Web YONGHU SHIYONG MOSHI YU XINGQU WAJUE FANGFA YANJIU

Web用户使用模式与兴趣挖掘方法研究

朱志国◎著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

教育部人文社会科学研究青年项目（编号：12YJCZH321）阶段性研究成果
辽宁省教育厅高校学术专著出版基金资助

Web YONGHU SHIYONG MOSHI YU XINGQU WAJUE FANGFA YANJIU

Web用户使用模式与兴趣挖掘方法研究

朱志国◎著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目(CIP)数据

Web用户使用模式与兴趣挖掘方法研究/朱志国著.—北京：北京师范大学出版社，2012.6
ISBN 978-7-303-14164-7

Ⅰ.①W… Ⅱ.①朱… Ⅲ.①主页制作－程序设计－研究 Ⅳ.①F393.092

中国版本图书馆CIP数据核字（2012）第019609号

营销中心电话 010-58802181 58805532
北师大出版社高等教育分社网 http://gaojiao.bnup.com.cn
电子信箱 beishida168@126.com

出版发行：北京师范大学出版社 www.bnup.com.cn
北京新街口外大街19号
邮政编码：100875
印　　刷：北京京师印务有限公司
经　　销：全国新华书店
开　　本：170 mm × 230 mm
印　　张：8.5
字　　数：153千字
版　　次：2012年6月第1版
印　　次：2012年6月第1次印刷
定　　价：16.00元

策划编辑：胡　宇　　责任编辑：岳庆昌　胡　宇
美术编辑：毛　佳　　装帧设计：李尘工作室
责任校对：李　菡　　责任印制：李　啸

前 言

随着互联网和万维网的迅速发展，用户访问信息广泛、呈海量式增长．这些信息从用户维、时间维、空间维、访问对象维等方面详尽地反映出用户的访问细节．对这些细节信息再进一步挖掘之后，就可以发现隐藏其中的一些更深层次的知识和规律——用户(用户群)的使用模式和访问兴趣．这些知识可以广泛应用于 Web 个性化服务、系统改进以及商业智能等领域．针对这个具有广泛而深远意义的研究课题，本书借鉴国内外相关学科知识，系统深入地介绍了从 Web 访问信息的收集分析到用户兴趣偏好访问模式挖掘的相关理论、方法和技术．本书从第 3 章至第 5 章的最后部分，对于提出的模型算法均进行了实例验证及讨论．主要内容有：

1．首先从用户使用模式挖掘过程中的四个主要阶段，即数据采集、数据预处理、模式发现以及模式分析，宏观综述了国内外学者一些经典和最新的研究进展，并对这些研究成果进行详细的整理、归纳与分析，力求展现出这个研究领域的全貌．

2．在 Web 使用数据预处理技术层面，Web 用户会话的识别与构建是其中一个非常关键的步骤．针对于此，本书在第 3 章提出了一个基于用户访问 URL 语义分析的会话识别方法．这个方法借助 Web 目录服务对 URL 记录进行概念化，为 Web 日志中的每一条 URL 访问记录赋予一定的语义信息，在此基础上再根据一些测度指标定义对 URL 之间的语义相似度进行评价，并建立预设时间间隔内的 URL 间语义距离矩阵．然后在静态和动态的 Web 日志情况

下，分别给出了两类日志数据的语义奇异值鉴别方法：SOA_S和SOA_D.

3. 在Web使用模式与兴趣挖掘方法层面上，本书第4章以Web用户访问信息的历史变化特性为视角，给出了一个Web用户聚类方法. 在这个方法体系中，首先需要依次构造出每个用户的历史访问序列树：E-WAS树和H-WAS树. 然后从H-WAS树中抽取出持久偏爱的Web访问模式PP-WAP作为Web用户的聚类特征. 接下来，根据本书定义的一些PP-WAP的相似度判定方法对用户的相似性进行度量，并且选用重要的划分聚类方法K-Medoid算法对用户相似度矩阵进行聚类计算.

4. 在Web使用模式与兴趣挖掘方法层面上，本书第5章以用户的访问兴趣为出发点，基于经典的隐马尔可夫模型建立了两个Web用户兴趣浏览路径模型：INPM和SINPM^{Pe}，并给出了从这两个模型中发现用户兴趣关联模式的方法. 这些发现的用户兴趣关联模式不仅可以反映出用户访问路径上的时间特性，而且更多的是反映了带有用户访问兴趣特性的最佳关联路径信息. 实验结果表明，提出这个兴趣关联模式发现方法的确是一个高效、扩展性良好的用户兴趣路径序列挖掘方法.

本书突出理论方法的系统性，强调方法的实用性和创新性，理论重点突出，详略得当. 知识管理、Web数据挖掘以及智能电子商务等方向研究工作的科研人员和工程技术人员在研究工作中，可以使用本书作为参考借鉴. 此外，本书也可以作为大专院校计算机应用、管理科学与工程、知识管理等学科的研究生及教师的教学用书或参考书.

目 录

第1章 引 言

Web用户使用模式与兴趣挖掘技术是随着互联网和万维网的迅速普及、发展而产生的一个新兴的、重要的研究领域. 它是利用数据挖掘方法的原则和思想，针对Web用户访问信息的新特性，对传统数据挖掘方法进行扩展和改进，力图从Web访问信息中发现一些有用的知识和规律. 本书正是从现实的需要和数据的特性出发，研究如何根据Web访问信息的特性，提出一些新颖高效的用户(用户群)使用模式和访问兴趣挖掘方法. 本章首先阐明相关的研究背景、意义与价值，并对本书研究工作的新特点进行分析. 最后介绍本书的整体内容章节安排.

1.1 相关研究背景

随着互联网在流量、规模和复杂度等方面的飞速增长，万维网已经成为一个涉及诸如新闻、广告、消费信息、金融管理、远程教育、政府网站、电子商务等信息的分布广泛的巨大知识宝库和信息海洋，而且万维网提供网络信息服务的竞争也日益激烈.

与此同时，一方面互联网上存储了大量文档、图形、图像等类型的数据，表现出了Web数据的多样性；另一方面，具有不同背景、兴趣和使用目的的用户可以随意地链接到大部分的互联网站点，Web用户也呈现出多样性的特点. 因此，互联网在给人们工作生活带来极大便利和丰富信息资源的同时，也产生了以下一些亟待解决的问题：

(1)难以准确获得所需要的信息. 虽然互联网上存储了海量数据，但由于Web是无结构的、动态的，并且Web页面的复杂程度远远超过了文本文档，给人们准确查找和定位所需要的信息带来了极大的困难. 虽然当前Web用户可以在一些搜索引擎工具中输入关键词来进行Web信息的检索，但由于检索结果的低精度和低召回率，检索的效果仍然不够理想.

(2)难以获得信息中潜藏的知识. 多样的、海量Web数据中蕴涵着许多有用的、潜在的且不容易被发现的知识和模式，正如John Naisbitt在《大趋势》一书中所说："人类正被数据淹没，但却渴望知识". 的确如此，在浩瀚的万维网信息海洋中，人们迫切需要一些发现知识和模式的方法及工具.

(3)欠缺个性化的信息服务. 不同层次、不同爱好和使用目的的浏览者需要个性化的信息服务. 这个问题涉及Web站点的管理、组织和经营. Web站点的经营和管理者为提高网站的声誉和效益，需要了解其用户究竟需要什么，其中包括根据大多数用户的共同兴趣，开展带有共性的优质Web服务以及针对特定用户开展的个性化服务. 一方面，从站点经营者来说，他们需要好的自动辅助设计工具，可以根据用户的访问兴趣、访问频度、访问时间动态地调整页面结构，改进服务，开展有针对性的电子商务以更好地满足访问者的需求；另一方面，从访问者来说，他们希望看到的是切合自身需求的个性化页面，希望从其他具有类似访问兴趣的用户访问行为中得到启发，从而得到更好的满足各自需求的服务，但是这些需求从某种意义上说，访问者本身也未必十分清楚.

要解决上述问题，必须根据Web数据的特性，对传统的数据挖掘方法进行扩展和改进，并结合统计学、计算机网络、数据库与数据仓库、可视化等众多领域的技术来对Web中蕴涵的多种信息进行挖掘. 因此，自动从Web文档和服务中发现知识的Web挖掘技术(Web Mining，具体见定义1.1)[1~3]已经成为数据挖掘领域中的一个研究热点.

定义1.1 Web挖掘是指从大量Web文档结构和使用的集合C中发现隐含的模式p. 如果将C看做输入，p看做输出，那么Web挖掘的过程就是从输入到输出的一个映射ξ：$C\rightarrow p$.

如图1-1所示，一般地，从Web挖掘的对象(Web中包含的Web页面的内容信息、丰富的超链接信息以及Web页面的访问和访问信息)可以将Web挖掘分为三类[4]：Web内容挖掘、Web结构挖掘和Web使用挖掘.

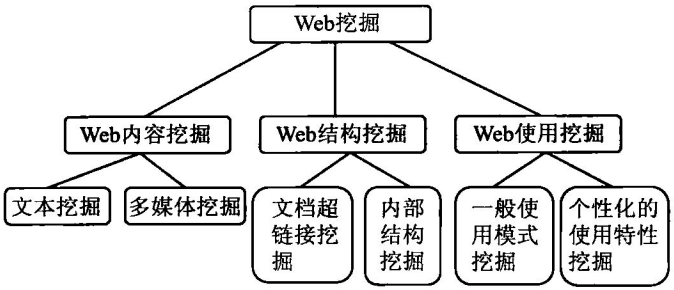

图 1-1　Web 挖掘分类图

(1)Web 内容挖掘(Web Content Mining). 从 Web 上的网页内容及其描述信息中获取潜在的、有价值的知识模式，以实现 Web 资源的自动检索，提高 Web 数据的利用率. 这项技术可以细分为 Web 文本挖掘[5]和 Web 多媒体挖掘[6](Multimedia Data Mining). Web 文本挖掘是对 Web 上大量文档集合的内容进行总结、分类、聚类和关联分析等. Web 多媒体挖掘是从 Web 多媒体数据(如音频、视频和图像等)中抽取事先未知的、隐藏的、完整的和新颖的知识.

基于最新的人工智能技术从万维网上更智能地提取信息的搜索工具，包括 Intelligent Web Agent Information Filtering/Categorization[7], Personalized Web Agents[8]. 它们可以对 HTML 页面内容进行挖掘、对页面中的文本进行文本挖掘以及对页面中的多媒体信息进行挖掘，也包括对页面内容作摘要、分类、聚类、过滤、信息提取、关联规则发现以及某种程度的个性化.

(2)Web 结构挖掘(Web Structure Mining). 主要通过对 Web 站点的超链接结构进行分析、变形和归纳，将 Web 页面进行分类，以利于信息的搜索. Web 不仅由页面组成，而且还包含了从一个页面指向另一个页面的超链接. 超链接信息包含了人类潜在的注释，大量的 Web 超链接信息提供了关于 Web 页面内容相关性、质量和结构方面的信息，有助于自动推断出页面的权威性(authority). 当一个页面包含指向另一个页面的超链接时，可以认为是对另一个页面的认可. 把对一个页面的不同注解收集起来，就可以用来反映该页面的重要性. 这就类似于信息检索领域中使用论文引用情况来评估该论文的质量. 发现的这种知识可以被用来改进传统的搜索引擎. 目前的主要方法是 PageRank[9]和 HITS[10].

(3)Web使用挖掘(Web Usage Mining)[11,12]. 对用户和Web站点交互过程中留下的访问记录数据进行挖掘. 这些数据包括：Web服务器的访问记录、代理服务器日志文件、浏览器日志记录、用户注册信息、用户对话或交易信息等. 这些Web访问信息体现了用户使用Web资源的行为特性，以及隐藏在行为背后的更深层次的动因和规律. Web访问信息的挖掘作为Web挖掘的一个重要组成部分，有其独特的理论和实践意义. 其研究主要有两个方向：一般使用模式挖掘和个性化的使用特性挖掘.

①一般使用模式挖掘. 主要是面向群体Web用户. 通过Web使用挖掘，对总的用户访问行为、频度、内容等进行分析，可以得到关于群体用户访问行为和访问方式的普遍知识，用以改进Web服务方设计. 更重要的是，通过对这些用户特征的理解和分析，可以有助于开展有针对性的电子商务活动，例如改进站点的组织结构提供更高效的访问；吸引用户，保持用户；用户的群体推荐；用户群体的地区、行业、阶层分析.

②个性化的使用特性挖掘. 面向群体中的每个用户. 通过Web使用挖掘，对每个用户访问行为、频度、内容等进行分析，提取出每个用户或用户群的特征，给每个用户或用户群提供个性化的电子商务服务. 比如个性化推荐，即如果用户的访问兴趣与其他一些用户很相似，那么就考虑将那些用户也感兴趣的一些东西推荐给该用户；用户建模，即根据用户的曾经访问记录，推断当前访问用户的特征；个性化推销(Direct Marketing)，即识别出对某种产品或服务的可能购买者，然后对其推荐相应的产品或服务.

本书正是要深入研究Web使用挖掘的这两个方面：群体用户的一般使用模式挖掘和用户个性化的使用特性挖掘.

最后，在表1-1[13]中对Web内容挖掘、Web结构挖掘和Web使用挖掘中的数据特征、表现形式、挖掘方法以及应用领域等方面进行了总结与比较.

表 1-1　Web 内容挖掘、结构挖掘和使用挖掘比较

	Web 挖掘			
	Web 内容挖掘		Web 结构挖掘	Web 使用挖掘
	信息检索领域	数据库领域		
数据	—文本文档 —超文本文档	—超文本文档	—链接结构	—Web 服务器日志 —Proxy 日志 —浏览器日志
数据特征	—非结构化 —半结构化	—半结构化 —视 Web 站点为一个数据库	—链接结构	—交互式数据
表现形式	—无序/有序的单词集合 —术语和短语 —概念/实体 —关系曲线	—对象交换模型(OEM) —关系曲线	—图	—关系表 —图
挖掘方法	—TFIDEF —机器学习 —统计	—专利算法 —关联规则及变形	—专利算法 —HITS 算法	—机器学习 —统计 —关联规则及变形 —聚类 —序列模式
应用领域	—分类 —聚类 —寻找抽取规则 —寻找文本模式 —用户建模	—发现频繁子结构 —提取 Web 站点大纲	—分类 —聚类	—站点结构管理及优化 —网络销售 —用户建模 —推荐系统

1.2 研究意义与价值

本书的研究目标是挖掘隐藏在Web访问信息中的一些用户使用模式和兴趣偏好. 基于Web访问信息自身的特点与内涵以及挖掘方法层面的考虑，使得本书的研究工作具备以下一些意义与特点[14]：

(1)Web访问信息数据是大规模的海量数据信息，其分布广泛，形态多样，具有丰富的内涵而且结构化程度高.

①数据是大规模且海量的. 一个中等规模的网站每天可以记载几兆比特的用户访问信息、数万次用户的访问. 随着时间的推移，所记载的用户访问信息量更是非常庞大.

②数据广泛分布于世界各处. 世界上每台Web服务器或Web代理服务器都会遵循W3C[15]的Web访问信息标准，记录来自不同地区、种族、阶层等访问者的浏览信息.

③数据时时刻刻地产生. 只要用户对站点进行访问，用户访问信息就会被记录；只要用户访问互联网，就必然至少有一个服务器记录其访问行为.

④访问信息形态多样. 在W3C标准的基础上，各个服务器可以根据各自特定的需求，制定新的扩展格式，以记载更加详细的用户访问信息. 访问信息格式的扩展，是当前Web服务发展的一个新趋势.

⑤访问信息具有丰富的内涵. 访问信息记载了来访者、被访问的页面、访问时间等一系列信息. 当这些信息被事务化后，可以从中提取出访问页面、访问路径和访问时间等特性，并将这些特性和网站的拓扑结构和内容分布信息结合起来. 这样，这些信息就具有了非常丰富的内涵.

⑥结构化程度高. 访问信息一般都按照确定的数据格式由系统自动记录，遵循W3C标准的访问信息记录格式，可以很方便地转化成关系式数据库进行结构化的处理，便于进行分类、聚类、统计分析和深层次的挖掘.

(2)Web访问信息记录的是每个用户的访问行为，代表每个用户的个性；或者是记录了同一类用户的访问，代表同一类用户的特性. 同时，一段时期的访问数据记载的又是群体用户的访问行为和群体用户的共性.

①每个用户的访问特点可以用来辨识该用户的特性.

②群体用户的访问行为可以被分割为不同的类别，以此来体现各个类别用户的共同特性.

③基于同一类用户的特性，可以给该类中的每个用户提供相应的推荐服务.

④群体用户特性可以用来改变站点的设计结构，方便群体用户的访问.

(3)Web访问信息中挖掘用户使用模式方法与传统数据挖掘方法相比，也面临着一些需要解决的新问题.

①在传统的关系数据库中，一条记录的各个字段之间不存在顺序关系，传统的基于事务的挖掘方法也无法处理事务内部各元素之间的顺序关系. 而在本书的研究工作中，事务元素之间存在着丰富的顺序信息，能反映用户的访问习惯和兴趣.

②传统的关系数据库中，一条记录的各个字段之间不可再分，传统的基于事务的挖掘方法也无法处理事务内部各元素内部之间的关系. 而在本书中，访问事务的元素是Web页面，每个页面的内容可以被抽象出不同的关键字，用户实际上是对这些关键字发生兴趣. 用户对页面的访问顺序、访问量、访问时间等特性，实际是表现在这些关键字上. 因此，挖掘对象进一步转化为由这种丰富的关键字访问顺序、访问量以及访问时间等所组成的新数据.

③在传统的关系数据库中，一条记录的各个字段之间不存在时间关系，事实上，传统数据挖掘方法也无法处理事务内部各元素之间的时间关系. 而在本书中，访问事务的元素是Web页面，用户对每个页面存在一个不同的访问时长，而访问时间的长短在某种程度上代表了用户对该页面的访问兴趣程度.

因此，传统的基于关系数据的挖掘方法(如分类、聚类、关联规则)发现，统计方法等需要在结合Web访问信息数据特性的基础上，进行扩展、改进，以适应新的要求. 研究的侧重点在于针对访问信息本身的特点，结合现实需要，提出新的思路和方法.

从目前发展状况来看，Web使用挖掘在Web知识管理及其相关领域内具有广泛的用途. 主要的用途可以分为以下四类：

(1)Web个性化服务(Web Personalization)

Web个性化服务是指Web站点能够根据用户的喜好和需求自动调整Web站点的信息组织和表示. 一方面用户能够在它的帮助下迅速找到需要的信息，另一方面每个不同的用户(或用户群)可以获得不同的访问体验，使得Web站点更具吸引力. 该服务的关键问题是用户建模和行为预测. 实现此类功能的一个捷径是使用Web使用挖掘技术从使用数据中获取有价值的用户行为信息.

(2)系统改进(System Improvement)

系统改进包括两个方面：一是改进系统的运行性能，例如使用 Web 缓冲技术，改善网络交通状况，提高站点响应速度；二是优化 Web 站点的设计，例如调整页面间的连接结构，使其更加符合用户的使用习惯，防止用户在访问过程中迷航. Web 使用挖掘技术能够帮助 Web 站点管理者及时了解用户的访问行为，发现系统运行和设计中存在的问题，优化系统的性能和设计.

(3)商业智能(Business Intelligence)

了解客户的行为和需求一直是商业领域中的一项很重要的工作. 在 Web 站点(特别是电子商务站点)可以收集大量的客户行为数据. 这些数据包含着许多可能影响商业决策的用户行为信息. 例如，哪些用户经常访问你的站点？如何吸引更多的访问者？如何保持这些新老用户？由于数据量很大，手工分析几乎变成了不可能的任务. 适合处理大量用户访问信息的 Web 使用挖掘技术便在此处体现出了它的商业价值. 它能够帮助企业从海量数据中迅速发现与商业决策相关的信息，提高应变能力.

(4)使用描述(Usage Characterization)

这项应用属于科学研究的范围，其目的是观察 Web 用户的使用行为，发现一些内在规律. 研究的成果可以用于改进与 Web 相关的工具和协议. 此类研究需要收集和分析大量的使用数据，Web 使用挖掘技术自然成了一个非常合适的研究工具.

综上所述，由于 Web 访问信息能够从各方面详尽反映出用户访问 Web 的细节，是取之不尽、用之不竭的宝贵资源. 为了更加有效地开发利用这些宝贵而丰富的资源，开展 Web 使用挖掘这项研究工作就具有了广泛而深远的意义. 本书的主要研究目标就是要通过对这些具有海量、广泛、形态多样、内涵丰富、结构化程度高特性的 Web 访问信息数据进行收集、分析和处理，抽取和挖掘出一些群体用户一般访问模式以及个性化的用户兴趣偏好访问模式，以此更好地理解用户行为，改进站点结构，从而为用户提供更为优质的、智能的 Web 信息服务.

1.3　研究思路与内容安排

1.3.1　研究思路

如图 1-2 所示，本书的研究思路基于 Web 使用挖掘的两个关键技术层面展开，Web 使用数据预处理层面和 Web 使用模式与兴趣挖掘方法层面. 在访问信息预处理层面，在对现有的基于时间和基于引用的两种经典会话启发式识别方法进行深入研究的基础上，发现对于像用户同时使用多个浏览器发出 Web 请求这样的情况，这两种经典方法存在一定的局限性. 本书拟定提出一种基于 URL 语义分析的用户会话识别方法. 该方法力图从 Web 日志中找到访问日志中的语义奇异值来对会话进行分割，从而达到会话识别和构建的目的. 在 Web 使用模式的挖掘层面，本书拟定提出两类模式的挖掘方法：Web 用户聚类方法和用户兴趣关联模式发现方法. Web 用户聚类方法拟定的研究思路是将 Web 用户聚类特征的选取着眼于从 Web 用户访问数据的历史变化特性. 首先从访问数据的历史演变过程中提取一种新的知识——持久偏爱的 Web 访问模式 PP-WAP；然后将具有相似 PP-WAP 的用户聚集在一起. 用户兴趣关联模式发现方法拟定的研究思路是首先给出一些用户访问站点的兴趣的相关定义；接下来基于隐马尔可夫模型，建立两个用户兴趣浏览模型；最后根据这两个模型给出用户兴趣关联模式的发现方法. 这些兴趣关联模式实际上也就是带有某些或某个兴趣的用户访问站点的最佳路径序列.

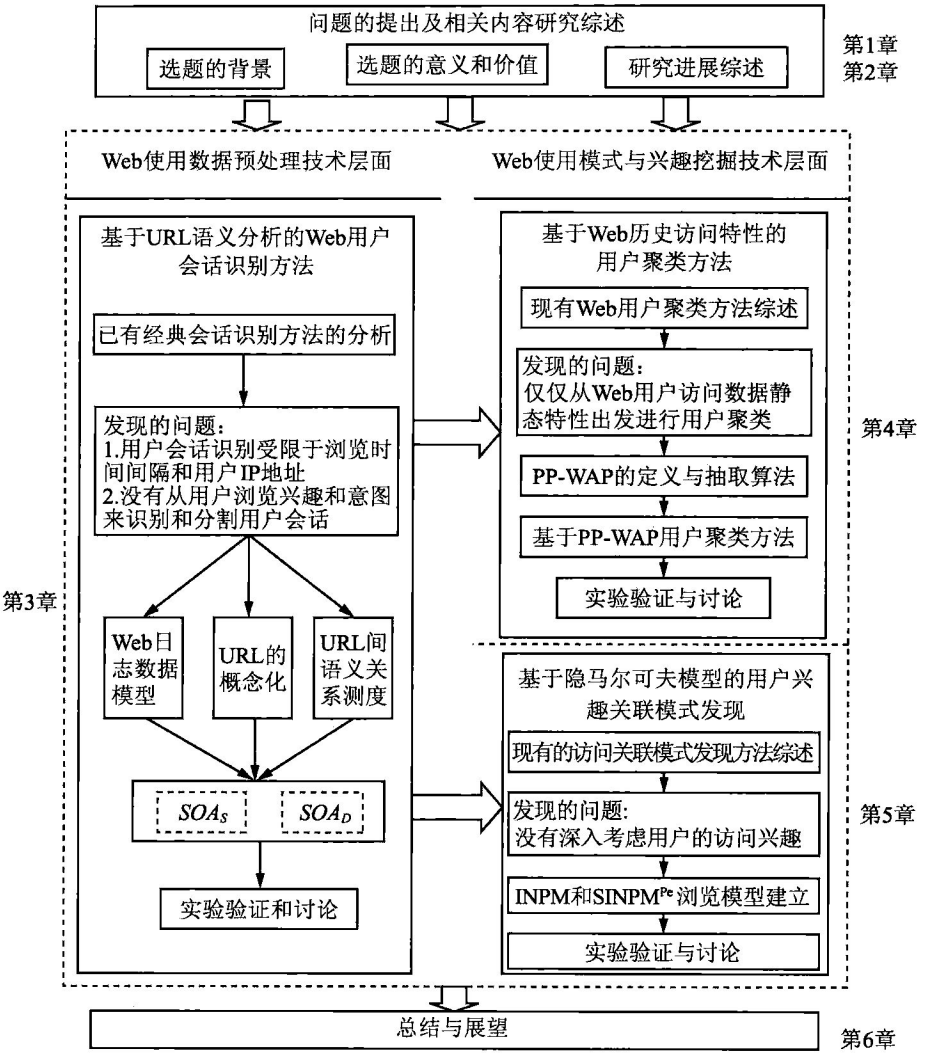

图 1-2　本书的研究思路框架图