

# 时间序列移动接入模式的挖掘研究<sup>①</sup>

## Mining Research of Temporal Sequence Mobile Access Pattern

李海强 吕廷杰 (北京邮电大学经济管理学院 北京 100876)

**摘要:** 通过对移动用户的位置、时间和业务需求信息进行组合分析,设计了一种时间序列移动接入模式的挖掘方法;它可以有效的挖掘出时间频繁序列移动接入模式,有利于业务提供商在相应的时间和位置重点推广对其敏感的业务;该挖掘方法采用了简洁表头的形式储存位置和业务,数据库可以对其进行快速扫描,降低了内存的使用并提高了效率。

**关键词:** 时间序列 频繁序列 移动接入模式 数据挖掘 时间敏感性

### 1 引言

在环境感知泛在网络 AUN (AmBint Ubiquity Network) 的生态环境中,网络就如同空气和水一样,自然而深刻地融入了人们的日常生活及工作中。网络不再被动地满足用户需求,而是主动感知用户场景的变化并进行信息交互,通过分析人的个性化需求主动提供服务。相应地,终端设备具备智能型接口及环境感知能力,使人们使用起来更加简单和方便。无线网络的发展允许移动用户在任何时间任何地点都能得到感兴趣的移动业务,网络能主动感知移动用户场景的变化,从而提供将要需要的业务及信息。因此,如何帮助用户获取感兴趣的业务信息是一件急需解决的问题。由于移动用户的行为模式与所处的位置和相应的业务有密切的关系,这就需要研究移动用户的序列接入模式。

目前,一些国际上的学术研究团队已经开始研究在宽带通信网络环境下的移动用户的序列接入模式,但他们大多是只侧重在与移动用户行为有关的一个或者两个属性来研究<sup>[1]</sup>,比如:移动用户位置和相应的需求业务,而缺少对时间属性的综合考虑。本文提出了一种新颖的基于时间的移动序列接入模式 T-SMAP (Temporal Sequence Mobile Access Pattern) 挖掘方法,综合考虑了位置业务的时间敏感性和用户需求性,并分析了连续时间内的频繁移动序列模式特点,使移动用户能更好的在当前位置体验感兴趣的移动业务。本

文主要阐述了时间序列移动接入模式的挖掘方法,第二部分描述了时间序列移动接入模式挖掘的系统架构,第三部分分析了时间序列移动接入模式的挖掘方法,最后一部分是结论。

### 2 T-SMAP 挖掘的系统架构

在这一部分,我们首先提出了 T-SMAP 挖掘的系统架构,并说明了各种场景信息经过 T-SMAP 挖掘系统输出移动用户感兴趣的业务,并推荐给终端用户。

最近,很多研究机构对基于位置和业务需求相关的序列移动接入模式进行了研究,但是很少考虑业务的时间敏感性,而 T-SMAP 是一种考虑时间敏感性的序列移动接入模式挖掘系统,对移动位置、需求业务和时间信息进行数据融合,充分考虑需求业务在相应位置的时间敏感性。T-SMAP 首先将相应的频繁序列移动接入模式 F-SMAP (Frequent Sequence Mobile Access Pattern) 挖掘出来,然后对其进行时间敏感性分析,发现相应的时间频繁序列移动接入模式 T-F-SMAP (Temporal Frequent Sequence Mobile Access Pattern), T-F-SMAP 的挖掘顺序为: SMAP, F-SMAP, T-F-SMAP。

这样业务提供商就可以在业务敏感的时间段重点推广该业务。图 1 给出了 T-SMAP 挖掘的系统架构。

① 基金项目:信息产业部基金项目(2007-R-152)

### 3 T-SMAP 的挖掘方法

在这一部分,本文首先给出了时间关联的序列移动接入模式挖掘的定义<sup>[2]</sup>,并提出了一种新颖的数据挖掘方法:T-SMAP Mine,它可以有效的发现与时间相关的频繁序列移动接入模式,该模式与位置和需求业务的时间敏感性相关。

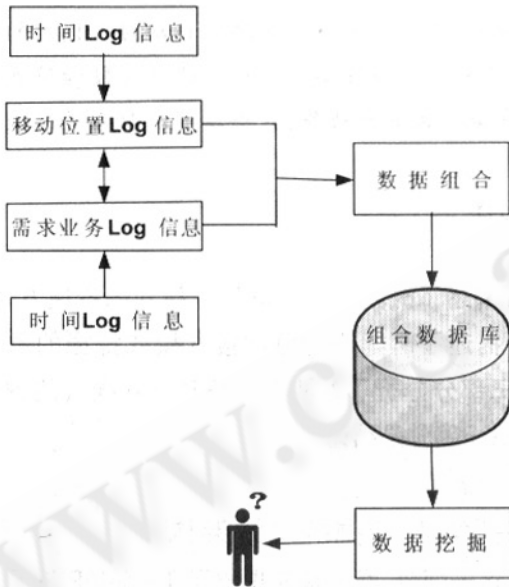


图1 T-SMAP 挖掘的系统架构

#### 3.1 基本定义

在组合数据库里储存与时间相关的各种数据,其中包括与序列移动接入模式相关的三个属性:用户位置、用户的需求业务和时间<sup>[3]</sup>,它们分别对应位置集合、需求业务集合和时间集合。假设  $p = (l, s, t)$ , 其中  $p$  是移动序列接入模式集合  $M$  的元素,  $l$  是位置集合  $L$  的元素,  $s$  是需求业务集合  $S$  的元素,  $t$  是时间集合  $T$  的元素,则移动接入模式可表示为:

$$P = L \times S \times T = \{ (l, s, t) \mid l \in L, s \in S, t \in T \} \quad (1)$$

则移动序列接入模式可以表示为  $M = \langle p_1, p_2, p_3, \dots, p_n \rangle$ , 其中  $p_i = (l_i, s_i, t_i)$ ,  $n$  是移动序列接入模式的维数。另外,假设  $M' = \langle p'_1, p'_2, p'_3, \dots, p'_m \rangle$ , 并且  $m < n$ , 则有如下,

$$\langle p'_1, p'_2, p'_3, \dots, p'_m \rangle \subset \langle p_1, p_2, p_3, \dots, p_n \rangle$$

即  $M' \subset M \quad (2)$

定义1 假设数据库  $D = \langle M_{s1}, M_{s2}, \dots, M_{sN} \rangle$ , 其中包括种移动接入模式, 则接入模式  $M_s$  归一化之后的

支持度定义为:

$$\text{sup}(M_s) = \frac{|\{M_{s_i} \mid M_s \subset M_{s_i}, 1 \leq i \leq N\}|}{N} \quad (3)$$

定义2 假设支持度  $\text{sup}(M_s)$  大于用户本身定义的阈值  $\delta$ , 则移动接入序列  $M_s$  被称为频繁序列移动接入模式(F-SMAP)。

由上面的定义可知,如果数据库包含移动用户接入模式的 log 信息和用户定义的阈值  $\delta$ , 那么就可以在该数据库挖掘出频繁序列移动接入模式 F-SMAPS, 然后根据 F-SMAPS 在时间域的分布特性, 过滤出对时间敏感的序列移动接入模式 T-SMAPS, 这样业务提供商就可以在一定的时间段内重点推广相应位置的数据业务。

#### 3.2 T-SMAP 的具体挖掘方法

本文涉及到的与时间序列移动接入模式相关的时间戳见表1, 移动用户在相应的时间  $t_i$  及位置  $l_i$  激活所需求的移动业务  $s_i$ , 形成了移动接入模式  $p_i = (l_i, s_i, t_i)$ 。移动用户在不同位置 and 不同时间激活了相应的移动业务, 这样在组合数据库  $D$  中记录了移动序列接入模式  $M$ 。

表1 移动接入模式所用的时间戳

序号	时间戳	序号	时间戳
$t_1$	07:35:40	$t_8$	11:25:30
$t_2$	08:20:35	$t_9$	12:35:35
$t_3$	09:30:30	$t_{10}$	12:45:45
$t_4$	09:40:30	$t_{11}$	13:10:23
$t_5$	09:55:45	$t_{12}$	13:12:20
$t_6$	10:20:35	$t_{13}$	14:10:20
$t_7$	10:30:35		

移动业务提供商要想更好的推广自己的业务, 需要知道移动用户在什么位置、什么时间需要什么样的业务。为了有效解决这一问题, 本文设计了一种频繁序列移动接入模式的挖掘方法 F-SMAP, 并在时间域内对频繁序列移动接入模式 F-SMAP 进行分析<sup>[4]</sup>, 根据业务的时间敏感性, 设计出一种时间频繁序列移动接入模式 T-F-SMAP 的挖掘方法, 这样移动业务提供商就可以在相应位置的某一段时间内重点推广用户感兴趣的移动业务。对时间频繁序列移动接入模式 T-F-SMAP 的挖掘方法, 可以分为以下三部分: 第一,

对序列移动接入模式树 SMAP - Tree 的构建; 第二, 在组合数据库 D 对频繁序列移动接入模式 F - SMAP 的挖掘; 第三, 针对 F - SMAPS 的时间敏感性分析。

### 3.2.1 SMAP - Tree 的构建

构建 SMAP - Tree 的主要目的是将移动序列接入模式以简洁的表格形式储存于数据库<sup>[5]</sup>, 以便更有效的挖掘频繁序列移动接入模式 F - SMAP 和时间频繁序列移动接入模式 T - F - SMAP, 下面以组合数据库 D 中的移动序列组合 log 信息表 2 为例来说明该挖掘方法;

表 2 序列移动接入模式 log 信息

用户序号	接入模式
100	$\langle (e, 1, t_1) (f, 2, t_2) (g, 5, t_9) (h, 8, t_{11}) \rangle$
101	$\langle (e, 1, t_2) (f, 3, t_7) (g, 5, t_{10}) (h, 8, t_{12}) \rangle$
102	$\langle (e, 3, t_1) (f, 2, t_3) (h, 7, t_5) \rangle$
103	$\langle (g, 6, t_2) (f, 2, t_4) (h, 7, t_6) \rangle$
104	$\langle (g, 8, t_8) (f, 1, t_{10}) \rangle$
105	$\langle (e, 3, t_1) (f, 6, t_8) (g, 8, t_{12}) (h, 7, t_{15}) \rangle$

表 2 中, 以用户 100 的移动序列接入模式  $\langle (e, 1, t_1) (f, 2, t_2) (g, 5, t_9) (h, 8, t_{11}) \rangle$  为例来说明 SMAP - Tree 的构建, 首先从  $\langle (e, 1, t_1) (f, 2, t_2) (g, 5, t_9) (h, 8, t_{11}) \rangle$  中提取移动序列  $\langle e, f, g, h \rangle$ , 然后插入 SMAP - Tree, 移动序列中每个节点的计数在 SMAP - Tree 中自动加 1; 在 SMAP - Tree 中, 移动序列  $\langle e, f, g, h \rangle$  的尾节点构建需求业务树 SR - Tree (Service Request Tree), 用来记录在相同移动序列下的不同业务需求, 因而在每个移动序列的尾节点都将有一个业务需求树 SR - Tree; 依据该方法我们从  $\langle (e, 1, t_1) (f, 2, t_2) (g, 5, t_9) (h, 8, t_{11}) \rangle$  中提取需求业务  $\langle 1, 2, 5, 8 \rangle$ , 并将它插入相应的 SR - Tree, 如图 2 所示, 给出 SMAP - Tree 的构建原理。

### 3.2.2 F - SMAP 的挖掘

我们使用 SMAP - Tree 来说明 T - SMAP 的挖掘方法, 如图 2 所示, 假设用户规定的支持度阈值是 30%, 则只有当每个位置和业务节点的计数大于或等于 2 时才能符合要求。

通过扫描图 3 的节点结构, 我们发现移动序列表头中的所有节点计数都满足最小的支持度阈值 2, 我们随机的选择移动序列表头中的 h 节点来构建

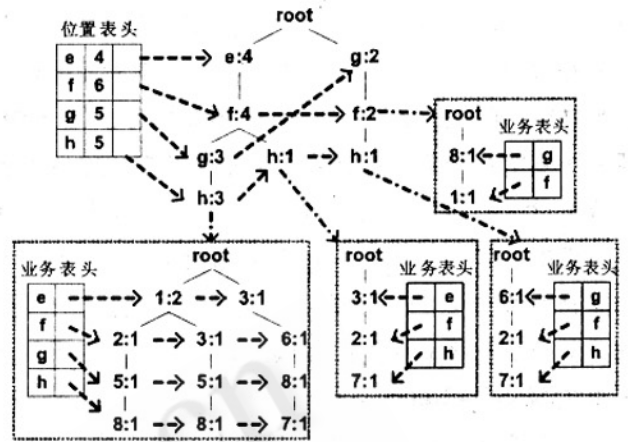


图 2 SMAP - Tree 的构建

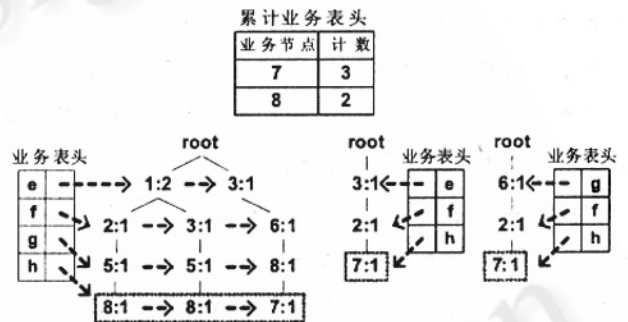


图 3 组合接入模式 (h, \*, \*) 的 SR - Tree

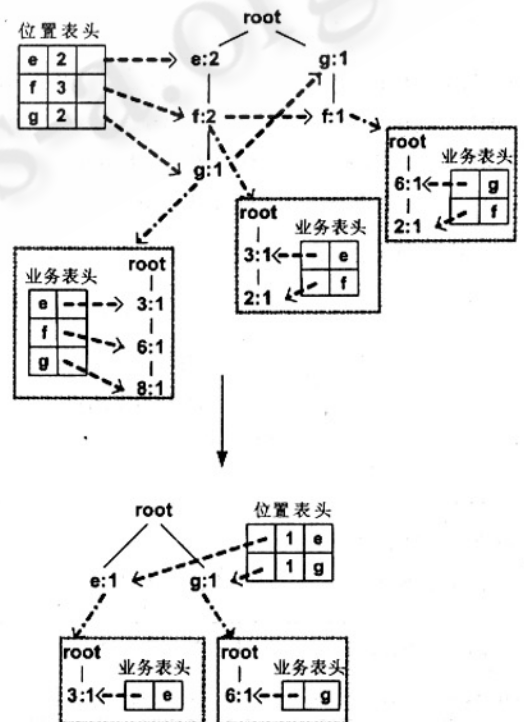


图 4 组合接入模式 (h, 7, \*) 的 SMAP - Tree 重构

(h, \*, \*) 的 SR-Tree。通过扫描 SMAP-Tree 中包含的移动序列各个节点,将这些节点的需求业务进行计数,并构建相应的 SR-Tree,如图 3 所示。图 3 中,SR-Tree 的每个节点都记录了需求业务在移动序列接入模式(h, \*, \*)中的计数,我们由此可以得到业务尾节点的业务累计表。在该业务累计表中,我们将小于支持度阈值 2 的业务节点序列删除,并把符合要求的业务节点和移动序列接入模式(h, \*, \*)进行组合来重构 SMAP-Tree。

如图 4 所示,箭头上是组合前缀接入模式(h, 7, \*)的 SMAP-Tree 重构例子,对图 4 中的最新业务累计表进行扫描,将满足支持度阈值 2 的移动接入模式(f, 2, \*)选出,通过使用(h, 7, \*) (f, 2, \*)作为前缀接入模式重构 SMAP-Tree。图 4 中,箭头下方由于最新业务累计表的计数都小于 2,所以作为频繁序列移动接入模式 F-SMAP 而被选出。

我们对以上处理流程进行递归处理,针对位置和需求业务的频繁程度,直到所有的 F-SMAPS 是被选出,表 3 列出了所有已发现的 F-SMAPS。

表 3 F-SMAPS

用户	F-SMAPS	SMAPS
100	$\langle (e, 1, *) (h, 8, *) \rangle$ ; $\langle (e, 1, *) (g, 5, *) \rangle$ ; $\langle (e, 1, *) (h, 8, *) \rangle$ ; $\langle (g, 5, *) (h, 8, *) \rangle$	$\langle (e, 1, t_1) (g, 5, t_9) (h, 8, t_{11}) \rangle$ ; $\langle (e, 1, t_1) (g, 5, t_9) \rangle$ ; $\langle (e, 1, t_1) (h, 8, t_{11}) \rangle$ ; $\langle (g, 5, t_9) (h, 8, t_{11}) \rangle$
101		$\langle (e, 1, t_2) (g, 5, t_{10}) (h, 8, t_{12}) \rangle$ ; $\langle (e, 1, t_2) (g, 5, t_{10}) \rangle$ ; $\langle (e, 1, t_2) (h, 8, t_{12}) \rangle$ ; $\langle (g, 5, t_{10}) (h, 8, t_{12}) \rangle$
102	$\langle (f, 2, *) (h, 8, t_{11}) \rangle$ ; $\langle (e, 3, *) (h, 7, *) \rangle$	$\langle (f, 2, t_3) (h, 7, t_5) \rangle$ ; $\langle (e, 3, t_1) (h, 7, t_5) \rangle$ ;
103	$\langle (f, 2, *) (h, 7, *) \rangle$	$\langle (f, 2, t_4) (h, 7, t_6) \rangle$
105	$\langle (e, 3, *) (h, 7, *) \rangle$	$\langle (e, 3, t_1) (h, 7, t_{15}) \rangle$

### 3.2.3 F-SMAPS 的时间敏感性分析

针对位置和业务的时间敏感性,业务提供商如何在最佳的时间段内重点推出移动用户感兴趣的业务已经成为业务提供商需要解决的一个主要问题<sup>[6]</sup>,这就需要我们对表 3 的 F-SMAPS 进行时间敏感性分析,

定义 3 在 F-SMAPS 中,如果一个频繁序列移动接入模式 F-SMAP 中的  $(l, s, t_i)$  之间的时间间隔在规定的时间内,则该 F-SMAP 被认为是在一定时间段内具有时间敏感性,即时间频繁序列移动接入模式(T-F-SMAP)。

对表 3 中的 F-SMAP 进行分析,假设规定的时间阈值是 1 小时,如图 5 所示,  $t_3, t_4, t_5, t_6$  的时间取值在时间范围 [9:30:00, 10:30:00] 内,而  $t_9, t_{10}, t_{11}, t_{12}$  的时间取值在时间范围 [12:30:00, 13:30:00] 内,并都在时间阈值  $\delta_t$  内分布,则  $(f, 2, *) (h, 7, *)$  和  $(g, 5, *) (h, 8, *)$  是在这两个时间段内的频繁移动序列接入模式 T-F-SMAP,因此业务提供商可以在 [9:30:00, 10:30:00] 和 [12:30:00, 13:30:00] 两个时间段内重点关注这两个移动序列接入模式。

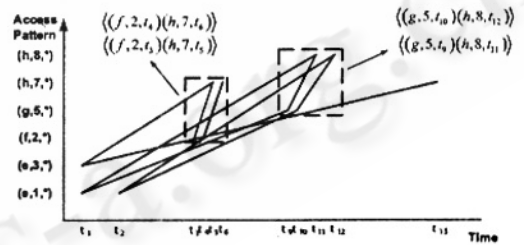


图 5 F-SMAPS 的时间敏感性分析

## 4 结论

在本文中,我们提出了一种新颖的数据挖掘方法:时间序列移动接入模式挖掘,它能有效的发现与位置和需求业务有关的移动用户时间行为模式;它的优点是能有效的按照位置和业务信息储存用户的行为模式,由于使用了简洁的数据结构表,所以该数据结构可以占用很少的存储空间;该挖掘方法可以有效的处理基于场景感知的数据信息。我们下一步将研究该挖掘方法的预测算法以及如何评估该方法的性能。

(下转第 14 页)

### 参考文献

- 1 Vincent S. Tseng, Kawuu W. Lin. Efficient Mining and prediction of user behavior patterns in moBie web systems. Information and Software Technology, 2005.
- 2 R. Agrawal and R. Srikant. Mining sequential Patterns. 11th International Conference on Data Engineering, TAiei, TAian, Mar. 1995.
- 3 S. Tseng, C. F. Tsui. An efficient method for mining associated service patterns in moBie web environments. Proceedings of the ACM Symposium on Applied Computing, 2003. 455 - 459.
- 4 Jun Wook Lee, Ok Hyun Paek, Keun Ho Ryu. Temporal moving pattern mining for location - based service. The Journal of Systems and Software, 2004.
- 5 王丽珍,周丽华,陈红梅. 数据仓库与数据挖掘原理及应用[M]. 北京: 科学出版社, 2005. 191 - 194.
- 6 Wang Lizheng, Zhou Lihua, Chen Hongmei. The principle and application for data warehouse and data mining[M]. Beijing: Science Press, 2005. 191 - 194.
- 7 Jordi Docter, Carlo Alberto Licciardi, Marco Marchetti. The telecom Industry and Context Awareness. Sixth International Conference on the Management of Mobile Business, 2007.