Appl Intell (2011) 35:32–40
DOI 10.1007/s10489-009-0201-z

Mining periodic movement patterns of mobile phone users based on an efficient sampling approach

Yao-Te Wang · Ju-Tzu Cheng

Published online: 9 December 2009

Abstract In m-commerce services, the periodic movement trends of customers at specific periods can be adopted to allocate the resources of telecommunications systems effectively and offer personalized location-based services. This study explores the mining of periodic maximal promising movement patterns. A detailed process for mining periodic maximal promising movement patterns based on graph mapping and sampling techniques is devised to enhance mining efficiency. First, a random sample of movement paths from time intervals is taken. Second, a unique path graph structure is built to store the movement paths obtained from the sample. Third, a graph traversal algorithm is developed to identify the maximal promising movement patterns. Finally, vector operations are undertaken to examine the maximal promising movement patterns in order to derive the periodic maximal promising movement patterns. Experimental results reveal that the sampling approach with mining has excellent execution efficiency and scalability in the investigation of periodic maximal promising movement patterns.

Keywords Data mining · Sampling · Sequential pattern · Periodic maximal promising movement pattern

Abbreviations

GSM	Global System for Mobile Communications;
PMP	Promising Movement Pattern;
MPMP	Maximal Promising Movement Pattern;
PMPMP	Periodic Maximal Promising Movement Pattern

Y.-T. Wang (✉)
Department of Computer Science and Information Management, Providence University, Taichung 433, Taiwan, ROC
e-mail: ytwang@pu.edu.tw

J.-T. Cheng
Department of Accounting Information, National Taichung Institute of Technology, Taichung 404, Taiwan, ROC
e-mail: jtcheng@ntit.edu.tw

1 Introduction

The integration of information and mobile communication technologies has recently made mobile computing a very popular subject of research and one with vast commercial potential. Two important factors, namely, the efficiency of communication system utilization and the services provided by telecommunications operators, play an essential role in determining the capability of telecommunications operators to make a profit. If various limited system resources, such as bandwidth, memory capacity, and data caches, could be more effectively deployed and utilized, service providers would be able to improve their competitive advantage and therefore become more profitable.

Apart from personal communication system technologies, which have turned voice services into prominent application services used extensively in daily lives, telecommunications operators are also aggressively developing the mobile broadband service sector (such as 3.5G, 4G) [11]. In light of such trends, a method that could identify trends in mobile service subscribers' movements while they make calls on their mobile phones would enable telecommunications operators to make arrangements and deploy relevant system resources and information sought by these users at prior [6, 9]. In addition to improving the operational performance of the entire communication system, such a scheme would enhance the ability of service providers to offer personalized services (e.g., deployment of base stations, hands-off strategies, dispatch of user access data, pricing

strategies, advertising and marketing and behavior analysis).

Furthermore, many mobile phone users engage in periodic activities such as commuting daily to the office and returning home, going to and returning from school, delivering goods and visiting clients. A reliable approach that can efficiently predict periodic movement trends of service subscribers when they use mobile phones would thereby enable system resources to be more effectively used by telecommunications operators. Periodic movement trends could also be applied to provide appropriate location-based mobile commerce services.

For instance, if a subscriber S_1 periodically moves along a fixed path, then system operators can transmit location information related to S_1 at specific time periods in advance. For example, service providers can notify S_1 of traffic congestion on specific road sections, or provide a list of stores with promotions ahead of time, for S_1 to receive when passing by them. These personalized services enable service providers to not only improve customer satisfaction with their mobile phone clientele and fortify customer loyalty, but also to better reach out to advertisers or even to establish cross-sector alliances to create opportunities for new businesses.

Clearly, research into trends of periodic movements for mobile phone users has definite practical values not only as a topic of academic research, but also as a subject with practical field applications in related sectors. To obtain a circumstantial account of periodic movement trends, a more discriminatory approach to the mining of movement paths that appear repeatedly in cyclic time periods should be explored. Previous research works in periodic patterns have concentrated on mining the repeated occurrence of events in time series or sequential data, and this is not directly applicable to the mining of periodic movement patterns. Anwar et al. [2] recently introduced an interval validation process to mine the periodicity of patterns after all sequential patterns were obtained. However, their mining processes are inefficient in finding all sequential patterns and testing the periodicity of these patterns, as indicated in the experiment results (see Fig. 2). This study describes the practicability of mining periodic movement patterns for mobile commerce, and presents a novel approach that involves a sampling technique in pattern mining processes to improve mining efficiency.

This work makes the following major contributions. Initially, the value of the periodic maximal promising movement pattern is clarified (as in Definition 5). Secondly, a highly efficient mining algorithm based on sampling is introduced to discover the periodic maximal promising movement patterns of mobile phone users.

2 Related work

Sequential pattern mining [1, 7] has received extensive attention. However, most investigations on movement pattern mining have been based on sequential pattern mining methodologies, while few works have presented mining schemes designed according to the movement patterns of mobile phone users.

The Apriori algorithm [1] employs a level-wise iterative search to identify frequent sequential patterns. It generates the candidate sequences of length k from prior known frequent sequences of length $k-1$ and then scans the database once to determine the frequent sequences of length k. Based on the Apriori algorithm, Peng and Chen [9] devised a method of mining movement patterns in a mobile computing environment to enhance the operational performance of a mobile computing system through data allocation. Yavaş et al. [12] proposed a three-phase model based on the Apriori algorithm to predict movements—identification of movement patterns precedes the generation of rules, which are then utilized to predict movements. Tseng and Lin [10] introduced *sequential mobile access patterns* (SMAP) and presented a SMAP-Mine algorithm that can mine user movement patterns according to the contents of services requested by users. The SMAP-Tree data structure stores movement sequences, allowing all further mining tasks to be performed by the given SMAP-Tree.

However, sequential patterns derived from traditional data mining methodologies such as Apriori-like algorithms, PrefixSpan algorithms, and other graph-based mining algorithms tend to be tedious and inefficient when adopted in applications designed to predict the movement paths of mobile phone users. The structure of telecommunications networks is not taken into account when the sequential patterns are mined; hence, too many candidate sequences which will never become frequent, are created and tested. Moreover, the movement trends of mobile phone users can be forecast more efficiently by potential maximal frequent calling path patterns than by sequential patterns [6].

Özden et al. [8] attempted to identify cyclic association rules through an extended Apriori algorithm. Their proposed algorithm divides a database into N time intervals and applies the Apriori algorithm to each time interval in order to obtain the frequent patterns of each time interval. The results are then processed through pattern matching to identify frequent cyclic patterns within the time intervals. Elfeky et al. [5] presented an approach to identify the periodicity of events from a given time sequence. Anwar et al. [2] presented an algorithm based on periodicity mining of sequential patterns in a post-mining environment, in which periodicity mining is performed after finding the sequential patterns.

The methods of Özden et al. and Anwar et al. can be extended, to locate movement patterns in every time interval, and then to identify the periodicity of the movement patterns. However, this method of identifying all movement patterns for every time interval leads to the massive generation of non-periodic patterns, and would require additional time and spatial costs to compare patterns. Additionally, these periodic pattern mining techniques do not improve the efficiency of periodic movement pattern mining, due to the need to have movement patterns in each time interval for a given period and the inability to adopt the highly specific restrictions of telecommunications networks. A particularly characteristic feature of telecommunications networks is that each hexagonal cell (the radius of a base station) theoretically is surrounded by at most six neighboring cells [4]. According to the connective structure of telecommunications networks, mining tasks can be simplified and improved.

In the field of periodic mobile movement pattern mining, the results form a set of sequences that emerge frequently and periodically. It is effective to predict movement trends with the observed mobile movement patterns instead of sequential patterns [6], and in this study the proposed mining approach is highly efficient at discovering periodic movement patterns.

3 Mining of periodic maximal promising movement pattern

3.1 Periodic maximal promising movement patterns

A GSM (Global System for Mobile Communications) network is based on cellular radio technology [4]. The radio coverage of the neighboring base stations in a GSM network overlaps in order to maintain phone connections. Two base stations are termed *adjacent* if they form neighboring cells in a GSM network. When a GSM network is deployed, the neighbors of each base station are determined. The positions of the base stations remain fixed until some new base station is included or the GSM network is reformed. Furthermore, the path graph (Definition 4) is constructed according to the movement paths; meanwhile, the reference set of the base stations is known and fixed.

This section formally defines the periodic maximal promising movement pattern.

Definition 1 The *movement path* is shown as the path formed by the trail of base stations $(bs_1, bs_2, \ldots, bs_n)$ which a user is traversing while using his mobile phone. The movement path is denoted as $\langle bs_1, bs_2, \ldots, bs_n\rangle$, where bs_i and bs_{i+1} represent adjacent base stations, $1 \leq i < n$.

Definition 2 An *edge* $\langle bs_i, bs_{i+1}\rangle$ is a movement path formed by adjacent base stations. A *link* $\langle bs_i, bs_{i+1}, bs_{i+2}\rangle$ is a movement path formed by an incoming edge $\langle bs_i, bs_{i+1}\rangle$ and an outgoing edge $\langle bs_{i+1}, bs_{i+2}\rangle$. An edge or link is frequent if it occurs at a frequency not less than the minimum support threshold. A movement path $\langle bs_1, bs_2, \ldots, bs_n\rangle$ can be decomposed into a set of edges $\{\langle bs_a, bs_b\rangle | 1 \leq a < b \leq n\}$ and a set of links $\{\langle bs_i, bs_j, bs_k\rangle | 1 \leq i < j < k \leq n\}$.

Definition 3 A movement path $P = \langle bs_1, bs_2, \ldots, bs_n\rangle$ is recognized as a *promising movement pattern* (PMP) if all edges $\{\langle bs_a, bs_b\rangle | 1 \leq a < b \leq n\}$ and links $\{\langle bs_i, bs_j, bs_k\rangle | 1 \leq i < j < k \leq n\}$ in P have been qualified as frequent. A PMP is *maximal* (denoted by MPMP) if it is not contained in any other PMP. If $P_{sub} = \langle bs_s, bs_{s+1}, \ldots, bs_t\rangle$, $1 \leq s < t \leq n$, then P_{sub} is called a *subpattern* of P, and it is denoted as $P_{sub} \subseteq P$.

Two PMPs $\langle bs_i, bs_{i+1}, \ldots, bs_{j-1}, bs_j\rangle$ and $\langle bs_s, bs_{s+1}, \ldots, bs_{t-1}, bs_t\rangle$ can be joined to form another, longer, PMP $\langle bs_i, \ldots, bs_{j-1}, bs_j, bs_{s+2}, bs_{s+3}, \ldots, bs_t\rangle$ if $bs_{j-1} = bs_s$ and $bs_j = bs_{s+1}$. Restated, if the outgoing edge for a promising movement pattern is consistent with the incoming edge for another promising movement pattern, then the two patterns can be joined to form a longer, promising movement pattern.

Definition 4 A *path graph G(V,E,L)* comprises a set of vertices $V = \{v_i | 1 \leq i \leq n\}$, a set of edges $E = \{\langle v_a, v_b\rangle | 1 \leq a, b \leq n, a \neq b, v_a, v_b \in V\}$ and a set of links $L = \{\langle v_i, v_j, v_k\rangle | 1 \leq i, j, k \leq n, i \neq j, j \neq k, v_i, v_j, v_k \in V\}$. Vertices $v_1, v_2, \ldots, v_n$ in a path graph represent the base stations that the user has passed while moving. The path graph holds information about movement paths during MPMP mining.

Definition 5 Assume that the time frame of a set of collected movement paths spans the time intervals T_0 to T_{n-1}. If an MPMP $P = \langle bs_1, bs_2, \ldots, bs_n\rangle$ starts to show cyclic occurrence beginning from time interval T_o once every l time intervals, that is $T_o, T_{o+l}, T_{o+2\times l}, \ldots, T_{o+(\lceil (n-o)/l\rceil -1)\times l}$, then P is a *periodic maximal promising movement pattern* (PMPMP). PMPMP is denoted as $P_{\pi_{l,o}}$, where $\pi_{l,o}$ represents the period of the pattern; l indicates the length of the period, and o is the offset of the period (the starting time interval).

Given a PMPMP $= \langle bs_1, bs_2, \ldots, bs_n\rangle$, telecommunications operators can predict the path that a user is likely to take during a given time period. When the user moves from bs_{i-1} to bs_i at a specific time, the PMPMP indicates that the user will head towards bs_{i+1}, and will then further move to bs_{i+2}. Mining PMPMP is a nontrivial task, and cannot

be performed well by traditional sequential pattern mining methods. The following section presents a highly efficient algorithm for mining PMPMP.

3.2 The sampling algorithm for mining PMPMP

The following property is adopted to facilitate the sampling procedure when mining PMPMP.

Property 1 *Assume that the mining period is $\pi_{l,o}$ and the time frame spans n time intervals. Given a set of MPMP, namely $S_{o+k\times l}$ that is obtained from the time interval $T_{o+k\times l}$ of the period $\pi_{l,o}, 0 \le k \le \lceil (n-o)/l \rceil - 1$, for any PMPMP $P_{\pi_{l,o}}$, if a specific MPMP $= P_{(o+k\times l)_i}$ is definitely from $S_{o+k\times l}$, then $P_{\pi_{l,o}} \subseteq P_{(o+k\times l)_i}$.*

Proof For $P_{\pi_{l,o}}$ to have the periodicity of $\pi_{l,o}$ in $n(T_0, T_1, \ldots, T_{n-1})$ time intervals, $P_{\pi_{l,o}}$must occur in all time intervals $(T_o, T_{o+l}, \ldots, T_{o+(\lceil (n-o)/l \rceil - 1)\times l})$ within the period. Namely, the MPMP $P_{(o+k\times l)_i} \in S_{o+k\times l}$ mined from time interval $T_{o+k\times l}$ must exist, and hence, $P_{\pi_{l,o}} \subseteq P_{(o+k\times l)_i}, 0 \le k \le \lceil (n-o)/l \rceil - 1$. □

From Property 1, all PMPMP must be included in a set of MPMP identified from every time interval within the period. Therefore, sampling improves the efficiency of PMPMP mining. Without loss of generality, in order to identify all PMPMPs quickly, the first time interval T_o is selected for MPMP mining before determining whether these MPMPs or their subpatterns provide periods that should be investigated. The PMPMP mining process is summarized as follows:

Suppose that D is a movement path database with a time span of n time intervals, and the period to be mined for PMPMP is $\pi_{l,o}$. Assume that a MPMP $P = \langle bs_1, bs_1, bs_2, \ldots, bs_k \rangle$ with a length of k that occurs at time interval $T_s, o \le s \le o + (\lceil (n-o)/l \rceil - 1) \times l$.

To determine whether P or its subpattern contains $\pi_{l,o}$, P is first decomposed into edges $\{\langle bs_a, bs_b \rangle | 1 \le a < b \le k\}$ and links $\{\langle bs_x, bs_y, bs_z \rangle | 1 \le x < y < z \le k\}$. Next, a movement path is retrieved from another time interval T_t $(t \ne s, o \le t \le o + (\lceil (n-o)/l \rceil - 1) \times l$, and decomposed into edges and links. Edge vectors $V_{\langle bs_1,bs_2 \rangle}$, $V_{\langle bs_2,bs_3 \rangle}, \ldots, V_{\langle bs_{k-1},bs_k \rangle}$ and link vectors $V_{\langle bs_1,bs_2,bs_3 \rangle}$, $V_{\langle bs_2,bs_3,bs_4 \rangle}, \ldots, V_{\langle bs_{k-2},bs_{k-1},bs_k \rangle}$ corresponding to time intervals $T_o, T_{o+l}, T_{o+2l}, \ldots, T_{o+(\lceil (n-o)/l \rceil - 1)\times l}$ are then created from the edge and link information obtained from the decomposition of P. The logical AND operation (represented as "&") to the vectors in the sequence of $V_{\langle bs_1,bs_2 \rangle}$, $V_{\langle bs_1,bs_2,bs_3 \rangle}$, $V_{\langle bs_2,bs_3,bs_4 \rangle}, \ldots, V_{\langle bs_{k-2},bs_{k-1},bs_k \rangle}$ until the intermediate operating result of $V_{\langle bs_{m-2},bs_{m-1},bs_m \rangle}$ becomes inconsistent with the vector $V_{one} = [1, 1, \ldots, 1]$. Namely, if $V_{\langle bs_1,bs_2 \rangle} \& V_{\langle bs_1,bs_2,bs_3 \rangle} \& V_{\langle bs_2,bs_3,bs_4 \rangle} \& \cdots \& V_{\langle bs_{m-3},bs_{m-2},bs_{m-1} \rangle} = V_{one}$, and $V_{\langle bs_1,bs_2 \rangle} \& V_{\langle bs_1,bs_2,bs_3 \rangle} \& \cdots \& V_{\langle bs_{m-2},bs_{m-1},bs_m \rangle} \ne V_{one}$, then the subpattern $\langle bs_1, bs_2, \ldots, bs_{m-1} \rangle$ of P has the period $\pi_{l,o}$. The next step is to perform a logical AND operation on the vectors in the sequence of $V_{\langle bs_{m-1},bs_m \rangle}, V_{\langle bs_{m-1},bs_m,bs_{m+1} \rangle}, \ldots, V_{\langle bs_{k-2},bs_{k-1},bs_k \rangle}$ to examine whether other subpatterns have the period $\pi_{l,o}$.

The sampling approach adopted in this study is divided into two phases. The first phase focuses on mining the MPMP for the first time interval in a given period. The second phase involves cyclic verification of mining results obtained from the first phase in order to determine PMPMP. Assume that the time frame spans n time intervals. For user specified period $\pi_{l,o}$ and minimum support ζ, the mining procedures are outlined as follows:

Input: a set of movement paths, n time intervals, period $\pi_{l,o}$, minimum support ζ
Output: all PMPMPs
Method:
Phase I: find MPMPs from the first time interval T_o of the period $\pi_{l,o}$

(1) repeat
(2) retrieve a movement path $\langle bs_s, bs_{s+1}, \ldots, bs_t \rangle$ from the first time interval T_o
(3) decompose $\langle bs_s, bs_{s+1}, \ldots, bs_t \rangle$ into a set of edges $\langle bs_a, bs_b \rangle$ and a set of links $\langle bs_i, bs_j, bs_k \rangle, s \le a < b \le t, s \le i < j < k \le t$; record their supports
(4) add $\langle bs_a, bs_b \rangle$ and $\langle bs_i, bs_j, bs_k \rangle$ to path graph G
(5) until all movement paths in T_o are retrieved and processed
(6) delete those edges and links whose supports are less than ζ
(7) traverse G in depth-first-search order to generate MPMPs in T_o

Phase II: find all PMPMPs in $\pi_{l,o}$

(8) for each MPMP $P = \langle bs_1, bs_2, \ldots, bs_k \rangle$ found in phase I
(9) decompose P into sets of edges E and links L
(10) for each edge $\langle bs_a, bs_b \rangle \in E$ and link $\langle bs_x, bs_y, bs_z \rangle \in L$
(11) create vectors $V_{\langle bs_a,bs_b \rangle}$ and $V_{\langle bs_x,bs_y,bs_z \rangle}$ of length $\lceil (n-o)/l \rceil$ based on the time intervals $T_o, T_{o+l}, T_{o+2l}, \ldots, T_{o+(\lceil (n-o)/l \rceil - 1)\times l}$
(12) endfor
(13) retrieve the movement paths from the time intervals $T_{o+l}, T_{o+2l}, \ldots, T_{o+(\lceil (n-o)/l \rceil - 1)\times l}$
(14) decompose the movement paths into edges and links, and keep the counts of the edges and links in the corresponding elements of $V_{\langle bs_a,bs_b \rangle}$ and $V_{\langle bs_x,bs_y,bs_z \rangle}$

(15) transform $V_{\langle bs_a, bs_b\rangle}$ and $V_{\langle bs_x, bs_y, bs_z\rangle}$ to Boolean vectors; if the element value is less than ζ, convert it to 0; otherwise, convert it to 1
(16) apply the logical *AND* operation to each edge vector $V_{\langle bs_i, bs_{i+1}\rangle}$ and its following link vectors $V_{\langle bs_i, bs_{i+1}, bs_{i+2}\rangle}$, $V_{\langle bs_{i+1}, bs_{i+2}, bs_{i+3}\rangle}$, ... to verify that the subpatterns of P are PMPMPs
(17) output resulting PMPMPs
(18) endfor

For instance, assume that the period includes n time intervals. Consider the problem of determining the PMPMP for a user *Usr* in a period $\pi_{3,1}$. Table 1 shows the movement paths of *Usr* in time interval T_1. The minimum support is set to 20%. The process of mining PMPMP is described as follows.

Phase I: In step (3), decompose the movement paths listed in Table 1 into edges and links, as shown in Table 2, and record the corresponding support. In step (4), build a path graph according to the edges and links presented in Table 2, while deleting edges and links with support counts below the minimum support in step (6), as illustrated in Fig. 1.

In step (7), continue by applying the depth-first-search to traverse the paths comprising the edges and links in the path graph. In Fig. 1, edge $\langle bs_a, bs_b\rangle$ is the incoming edge of a MPMP, since it is not the outgoing edge of any link in the graph. Begin the trace of MPMP from node a to node b and arrive at node b. The link $\langle bs_a bs_a, bs_b, bs_c\rangle$ leads to node c. Another link $\langle bs_b, bs_c, bs_d\rangle$ is discovered at the new node, leading the trace further to node d. The leaf node f is eventually reached, resulting in an MPMP of $\langle bs_a, bs_b, bs_c, bs_d, bs_e, bs_f\rangle$.

Phase II: In steps (8) and (9), take the MPMP obtained from Phase I, and decompose the pattern into edge set E and link set L. In steps (13) and (14), retrieve the movement paths from other time intervals $T_{1+k\times 3}$ $(0 < k \leq \lceil (n-1)/3\rceil - 1)$ during the same period $\pi_{3,1}$ in the database D, and decompose them into edges and links. In addition, two sets of vectors V_e and V_l are derived from E and L in step (14), where $e \in E, l \in L$, and the occurrence of every edge and link within each time interval is recorded in the corresponding elements of the vectors. In step (15), convert V_e and V_l into Boolean vectors, assigning "1" to corresponding elements with counts greater than or equal to the minimum support, and "0" to corresponding elements with counts below the minimum support.

Now refer to the example shown in Fig. 1. Decompose the MPMP $\langle bs_a, bs_b, bs_c, bs_d, bs_e, bs_f\rangle$ identified from time interval T_1 into edges $\langle bs_a, bs_b\rangle$, $\langle bs_b, bs_c\rangle$, $\langle bs_c, bs_d\rangle$, $\langle bs_d, bs_e\rangle$, $\langle bs_e, bs_f\rangle$ and links $\langle bs_a, bs_b, bs_c\rangle$, $\langle bs_b, bs_c, bs_d\rangle$, $\langle bs_c, bs_d, bs_e\rangle$ and $\langle bs_d, bs_e, bs_f\rangle$. Scan the database for other movement paths in the remaining time intervals $T_4, T_7, \ldots, T_{1+(\lceil (n-1)/3\rceil - 1)\times 3}$ during the same period. Decompose the movement paths into edges and links, and note the frequency of occurrence in the corresponding vector elements. Table 3 lists the final results of step (15), which generates nine vectors $V_{\langle bs_a, bs_b\rangle}$, $V_{\langle bs_b, bs_c\rangle}$, $V_{\langle bs_c, bs_d\rangle}$, $V_{\langle bs_d, bs_e\rangle}$, $V_{\langle bs_e, bs_f\rangle}$, $V_{\langle bs_a, bs_b, bs_c\rangle}$, $V_{\langle bs_b, bs_c, bs_d\rangle}$, $V_{\langle bs_c, bs_d, bs_e\rangle}$ and $V_{\langle bs_d, bs_e, bs_f\rangle}$.

The step (16) is to apply the AND operation to these vectors in the given sequence to derive the PMPMPs that feature in the period $\pi_{3,1}$. If a vector $V_{\langle bs_a, bs_b\rangle}$ is equivalent to the

Table 1 Movement paths for *Usr* in T_1

Path ID	Movement path
P001	$\langle bs_a, bs_b, bs_c\rangle$
P002	$\langle bs_a, bs_b, bs_g\rangle$
P003	$\langle bs_a, bs_b, bs_c, bs_d\rangle$
P004	$\langle bs_b, bs_c, bs_d\rangle$
P005	$\langle bs_b, bs_c, bs_d\rangle$
P006	$\langle bs_c, bs_d, bs_e\rangle$
P007	$\langle bs_c, bs_d, bs_e, bs_h\rangle$
P008	$\langle bs_b, bs_c, bs_d, bs_e, bs_f, bs_i\rangle$
P009	$\langle bs_c, bs_d, bs_e, bs_f\rangle$
P010	$\langle bs_b, bs_c, bs_d, bs_e\rangle$

Table 2 Edges, links, and the support counts

Edge/Link	Support
$\langle bs_a, bs_b\rangle$	30%
$\langle bs_b, bs_c\rangle$	60%
$\langle bs_b, bs_g\rangle$	10%
$\langle bs_c, bs_d\rangle$	80%
$\langle bs_d, bs_e\rangle$	50%
$\langle bs_e, bs_f\rangle$	20%
$\langle bs_e, bs_h\rangle$	10%
$\langle bs_f, bs_i\rangle$	10%
$\langle bs_a, bs_b, bs_c\rangle$	20%
$\langle bs_a, bs_b, bs_g\rangle$	10%
$\langle bs_b, bs_c, bs_d\rangle$	50%
$\langle bs_c, bs_d, bs_e\rangle$	50%
$\langle bs_d, bs_e, bs_f\rangle$	20%
$\langle bs_d, bs_e, bs_h\rangle$	10%
$\langle bs_e, bs_f, bs_i\rangle$	10%

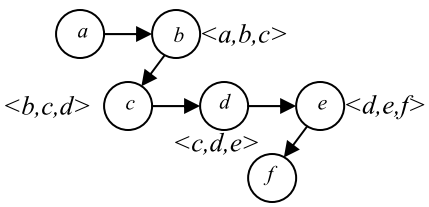

Fig. 1 Path graph

Table 3 Edge and link vectors from the decomposition of the MPMP $\langle bs_a, bs_b, bs_c, bs_d, bs_e, bs_f \rangle$

Time interval	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
T_1	1	1	1	1	1	1	1	1	1
T_4	1	1	1	1	1	1	1	1	1
T_7	1	1	1	1	1	1	**0**	1	1
T_{10}	1	1	1	1	1	1	1	1	1
T_{13}	1	1	1	1	1	1	1	1	1
T_{16}	1	1	1	1	1	1	1	1	1
T_{19}	1	1	1	1	1	1	1	1	1
T_{22}	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$T_{1+(\lceil (n-1)/3 \rceil - 1)\times 3}$	1	1	1	1	1	1	1	1	1

Note: V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, and V_9 represent $V_{\langle bs_a,bs_b \rangle}$, $V_{\langle bs_b,bs_c \rangle}$, $V_{\langle bs_c,bs_d \rangle}$, $V_{\langle bs_d,bs_e \rangle}$, $V_{\langle bs_e,bs_f \rangle}$, $V_{\langle bs_a,bs_b,bs_c \rangle}$, $V_{\langle bs_b,bs_c,bs_d \rangle}$, $V_{\langle bs_c,bs_d,bs_e \rangle}$, and $V_{\langle bs_d,bs_e,bs_f \rangle}$ respectively

vector V_{one}, then $\langle bs_a, bs_b \rangle$ has occurred for all time intervals in the period $\pi_{3,1}$. Continue by performing the AND operation on $V_{\langle bs_a,bs_b \rangle}$ and $V_{\langle bs_a,bs_b,bs_c \rangle}$ to calculate vector V_a; if $V_a = V_{one}$, then $\langle bs_a, bs_b, bs_c \rangle$ has the periodicity of $\pi_{3,1}$. Perform the AND operation again for V_a and $V_{\langle bs_b,bs_c,bs_d \rangle}$ to obtain V_b; if $V_b = V_{one}$, then $\langle bs_a, bs_b, bs_c, bs_d \rangle$ has the periodicity of $\pi_{3,1,}$ and so forth.

The first entry in the example in Table 3 is $V_{\langle bs_a,bs_b,bs_c \rangle} = V_{one}$, meaning that the period $\pi_{3,1}$ has the pattern $\langle bs_a, bs_b, bs_c \rangle$. The next entry has $V_{\langle bs_a,bs_b,bs_c \rangle} \& V_{\langle bs_b,bs_c,bs_d \rangle} \neq V_{one}$, implying that the period $\pi_{3,1}$ does not contain $\langle bs_a, bs_b, bs_c, bs_d \rangle$. Moving on, the next entry shows $V_{\langle bs_c,bs_d,bs_e \rangle} = V_{one}$, and therefore period $\pi_{3,1}$ contains $\langle bs_c, bs_d, bs_e \rangle$. Finally, because $V_{\langle bs_c,bs_d,bs_e \rangle} \& V_{\langle bs_d,bs_e,bs_f \rangle} = V_{one}$, the period $\pi_{3,1}$ must also contain the pattern $\langle bs_c, bs_d, bs_e, bs_f \rangle$. Therefore, two PMPMPs, namely $\langle bs_a, bs_b, bs_c \rangle$ and $\langle bs_c, bs_d, bs_e, bs_f \rangle$, can be identified for *Usr* in the period $\pi_{3,1}$.

4 Experimental results

Experiments were conducted to indicate the efficiency and effectiveness of the proposed sampling method. The parameter representation format is described as follows: Ta denotes the average movement path length of a; Bb represents a telecommunications network involving b base stations; Dc stands for a database of c movement paths; TId signifies that all movement paths are divided into d time intervals; $\pi_{e,f}$ is a mining period of e unit duration with an offset of f; Sg indicates that the minimum support is g.

The experiments were performed on an ASUS V6800V notebook PC equipped with an Intel Pentium M 1.73 GHz CPU, 1.25 GB of system memory and a 60 GB 5400 rpm hard disk. The OS installed on the system was Microsoft Windows XP Professional. The algorithm for PMPMP mining was coded in C++, and compiled by Microsoft Visual C++ Studio 6.0.

The experiment data of synthetic movement paths were created by path construction, as described by Lee and Wang [6], and were annotated with periodicity parameters, as described by Özden et al. [8]. Table 4 lists the parameters and their default values for the generation of movement paths.

The first set of experiments was undertaken to demonstrate the effectiveness of the proposed sampling technique. The experiment parameters were T5.0B100D1kTI10$\pi_{3,0}$S5%. The experiments compared two mining approaches: pattern matching and the proposed sampling scheme. The first stage of pattern matching was to mine the movement paths from every time interval (T_0, T_3, T_6 and T_9) individually, in order to derive the MPMPs as shown in Table 5. The second stage was to perform pattern comparison of the mined MPMPs for time intervals T_0, T_3, T_6 and T_9 to obtain two PMPMPs (namely $\langle 47, 57, 67, 57 \rangle$ and $\langle 56, 45, 56 \rangle$ marked bold in Table 5). In contrast, the proposed sampling algorithm identified the two PMPMPs directly by mining the movement paths from T_0 and using vector computations.

Figure 2 compares the execution time of the proposed sampling approach and the pattern-matching method. It can be noted that the execution time for the pattern matching shown in Fig. 2 only counts the time spent mining the MPMPs for time intervals T_0, T_3, T_6 and T_9 but does not include that spent matching the MPMPs in time intervals T_0, T_3, T_6 and T_9. According to Fig. 2, both methods had similar I/O times for accessing movement paths, since both techniques needed to access movement paths for time intervals T_0, T_3, T_6 and T_9. However, the sampling approach only required sample mining for the first time interval, while the pattern-matching method performed mining for every time interval. Naturally, the sampling method had higher execution efficiency than the pattern-matching method.

The aim of the second set of experiments was to validate the execution efficiency of the proposed sampling method on various minimum support settings. The experiment parameters were T5.0B1kD100kTI100$\pi_{3,0}$. Figure 3 depicts the experimental results. The minimum support was set to 0.1%,

Table 4 Parameters for generating movement paths

	Default value	Description
Path parameter		
D	100,000	Number of movement paths
T	5.0	Average length of movement paths
I	4.0	Average length of MPMP
L	50	Number of MPMP
B	1,000	Number of GSM base stations
TI	100	Number of time intervals
Periodicity parameter		
P_{num}	10	Average number of PMPMP
P_{min}	3	Minimum length of PMPMP
P_{max}	8	Maximum length of PMPMP
P_{den}	0.4	Average density of PMPMP

Table 5 MPMPs in each time interval during period $\pi_{3,0}$

T_0	T_3	T_6	T_9
⟨10, 11⟩	⟨1, 11⟩	⟨27, 17⟩	⟨6, 5⟩
⟨11, 0⟩	⟨11, 22⟩	⟨40, 51⟩	⟨11, 12⟩
⟨11, 10⟩	⟨29, 39⟩	⟨**47**, **57**, **67**, **57**⟩	⟨12, 22⟩
⟨22, 11⟩	⟨**47**, **57**, **67**, **57**⟩	⟨51, 61⟩	⟨27, 16⟩
⟨39, 29⟩	⟨53, 64⟩	⟨**56**, **45**, **56**⟩	⟨29, 18⟩
⟨**47**, **57**, **67**, **57**⟩	⟨**56**, **45**, **56**, 46⟩	⟨59, 58⟩	⟨36, 25⟩
⟨63, 64⟩	⟨57, 46⟩	⟨61, 60⟩	⟨**47**, **57**, **67**, **57**, 68⟩
⟨67, **56**, **45**, **56**⟩	⟨64, 63⟩	⟨68, 57⟩	⟨**56**, **45**, **56**⟩
⟨73, 63⟩	⟨71, 72⟩	⟨78, 89⟩	⟨85, 96⟩
⟨85, 75⟩	⟨72, 82⟩	⟨79, 78⟩	⟨95, 84⟩
⟨94, 95⟩	⟨77, 88⟩	⟨80, 69⟩	⟨96, 85⟩
⟨95, 96⟩	⟨81, 82⟩	⟨82, 81⟩	
	⟨82, 83⟩		
	⟨82, 93⟩		
	⟨84, 94⟩		
	⟨89, 90⟩		
	⟨90, 79⟩		
	⟨93, 92⟩		

leading to frequent movement paths in all time intervals. The proposed sampling method performed very well at the low minimum support settings.

The third set of experiments used various volumes of movement paths to test the scalability of the sampling method. The experiment parameters were T5.0B1kTI100$\pi_{3,0}$S0.1%, with the data volume increased from 100k to 600k. Figure 4 illustrates the experimental results. The linear rise in execution time indicates that the sampling technique indeed has decent scalability.

The experimental results adequately demonstrate that the proposed sampling mining approach has excellent execution performance and scalability. The sampling mining approach performs significantly better than traditional mining methods for mining PMPMPs, in three ways: (1) it does not repetitively scan the database while mining PMPMPs, (2) it needs only one sample mining process in a given period, and (3) it requires no pattern-matching operations. Since the sampling approach has an execution time directly proportional to the actual number of MPMPs identified in the mining process for a time interval, it has excellent execution performance.

5 Conclusions and future work

This study addresses mining PMPMP for the prediction of movement trends. The PMPMP is extremely useful information for telecommunications service providers, as the pat-

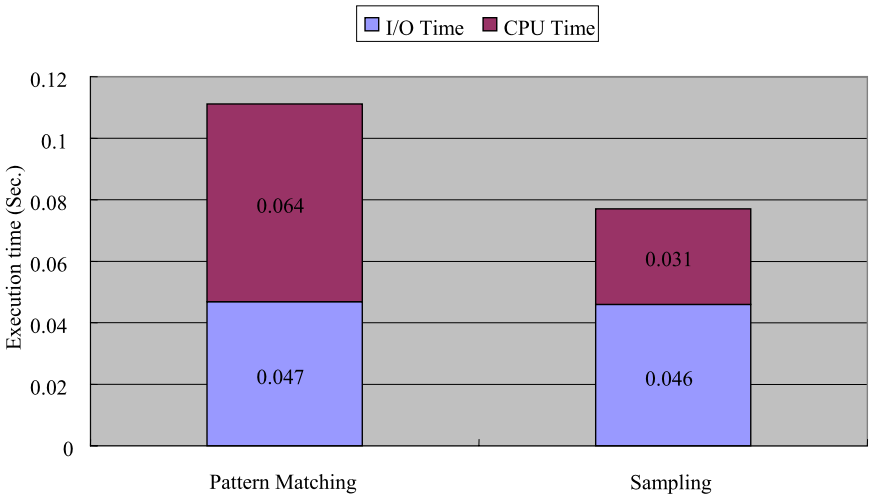

Fig. 2 Comparison of efficiency between sampling and pattern matching techniques

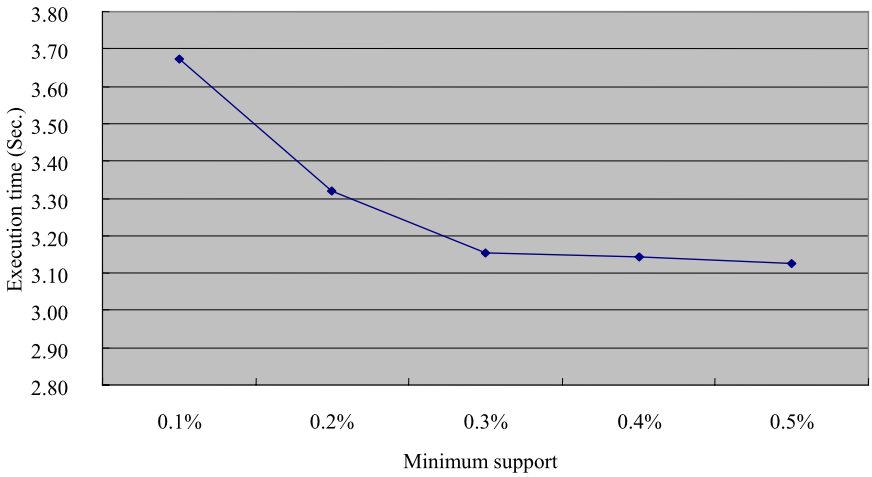

Fig. 3 Execution efficiency with various settings of minimum supports

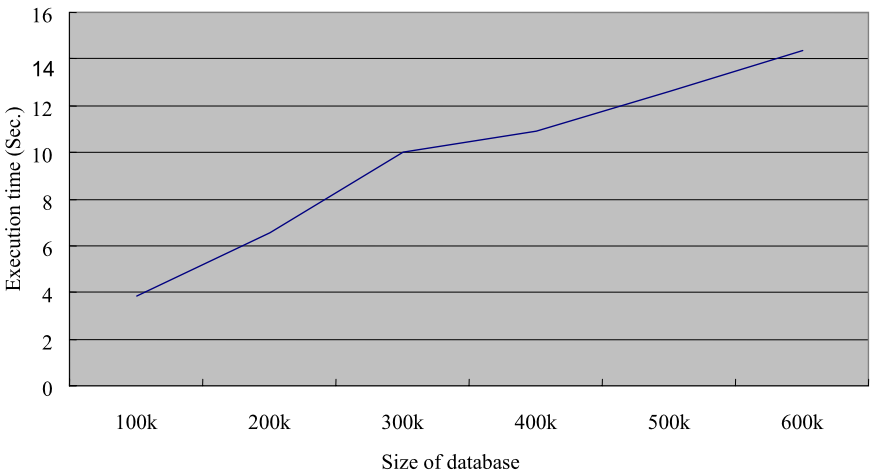

Fig. 4 Scalability verification for the sampling technique

tern may be used not only to enhance system operation performance, but also to provide quality value-added services. Therefore, this work develops a sampling algorithm applicable to the identification of PMPMP. The proposed sampling algorithm adopts two methods that improve operational efficiency. (1) A unique path graph that can track the movement of a user is created. The MPMP can thus be located by simply scanning, once, the movement paths recorded in the

first time interval of the period. (2) Vector computations are employed to prevent repetitive graph traversing and pattern matching. As the sampling algorithm has very high execution efficiency, it is a feasible solution that enables telecommunications operators to predict user movement trends in real time, and to dynamically allocate available resources to provide location-based personalized services.

The proposed scheme for mining PMPMP is applicable to many research domains, including the mining of periodic Web navigation patterns, the scheduling of periodic transportation routes and security portfolio management [3, 13]. The efficient sampling mining approach devised herein has the potential to be utilized to solve the problems of periodic sequential pattern mining in other research fields.

Partial periodic movement pattern mining is also an interesting issue. Future work will consider the mining of partial PMPMP by using the proposed sampling method.

Acknowledgements The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 92-2416-H-126-012-.

References

1. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the 11th international conference on data engineering, pp 3–14
2. Anwar F, Petrounias I, Kodogiannis VS, Tasseva V, Peneva D (2008) Efficient periodicity mining of sequential patterns in a post-mining environment. In: Proceedings of the 4th international conference on intelligent systems, pp 16-2 to 16-11
3. Bao D (2008) A generalized model for financial time series representation and prediction. Appl Intell 29:1–11
4. Eberspächer J, Vögel HJ (1999) GSM: switching, services and protocols. Wiley, Chichester
5. Elfeky MG, Aref WG, Elmagarmid AK (2005) Periodicity detection in time series databases. IEEE Trans Knowl Data Eng 17(7):875–887
6. Lee AJT, Wang YT (2003) Efficient data mining for calling path patterns in GSM networks. Inf Syst 28:929–948
7. Lee CH (2007) IMSP: an information theoretic approach for multi-dimensional sequential pattern mining. Appl Intell 26:231–242
8. Özden B, Ramaswamy S, Silberschatz A (1998) Cyclic association rules. In: Proceedings of the 14th international conference on data engineering, pp 412–421
9. Peng WC, Chen MS (2003) Developing data allocation schemes by incremental mining of user moving patterns in a mobile computing system. IEEE Trans Knowl Data Eng 15(1):70–85
10. Tseng VS, Lin KW (2006) Efficient mining and prediction of user behavior patterns in mobile web systems. Inf Softw Technol 48:357–369
11. Velez FJ, Correia LM (2002) Mobile broadband services: classification, characterization, and deployment scenarios. IEEE Commun Mag 40(4):142–150
12. Yavaş G, Katsaros D, Ulusoy Ö, Manolopoulos Y (2005) A data mining approach for location prediction in mobile environments. Data Knowl Eng 54:121–146
13. Zhou B, Hui SC, Fong ACM (2006) An effective approach for periodic Web personalization. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence, pp 284–292

Yao-Te Wang was born on 2 January 1968 in Taiwan ROC He received the B.S. degree from Tatung University, Taiwan, in 1990, and the M.S. degree from Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, in 1992. He got the Ph.D. degree from Department of Information Management, National Taiwan University in May 2003. In August 2003, he joined the Department of Computer Science and Information Management at Providence University and he is now an assistant professor. His current research interests include information retrieval, data mining, and grid computing.

Ju-Tzu Cheng was born on 7 June 1970 in Taiwan ROC She received the B.S. and MBA degree in accounting from National Taiwan University in 1992 and 1994 respectively and the Ph.D. degree from Department of Accounting, National Chengchi University, Taiwan, in 2000. Dr. Cheng is currently an associate professor in the Department of Accounting Information, National Taichung Institute of Technology, Taiwan. Her research interests include government accounting, accounting information systems, and financial data mining.