniques for Mobile Sales and Inventory Applications. *ACM Journal of Wireless Network*, 3(3):235-246, July 1997.
- [14] D. L. Lee. Data Management in a Wireless Environment. *Tutorial of International Conference on Database System for Advance Applications*, April 1999.
- [15] C.-R. Lin and M.-S. Chen. On the Optimal Clustering of Sequential Data. In *Proceeding of SIAM International Conference on Data Mining*, 2002.
- [16] Y.-B. Lin. GSM Network Signaling. *ACM Mobile Computing and Communications*, 1(2):11-16, 1997.
- [17] Y.-B. Lin. Reducing Location Update Cost in a PCS Network. *IEEE/ACM Transactions on Networking*, 5(1):25-33, February 1997.
- [18] W.-C. Peng and M.-S. Chen. Developing Data Allocation Schemes by Incremental Mining of User Moving Patterns in a Mobile Computing System. In *Proceeding of IEEE Transactions on Knowledge and Data Engineering, Volume 15*, pages

70-85, 2003.

- [19] E. Sami Tabbane. Location Management Methods for Third-Generation Mobile Systems. In *Proceeding of IEEE Communication Magazine*, pages 72-84, August 1997.
- [20] M. Satyanarayanan. Mobile Information Access. *IEEE Personal Communication*, pages 26-33, February 1996.
- [21] S. M. Tseng and C. F. Chiu. An Efficient Method for Mining Associated Service Patterns in Mobile Web Environments. In *Proceeding of ACM Symposium on Applied Computing*, March 2003.