2009年10月23-25日
中国 武汉

信息系统协会中国分会第三届学术年会
The Third CNAIS National Congress, CNAIS2009

中国信息系统研究：新兴技术背景下的机遇与挑战

（下册）

陈国青 马费成 主编

Information Systems Research in China: Opportunities and Challenges in the Context of Emerging Technologies

WUHAN UNIVERSITY PRESS
武汉大学出版社

2009年10月23-25日
中国 武汉

信息系统协会中国分会第三届学术年会
The Third CNAIS National Congress, CNAIS2009

中国信息系统研究：新兴技术背景下的机遇与挑战

陈国青 马费成 主编

（下册）

Information Systems Research in China: Opportunities and Challenges in the Context of Emerging Technologies

WUHAN UNIVERSITY PRESS
武汉大学出版社

利用频繁事件组挖掘用户上网行为模式*

周文志[1]，刘红岩[1]

（1. 清华大学经济管理学院，北京 100084）

摘 要：研究不同用户群上网行为的异同模式具有多方面的实际意义。频繁事件组挖掘常用于发现潜藏在事件序列中的频繁模式。本文利用该技术来研究某一类用户的上网特点，设计了一种基于最小发生域的频繁事件组挖掘算法 MOFE，并结合最小发生域的特性，提出了若干优化方法。MOFE 同时适用于连续时间和离散时间的事件序列，且不会出现重复计数问题。应用 MOFE 挖掘老、中、青三类用户的上网习惯，发现了许多有趣的行为模式。MOFE 也可用于其他具有连续或离散时间事件序列背景的实际应用研究。

关键词：事件序列；频繁事件组；最小发生域；上网模式

中图分类号：TP182

在数据挖掘应用中，实际收集到的数据往往具有内在的有序性，一般可以抽象为时间上具有先后顺序的事件序列。事件组是指序列中满足一定偏序关系、在一定时间内发生的一组事件。挖掘频繁事件组（frequent episode）就是从输入序列中发现经常发生的事件组（episode）。通过频繁事件组可以导出序列中的潜在规则，帮助描述和预测该序列的行为模式。

频繁事件组的概念最早由 Mannila 等研究者提出[1]，随后有一些相关的研究[2][3][4][5][6]。Mannila 等人提出了基于滑动时间窗的挖掘算法 WINEPI，但该方法存在一定的缺陷。首先，事件组 α 的同一个实例可能出现在多个连续的时间窗中，由此引发重复计数问题，使得部分实际发生次数较少的事件组被不恰当的评估为频繁事件组。其次，给定基本时间单位，WINEPI 算法的设计模式只适用于时间严格连续的输入事件序列，对实际应用中更常见的离散时间事件序列，通常需要先进行转换。另外，WINEPI 算法在每次迭代生成频繁事件组的过程中，都需要扫描整个输入序列，算法效率显著受到数据总量的影响。

本文将要探讨不同网络用户群体上网行为的相同和不同之处。通过对用户上网模式的分析，网站经营者可以更好地掌握客户需求，更合理地设计页面布局，更有效地给每个用户提供个性化链接推荐服务，进而提高网站的知名度和访问流量，为网站带来更多的商业契机。实际用户的上网活动往往具有周期性和集中性的特点，单个用户访问网站的数据流通常属于离散时间事件序列，采用 WINEPI 算法来挖掘频繁的上网模式将会导致诸多不便。

Mannila 等研究者通过引进最小发生域（minimal occurrence）的概念，提出了基于最小发生域发现频繁事件组的基本设想（算法称为 MINEPI 算法）[1]，并给出了算法框架[2]，希望可以解决 WINEPI 的上述不足，但没有给出该算法的具体实现方法。本文通过借鉴最小发生域的基本概念，设计并实现了从离散时间事件序列中发现频繁事件组的新算法（我们称之为 MOFE），并将该算法应用于网络用户上网行为的实际研究中，发现了一些有意义的模式。

1 模型描述和基本性质

给定事件类型集合 E，一个离散时间事件序列可以表示成一个二元组(S, T)。其中：

$S=\{(s_i, i) \mid i=1,2,\cdots,N; s_i=\{A_{i1}, A_{i2}, \cdots, A_{ik}\}, A_{ij} \in E\}$；$T=\{(t_i, i) \mid i=1,2,\cdots,N; t_i < t_{i+1} (i=1,2,\cdots,N-1)\}$；$N$ 为实际有事件发生的时间点个数。

事件为事件类型—时间对(A_{ij}, t_i)。

本文所研究的事件组为串行事件组[1]，即所有事件都有严格的时间先后顺序的事件组（下文统

* 基金项目：国家自然科学基金项目（70871068，70621061，70890083）

通讯作者：刘红岩，liuhy@sem.tsinghua.edu.cn

一简称为事件组)。一个串行事件组 α 可用如下一个序列表示：$\langle A_1, A_2, \cdots, A_n\rangle$。例如，事件组$\langle A, B\rangle$和$\langle B, A\rangle$是两个不同的串行事件组。在$\langle A, B\rangle$中，事件 B 在事件 A 之后发生，而在$\langle B, A\rangle$中，事件 B 则发生在事件 A 之前。

下面是设计 MOFE 算法将要使用的一些重要的定义和性质。

定义 1[1] 如果一个时间区间$[t_s, t_e)$满足：(1)在$[t_s, t_e)$内，包含事件组 α 的一个实例；(2)不存在任何一个更小的时间区间$[t'_s, t'_e) \subset [t_s, t_e)$(即$t_s \leqslant t'_s$，$t'_e \leqslant t_e$且$t'_e - t'_s < t_e - t_s$)，包含该实例；则称$[t_s, t_e)$是 α 的一个最小发生域(minimal occurrence)。其中 t_s、t_e 分别为最小发生域的起始时间、结束时间。$t_e - t_s$ 为最小发生域的时间跨度。

由定义 1 可知，(1)事件组 α 在序列(S, T)中的每一个实例都对应于一个最小发生域；(2)事件组 α 的每一个最小发生域都精确记录了某个实例发生的具体时间，同时还确定了该实例首、尾两事件发生的具体时间点。

引理 1 如果$[t_s, t_e)$是事件组 $\alpha = \langle A_1, A_2, \cdots, A_n\rangle$的一个最小发生域，那么在输入事件序列$(S, T)$中，必定存在$(A_1, t_s)$和$(A_n, t_e - 1)$两个事件。

一个事件组在事件序列中往往存在多个实例，这些实例之间具有先后关系，先发生的实例也必然先结束。结合事件组实例和最小发生域之间的一一对应关系，可得如下引理。

引理 2 假设时间区间$[t_s, t_e)$、$[u_s, u_e)$是事件组 α 的两个最小发生域，如果 $t_s < u_s$ 则 $t_e < u_e$；反之亦然。

给定事件序列(S, T)和时间窗阈值 win，时间跨度在时间窗阈值范围内$(t_e - t_s \leqslant \text{win})$的最小发生域称为事件组 α 的**有效最小发生域**。下面是 α 在序列(S, T)中的**有效最小发生域有序集合**的定义。

定义 2 给定事件序列(S, T)和时间窗阈值 win，如果事件组 α 的某个最小发生域集合具备如下三个性质：

(1)有效性，集合中所有元素均为 α 的有效最小发生域；

(2)完备性，集合中包含 α 在序列(S, T)中的所有有效最小发生域；

(3)有序性，集合中所有元素按照起始时间升序(或降序)排列。

则该集合为事件组 α 在事件序列(S, T)中的有效最小发生域有序集合(下文简称为最小发生域集合)，记做 $\text{mo}(\alpha)$。

显然，事件组 α 最小发生域集合的大小——$|\text{mo}(\alpha)|$，即为该事件组在序列(S, T)中发生的次数，可以作为衡量 α 的出现频率的指标。基于以上的性质我们设计了 MOFE 算法。

2 MOFE 算法

MOFE 算法的基本思路仍然是先发现长度为 k 的频繁事件组，然后根据长度为 k 的频繁事件组生成长度为 $k+1$ 的候选事件组，再通过特定的评估方式判断这些事件组是否频繁，进而得到长度为 $k+1$ 的频繁事件组，依次循环迭代，直至没有更长的频繁事件组生成为止。借助最小发生域集合，MOFE 算法在多次迭代过程中，总共只需要扫描一次原始输入序列，就可将其中所有有价值的事件信息完整地保存并传递下去。整个挖掘过程可以分为三个阶段——生成 F_1(频繁事件组集合，集合中每个频繁事件组的长度都为 1；生成候选事件组和评估候选事件组，分别实现了读入原始信息、传递有效信息和使用频繁信息的功能。

MOFE 算法

输入：事件序列(S, T)，时间窗阈值 win，绝对频率阈值 min_fr；

输出：所有频繁事件组的总集合 F；

过程：

1. 遍历输入事件序列 (S, T)，生成 F_1 和 MO_1；
2. while $(F_k != \phi)$ {
3. 调用 GenCandidate 函数；生成候选事件组集合 C_{k+1} 以及相应的 CMO_{k+1}；
4. 评估 $\alpha \in C_{k+1}$ 是否频繁；生成频繁事件组集合 F_{k+1} 和相应的 MO_{k+1}；
5. }

/* MO_i、CMO_i 均为 mo (α) 的集合，其中 α 的长度为 i */

2.1 生成 F_1

在这一阶段，MOFE 算法通过遍历输入事件序列，将事件 (A_i, t_i) 中的原始信息以 A_i 的最小发生域 $[t_i, t_{i+1})$ 的形式存储在集合 mo (A_i) 中；然后根据 | mo (A_i) | 的大小，判断 A_i 是否频繁，最终生成 F_1。

2.2 生成候选事件组

MOFE 算法通过最小发生域集合来保存和传递

有效信息，因此要求在产生候选事件组后，立即生成与之相应的最小发生域集合。前者用来确定信息传递的对象，后者完成信息传递的具体过程。

MOFE 采用与 WINEPI 相同的候选事件组生成方法。首先，通过组合同一数据块中的两个频繁事件组产生可能的候选事件组（对长度为 n 的两个频繁事件组，如果它们的前 $n-1$ 个事件相同，则认为它们在同一个数据块中），再根据文后参考文献［1］中的引理 1“任何频繁事件组的子事件组必定是频繁的”，来进一步确定可能的候选事件组。

MOFE 算法的核心是生成候选事件组的最小发生域集合。Mannila 等研究者分析了目标事件组（即长度为 n 的候选事件组）的最小发生域和它的子事件组（长度为 $n-1$）的最小发生域之间的关系，并由此提出了生成目标事件组的最小发生域的方法[1][2]。在此基础上，本文提出以下由两个特殊子事件组的最小发生域集合生成目标事件组最小发生域集合的方法。

假设目标事件组 $\alpha = \langle A_1, A_2, \cdots, A_n \rangle$，$\alpha_1$ 和 α_2 分别是 α 的两个特殊子事件组。其中 $\alpha_1 = \langle A_1, A_2, \cdots, A_{n-1} \rangle$，没有 α 的最后一个元素；$\alpha_2 = \langle A_2, \cdots, A_{n-1}, A_n \rangle$，没有 α 的第一个元素。mo（α_1）和 mo（α_2）分别是 α_1、α_2 的最小发生域集合。mo（α_1）［i］= ［t_s, t_e）表示 α_1 的第 i 个最小发生域，mo（α_2）［j］= ［u_s, u_e）表示 α_2 的第 j 个最小发生域，其中 $t_s < u_s$。

如果不存在任何 $k_i > i$ 使得 mo（α_1）［k_i］= ［t'_s, t'_e）且 $t'_s < u_s$，则称［t_s, t_e）是 α_1 在 u_s 之前“**最晚**”发生的实例。如果不存在任何 $k_j < j$ 使得 mo（α_2）［k_j］= ［u'_s, u'_e）且 $u'_s > t_s$，则称［u_s, u_e）是 α_2 在 t_s 之后“**最早**”发生的实例。如果［t_s, t_e）是 α_1 在 u_s 之前“最晚”发生的实例，同时［u_s, u_e）是 α_2 在 t_s 之后“最早”发生的实例，则称［t_s, t_e）和［u_s, u_e）为 α_1、α_2 的一组**相邻发生域**。

引理 3 由 α_1、α_2 的一组相邻发生域组合得到的时间区间［t_s, u_e）是 α 的一个最小发生域。

因篇幅所限，证明略。

函数 GenMO 通过交叉遍历 α_1、α_2 的最小发生域集合，寻找并组合所有可能的相邻发生域，生成了事件组 α 的一个有效最小发生域的集合 mo′（α）。

GenMO 函数

作用：生成目标事件组 α 的最小发生域集合

输入：最小发生域集合 mo（α_1）、mo（α_2），时间窗窗口阈值 win；

输出：目标事件组 α 的最小发生域集合 mo（α）；

过程：

```
1. i=1, j=1;
2. ω1 = mo(α1)[i]; ω2 = mo(α2)[j];
3. do{
4.    while(ω1.ts ≥ ω2.ts) do
        /*固定当前 α1 的最小发生域[t's, t'e)，在 mo(α2)中找出第一个有可能组成相邻发生域的时间区间，确定α2 在 t's 之后"最早"发生的实例[us, ue)*/
5.      j = j+1;
6.      if (j ≤ |mo(α2)|) then do ω2 = mo(α2)[j];
7.      else break to line 16
8.    while(ω1.te < ω2.ts) do
        /*固定当前 α2 的最小发生域[us, ue)，在 mo(α1)中找出 α1 在 us 之前"最晚"发生的实例[ts, te)*/
9.      i = i+1;
10.     if (i ≤ |mo(α1)|) then do ω1 = mo(α1)[i];
11.     else break to line 12;
12.   ω0 = [mo(α1)[i-1].ts, ω2.te);
13.   if ((ω2.te - mo(α1)[i-1].ts) ≤win) then do
14.     mo(α) = mo(α) + ω0;
        /*组合当前 α1、α2 的相邻发生域，生成目标事件组 α 的一个最小发生域*/
15. }while(true);
16. output mo(α);
```

定理 由 mo（α_1）和 mo（α_2）按照 GenMO 函数生成的集合 mo′（α）是目标事件组 α 的有效最小发生域有序集合，即 mo′（α）= mo（α）。

证明：我们分别从有效性、完备性和有序性进行证明。

（1）有效性。GenMO 中 Line 13 的判断确保了 mo′（α）中的元素是 α 的有效最小发生域。

（2）完备性。用反证法，不妨假设存在 α 的一个有效最小发生域［t_s, t_e）$\notin$ mo′（α）。依据参考文献［1］中的引理 3，必定存在 $t_s^1 > t_s$ 和 $t_e^1 < t_e$，使得［t_s, t_e^1）$\in$ mo（α_1）、［t_s^1, t_e）$\in$ mo（α_2）。按照 GenMO 交叉遍历的设计思路，由于存在 $t_s < t_s^1$，则必定可以找到 α_1 在 t_s^1 之前“最晚”发生的实例［t, t_e^2），使得 $t_s \leqslant t < t_s^1$、$t_e^2 < t_e$；同理，由于存在 $t_s^1 > t$，也必定可以找到 α_2 在 t 之后“最早”发生的实例［t_s^2, u），使得 $t < t_s^2 \leqslant t_s^1$、$u \leqslant t_e$；从而得到了 α_1、α_2 的一组相邻发生域［t, t_e^2）和［t_s^2, u）。

GenMO 通过组合这组相邻发生域可得 α 的一个最小发生域 $[t, u)$，且满足 $t_s \leqslant t$、$u \leqslant t_e$。显然，$[t, u)$的时间跨度 $u-t$ 小于或等于 $[t_s, t_e)$ 的时间跨度 $t_e - t_s$，即应有 $[t, u) \subset [t_s, t_e)$ 或者 $[t, u) = [t_s, t_e)$ 成立。如果 $[t, u) \subset [t_s, t_e)$，说明 $[t_s, t_e)$ 不是 α 的最小发生域，与假设矛盾。如果 $[t, u) = [t_s, t_e)$，说明 $[t_s, t_e) \in mo'(\alpha)$，也与假设矛盾。完备性得证。

（3）有序性。GenMO 交叉遍历 mo (α_1) 和 mo (α_2) 的设计方式可确保所生成的集合为有序集合。

下述函数 GenCandidate 实现了第二阶段生成候选事件组及其最小发生域集合的功能。

GenCandidate 函数

作用：生成候选事件组集合以及相应的 mo（α）集合

输入：事件组长度为 k 的频繁集合 F_k、最小发生域集合 MO_k，时间窗窗口阈值 win；

输出：事件组长度为 $k+1$ 的候选集合 C_{k+1}，以及相应的最小发生域集合 CMO_{k+1}；

过程：

```
1. count = 0;
2. if k == 1 then for h = 1 to |F_1|
       do F_1.block_start [h] = 1;
3. for i = 1 to |F_k| do
4.     current_block_start = count + 1;
5.     for (j = F_k.block_start [i]; F_k.block_start [j] ==
   F_k.block_start [i]; j++) do
6.         record α_1 = F_k [i];
       /*生成候选事件组*/
7.         for x = 1 to k do α [x] = F_k [i] [x];
8.         α [k+1] = F_k [j] [k];
9.         for y = 1 to k - 1 do
10.            for x = 1 to y - 1 do β [x] = α [x];
11.            for x = y to k do β [x] = α [x+1];
12.            if y == 1 then record α_2 = β;
13.            if β is not in F_k then continue with the next j at
   line 6;
14.        find mo (α_1) and mo (α_2) in MO_k;
           /*调用 GenMO，生成候选事件组的最小发生域
   集合 mo (α)*/
15.        mo (α) = GenMO (mo (α_1), mo (α_2), win);
16.        count = count + 1;
17.        C_{k+1} [count] = α;
18.    C_{k+1}.block_start [count] = current_block_start;
19.        CMO_{k+1} [count] = mo (α);
20. output C_{k+1} and CMO_{k+1};
```

2.3 评估候选事件组

候选事件组 α 的最小发生域集合的大小表明了该事件组在序列 (S, T) 中发生的次数，通过比较 | mo (α) | 和频率阈值 min_ fr 的大小即可判断 α 频繁与否。

2.4 简单比较 MOFE 和 WINEPI

1）解决重复计数问题

WINEPI 中，目标事件组 α 的发生频率由包含该事件组的时间窗的数目来表示。然而 α 的同一个实例可以出现在多个时间窗口当中，带来了重复计数问题，使得部分实际发生次数较少的事件组被不恰当的评估为频繁事件组。如表 1 中事件序列：

表 1 事件序列 Test（win = 5）

Time（min）	1	2	3	4	5	6	…	86	87
Events	A	B	C	D	A	C	NULL	A	B

考虑事件组〈A，B〉。WINEPI 算法统计的发生次数为 8 次；而〈A，B〉实际发生的次数仅为 2 次。如果频率阈值设定为 3 ~ 7，事件组〈A，B〉将被不恰当的评估为频繁事件组。

MOFE 算法通过最小发生域可以精确记录事件组每一个实例发生的具体时间区间，并由此建立事件组实例和最小发生域之间的一一对应关系，有效地解决了重复计数问题。例如，事件组〈A，B〉的最小发生域集合 mo(〈A,B〉) = {[1,3), [86,88)}，其中只有两个元素，分别对应了〈A,B〉的两个实例。

2）适用于连续时间和离散时间序列

WINEPI 算法的时间窗一次只能向后滑动一个时间单位，为了保证滑动过程中时间窗的大小不变，在给定基本时间单位后，读入内存的事件序列必须为严格的连续时间事件序列。然而在实际应用中，具有周期性集中生成特点的离散数据流往往更常见（如用户上网活动的记录）。由这类数据流转换而得的连续时间事件序列，通常会包含大量不发生任何事件的空时间点。这不仅会造成内存资源的浪费，还将使得 WINEPI 算法在运行过程中产生大量无效的空时间窗口，影响算法的效率。以事件序列 Test 为例，在一次遍历过程中，WINEPI 算法总共需要访问 87 个时间点，其中有 80 个为空时间点；总共产生了 91 个时间窗口，其中有 75 个为空时间窗口。

因此，WINEPI 算法在某些情况下不适用于离

散时间事件序列的频繁事件组挖掘。而 MOFE 算法所针对的数据序列(S, T)中,T 可以是连续时间,也可以是离散时间。考虑 Test 序列,MOFE 只需要访问其中的 8 个时间点,且只访问一次。与 WINEPI 相比,MOFE 对离散时间数据序列的处理将更加高效、方便。

3 不同客户群上网行为挖掘实验

网络用户的上网行为因人而异。不同的用户经常访问的网站不同,浏览网页的顺序也不同。本文将网络用户按照年龄大小分为老、中、青三个群体,试图达到以下两个目的:a)找出每个用户的上网模式,归结单个群体的上网特点;b)发现不同群体之间用户上网模式的差别,分析其中原因。(其中,青年用户是指年龄在 18 ~ 29 岁之间的用户,中年用户是指 30 ~ 49 岁之间的用户,而老年用户则是指年龄在 50 岁以上的用户。)

本文所采用的数据来自 COMSCORE 商用数据集(记录用户上网数据的数据集)。经过对原始数据的分析、抽象、整理,最终确定 400 名网络用户作为实验样本(其中,中年用户 200 名、青年用户 100 名、老年用户 100 名),并选取这些用户在同一个月内的上网记录作为输入数据。

实验的基本过程如下:

(1)应用 MOFE 算法,分别挖掘出每个样本用户的频繁上网模式(即频繁事件组),得到三张记录用户编号和频繁上网模式的事务表 D_i, $i=1, 2, 3$;

(2)从 D_i 中挖掘频繁一项集,发现相应用户群的上网习惯;

(3)比较不同用户群的上网模式,发现它们之间的相同和不同之处并总结规律。

实验参数设定如下:

(1)在步骤 2 中,挖掘频繁一项集的相对频率阈值 min_sup 设为 0.2 (D_i 为输入数据);

(2)在步骤 1 中,所有输入事件序列的基本时间单位都设定为分钟,绝对频率阈值 min_fr = 30(即认为平均每天出现一次的事件组为频繁事件组);

(3)分别设时间窗阈值 win = 15 和 win = 20,进行两次实验。

实验结果和分析如下。

3.1 三个用户群共有模式

老、中、青三个用户群大部分上网模式都相同,反映了现代网络用户之间的通性。如表 2 所示。

表 2 老中青三个用户群共有的上网模式

事件组长度	win = 15 min_ fr = 30	win = 20 min_ fr = 30
1	A, D, F, H, O, P, Q, T, U, V, Z	A, D, F, H, O, P, Q, T, U, V, Z
2	⟨A, A⟩ ⟨Q, A⟩ ⟨Q, P⟩ ⟨Q, Q⟩ ⟨Q, T⟩ ⟨Z, Z⟩	⟨A, A⟩ ⟨Q, A⟩ ⟨Q, D⟩ ⟨Q, H⟩ ⟨Q, P⟩ ⟨Q, Q⟩ ⟨Q, T⟩ ⟨T, Q⟩ ⟨Z, Z⟩
3	⟨Z, Z, Z⟩	⟨Z, Z, Z⟩
4	⟨Z, Z, Z, Z⟩	⟨Z, Z, Z, Z⟩

注:表中显著标识的事件组为两次实验中均发现的三个用户群共有的上网模式

从表 2 中不难发现,在老、中、青三个用户群所共有的上网模式中,大部分都包含 Q 类网站,即 Portals 类网站。

表 3 老中青用户群中经常访问 Portals 类网站的用户比例

用户群	win = 15 min_ fr = 30	win = 20 min_ fr = 30
Old	0.96	0.96
Middle	0.945	0.945
Young	0.97	0.97

结合表 3 以及 Q 类网站的主要特征(Q 类网站主要是导航网站或者门户网站),可以得出大部分网络用户所共有的上网习惯:先访问导航网站或者门户网站;然后通过该类网站中提供的链接访问其他类型的网站。可能原因:网络用户更偏好于使用以网站名称简单标识的链接而不是具体、繁杂的网络地址。

3.2 单个用户群特有模式

存在部分上网模式,只在单个群体中频繁出现,体现了该类客户的特点。如表 4 所示。

表 4 单个用户群特有的上网模式

用户群	win = 15 min_ fr = 30	win = 20 min_ fr = 30
Only-Young	⟨A, Q⟩, ⟨G⟩, ⟨T, Q⟩	⟨A, Q⟩, ⟨G⟩, ⟨H, Q⟩
Only-Middle	⟨I⟩, ⟨Z, Z, Z, Z, Z, Z⟩	⟨I⟩, ⟨Z, Z, Z, Z, Z, Z⟩, ⟨Z, Z, Z, Z, Z, Z, Z⟩
Only-Old	⟨H, H⟩, ⟨R⟩	⟨D, D⟩, ⟨R⟩, ⟨T, T⟩

〈G〉和〈A，Q〉为青年用户特有的上网模式；〈I〉和〈Z，Z，Z，Z，Z，Z〉为中年用户特有的上网模式；老年用户特有的上网模式为〈R〉。结合各类网站的主流信息，发现：

（1）青年人更多地访问 G：Education 类网站。可能原因：青年用户普遍处于求学阶段。

（2）中年人更加关注 I：Finance &Investing 类型的网站。可能原因：中年人需要承担更多家庭责任。

（3）部分中年人（23.5%）会在一段时间内频繁访问成人类网站（Z：XXX/Adult Content）。

3.3 两个用户群共有模式

存在部分上网模式，只由两个用户群共有，反应了它们之间在某些方面的相似性。

表 5 只由两个用户群中共有的上网模式

用户群	win =15 min_ fr =30	win =20 min_ fr =30
Young-Middle	〈Z，Z，Z，Z，Z〉	〈Z，Z，Z，Z，Z〉
Young-Old	〈H，A〉，〈Q，D〉，〈Q，H〉	〈H，A〉，〈H，H〉
Middle-Old	NULL	NULL

从表 5 中可知：部分青年人和中年人会在短时间内频繁访问成人类网站；部分青年人和老年人会在访问完娱乐类型（H：Entertainment）的网站后，接着访问类似网上商城的网站（A：Ad Banner Networks）。可能原因：老年人和青年人有更多空闲时间关注娱乐新闻和进行网上购物。

4 结 论

本文通过借鉴最小发生域的基本概念，设计并实现了基于最小发生域挖掘频繁事件组的 MOFE 算法。该算法可以从事件序列中有效地挖掘频繁事件组，解决了重复计数问题和离散时间序列带来的问题，并且只需要遍历一次原始事件序列，降低了原始数据总量对算法效率的影响。在不同客户群上网行为的实际研究中，应用 MOFE 算法，可以快速地发现每个样本用户的上网习惯，为进一步分析群体的上网模式准备数据。不同客户群的上网模式既有相同之处，又存在差别。两个客户群相同的上网模式是二者在某些方面相似性的体现。而由某个用户群所独有的上网模式则反应了该类用户在某方面与其他用户群的显著不同。

参 考 文 献

[1] Mannila H，Toivonen H，Verkamo A I. Disco Very of Frequent Episodes in Event Sequences [J]. Data Mining and Knowledge Discovery，1997，1 (3)：259-289.

[2] Mannila H，Toivonen H. Discovering Generalized Episodes Using Minimal Occurrences [M]. KDD Conference，1996.

[3] Laxman S，Sastry P S，Unnikrishnan K P. A Fast Algorithm For Finding Frequent Episodes In Event Streams [M]. KDD Conference，2007.

[4] Anny Ng，Ada Wai-chee Fu. Mining Frequent Episodes for Relating Financial Events and Stock Trend. Pacific-Asia on Knowledge Discovery and Data Mining，2003.

[5] 邓勇，施文康．发现频繁情节的改进算法 [J]．上海交通大学学报，2005，39 (3)：405-408.

[6] 王云岚，周兴社，侯正雄．频繁情景并行挖掘算法研究．西北工业大学学报，2007，25 (2).

Finding User's Online Browsing Patterns by Frequent Episode Mining

ZHOU Wenzhi[1]，LIU Hongyan[1]

(1. School of Economics and Management，Tsinghua University，Beijing 100084，China)

Abstract：Internet users have their own way to browse websites. The patterns behind their action are useful for internet companies. So how to find each user's browsing pattern from his (her) website-browsing sequence quickly and efficiently is a challenging task. Frequent episode discovery is an important data mining technique for finding patterns from event sequence. There are two existing algorithms，WINEPI and MINEPI，for mining frequent episodes. In this paper，we present a new algorithm for frequent episodes mining under the definition of minimal occurrence. This algorithm can be applied to both consecutive and discrete events sequence. In order to improve its performance，we found some valuable properties and proposed several useful methods. Before its application on website-browsing pattern analysis，we first group users into three parts，young，middle and old. Using our methods we get some interesting and reasonable results.

Key words：events sequence；frequent episode；minimal occurrence；online browsing pattern

网络广告位置对点击率 CTR 影响的实证研究

卫强[1]，阮楠[2]，单艺[3]

（1. 清华大学 经济管理学院管理科学与工程系，北京 100084
2. 清华大学 经济管理学院市场营销系，北京 100084
3. 北京传智锐媒广告有限公司，北京 100005）

摘 要：网络广告作为一种新兴广告形式正在高速发展，而点击率（CTR）是影响点击量、广告效果、广告位定价及用户竞拍意愿的关键指标之一。其中，研究广告位置如何对 CTR 产生影响具有理论意义和实践意义。本文针对某网络广告服务商提供的 27 天广告浏览和点击日志数据，采用非参数检验，分别按行、列与九宫格对广告位对 CTR 影响进行实证分析并发现：（1）广告位置对 CTR 的影响显著；（2）右列 CTR 显著低于前两列 CTR；（3）中间行 CTR 显著高于上方行。在论文最后，作者进行了归纳分析和展望。

关键词：网络广告；CTR；非参数检验

中图分类号：TP399

1 背景简介

随着互联网的快速发展和普及，互联网经济规模和影响力越来越大。而支撑大量免费互联网服务的主要经济形式之一是网络广告。网络是广告主以付费方式运用互联网对公众进行劝说的一种信息传播活动。虽然互联网广告占整个广告市场的份额较小，但它以高精准性、相对低成本、内容丰富性、时空灵活性取得了高速发展[1][2][3][4][5][6][7][8]。据艾瑞咨询统计显示，2008 年我国网络广告市场规模达 180.6 亿元人民币，近些年以 70% 以上增速发展。占整体广告市场比重从 2006 年的 3.8% 增至 2008 年的 10% 左右[2]。特别在目前全球经济危机背景下，网络广告更具独特优势。

网络广告主要形式是在网页相关位置上添加各种形式的文本广告、图片广告及富媒体广告等。不同网络广告有多种不同收费模式[5][6]，如每千人广告展示数 CPM（Cost Per Thousand Impressions），每点击成本 CPC（Cost Per Click），每行动成本 CPA（Cost Per Action），每购买成本（Cost Per Purchase）等。这些不同收费模式（除 CPM）都跟网络广告被点击次数相关。而点击率（CTR，Click Through Rate）=广告点击量/广告展示量，是最能广泛应用的网络广告质量评估指标之一。它反映在一定时间内，网络广告受关注的概率估计值。1995 年，美国网站横幅广告点击率能达到 30% ~40%，这反映了互联网时代初期的用户特点。然而随着网络广告越来越多，同时用户逐渐对广告产生了“免疫”，近年来 CTR 大幅下降。2008 年，网络广告的平均 CTR 不到 0.1%[4][9]（尚未剔除点击欺诈的影响）。

因此，网络广告运营商和厂商都希望能有效提高广告的点击率，从而可赚取更多利润（广告运营商）或吸引更多眼球（厂商）。这就需要找到影响 CTR 的关键因素并采取针对性措施。业界和学界普遍认为影响 CTR 的主要因素包括：（1）广告的一些基本特征（生动性、大小、样式）；（2）广告内容，以及与网页/网站内容的相关程度；（3）广告位置的影响；（4）时间因素的影响[17][18][19][20][21]，等等。

但是，上述因素的有效性和合理性尚未得到严格证明和分析。许多分析是定性说明和解释，而缺乏数据分析的支撑。而一些数据分析结果也更多的是网络公司的分析报告。出于竞争的需要，这些公司一般不提供技术细节。在我国，针对我目前的网络广告 CTR 的影响因素和效果分析的研究和行业报告还非常不够。

本论文通过与某专业网络精准广告服务商进行合作，获得了 27 天的网络广告浏览和点击日志数据，并基于此进行实证分析，得到一些具有一定指导意义的分析结果。不但填补了我国这方面研究不足，也为互联网广告行业发展提供一定指导。

2 网络广告位置与 CTR 的关系

网络广告行业在十几年高速发展中一直受到 CTR 下降的困扰，网络用户越来越趋于实用和理性，早在 20 世纪 90 年代中期广告 CTR 很高的时候，就已经开始出现了明显下降的趋势。2000 年，横幅广告平均 CTR 已下降到约 0.5%[10]。2007 年，eMarketer 报告显示横幅广告平均 CTR 只有约 0.2%[11]。中国的情况也如此。2007 年 AC 尼尔森的《2007 年 10 月中国网络广告市场研究报告》显示，横幅、弹出、按钮等广告形式的平均 CTR 不足 0.18%，搜索广告 CTR 也出现不同程度下降[12]。comScore 公司数据表明 2008 年展示广告平均 CTR 不足 0.1%[4]。

从用户点击网络广告的原因来看，Rubin 的使用与满足理论将媒体使用动机分为工具性行为（instrumental behavior）和仪式化行为（ritualistic behavior）。前者表现为寻找信息，如搜索等；后者表现为寻找娱乐，如随意浏览冲浪行为等[13]。Hirshman 和 Holbrook 将人们消费行为划分为实用型消费和享乐型消费。在网络广告上，前者更注重内容，后者更追求广告颜色、图像、动画等带来的感官感受[14]。1996 年，Hoffman 和 Novak 发展了使用满足理论，提出两种网络导航模式：目标导向（Goal-directed）和体验性（Experiential）网页浏览行为[15]。

人们点击网络广告主要也是基于这两种动机，第一种是目标导向，即用户有意搜索信息，慎重考虑广告提供的信息；第二种是在网上冲浪或进行娱乐时，用户未经主动探究和思考就受到了周围广告的暗示及影响而点击了广告[16]。

关于广告位置对 CTR 的影响，国内外研究都相对匮乏，以定性研究居多。相关研究中，一般将网页上位置划分成九宫格，如图 1 所示。这种划分方法比较简单实用，也是实际应用中网络广告摆放时所考虑的位置[22]。

在相关研究中，Doyle 通过对几个网页进行了一周的数据统计，对三对位置进行比较发现[21]：（1）C3 的 CTR 要比 A 行的 CTR 高 228%，B 行的 CTR 比 A 行的 CTR 也显著高。其原因主要在于用户对横幅广告产生了相当的“免疫”而直接跳过。（2）在同一页面放两个相同广告的 CTR 要比一个广告 CTR 略高，但不显著。然而，Doyle 的研究数据量较小，同时没有采用合适统计方法进行检验，结果也过于简单，缺乏代表性。在 Google 的统计

	1 列	2 列	3 列
A 行	A1	A2	A3
B 行	B1	B2	B3
C 行	C1	C2	C3

图 1 网络广告九宫格位置示意图

报告中[22]，首页上方的 A 行横幅广告，内容页面 A1 的广告或第 3 列的广告 CTR 较高，论坛页面下方 C 行横幅广告的 CTR 较高。但是 Google 的报告没提供必要技术细节。

总的来说，广告位置对广告 CTR 的影响的研究非常有限。且由于数据获得比较困难，严谨的统计分析和实证研究难以进行。而本文通过与某专业网络精准广告服务商的合作，获得了 27 天的网络广告浏览和点击的日志数据，并基于此进行实证分析和定量分析，得到了一些具有一定指导意义的分析结果。

3 数据描述与分析方法说明

提供数据的是一家网络精准广告联盟平台服务商。该服务商提供精准定向和智能化平台来提供互联网广告发布、监测与效果优化服务。其客户包括很多国际国内知名企业，合作媒体包括各个大型行业网站。该公司从网络媒体上买下部分广告位，再卖给广告主，并提供广告创意设计、内文匹配广告发布、广告布局优化等服务，致力于达到网络广告精准投放的目的，使得广告主、网站和它自己三方获益。该平台具有自动识别和匹配功能模块，可将广告投放到与之相关网页上。

该公司对广告主收费大多采用点击付费 CPC 方式，因此如何有效提高 CTR 非常重要。该公司合作网站超过 100 个，90%是垂直门户网站，涉及行业很广。该公司进行广告投放等操作时，也采用图 1 中九宫格方式进行。网络广告形式大多为包含文字和图像的横幅广告以及少量小规模按钮广告。该公司购买广告位大多不在同一网页上。同一个广告位一天会自动滚动多个广告。

该公司所提供数据包括两种网络日志文件，一种为展示日志，一种为点击日志。数据整理发现，每天展示日志有 8000 多万条记录，点击记录只有几万条，一天总数据规模超过 20G。本文分析了 27 天的数据（从 2009 年 3 月 30 日至 4 月 26 日），数据总规模约有 500G。前期数据处理工作量较大，采用 MySQL 进行按不同网站广告位进行统计，然后对应九宫格位置进行提取和汇总，计算得到每个位置每天浏览量、点击量和 CTR。

经过前期处理发现数据具有以下几方面特点：（1）每个网站广告位非常少，大多数只有 1 个；（2）各个广告位上浏览次数分布极不平均。如 A3，C2 和 C3 浏览量非常少；（3）点击率都非常低，平均各个位置点击率约不足 0.05%。最多的位置，如 B2，C1，能达到千分之几。

由于影响 CTR 因素非常多。如考虑广告位置对 CTR 影响，最理想情况应在同一个网页的九个位置上都有相同广告，且网页数量较多。这样才能剔除其他因素进行统计。显然这种情况在实际数据中不存在。因此，为了构建合理统计量，并针对广告位置的影响因素进行实证分析，本文将 27 天中每天九个位置的 CTR 作为一个样本，并将九个位置上的每天总 CTR（每个位置的一天内的总点击量/总浏览量）分别作为一组样本。这样的数据量较大，都能达到数十万级以上的浏览量，则其他因素，如广告内容，广告时间等影响相对而言可忽略。表 1 所示的就是处理后所得到的用于统计的部分数据。

表 1　27 天的 9 个位置的 CTR 数据

（单位：万分之一）

CTR	3 月 30 日	4 月 1 日	4 月 2 日	…	4 月 26 日
A1	3.1975	3.0364	5.7203	…	7.2023
A2	5.1235	5.7924	4.8289	…	7.795
A3	2.9399	2.5914	2.2435	…	4.2179
B1	6.7899	11.997	8.6206	…	81.756
B2	16.782	19.453	13.216	…	16.542
B3	2.4842	1.8687	2.0157	…	2.0012
C1	17.0933	20.3098	17.2739	…	23.3114
C2	7.8138	10.559	10.3718	…	34.1413
C3	6.6252	4.2022	5.4086	…	13.5501

进一步，在统计分析方法选择上，常用的 T 检验要求总体符合正态分布，F 检验要求误差呈正态分布且各组方差整齐等约束太强。而 CTR 数据分布的理论分析匮乏，无法保证满足这些要求。因此采用非参数秩检验方法，这些方法不要求总体服从正态分布[22][23][24]。针对多组样本非参数检验，采用 Friedman 检验方法。针对两组样本非参数检验，采用 Wilcoxon 检验方法。使用的统计软件是 SPSS。

4　假设检验与实证分析

首先，需要判断位置是否是影响广告 CTR 的重要因素之一，则第一个假设是：

H1：广告位置对 CTR 有影响

根据表 1 中的数据，采用 Friedman 方法进行研究得到结果如图 2 所示。

Ranks

	Mean Rank
A1	4.85
A2	4.11
A3	2.63
B1	8.22
B2	7.63
B3	1.63
C1	5.22
C2	7.33
C3	3.37

Test Statistics[a]

N	27
Chi-Square	155.654
df	8
Asymp. Sig.	.000

a. Friedman Test

图 2　Friedman 检验结果

检验结果显示拒绝零假设，即可以接受认为位置对 CTR 的影响是显著的。更为准确的说法是，至少有两个位置 CTR 差别显著不同。按照顺序大小排列，9 个位置 CTR 顺序是 B3 < A3 < C3 < A2 < A1 < C1 < C2 < B2 < B1。需要说明的是，由于在 C1，C2 和 C3 的浏览量非常低（许多天的浏览量低于 10 万），因此 C1，C2 和 C3 点击率会收到其他因素扰动较大，所以剔除 C 行位置。但是，从此结果仍可以得到一个粗略结果，即第 3 列 CTR 是最低的，而第 1、2 列 CTR 相对较高。进一步统计分析假设不同列 CTR 之间差异显著，因此提出以下三个假设：

H2a：第 1 列的 CTR 和第 2 列的 CTR 不同

H2b：第 1 列的 CTR 和第 3 列的 CTR 不同

H2c：第 2 列的 CTR 和第 3 列的 CTR 不同

各列 CTR 汇总后则数据量足够大（每行每天浏览量超过 10 万），可进行统计分析。采用 Wilcoxon 检验发现：（1）H2a 不成立，即第 1 列与第 2 列 CTR 没有显著差别；（2）H2b 和 H2c 的零假设被拒绝，即第 3 列 CTR 明显少于 1，2 列。

最右侧广告 CTR 很低的原因大体有两个：（1）用户浏览网页习惯通常从左向右，对右部页面关注较少，则左面广告 CTR 会比较高。（2）由于技术习惯，很多网站页面是从左向右逐渐载入，则右侧的广告也最晚展示而被用户忽略。因此，对广告主来说，可能更愿意购买左侧广告位。而对广告运营商来说，一方面应对右侧广告位定价进行调整，另一方面需考虑是否采取技术手段来平衡页面载入。

进一步，需要对各行 CTR 进行分析比较。由于用户浏览习惯是从上至下，因此提出假设认为各行 CTR 也有所不同。但由于 C1，C2 和 C3 的浏览量过低，因此受到其他因素扰动过大，因此不对此进行检验（但由于 C 行的浏览量都非常低，则 C 行广告不适宜采用 CPM 的计价方式）。进一步提出假设如下：

H3：A 行的 CTR 与 B 行的 CTR 不同

将各行 CTR 进行汇总，采用 Wilicoxon 检验发现拒绝零假设，则认为 H3 假设成立，即 B 行 CTR 要显著高于 A 行 CTR。这与 Doyle 的结果也一致[21]。这可解释为，随着网上用户越来越成熟，好奇型点击行为很少，过于明显的广告虽容易被看到，但往往被用户忽略，被点击几率不高[17][18][19][21]，而和内文广告靠近广告（通常在 B 行）很可能在用户浏览内容时被看到，而由于内容相关获得点击（或遭到误点），因此 B 行 CTR 较高。这对于实践来说具有重要意义。即对于广告服务商而言，内文广告倾向于采取 CPC 模式，而横幅广告倾向于采取 CPM 方式，以获得利益的最大化。实际上，许多网站广告运营商已采取这种策略。

进一步对 9 个广告位 CTR 进行两两比较分析，同样采用 Wilicoxon 检验方法。由于 C1，C2 和 C3 的浏览量过低，从统计上来说缺乏可靠性，故略去。提出的假设为“X 的 CTR 与 Y 的 CTR 不同”等，其中 X 和 Y 分别为 A1，A2，A3，B1，B2 和 B3 中的任一位置。统计结果如表 2 所示。

27 天的广告位（A1，A2，A3，B1，B2 和 B3）

表 2 CTR 两两比较结果

	A1	A2	A3	B1	B2	B3
A1		显著	显著	显著	显著	显著
A2			显著	显著	显著	显著
A3				显著	显著	显著
B1					显著	显著
B2						显著
B3						

表 2 中的结果显示，任意两个广告位 CTR 之间差别都显著。进行统合后得到 B3 < A3 < A2 < A1 < B2 < B1，该排序中每两个之间差别都显著。这也与 H1 的假设检验分析的结果一致。

5 总 结

本文针对广告位置对网络广告点击率 CTR 的影响进行了实证分析。通过对某网络广告运营商 27 天网络广告浏览和点击日志数据进行整理，并采用通行九宫格形式，采用 Friedman 和 Wilicoxon 非参数检验方法进行了统计分析。通过假设检验分析发现：（1）广告位置对 CTR 的影响显著；（2）右列的 CTR 显著低于前两列 CTR；（3）中间行的 CTR 显著高于上方行。并进一步对造成这些情况的原因进行了解释和分析。

本研究的主要贡献体现在：（1）目前国内外对影响 CTR 关键因素的研究较少，特别是对于广告位置影响更为匮乏。因此本文做出了有益尝试。（2）由于缺乏第一手数据，相关定量研究非常有限。由于广告运营商出于保护竞争优势目的也不愿公布数据。因此本研究通过与广告运营商的合作获取了第一手数据，并进行了统计分析，具有一定新意。

本研究也存在一些不足。首先，数据来源是一个公司，则不可避免地受到该公司选择网站和行业的偏好影响，会存在偏差。其次，数据规模虽大，但无法确定数据分布，因此难以采用更为可靠统计方法。进一步的工作主要包括：（1）进一步对数据进行整理并深入分析，并结合其他因素来构建影响 CTR 指标体系。（2）结合实际应用背景，分析结果的管理含义和价值。

参考文献

[1] CNNIC,《第二十三次中国互联网络发展状况报告》[R]. 2009.

[2] 艾瑞咨询 (iResearch),http: //www. iresearch. com. cn/.

[3] 艾瑞咨询 (iResearch),《2009 年第一季度中国网络广告市场监测报告》[R]. 2009.

[4] Gian M. Fulgoni, Marie Pauline Mörn. How online advertising works, Empirical Generalizations in Advertising Conference for Industry and Academia [C]. December 4-5, 2008.

[5] 邓文峰, 周朝民 .《浅谈网络广告效果评价方法》[J].《上海管理科学》2005 (4).

[6] 孟今圣 .《门户网站网络广告定价策略及定价模型研究》[R]. 吉林大学, 2008.

[7] 吕学良 .《基于视觉信息的上下文广告关键词提取算法研究》[R]. 浙江大学, 2007.

[8] 艾瑞咨询,《网络广告的市场现状及发展趋势》[R]. 2009.

[9] IAB, 2007 Internet Advertising Revenue Report [R]. http: //www. iab. net/media/file/IAB_PwC_2007_full_year. pdf.

[10] Sweeney T. Advertisers Seek More Bang for Their Web Bucks [J]. Informationweek. com, October 2, 2000, 130.

[11] eMarketer. The E-Mail Marketing Report [R]. [2007-02-25] eMarketer, January 2007, http: //www. emarketer. com/.

[12]《尼尔森 2007 年 10 月中国网络广告市场研究报告》[R]. http: //www. inghe. net/ingheblog/7. html.

[13] Rubin A M. Ritualized and Instrumental Television Viewing [J]. The Journal of Communication, 1984.

[14] Hirschman E C, Holbrook M B. Hedonic Consumption: Emerging Concepts, Methods and Propositions [J]. Journal of Marketing, Vol. 46 No. 3, 1982: 92-101.

[15] Donna L. Hoffman and Thomas P. Novak. Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations [J]. Journal of Marketing, Vol. 60, No. 3 (1996), pp. 50-68.

[16] Chang-Hoan, Hongsik John Cheon. Why Do People Avoid Advertising on the Internet [J]. Journal of Advertising, 33 (4): 89-97.

[17] Cho, Chang-Hoan. Factors Influencing Clicking of Banner Ads on the www [J]. Cyberpsychology & Behavior, 2003, Vol 6, No 2, 201-215.

[18] Moore R S, Stammerjohan C A. Coulter R A. Banner advertiser-web site context congruity and color effects on attention and attitudes [J]. Journal of Advertising, 2005, Vol 34, No 2, 71-84.

[19] Cho Chang-Hoan. The Effectiveness of Banner Advertisements: Involvement and Click-Through [J]. I&MC Quarterly, 2003, Vol. 80, No. 3, 623-645.

[20] Chtourou M S, CHandon J L, Zollinger M. Effect of Price Information and Promotion on Click-Through Rates for Internet Banners [J]. Joural of Euromarketing, 2001, Vol. 11, Issue 2, 23-41.

[21] Doyle, K, Minor A and Weyrich C. Banner Ad Placement Study [R/OL] . University of Michigan. 1997. http: //www. webreference. com/dev/banners/ (accessed July 2007).

[22] Google Adsense Blog [N/OL], http: //adsense. google-chinablog. com/.

[23] 薛薇 .《统计分析与 SPSS 的应用》[M]. 北京: 中国人民大学出版社, 2002.

[24] 袁卫, 庞皓, 曾五一, 贾俊平 .《统计学》[M]. 北京: 高等教育出版社, 2006.

Empirical Studies on Internet Ads Positions Influencing CTR

WEI Qiang[1], RUAN Nan[2], Shan Yi[3]

(1. Department of Management Sciences and Engineering,
School of Economics and Management, Tsinghua University, Beijing 100084, China
2. Department of Marketing, School of Economics and Management,
Tsinghua University, Beijing 100084, China
3. ITOP 24/7, Beijing 100005, China)

Abstract: Internet Ads have been rapidly developing for decades of years. CTR (Click Through Rate) is one of the key factors that influencing ads' clicks, ads' effectiveness, ads pricing, etc. Ads psotioning is one of the important issues, which is meaningful not only theoretically but applicably. This paper empirically studies how rows, columns and 9 positions influence CTRs using non-parameter test methods, based on the collected 27-days browsing and clicking data from a online-ads company. The results show that:

① positions significantly influence CTR; ② Right column's CTR is significantly lower than the other two columns; ③ Middle row's CTR is significantly higher than the upper row. The conclusion is presented in the last section.

Key words: internet ads; CTR; non-parameter test

决策支持系统研究的设计科学研究方法探讨*

尚维[1]，徐山鹰[1]，汪寿阳[1]

（1. 中国科学院数学与系统科学研究院，北京 100190）

摘　要：为提高决策支持系统研究的科学性与规范性，本文提出了一个决策支持系统研究的设计科学研究方法框架。该框架从决策问题发现到模型和系统的开发、评估与结论得出的循环递进保证了研究在理论上的规范性和与决策实践的紧密联系。文献分析表明：目前我国决策支持系统研究在问题提出、方案设计、模型研发与系统设计方面比较规范科学，但在对整个研究产出进行评估和结论得出部分略显不足。在评估方面应用实验、调查等方法，将有助于提高决策支持系统研究的完整性、规范性和科学性。

关键词：决策支持系统；研究方法；设计科学

中图分类号：C931.6

决策支持系统研究是信息系统研究的一个重要分支。20世纪40年代第一台现代计算机出现，人们就开始不断探索如何利用电子计算机技术来辅助进行管理决策。最早的用计算机技术进行管理决策支持方式是电子表格及一些优化算法的求解。20世纪60年代，为了应用计算机技术解决更加复杂的管理决策问题，研究人员们开始系统地针对决策的制定、规划等管理任务在计算机系统上的实现进行研究。于是决策支持系统（Decision Support Systems）的概念出现了。早期的决策支持系统是利用定量模型计算来为决策提供支持的计算机系统[1][2]。为了更深入地研究决策支持系统相关的理论问题和应用问题，从20世纪70年代开始，决策支持系统逐渐作为一个研究课题出现在管理科学与计算机科学等领域的重要学术会议中。到了80年代，决策支持系统的应用领域不断拓展，应用规模不断增大。对决策支持系统理论及规范性框架等方面的研究也逐渐开展，决策支持系统的数据库、模型库和对话生成等独有的特性被普遍接受，Sprague和Carlson的著作[3]成为决策支持系统研究的重要标志。进入21世纪之后，随着信息通信技术的发展和决策支持系统研究设计的领域更加广泛，应用也日趋复杂。迄今为止，决策支持系统的研究涉及运筹学、系统科学、人工智能、软件工程、信息技术，决策与对策理论和组织行为学等多个基础学科与应用学科领域。

决策支持系统的发展过程是决策支持系统不断为其他领域研究提供支持与服务的过程，同时，也是决策支持系统吸收其他研究成果的过程。决策支持系统研究作为一类特殊的具有极强交叉性的研究课题，其研究进展与其他学科的发展密不可分。从应用领域角度，决策支持系统研究多集中于项目或投资评价、与地理信息系统相结合的水资源管理或土地资源管理、供应链协调与管理等领域。从技术角度，决策支持系统研究主要涉及数据仓库的应用、SOA架构的应用、智能算法与新的推理技术应用等。

决策支持系统研究可针对决策支持系统设计或使用中的某个问题的研究，包括系统采用的决策模型的研究、决策过程行为模型的研究、决策支持效率和效果的研究等[4][5][6]；也可对某一特定应用或特定类型的决策支持系统的研究，包括设计、应用及评价等内容[7][8]。决策支持系统的研究方法有规范性研究、实证研究、行动研究、案例分析和设计科学研究等。鉴于决策支持系统超强的实用性，设计科学研究可以说是决策支持系统研究的最重要方法。本研究提出了面向决策支持系统研究的设计科学研究方法框架，对我国近年来的决策支持系统研究从研究方法层面进行分析和评述，给出了加强决策支持系统研究科学性的建议。

* 基金项目：国家自然科学基金（70801059）

通讯作者：尚维，E-mail：shangwei@amss.ac.cn

1 科学研究方法简述

作为一项严肃、神圣的工作，一般来说，科学研究的任务包括发现事实真相、创建新的知识和改进已有知识。科学研究所采用方法的科学性直接决定了科学研究工作成果的正确性。因此在任何科学研究工作之前都有必要对所采用的研究方法作出谨慎的选择。科学研究的方法论（Methodology）指导基于对客观事实的认知（Epistemology）方式，而对客观事实的认知则基于对客观事实本身的定义，即本体论（Ontology）。如图 1 所示，研究客观事实本身的哲学属于本体论范畴，在本体论的基础上对认知的研究属于认识论范畴，在认识论基础上所建立的对如何去认知的研究属于方法论范畴。

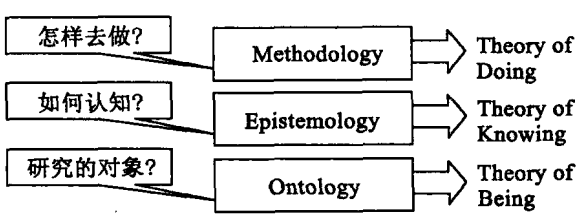

图 1 本体、认识论与方法论

本体论层面涉及对世界本源认识的哲学问题，基本上可分为唯物主义和唯心主义两大基础阵营，随着哲学研究的发展，这两大阵营分别有了繁多的流派及交叉，在此不再赘述。在唯物主义本体论基础上有实证主义认识论方法（Positivistic）；在唯心主义本体论基础上有解释主义认识论方法（Interprestic）。实证主义认为客观世界可以被唯一认知，人对客观世界的观察可以是无偏见的；而解释主义认为客观世界不能被唯一认知，人对客观世界的观察都是有偏见的。实证主义科学研究与解释主义科学研究的任务、态度、方法、形式及起源如表 1 所示。

表 1 实证主义研究与解释主义研究含义的对比

	实证主义	解释主义
	Positivism	Interpretivism
任务 Mission	发现事实 To find the truth	解释事实 To interpret
态度 Attitude	客观 Objective	主观 Subjective
方法 Method	定量方法 Quantitative	定性方法/定量方法 Qualitative/Quantitative
形式 Style	数字统计与度量 Number crunchers	论述 Story telling
起源 Origin	自然科学 Natural Sciences	行为科学 Behavioral Sciences

2 信息系统研究方法论

2.1 信息系统研究方法

信息系统学科产生于 20 世纪 70 年代。Keen 在第一届信息系统国际会议（International Conference on Information Systems，简称 ICIS）上指出，信息系统是一个综合借鉴许多其他交叉学科的应用学科[9]。信息系统研究围绕信息技术制品（IT Artifacts）展开。信息系统学科的研究对象包括：表现为软硬件产品的信息技术制品、应用信息技术制品的各项任务、服务于信息技术制品应用任务的组织结构以及整个以信息技术制品为核心的环境[10]。分析（Analytical）、实证（Empirical）和技术（Technical）是信息系统研究的三个主要方面。信息系统学科的研究方法，从以认识论为基础的科学研究方法论角度可分为实证主义方法（Positivistic）和解释主义研究方法（Interpretistic）。但随着各种研究方法不断融合发展，其界限开始变得不那么明显。信息管理学科的研究方法体系见图 2。严格遵循理论构建与统计检验原理的调查（Survery）和实验研究（Lab

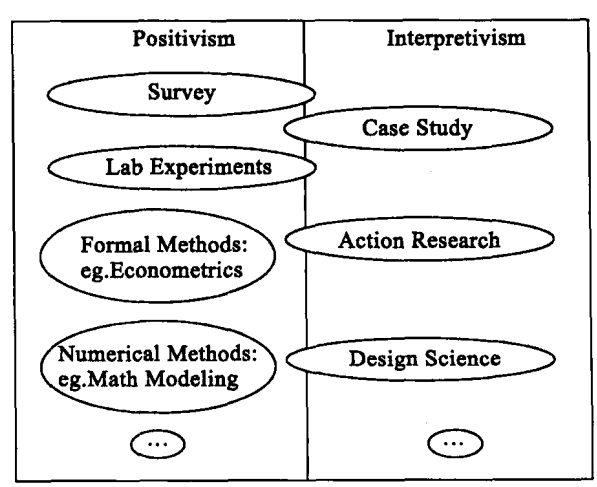

图 2 信息管理学科的研究方法体系

Experiments）也经常辅以解释性的案例和阐述。而原本基于解释主义认识论的案例研究（Case Study）、行动研究（Action Research）和设计科学（Design Science）也经常包含一些规范化的统计测度，从而带上了实证主义的色彩。

2.2 信息系统的设计科学研究方法

设计科学（Design Science）是近年来信息管理研究中的重要研究方法流派[11]。设计科学起源于计算机和工程学科。Smith 在文献[12]中区分了自然科学（Natural Science）和人工科学（Science of the Artificial）的概念，并将作为人工科学的重要研究方法的设计科学定义为：科学研究者参与的以创造概念、模型、方法和实例为目的的研究方法。经过众多学者的研究与实践[13][14]，设计科学研究方法在信息系统学科领域的应用逐渐成熟[11][15]。Vaishnavi 总结了设计科学研究方法体系。

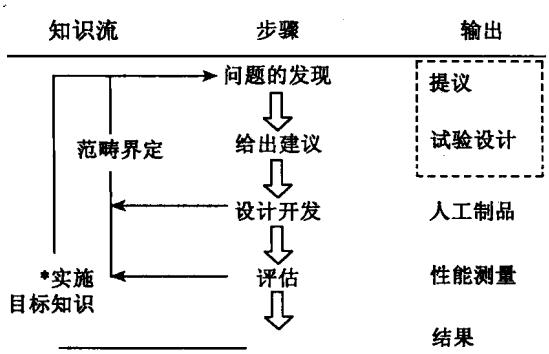

图 3 设计科学研究的方法体系[16]

如图 3 所示，设计科学研究方法的核心过程包括识别问题、提出建议与假设、设计开发、评估和结论得出。这些过程通过一个或几个知识产生循环，产生相应的信息技术制品产生。

设计科学的特点是通过模型、构架或系统的设计、开发与实施，实现从实践问题到解决方案的循环推进。其优势是研究过程不仅包括理论模型的构建，还包括方法、概念的实现以及实证检验，能够做到从多方位多角度对目标问题进行研究。设计科学方法适用于已有一定理论研究基础，亟须寻找如何进行实践应用的研究问题。

3 决策支持系统的设计科学研究方法

决策支持系统与一般信息系统相比，更加侧重于通过提供信息的方式对决策者决策过程进行支持。决策支持中，最为重要的是通过决策分析模型的实现来为决策提供依据。因此，在很多决策支持系统的研究中，决策模型的实现和决策支持系统实现被作为研究过程迭代的两个重要步骤[17]。决策模型一般为优化、评价等规范性模型，作为对决策问题建模及求解的解决方案，使决策支持系统的一项重要产出。而决策支持系统的建立，一般是在决策模型的基础上进行设计和分析，作为最终对决策者决策进行辅助的产品。因此，本研究将图 3 所示的信息系统的一般设计科学方法体系发展为图 4 的决策支持系统的设计研究方法体系。

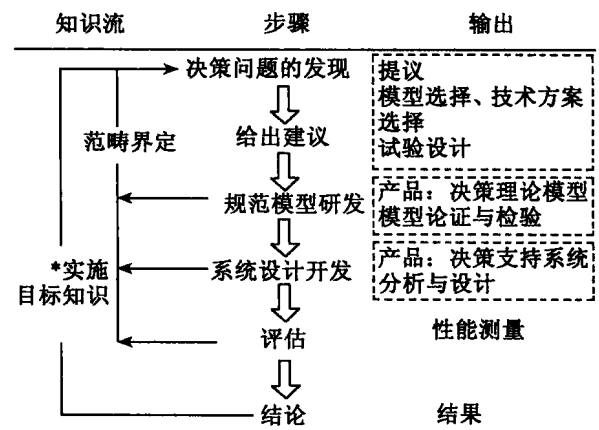

图 4 决策支持系统的设计科学研究的方法体系

决策问题的发现包含决策问题的界定、决策主体的识别和决策目标确定三个方面。确定了决策问题后，要根据已有的经验给出解决决策问题的初步方案，这里的研究输出为决策方法模型和对决策效果评价的试验设计。常用的决策方法模型有最优化模型、组织行为学模型、经济学模型等。评价决策效果的试验设计主要侧重于如何度量决策的效果，一般的试验有现场试验、模拟场景试验和数值仿真试验。针对决策问题本身的特点，选择了决策模型方法和试验之后，就需要利用所选择的决策方法模型工具对决策问题进行建模和求解。在这里，如果决策模型无法得到解析解，就需要补充设计数值仿真试验对模型求解的稳定性进行验证。如果所采取的模型是行为模型和经济计量模型，则需要利用试验数据或实证经济数据对模型进行检验。完成规范模型开发之后一般可继续进行决策支持系统设计开发。决策支持系统设计开发的产品一般是一个支持利用已论证的决策模型，针对所提出的决策问题进行决策过程支持的软件系统。同时，产出还包括系统分析与设计的相关文档。一般决策支持系统研究的学术文章中用系统设计的一些关键图来代替无法