学术专著
Academic Monograph

不确定理论与Web挖掘

吴　瑞◎著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

不确定理论与 Web 挖掘

吴 瑞 著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

前　言

www 给网站设计者和运营者带来巨大商机的同时，也带来了巨大的挑战。如果一个特定的网站不能在短时间内满足用户的需要，这些用户就会很快转向其他网站，因此了解网络用户的浏览行为和特征是非常有必要的。人们迫切需要开发自动挖掘技术从大量的 www 数据中发现人们感兴趣的模式（知识），因此，Web 挖掘成为一个越来越热门的研究领域。Web 挖掘可分为三类：Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。目前国际上对 Web 使用挖掘的研究比较多，Web 挖掘的国际权威会议 WebKDD 也把 Web 使用挖掘作为重点。

不确定理论是处理不确定性问题的一种数学工具，它包括可信性理论、模糊随机理论、随机模糊理论、粗糙理论和模糊粗糙理论等。不确定理论作为软计算技术的一种，已经被越来越频繁地应用于智能系统中，它为处理不确定信息提供了一种自然的方法。越来越多的研究者逐渐把不确定理论应用到 www 知识发现和分析中，Pal 等人介绍了“软 Web 挖掘”的定义以及各种软计算工具在 Web 挖掘中的应用。

本书在 Web 挖掘和不确定理论的基础上，系统阐述了不确定理论在 Web 挖掘中的应用。全书共分为九章：第 1 章介绍了不确定理论的基础知识；第 2 章介绍了 Web 挖掘的基本知识和挖掘方法；第 3 章探讨了 Web 日志中泛化关联规则的挖掘方法；第 4 章研究了基于 FLaAT 的 Web 浏览模式的挖掘；第 5 章讨论了模糊环境下的用户偏爱浏览模式的挖掘；第 6 章提出了模糊环境下的 Web 用户浏览模式集的粗糙聚类方法；第 7 章对基于 Leader 算法的用户浏览模式的聚类方法进行了讨论；第 8 章探讨了综合策略下用户浏览模式的聚类；第 9 章对本书的研究进行了总结。

值本书出版之际，我要特别感谢我的博士导师唐万生教授，他不仅在学术上给予我悉心的指导，在工作和生活方面也给了我无私的帮助，在这里谨向恩师表示由衷的感谢和诚挚的敬意！我还要特别感谢赵瑞清老师孜孜不倦的教诲和激励！

本书的出版得到了国家自然科学基金（No. 70802043）的特别资助，并得到山西省自然科学基金（No. 2008011029—2）的部分资助，在此一并表示感谢！

本书的内容中有一部分反映了不确定理论应用于 Web 挖掘的最新研究成果、研究方法和研究动向，在理论体系和方法上均有所创新，构建了不确定理论与 Web 挖掘研讨的平台。本书可作为管理科学、计算机科学、系统科学、信息科学与工程技术等专业高年级大学生和研究生的教材，对相关领域的研究人员也有重要的参考和使用价值。

由于作者才疏学浅，书中难免有所疏漏，恳请各位专家学者批评指正，提出宝贵意见。

吴　瑞

2011 年 3 月

目　录

第 1 章　不确定理论基础

在运筹学、管理科学、信息科学、计算机科学等专业领域以及众多工程中都存在客观的或人为的不确定性，这些不确定性的表现形式是多种多样的，如随机性、模糊性以及粗糙性等。伴随着这些不确定性必然会产生很多不确定问题。然而对于包含这些不确定性的决策问题，经典理论通常是无能为力的，因而在解决这类问题时，引入不确定理论是非常必要的。

1.1　模糊理论

L. A. Zadeh[113]首先引入了模糊集（Fuzzy Set）的概念，其基本思想是把普通集合中的决定隶属关系灵活化，使元素对“集合”的隶属度从集合{0,1}中的值扩充为[0,1]中。本节将介绍一些有关模糊集及模糊变量的相关理论知识。

1.1.1　模糊变量

自从 Zadeh[113, 114]于 1978 年提出模糊集概念以来，模糊理论得到了充足的发展。Nahmias[60]在 1978 年提出模糊变量的概念。此后 Liu 等

人[49~51]提出了置信度的概念，并在此基础上给出了模糊变量的一系列理论体系。这里从模糊集的概念逐渐引入模糊变量的知识。

在经典集合论中，论域 U 上的一个普通集合 A 定义为 U 中某些元素 x 组成的群体。每个元素或者属于集合 A，或者不属于集合 A。然而在很多情形下这种隶属关系并不是明确的。例如，“强壮”、“著名”、“年轻”，等等，这些概念所表达的含义并不是具体、明确的。在这种情况下，经典集合论并不适用。为了处理这类问题，首先引入模糊集的概念。

定义 1.1[49,50] 设 U 为论域。$\tilde{A}$ 为 U 的一个子集，对任意元素 $x \in U$，函数

$$\mu_{\tilde{A}}: U \to [0,1] \qquad (1\text{-}1)$$

指定了一个值 $\mu_{\tilde{A}} \in [0,1]$ 与之对应，$\mu_{\tilde{A}}(x)$ 在元素 x 处的值反映了元素 x 属于 $\tilde{A}$ 的程度，称集合 $\tilde{A}$ 为模糊子集，而 $\mu_{\tilde{A}} \in [0,1]$ 称为 $\tilde{A}$ 的隶属函数。也就是说，$\mu_{\tilde{A}}(x)$ 的值越大，元素 x 属于 $\tilde{A}$ 的程度也就越高。

当前，模糊集理论发展很快，模糊技术几乎渗透到了所有领域。Zadeh(1965)[97]首先提出了模糊集和它的隶属函数的概念，Zadeh(1978)[114]提出了可能性理论，之后，许多研究者如 Nahmias[60]，Dubios 和 Prade[25]，Klir[40]，Liu[49, 50]，以及 Liu 和 Liu[51]等人丰富了模糊理论。尤其 Liu 等人[50]发展了一套完善的类似于概率论的，研究模糊性的公理体系，称之为可信性理论。

在模糊理论中，$\mathrm{Pos}\{A\}$ 描述了事件 A 发生的可能性。为了保证 $\mathrm{Pos}\{A\}$ 在实际中的合理性，它需要满足一些数学性质，其中有 4 条公理是必须满足的[49,50]。

假定 Θ 是一个非空集合，$P(\Theta)$ 是 Θ 的幂集。

公理 1 $P\{\Theta\}=1$。

公理 2 $P\{\Phi\}=0$。

公理 3 对于 $P(\Theta)$ 中的任意集合 $\{A_i\}$，$\mathrm{Pos}\{\cup_i A_i\} = \sup_i \mathrm{Pos}\{A_i\}$

公理 4 如果 Θ_i 是非空集合，其上定义的 $\mathrm{Pos}\{\bullet\}$，$i=1,2,\cdots,n$，满足

前三条公理，并且 $\Theta=\Theta_1\times\Theta_2\times\cdots\times\Theta_n$，且对于每个 $A\in P(\Theta)$，

$$\mathrm{Pos}\{A\}=\sup_{(\theta_1,\theta_2,\cdots,\theta_n)\in A}\mathrm{Pos}_1\{\theta_1\}\wedge\mathrm{Pos}_2\{\theta_2\}\wedge\cdots\mathrm{Pos}_n\{\theta_n\}$$

记做 $\mathrm{Pos}=\mathrm{Pos}_1\wedge\mathrm{Pos}_2\wedge\cdots\wedge\mathrm{Pos}_n$。

为了定义可能性测度，Nahmias 在文献[60]给出了前三条公理，为了定义乘积可能性测度，Liu 在文献[49,50]给出了第四条公理，并且证明了 $\mathrm{Pos}=\mathrm{Pos}_1\wedge\mathrm{Pos}_2\wedge\cdots\wedge\mathrm{Pos}_n$ 满足前三条公理。

定义 1.2　假定 Θ 是一个非空集合，$P(\Theta)$ 是 Θ 的幂集，如果 Pos 满足前三条公理，则称为可能性测度。

定义 1.3　假定 Θ 是一个非空集合，$P(\Theta)$ 是 Θ 的幂集，Pos 是定义在 $P(\Theta)$ 上的可能性测度，则三元组（$\Theta,P(\Theta),\mathrm{Pos}$）称为可能性空间。

定义 1.4　假定（$\Theta,P(\Theta),\mathrm{Pos}$）是可能性空间，$A$ 是幂集 $P(\Theta)$ 中的一个元素，则 A 的必要性

$$\mathrm{Nec}\{A\}=1-\mathrm{Pos}\{A^c\}\qquad(1\text{-}2)$$

其中，A^c 是 A 的对立事件。

定义 1.5[49, 50]　假定 $(\Theta,P(\Theta),\mathrm{Pos})$ 是可能性空间，A 是幂集 $P(\Theta)$ 中的一个元素，则 A 的可信性测度为

$$\mathrm{Cr}\{A\}=\frac{1}{2}(\mathrm{Pos}\{A\}+\mathrm{Nec}\{A\})\qquad(1\text{-}3)$$

定义 1.6[60]　假设 ξ 是一个从可能性空间（$\Theta,P(\Theta),\mathrm{Pos}$）到实直线 R 的函数，则称 ξ 是一个模糊变量。

假设 ξ 是可能性空间 $(\Theta,P(\Theta),\mathrm{Pos})$ 上的模糊变量，它的隶属函数可由可能性测度 Pos 导出，即

$$\mu(x)=\mathrm{Pos}\{\theta\in\Theta\mid\xi(\theta)=x\},\quad x\in R\qquad(1\text{-}4)$$

定义 1.7　假定 ξ 是可能性空间（$\Theta,P(\Theta),\mathrm{Pos}$）上的模糊变量，一个模糊事件 $\{\xi\geqslant r\}$ 的可能性、必要性和可测性测度分别为

$$\begin{aligned}\mathrm{Pos}\{\xi \geqslant r\} &= \sup_{u \geqslant r} \mu(u) \\ \mathrm{Nec}\{\xi \geqslant r\} &= 1 - \sup_{u < r} \mu(u) \\ \mathrm{Cr}\{\xi \geqslant r\} &= \frac{1}{2}(\mathrm{Pos}\{\xi \geqslant r\} + \mathrm{Nec}\{\xi \geqslant r\})\end{aligned} \tag{1-5}$$

这里 μ 是 ξ 的隶属函数。

可以在 Liu 文献[49,50]中找到模糊事件 $\{\xi \geqslant r\}$ 的必要性和可测性测度的定义。

定义 1.8[49, 50] 假定 ξ 是可能性空间 $(\Theta, P(\Theta), \mathrm{Pos})$ 上的模糊变量，则 ξ 的期望值可被定义成如下形式

$$E[\xi] = \int_0^{\infty} \mathrm{Cr}\{\xi \geqslant r\}\mathrm{d}r - \int_{-\infty}^{0} \mathrm{Cr}(\xi \leqslant r)\mathrm{d}r \tag{1-6}$$

右端的两个积分中至少有一个是有限的（为了避免出现 $-\infty \sim \infty$ 情形，要求上式右端中两个积分至少有一个有限）。

Liu 等人在文献[50]中分别给出了离散型模糊变量和连续型模糊变量期望值的计算方法。具体方法描述如下：

情形 1：设 ξ 为离散型模糊变量，其隶属函数为

$$\mu(x) = \begin{cases} \mu_1, & x = a_1 \\ \mu_2, & x = a_2 \\ \vdots & \vdots \\ \mu_N, & x = a_N \end{cases}$$

不失一般性，假设 $a_1 \leqslant a_2 \leqslant \cdots \leqslant a_N$。由定义 1.5 可知，模糊变量 ξ 的期望值为

$$E[\xi] = \sum_{i=1}^{N} \omega_i a_i$$

其中，权重 $\omega_i (i = 1,2,\cdots,N)$ 分别为：

$$\omega_1 = \frac{1}{2}(\mu_1 + \max_{1 \leqslant j \leqslant N} \mu_j - \max_{1 < j \leqslant N} \mu_j),$$

$$\omega_i = \frac{1}{2}(\max_{1 \leqslant j \leqslant i} \mu_j - \max_{1 \leqslant j < i} \mu_j + \max_{i \leqslant j \leqslant N} \mu_j - \max_{i < j \leqslant N} \mu_j), \quad 2 \leqslant i \leqslant N-1$$

$$\omega_N = \frac{1}{2}(+\max_{1\leqslant j\leqslant N}\mu_j - \max_{1\leqslant j<N}\mu_j - \mu_N)。$$

情形 2：设ξ为连续型模糊变量，其隶属函数为μ。按照 Liu 和 Liu[50]提出的方法，首先从模糊变量ξ的δ-水平集上均匀产生N个样本点$a_i(i=1,2,\cdots,N)$，就得到一个具有隶属函数$\mu_{\xi'}(a_i)=\mu_{\xi}(a_i)(i=1,2,\cdots,N)$的新离散型模糊变量$\xi'$，这样就可以根据情形 1 计算出模糊变量$\xi'$的期望值，然后用$\xi'$的期望值来估计模糊变量$\xi$的期望值，前提条件是$N$足够大。

Liu 和 Liu[50]也给出了一些特殊模糊变量期望值的求解。

例 1　$\xi=(a,b,c)$是一个三角模糊变量，且有$a\leqslant b\leqslant c$，则它的期望值为

$$E[\xi]=\frac{1}{4}(a+2b+c) \tag{1-7}$$

例 2　$\xi=(a,b,c,d)$是一个梯形模糊变量，且有$a\leqslant b\leqslant c\leqslant d$，则它的期望值为

$$E[\xi]=\frac{1}{4}(a+b+c+d) \tag{1-8}$$

定理 1.1[49,50]　假设ξ和 η是相互独立的模糊变量，且期望值有限，则对任意的实数a和b有

$$E[a\xi+b\eta]=aE[\xi]+bE[\eta] \tag{1-9}$$

在模糊环境下不存在一种自然的模糊排序方法，Liu 等人[49,50]给出了按模糊变量的期望值进行模糊排序的方法。

定理 1.2[49, 50]　当且仅当$E[\xi]>E[\eta]$，有$\xi>\eta$。

1.1.2　模糊模拟技术

在不确定理论中，涉及对不确定函数的期望值进行计算的问题。但在大多数现实问题中，由于函数结构的复杂性，使得我们无法通过解析

方法求得这些函数的期望值，本文将采用模拟技术对这些函数的期望值进行估计。下面介绍计算不确定函数的期望值的模糊模拟。

假设 $f: R^n \to R$ 是一个可测函数，$\xi = f(\xi_1, \xi_2, \cdots, \xi_n)$ 是可能性空间 $(\Theta, P(\Theta), \mathrm{Pos})$ 的模糊变量，则 $f(\xi)$ 也是一个模糊变量，它的期望值定义为

$$E[\xi] = \int_0^{+\infty} \mathrm{Cr}\{f(\xi) \geqslant r\}\mathrm{d}r - \int_{-\infty}^{0} \mathrm{Cr}(f(\xi) \leqslant r)\mathrm{d}r \qquad (1\text{-}10)$$

Liu 等人在文献[51]中给出了计算模糊变量 $f(\xi)$ 期望值的模糊模拟技术，利用模糊模拟技术可近似计算出 $f(\xi)$ 的期望值 $E[f(\xi)]$。

估计函数 $f(\xi)$ 的期望值 $E[f(\xi)]$ 的模糊模拟方法如下：

（1）令 $e=0$。

（2）从 Θ 中随机产生 θ_k 使得 $\mathrm{Pos}\{\theta_k\} \geqslant \varepsilon$（$k=1,2,\cdots,N$），这里 ε 是一个足够小的数。

（3）令 $v_k = \mathrm{Pos}\{\theta_k\}$。

（4）令 $a = f(\xi(\theta_1)) \wedge f(\xi(\theta_2)) \wedge \cdots \wedge f(\xi(\theta_N))$，

$b = f(\xi(\theta_1)) \vee f(\xi(\theta_2)) \vee \cdots \vee f(\xi(\theta_N))$。

（5）从 $[a,b]$ 随机产生 r。

（6）如果 $r \geqslant 0$，那么 $e \leftarrow e + \mathrm{Cr}\{f(\xi) \geqslant r\}$，这里

$$Cr\{f(\xi) \geqslant r\} = \frac{1}{2}\left(\max_{1\leqslant k\leqslant N}\{v_k \mid f(\xi(\theta_k)) \geqslant r\} + \min_{1\leqslant k\leqslant N}\{1 - v_k \mid f(\xi(\theta_k) < r\}\right);$$

如果 $r \leqslant 0$，那么 $e \leftarrow e - \mathrm{Cr}\{f(\xi) \leqslant r\}$，这里

$$Cr\{f(\xi) \leqslant r\} = \frac{1}{2}\left(\max_{1\leqslant k\leqslant N}\{v_k \mid f(\xi(\theta_k)) \leqslant r\} + \min_{1\leqslant k\leqslant N}\{1 - v_k \mid f(\xi(\theta_k) > r\}\right)$$

（7）重复步骤（5）～（6）N 次。

（8）计算 $f(\xi)$ 的期望值

$$E[f(\xi)] = a \vee 0 + b \wedge 0 + e(b-a)/N$$

（9）返回 $E[f(\xi)]$。

1.2　粗糙理论

Zazislaw Pawlak[69] 在 1980 年初引进了粗糙集（简称粗集）理论来处理数据表的分类分析，引入粗集可从数据中提取人们所感兴趣的模式，这一节首先介绍粗集理论的一些基本概念。

1.2.1　粗糙集

设 $U=\{u_1,u_2,\cdots,u_n\}$ 是一个论域，R 是 U 上的一个等价关系（也称为不可识别关系），那么 $A=(U,R)$ 就被称之为一个近似空间。对任一集合 $X\subseteq U$ 都可以被定义成一个粗集，这个粗集是由一个上近似和下近似组成的一个区间，其形式为 $(\underline{R}X,\overline{R}X)$。

根据 Pawlak 的定义[69]，$X\subseteq U$ 的上近似，下近似可表示为：

$$\underline{R}X=\{x\in U\mid [x]\subseteq X\},$$

$$\overline{R}X=\{x\in U\mid [x]\cap X\neq\varnothing\},$$

这里 $[x]$ 表示在关系 R 上的包含元素 x 的等价类。

X 是关于 R 粗糙的，当且仅当 $\overline{R}X\neq\underline{R}X$，否则 X 就是可识别的，因此 X 的粗集被定义成 $A_R(X)=(\underline{R}X,\overline{R}X)$。

$\underline{R}X$ 表示 X 的下近似，说明下近似中的所有元素均属于 X，$\overline{R}X$ 表示 X 的上近似，说明上近似中的元素可能属于也可能不属于 X。

任一集合 $X\subseteq U$ 的上近似和下近似需满足如下的条件。

(1) $\varnothing\subseteq\underline{R}X_i\subseteq X_i\subseteq\overline{R}X_i\subseteq U$

(2) $\underline{R}X_i\cap\underline{R}X_j=\varnothing$，$i\neq j$

(3) $\underline{R}X_i\cap\overline{R}X_j=\varnothing$，$i\neq j$

如果一元素 u_k 不属于任何一个粗集的下近似，那么它一定属于两个以上粗集的上近似。

等价关系 R 把集合 U 划分成若干个不相连的子集，这种划分可表示成如下形式：

$$U/R=\{X_1,X_2,\cdots,X_l\}\ ,$$

这里 X_i（$1\leqslant i\leqslant l$）是 R 上的一个等价类。如果有两个模式 $u,v\in U$ 属于同一个等价类 $X_i\subseteq U/R$，即认为 u 和 v 是不可识别的。由于不可能区分同一个等价类中的模式，因此不可能得到任意集合 $X\subseteq U$ 在近似空间 A 上的精确描述，然而，任意 X 都可用它的上近似和下近似 $(\underline{R}X,\overline{R}X)$ 来表示。下近似 $\underline{R}X$ 是 X 的子集的那些集合的并集，上近似 $\overline{R}X$ 是与 X 的交集不为空的那些集合的并集。

1.2.2 粗糙变量

Pawlak[69]在 1980 年初提出了粗糙集的概念，并应用到数据表的分类分析中。在处理具有模棱两可性质的对象时，它是一种非常好的数学工具。当利用描述某一对象时，如果无法准确地表示该对象，其中一种方法就是用其他集合来近似表示该对象集合。当前集合称为边界不确定的粗糙集，我们可以用清晰的集合来表示粗糙集，即它的上近似和下近似，它们都是定义在一种等价关系之上的。

在描述粗糙变量时，Liu 在文献[49]中给出了以下四条公理。

假定 Λ 是一个非空集合，$\mathcal{A}$ 是 Λ 子集上的一个 σ-代数，Δ 是 $\mathcal{A}$ 中的一个元素，π 是 $\mathcal{A}$ 上的一个实值集函数，则满足下面四条公理：

公理 1 $\pi\{\Lambda\}<+\infty$

公理 2 $\pi\{\Delta\}>0$

公理 3 $\pi\{A\}\geqslant 0\quad \forall A\in\mathcal{A}$

公理 4 对于彼此不相交的事件的可数序列 $\{A_i\}_{i=1}^{\infty}$，有

$$\pi\{\bigcup_{i=1}^{\infty}A_i\}=\sum_{i=1}^{\infty}\pi\{A_i\} \tag{1-11}$$

事实上，满足这四条公理的集合函数 π 显然是一个测度，三元组

$(\lambda,\mathcal{A},\pi)$是一个测度空间。

定义 1.8[49] 假定λ是一非空集合，$\mathcal{A}$是λ子集上的一个σ-代数，Δ是$\mathcal{A}$中的一个元素，π是$\mathcal{A}$上的一个实值集函数，且满足以上四条公理，则$(\lambda,\Delta,\mathcal{A},\pi)$被称为一个粗糙空间。

定义 1.9[49] 一个粗糙变量η是一个从粗糙空间$(\lambda,\Delta,\mathcal{A},\pi)$)到实数集的可测函数。即是说对$\mathfrak{R}$的任意 Borel 集$B$，有

$$\{\lambda\in\Lambda\,|\,\eta(\lambda)\in B\}\in \mathrm{A} \tag{1-12}$$

粗糙变量η的下近似和上近似可被定义成如下形式：

$$\underline{\eta}=\{\eta(\lambda)\,|\,\lambda\in\Delta\} \tag{1-13}$$

$$\overline{\eta}=\{\eta(\lambda)\,|\,\lambda\in\Lambda\} \tag{1-14}$$

引理 1[49] 由于$\Delta\subset\Lambda$，显然$\underline{\eta}\subset\overline{\eta}$。

$\underline{\eta}$是粗糙变量η的下近似，它是一个集合，其中的元素（模式）一定属于粗糙变量η所属集合。$\overline{\eta}$是粗糙变量η的上近似，它也是一个集合，其中的元素（模式）可能属于，也可能不属于粗糙变量η所在集合。

第 2 章　Web 挖掘

基于 www 的全球信息系统的发展使人们沉浸在信息的海洋之中，大量信息在给人们带来便利的同时也带来了诸多问题：如人们很难从海量数据中发现有用知识，数据丰富、知识贫乏已经成为一个典型问题。Data Mining（数据挖掘）的目的就是有效地从海量数据中提取出需要的答案，实现“数据—信息—知识—价值”的转变过程。

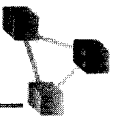
2.1　数据挖掘与 Web 挖掘

数据挖掘(Data Mining)，简单说来，就是从大量数据中挖掘或发现隐藏在其中的知识。它是在海量数据中探索，并从中发现并提取隐藏在其中的有用信息的一种处理分析过程，是一门综合了数据库技术、人工智能、神经网络、模式识别、统计学、决策树、贝叶斯分析、遗传算法、模糊集、粗糙集等领域基础并吸收了大量新颖思想的非常活跃的交叉学科，是 KDD(KnowledgeDiscovery in Databases)过程中的一个关键环节。

数据挖掘是知识发现过程中的一个关键环节，完整的知识发现过程由以下步骤组成：

（1）数据清洗。即去除噪声和不一致的数据。

（2）数据集成。把多个数据源中的数据集成到一起。



（3）数据选择。即从数据库中选择与分析任务相关的数据。

（4）数据变换。通过汇总、聚集等方法，把数据转换为适于进行挖掘的形式。

（5）数据挖掘。采用智能的方法来提取数据中的模式。

（6）模式评价。选取正确的有用的模式。

（7）知识展现。采用可视化的展现技术把知识展示给用户。

在过去的十年间，www 上的信息量呈爆炸式增长。今天，我们可以通过 Web 浏览器方便地存取 www 上的信息。超过 1 000 000 000 的网页被搜索引擎进行索引，然而，要发现我们真正所需的信息却不是一个简单的任务。人们迫切的需要开发一种自动从 www 挖掘的技术，因此人们提出了 Web 挖掘这个术语。

Web 是一个庞大的不受控制的异构文档集合，它具有庞大性、多样性以及动态性，由此带来了相关的度量性、异构性和动态性问题[64]。

数据挖掘就是从数据中识别合理的、新颖的、潜在有用的，而且最终能被理解的模式的一个复杂过程，因此人们认为 Web 挖掘是数据挖掘的一个研究领域。

Web 挖掘通常被定义为从 www 上发现和分析有用信息[64]。在 Web 挖掘中，数据可来自于服务器端、客户端、代理服务器端，或者来自一个机构数据库。

数据源不同，所收集的数据类型也不同，也就使得对一个特定任务所使用的挖掘技术不同。

Web 数据具有如下特点[64]：

（1）未标示；

（2）分布性；

（3）异构性；

（4）半结构化；

（5）随时间变化；