WEB YONGHU CHAXUN RIZHI WAJUE YU YINGYONG

Web用户查询日志挖掘与应用

王继民 著

知识产权出版社
全国百佳图书出版单位
www.laichushu.com

Web 用户查询日志挖掘与应用

王继民　著

知识产权出版社
全国百佳图书出版单位

图书在版编目（CIP）数据

Web 用户查询日志挖掘与应用/王继民著. —北京：知识产权出版社，2014.4

ISBN 978-7-5130-2658-1

Ⅰ. ①W… Ⅱ. ①王… Ⅲ. ①计算机网络—情报检索—研究 Ⅳ. ①G354.4

中国版本图书馆 CIP 数据核字（2014）第 055574 号

内容提要

本书介绍了互联网用户查询日志挖掘及其应用研究领域的主要技术、方法与实证研究成果。全书由 3 篇共 14 章内容组成。其中，上篇对搜索引擎用户日志与移动搜索用户日志的研究现状进行了系统的分析，给出了 Web 用户查询日志挖掘研究框架；中篇介绍了基于不同类型用户日志所开展的实证研究结果，包括大规模 Web 搜索引擎系统的用户日志、大型期刊数据库的用户日志、移动搜索的用户日志等；作为应用研究，下篇介绍基于用户日志进行查询推荐的方法与舆情监测实例。

本书可作为高等院校图书情报与档案管理、信息管理与信息系统、电子商务、计算机科学与技术等相关专业的学术研究与教学参考用书。

责任编辑：李德升　　**责任出版**：谷　洋

Web 用户查询日志挖掘与应用

WEB YONGHU CHAXUN RIZHI WAJUE YU YINGYONG

王继民　著

出版发行：知识产权出版社有限责任公司　　网　　址：http://www.ipph.cn

电　　话：010-82004826　　http://www.laichushu.com

社　　址：北京市海淀区马甸南村 1 号　　邮　　编：100088

责编电话：010-82000860 转 8355　　责编邮箱：lidesheng@cnipr.com

发行电话：010-82000860 转 8101/8029　　发行传真：010-82000893/82003279

印　　刷：知识产权出版社电子制印中心　　经　　销：各大网上书店、新华书店及相关专业书店

开　　本：720mm×960mm　1/16　　印　　张：12

版　　次：2014 年 3 月第 1 版　　印　　次：2014 年 3 月第 1 次印刷

字　　数：205 千字　　定　　价：39.00 元

ISBN 978-7-5130-2658-1

前　言

随着计算机网络技术的日益成熟与 Web 信息量的快速增长，用户可以利用网络在任何地点对各类 Web 检索系统进行信息查询，包括 Web 搜索引擎、电子商务站点、数字图书馆等。Web 检索系统的服务器日志记录了用户与系统交互的整个过程，主要包括用户的访问时间、所输入的查询词、点击的检索结果及点击时间、移动用户的终端设备信息等。这些日志文件所包含的查询或点击记录的规模一般都很大，尤其是大型商业搜索引擎，它每天能接受几千万甚至上亿次的用户查询。

开展大规模互联网用户查询日志挖掘及其应用研究，可以发现中文用户进行 Web 查询行为的特征与规律，改善 Web 检索系统的性能（效果与效率），实现个性化信息服务，在发现用户查询主题的变化及其与社会事件之间的关系等方面也具有重要的理论与实际意义。

本书介绍了互联网用户查询日志挖掘及其应用研究领域的主要技术、方法与实证研究成果，全书由 3 篇共 14 章内容组成，具体如下。

上篇是对用户查询日志挖掘及其应用研究的概括性分析与总论。首先利用文献计量与社会网络分析等方法，剖析了 Web 搜索引擎用户日志与移动搜索用户日志的国内外研究热点、主要科研团队等研究现状。然后给出了 Web 用户查询日志挖掘研究框架，分别针对搜索引擎用户日志和移动搜索用户日志，阐明了使用何种数据分析与挖掘的理论、技术与方法，归纳并总结了目前已有的研究成果，包括：日志挖掘的研究内容、数据集的选择方法、数据预处理的方法、不同地域用户行为的特征与比较、如何应用于系统性能的改善等内容。该框架的建立可以指导一般的 Web 检索系统、电子商务站点及其类似 Web 日志挖掘的研究等。上篇由 4 章内容组成。

中篇介绍基于不同类型日志所开展的实证研究。用户日志来自北大天网大规模 Web 搜索引擎系统的用户日志、国内某大型期刊数据库的用户日志、移动搜

索的用户日志，这三类日志分别涵盖搜索引擎的使用情况、学术期刊数据库的使用情况和移动搜索的使用情况，代表性较强。对这些日志数据集，我们开展了多维度、多方法的综合性试验研究。取得了许多有价值的研究成果，包括：基于时间序列的用户访问量模型、中文 Web 搜索引擎用户检索的一般特征与规律、多任务中文 Web 查询的特征、用户点击 URL 的局部性与自相似性、中文 Web 用户查询行为的演化趋势、高校用户检索策略的影响因素模型、国内移动搜索用户与传统 PC 搜索用户的比较研究等。中篇由 7 章内容组成。

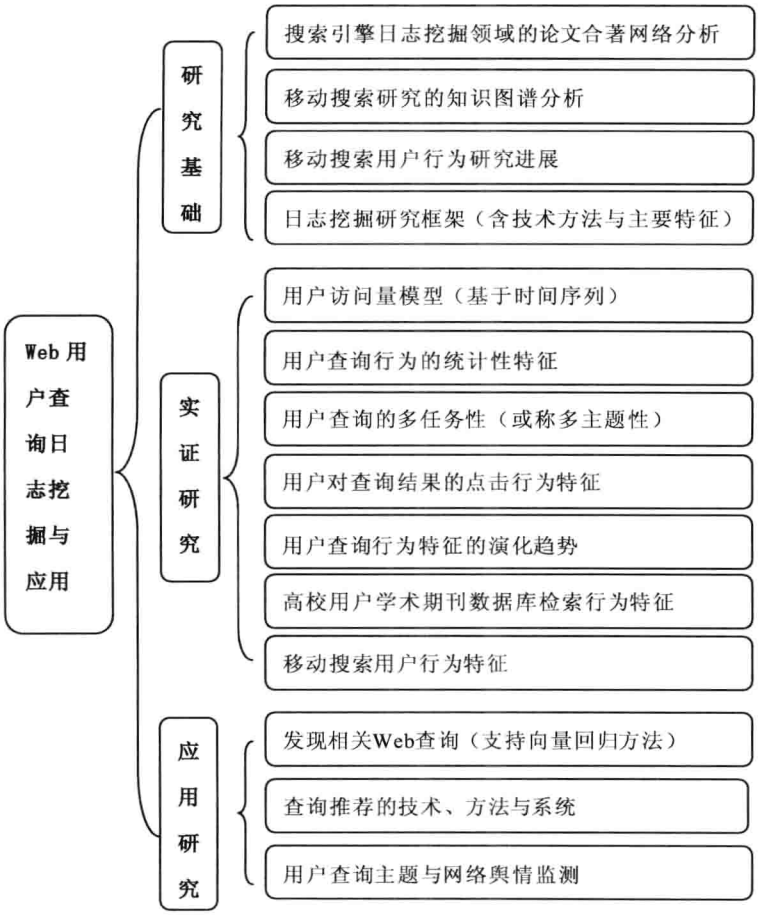

图 1　本书的基本结构

下篇介绍应用研究方面的工作，主要包括 3 部分内容。其一为基于用户日志进行查询推荐的一项实验性研究。其二由一项发明专利的内容构成，所介绍的查

询推荐方法及系统具有实际可操作性。它充分利用用户日志的数据项，为用户提供最可能反映其查询意向且系统具有最佳反馈结果的查询串。其三介绍基于用户查询的舆情监测实例与一个实证研究结果。下篇由3章内容组成。本书的整体逻辑框架如图1所示。

Web用户查询日志挖掘所使用的技术与方法主要有：中文信息处理技术、Web使用挖掘方法、建模分析与预测、社会网络分析方法、可视化技术、新事件探测技术、网络舆情分析分法等。整体研究思路与技术路线如图2所示。

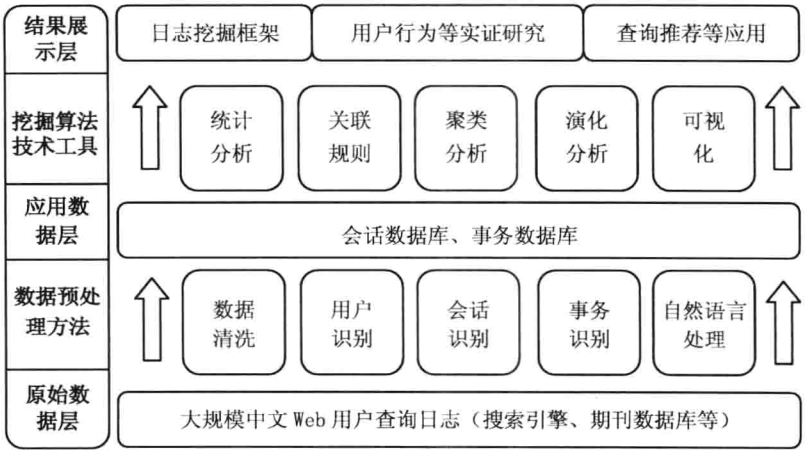

图2　Web用户查询日志挖掘的研究思路与技术路线

本书内容主要来自作者近年来在该领域所做的研究工作，多数章节的内容直接来自本人与他人合作发表的学术研究论文；其中，基于搜索引擎实证研究的数据主要来自北大天网前几年的用户日志（作为公益性中文搜索引擎，它始终没有商业化，近几年的用户访问量较少），而主流的商业搜索引擎如百度、谷歌等都不对外提供或不愿完整提供自己的用户日志，这使得我们无法进一步更新这些实证研究成果。本书的部分研究工作得到中国博士后科学基金、教育部人文社科基金、国家社会科学基金等科研基金的资助。作者早期的研究工作是在北京大学信息科学技术学院网络实验室做博士后时完成的，感谢合作导师李晓明教授的指导与帮助。本书中的搜索引擎日志分析的工作大多是与北大网络实验室的老师和研究生们共同完成的，包括闫宏飞、彭波、孟涛、陈翀、龚笔红等；移动搜索和期刊数据库检索日志挖掘的工作大多是与我指导的研究生们一起完成的，主要有王建冬、李雷明子、张鹏、王明星、郑玉凤、张玉涛等，本科生有孟凡、王一博

等；我系博士后化柏林老师和知识产权出版社的李德升老师对书稿的修订和完善提出了有益的建议。感谢所有与我进行过合作研究和为我提供帮助的老师和同学们。在写作过程中，我们参考或借鉴了大量的中外文参考资料，由于篇幅所限或工作疏忽，未能一一列出，在此特向所有的参考文献作者表示衷心的感谢。

本书的撰写工作虽几经努力，但限于能力和水平，难免有疏漏与错误之处；同时，Web 用户查询日志挖掘与应用属于一个新兴的研究领域，具有多学科交叉属性；随着互联网的快速发展，特别是移动互联网的普及，Web 用户的查询行为也将产生新的变化，本书中的一些理论、技术与方法也需要进一步完善和提高。因此，恳请各位专家和读者批评指正（E-mail：wjm@ pku. edu. cn）。

王继民

2013 年 10 月于北京大学静园三院

目　录

上　篇　Web 用户查询日志挖掘研究基础

中 篇 基于Web用户查询日志的实证研究

下　篇　基于 Web 用户查询日志的应用研究

上 篇

Web用户查询日志挖掘研究基础

为全面了解 Web 用户查询日志挖掘领域的研究现状，本篇首先选择了与该研究领域密切相关的两个研究主题开展计量学研究。所选择的文献数据源有 Web of Science、EI、知网（CNKI）等中外期刊论文数据库。所选择的主题是：搜索引擎日志挖掘与移动搜索。所使用的研究方法有：统计学、社会网络分析、知识图谱、科学计量学等。所分析的主要指标有：基本的文献计量学指标，共词网络、作者合著网络、机构合著网络等各种网络静态几何量。

然后，针对 Web 搜索引擎日志、移动搜索日志等特定数据集，提出了对其进行挖掘的一般流程（框架），既涉及所使用的理论、技术与方法，也归纳总结了目前已有的研究成果。该框架的建立可以指导一般的 Web 检索系统、电子商务站点及其类似 Web 日志挖掘的研究等。

本篇共包含 4 章内容，具体如下。

（1）第 1 章对搜索引擎日志挖掘领域进行了计量学研究。重点分析了论文合著网络，包括网络的中心性、小世界特性、连通性等基本网络特征，发现了该领域中最核心的科研合作团队、研究人员及其研究内容等。

（2）第 2 章对移动搜索领域进行了计量学研究。重点分析了国内外移动搜索领域的研究热点和科研合作网络，利用多种科学知识图谱方法和工具，对其进行了对比研究和可视化展现。

（3）第 3 章介绍了基于日志挖掘的移动搜索用户行为研究的最新进展，包括移动搜索日志挖掘的理论基础和核心文献；国外三类较为典型的研究成果；移动搜索用户日志分析的研究框架，以及改进移动搜索服务的基本方法等。

（4）第 4 章提出了一个 Web 搜索引擎日志挖掘的研究框架，包括：日志挖掘的研究内容、数据集的选择方法、数据预处理的方法、不同地域用户行为的特征与比较、如何应用于系统性能的改善等内容。

第1章　搜索引擎日志挖掘领域的论文合著网络分析

经过十余年的发展，搜索引擎日志挖掘已成为 Web 使用挖掘的一个重要研究分支。本章基于 Web of Science 和 EI 数据库中所收录的有关搜索引擎日志挖掘领域的研究论文，构建了作者合著网络，利用社会网络分析方法研究了合著网络的中心性、小世界特性、连通性等基本特征，发现了该领域中最核心的科研合作团队、研究人员及其研究内容等。

1.1　引言

搜索引擎系统的日志文件记录了用户与系统交互的所有信息。分析与挖掘系统的用户日志可以发现用户进行 Web 查询的特征与规律，进而改善搜索引擎的系统性能[1]。近十余年来有关搜索引擎日志挖掘的论文呈逐年增长的趋势，目前已成为 Web 使用挖掘的重要研究分支之一。

科研合作最显著的表现形式是科研人员之间合作发表论文，而对论文合著情况的研究是分析科研合作的一个重要切入点。合著论文总数是评价作者、地区或机构之间科研合作与学术交流水平的一个重要指标。一定时期内某领域作者合著论文的数量及合作状况，在一定程度上反映了这个领域科研合作与学术交流的发展速度和质量[2]。

以论文作者为结点，以两个作者共同发表论文为边，可以构建一个作者合著关系网络。利用社会网络分析方法对合著网络进行研究和分析，已成为国内外对此类网络进行研究的主流方法，目前已取得许多的研究成果[2-9]，如 Newman 曾对物理学、生物医学和计算机科学等自然科学领域的合著网络进行分析与对比，指出了不同学科之间合作的差异[3]；Liu Xiaoming 等对数字图书馆领域的合著网络进行了分析和研究，并借鉴网页排序的 PageRank 算法提出了作者排序的

Author Rank 方法[4]；Nuša Erman 等借助论文合著网络分析了电子政务研究领域里最活跃的作者[5]；国内的李亮和朱庆华从中心性、凝聚子群和核心—边缘结构等三个角度，对我国情报学领域的合著现象进行了分析[6]，等等。

为对搜索引擎日志挖掘这一新的研究领域的科研合作情况有一个较为概括和清晰的认识，进而了解该领域的主要科研团队、主要研究内容及其研究现状，本章利用社会网络分析方法对该领域的作者合著关系网络特征进行了研究和分析，其中，1.2 节介绍了论文数据的来源和所采用的数据预处理方法，1.3 节给出了所搜集数据的基本统计结果，1.4 节构建了合著关系网络，并研究了该网络的中心性、小世界特性等网络特征，1.5 节对该领域内的三个主要科研团队（凝聚子群）的情况进行了研究，1.6 节总结了全章内容。

1.2 数据准备

1.2.1 数据来源

为确保所分析论文的权威性和代表性，选取 Web of Science（包括 SCI、SSCI、A&HC）和 EI（The Engineering Index）作为论文检索数据库，检索范围为：主题（标题、摘要或者关键词）中同时包含“search engine”和“log”的论文，并选择“所有年份”作为时间段进行检索，共获得 1 036 篇论文的题录信息，包括论文的题目、作者、作者单位、关键词、发表时间及类型（期刊论文、会议论文）等信息。就“搜索引擎日志挖掘”这一特定研究领域而言，检索式的主题中同时包含“search engine”和“log”的论文，基本可以确定是与该研究主题相关的论文。

1.2.2 数据预处理方法

进行有效的数据预处理可以提高挖掘模式的质量，降低挖掘所需要的时间。由 Web of Science 和 EI 这两个数据库所导出的题录信息存在数据格式的不一致性，而且部分数据不完整甚至存在噪声数据。我们在数据分析与模式挖掘之前，先进行了数据的预处理工作，主要包括：剔除不相关的论文、去除重复的论文、拆分同一篇论文中的多个关键词和多个作者、归并同一作者的不同表示等工作，具体如下。

（1）主题去重。由于大规模搜索引擎的使用和普及是在 1995 年之后才开始的，所以在此时间点之前发表的论文予以剔除。通过人工筛查，我们也删除了几十篇与主题内容完全无关的论文。

（2）论文去重。在选取“作者”“期刊来源”“文章标题”“发表时间”“关键词”作为分析数据项时，着重检查了相同文献在不同数据库中出现的问题，包括标题大小写字母的不同、标点和空格间断的不同等问题，避免了同一论文重复出现的问题。

（3）作者归并。论文在被同一或不同数据库收录时，经常会出现同一作者的不同表示形式问题，如本书作者在此数据集中就同时存在 Wang Ji-min 和 Wang Jimin 两种形式，将来还有可能出现 Wang J M 等。我们对论文中所出现的作者进行了简单的归并处理，具体过程是：由论文作者数据构建一个作者合著网络，计算各结点的度值，然后按降序进行排列，去掉度值较小的节点（如删除度值小于 3 的结点），再按字母顺序进行作者排序，人工判断连续的两个或多个作者是否为同一作者，构造映射规则库（如 Wang Ji-min 映射为 Wang Jimin），在原数据集上进行作者姓名替换，即用一个统一的名称去表示同一个作者，然后重新构造作者合著网络。在处理本章的这批数据时，我们构建了近百条映射规则，很显然，这种做法并未合并度值较小的结点，这将对计算结果有微弱的影响。

在经过上述数据预处理后，我们得到符合“搜索引擎日志挖掘”研究的论文 887 篇，不同作者 1 969 个。如下我们将基于这一数据集进行展开研究。

1.3　基本统计结果

按时间顺序统计各年发表的论文总数，结果显示：论文数量呈逐年递增的趋势，近 4 年年均发文量为 150 篇左右。这 887 篇论文中会议论文和期刊论文的大致比例是 2 ∶ 1，其中，会议论文主要来自 International World Wide Web Conferences（WWW）、ACM-SIGIR Conference、International Conference on Information and Knowledge Management、Conferences for IEEE Computer 等互联网、信息检索、数据挖掘等重要的学术会议。期刊论文则主要刊载于 *Lecture Notes in Computer Science*、*Journal of the American Society for Information Science and Technology*、*Information Processing and Management*、*Journal of Computational Information Systems* 等。这些会议和期刊主要是计算机、信息检索、人工智能和信

息系统领域的核心会议和期刊。

总体来看，这些论文所涉及的内容既有关于搜索引擎日志挖掘的理论、技术、方法的研究，也有具体的实证研究。其中，已被分析的搜索引擎日志有 10 余个，包括美国的 Excite 和 AltaVista、智利的 TodoCL、德国的 Fireball、西班牙的 BWIE、韩国的 NAVER、中国大陆的北大天网和搜狗、中国台湾的 GAIS 等。这些论文所使用的日志挖掘技术和方法主要包括：统计分析方法、建模分析与预测、序列模式发现、关联规则挖掘、聚类分析等；挖掘的具体内容包括：词项级、查询级和会话级的数据分析、用户结果页面的查看和点击 URL 的特征、用户查询行为的演化趋势、不同地域用户查询行为的比较，以及如何利用日志分析改进搜索引擎系统的性能等。

统计每一作者的发文数量并进行排序，居前十位的作者如表 1-1 第 2 列所示。该领域的一些出色的研究人员都位列其中，包括：美国匹兹堡大学 Amanda Spink 和宾西法尼亚州立大学 Bernard J. Jansen、微软亚洲研究院的 Chen Zheng（陈正）和 Ma Wei-Ying（马维英）、智利大学的 Ricardo Baeza-Yates，以及清华大学的 Ma Shaoping（马少平）和 Zhang Min（张敏）等。

表 1-1　合著关系网络的中心性排序

序号	作者	发文量	点度中心度排序	度值（%）	介数中心度排序	度值（%）	接近中心度排序	度值（%）
1	Spink, Amanda	24	Chen, Zheng	49	Ma, Wei-Ying	2.24	Zhao, Qiankun	3.80
2	Chen, Zheng	22	Ma, Wei-Ying	35	Giles, C. Lee	2.14	Ma, Wei-Ying	3.80
3	Jansen, BJ	19	Yan, Jun	24	White, Ryen W.	2.11	Giles, C. Lee	3.75
4	Baeza-Yates, Ricardo	16	Baeza-Yates, Ricardo	19	Zhao, Qiankun	2.03	Cucerzan, Silviu	3.63
5	Ma, Wei-Ying	15	Li, Hang	19	Cucerzan, Silviu	2.01	Hoi, Steven C. H.	3.52
6	Ozmutlu, Seda	15	Murdock, Vanessa	19	Dumais, Susan T.	1.71	Chen, Zheng	3.51
7	White, Ryen W.	14	Silvestri, Fabrizio	18	Jones, Rosie	1.68	White, Ryen W.	3.50
8	Zhang, Min	14	Spink, Amanda	18	Dupret, Georges	1.06	Liu, Tie-Yan	3.50
9	Liu, Yiqun	12	White, Ryen W.	18	Baeza-Yates, Ricardo	0.99	Bhowmick, Sourav S.	3.49
10	Ma, Shaoping	12	Yu, Yong	18	Chen, Zheng	0.90	Lyu, Michael R.	3.49