知识管理与知识服务研究

王伟军 主编

基于Web挖掘的个性化信息推荐

Personalized Information Recommendation Based on Web Mining

易 明 著

科学出版社
www.sciencep.com

知识管理与知识服务研究

王伟军 主编

基于Web挖掘的个性化信息推荐

Personalized Information Recommendation Based on Web Mining

易 明 著

科学出版社
北京

内 容 简 介

基于 Web 挖掘的个性化信息推荐是解决当前互联网“信息过载”问题的重要手段之一。本书在继承国内外相关研究成果的基础上，建立了基于 Web 挖掘的个性化信息推荐模型，并构建了语法层次、语义层次和语用层次的个性化信息推荐方法体系。然后，从语法层次的角度，利用 Web 使用挖掘方法研究了 Web 用户偏好分析与推荐问题，并借鉴复杂网络中的社团结构划分方法，提出了基于网络书签的个性化信息推荐方法；从语义层次的角度，提出了基于 Web 文本挖掘的推荐规则获取与匹配方法，分析了基于 Web 领域本体的个性化信息推荐方法，研究了基于社会化标签的 Web 用户兴趣建模方法；从语用层次的角度，利用用户反馈和贝叶斯网络理论讨论了 Web 用户效用函数的构建方法。

本书内容丰富、应用性强，可供信息管理、计算机应用等领域从事相关研究的专家学者、工程技术人员及高等院校相关专业教师、研究生参考使用。

图书在版编目(CIP)数据

基于 Web 挖掘的个性化信息推荐 / 易明著. —北京：科学出版社，2010

（知识转移与知识服务研究）

ISBN 978-7-03-027446-5

Ⅰ. ①基… Ⅱ. ①易… Ⅲ. ①机器检索－检索系统 Ⅳ. ①TP391.3 ②G354.4

中国版本图书馆 CIP 数据核字（2010）第 081058 号

责任编辑：刘 鹏 / 责任校对：赵燕珍

责任印制：钱玉芬 / 封面设计：耕者设计

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2010 年 5 月第 一 版 开本：B5（720×1000）

2010 年 5 月第一次印刷 印张：12 3/4 插页：2

印数：1—1 500 字数：255 000

定价：48.00 元

（如有印装质量问题，我社负责调换）

《知识管理与知识服务研究》丛书
编委会

总　　序

知识，作为社会经济活动的基本要素，已成为社会经济发展的基本资源和根本动力，人类因此进入知识经济和知识社会的新时代。但是，新的知识环境在促进社会发展和人类进步的同时，也让我们置身于知识生态的重重矛盾之中：一方面知识存量激增，并呈爆炸性增长；另一方面知识稀缺严重，人们生活在知识的海洋中，却难以获得所需要的知识。一方面知识产生速度加快，新知识源源不断；另一方面知识老化加速，知识更新周期缩短。一方面知识广泛传播，互联网络提供了知识传播的新途径，跨越了知识扩散的时空障碍；另一方面数字鸿沟日趋明显，城乡差距、地区差异、人群差别影响知识的扩散。因此，如何有效地管理和开发利用知识资源，更好地满足人们日益增长和迫切的知识需求，是人类自我完善和自我发展的需要，更是推动知识创新与知识经济发展的前提和基础，是社会全面协调和科学发展的关键。

知识管理与知识服务诞生于知识经济逐渐兴起、信息技术飞速发展、商业竞争日益加剧的环境中，广泛融合了信息科学、管理学、图书情报学等多学科理论与方法，形成了以“知识”为核心和研究对象的一个新的跨学科研究领域。从管理学视角，知识管理是将组织可获得的各种来源的信息转化为知识，并将知识与人联系起来的过程，强调对显性知识和隐性知识的管理与共享，利用集体的智慧提高组织的应变和创新能力；而知识服务是知识管理领域的演变进化，是随知识管理发展而延伸的概念，是新兴的服务科学、管理和工程学科（SSME）的重要分支。从图书情报视角，知识管理是信息管理的进一步发展，知识服务是信息服务的深化与拓展，知识服务的功能应建立在信息管理和知识管理的基础之上，以满足用户的知识需求和实现知识增值为目标。因此，知识管理是知识服务的基础，知识服务是知识管理的延伸，也是知识管理实现知识创新目标的有效途径。知识管理与知识服务也逐渐成为图书情报学、管理学和信息科学等多学科关注的重要领域和研究热点。

华中师范大学信息管理系及其相关院所的部分教师，长期以来围绕“信息——信息资源——知识的组织与管理、服务与开发利用”等方面，展开积极的探索，从人、环境、信息及其交互关系的视角，运用图书情报学、心理学、管理学、信息科学等多学科的理论和研究方法，开展知识管理与知识服务基础理论、知识组织与检索、知识管理评价与优化、知识管理与知识服务系统及其关键技术、知识转移与知识创新等方面的研究。先后承担或参与了国家“863”计划、国家“十一五”科技攻关计划、教育部高等学校学科创新引智计划、教育部新世纪优秀人才支持计划、国家自然科学基金和国家社会科学基金等多个国家级项目和省部级课题，取得了一系列的研究成果，产生了一定的社会和学术影响，并有多位教师入选教育部新世纪优秀人才支持计划。通过这些重要项目的引领和驱动，华中师范大学逐渐显现出知识管理与知识服务方面的研究特色与发展潜力，基本形成了以信息管理系部分教师为主体的充满激情和活力的研究队伍。为了进一步凝聚学科发展方向，提升学科发展的核心竞争力，学校特成立知识管理与知识服务研究中心，定位于跨学科、创新性的研究平台，以更好地团结和组织相关研究人员开展跨学科联合攻关，服务于国家战略和区域经济与社会发展。

知识管理与知识服务研究中心的一项重要工作就是搭建一个开放式的学术交流平台，经常性地开展学术讲座、专题研讨和学术沙龙等活动，并及时精选研究团队中有价值的研究成果予以发展。现在将首次呈现在读者面前的《知识管理与知识服务研究》丛书共有 9 部著作：《基于 Web 2.0 的网络信息资源管理》（王伟军），《XML 文档全文检索的理论和方法》（夏立新），《网格知识管理与服务》（李进华），《基于 Web 挖掘的个性化信息推荐》（易明），《供应链中的知识转移与知识协同》（李延晖），《区域产业集群中的知识转移研究》（段钊），《知识交流中的版权保护与利益平衡研究》（刘可静），《数字图书馆评价方法》（吴建华），《知识流程服务外包》（王伟军、卢新元等）。这些著作都是从国家级项目的研究成果或博士学位论文中精选出来，经过进一步补充与完善而写成的学术专著。

以上选题涉猎虽广，但都聚焦于“知识”或“知识流”这一核心，置之于新一代互联网环境，关注知识的组织、交流与共享、转移与创新、评价与服务，分别立足于宏观基础、中观产业和微观组织层面展开相关研究。例如，宏观层面的基于 Web 2.0 的信息资源与知识管理变革、网格知识管理与服务的实现、知

识交流中的知识产权保护与利益平衡研究；中观产业层面的区域产业集群中的知识转移与知识创新、供应链中的知识转移与知识协同、知识流程服务外包研究；微观组织或具体应用层面的 XML 文档全文检索的理论与方法、基于 Web 挖掘的个性化信息服务、数字图书馆评价方法等。从中我们不难发现，这些研究都是针对现实中具体的理论与应用问题展开的积极探索，具有很强的跨学科性，显著的创新性和前沿性。

知识管理与知识服务仍是一个新兴的跨学科领域，需要我们大胆地探索。丛书是开放性的学术平台，今后还会不断推出优秀的研究成果，旨在促进我国知识管理与知识服务的理论创新与应用研究，形成有中国特色的知识管理与知识服务理论和方法体系，指导我国知识管理与知识服务的应用实践，为促进我国知识经济的发展和创新型国家建设做出积极的贡献。

本套丛书的出版得到了华中师范大学研究生处、社科处、科技与产业处和信息管理系的大力支持，也得到了科学出版社的鼎力相助。在此表示衷心的感谢！

王伟军

武汉桂子山

2009 年 3 月 28 日

序

在 Web 环境下，互联网已成为全球最大的信息资源库，它在给人类的生活和工作带来革命性变化的同时，也引发了“信息泛滥”、“信息迷航”等问题。个性化信息推荐能够依据 Web 用户的信息需求主动将合适的信息提供给 Web 用户，已经成为解决这些问题的重要手段之一。同时，Web 挖掘能够在海量 Web 信息中获取大量看似无关的信息之间的联系和规律，可以在个性化信息推荐的研究中得到广泛的应用。由此，基于 Web 挖掘的个性化信息推荐便成为信息管理领域和计算机应用领域的研究热点。

易明博士所著《基于 Web 挖掘的个性化信息推荐》一书是目前国内较系统、全面研究基于 Web 挖掘的个性化信息推荐问题的一部专著。该书坚持“继承与创新”相结合的原则，在总结国内外个性化信息推荐相关理论与方法的基础上，综合运用管理科学、信息科学、信息管理、计算机科学、数据挖掘、社会网络、复杂网络等多学科、跨学科的理论与方法，从“点击流”信息运动视角研究了基于 Web 挖掘的个性化信息推荐机理，架构了语法层次、语义层次和语用层次的个性化信息推荐方法体系，并利用定量推导实现不同层次的个性化信息推荐方法。

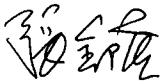
该书以作者的博士学位论文为基础进行纵深拓展，并融入作者最近三年的研究成果，应用性较强，反映出信息管理领域前沿发展趋势，具有十分重要的理论意义、学科价值和应用前景。其主要创新之处在于：

第一，研究视角的创新。该书将基于 Web 挖掘的个性化信息推荐视为一种典型的“点击流”信息资源开发与利用的过程。对于这一过程，作者利用全信息理论和信息运动过程模型进行剖析，进而架构了语法层次、语义层次和语用层次的个性化信息推荐方法体系。当前的相关研究主要局限于语法层次方法，少部分研究涉及了语义层次方法，语用层次方法的研究成果几乎没有。无疑，本书为个性化信息推荐问题的研究引入了全新的分析视角。

第二，研究内容的创新。作者采用“移植”借鉴的方法，通过科学“移植”信息科学领域的经典理论，对基于 Web 挖掘的个性化信息推荐机理进行深入研究，创新认识；借鉴 Web 使用挖掘、社会网络分析、Web 文本挖掘、社会化标签、复杂网络、本体、贝叶斯网络等理论与方法，分别从语法层次、语义层次和语用层次构建基于 Web 挖掘的个性化信息推荐模型，能够有效弥补当前个性化信息推荐方法的相关不足，具有优越性。

第三，研究方法的创新。该书在比较传统个性化信息推荐方法和基于 Web 挖掘的个性化信息推荐方法的研究现状与发展趋势的基础上，从大量个案方法中归纳出三种方法层次：语法层次、语义层次和语用层次。在此基础上，通过抽象演绎研究，构建基于 Web 挖掘的个性化信息推荐模型与方法体系。这种比较归纳研究与抽象演绎研究相结合的分析方法，突破了传统个性化信息推荐研究方法框架的束缚，具有创新性。

最后，需要说明的是，作者在力图创新和对问题的全面把握过程中，有些问题的探讨还有待深入，这是该书的不足之处。相信随着研究的不断推进，这些深层次问题将会在作者的其他论著中得到较好解决。



教授

华中科技大学管理学院院长

2010 年 3 月

前　言

随着科学技术的发展，尤其是20世纪八九十年代信息技术的飞速发展，人们积累了越来越多的数据。但是，如何开发利用这些海量数据，当时的技术手段对此束手无策，人们普遍感觉到自己处在“数据爆炸但知识贫乏”的境地。面对这个问题，1989年人们提出了“数据挖掘”的概念。数据挖掘技术能够从大量的、不完全的、模糊的、随机的数据中，提取隐含在其中的潜在的有用信息和知识。如今，数据挖掘技术经过二十多年的发展，已经取得了很大的成就，其中包括数理统计、人工智能、机器学习、神经网络、模式识别、数据库技术、知识获取和信息检索等。

Web挖掘是数据挖掘的一个重要分支，是随着数据库技术、人工智能技术和网络技术的发展而提出的。随着Web应用与Web信息服务的不断发展，信息总量不断增加，更迫切需要有效的信息分析工具，以便能提取有用的信息和知识，由此Web挖掘得到了广泛应用。当前，将Web挖掘运用到个性化信息推荐，为Web用户提供个性化信息服务已成为理论研究与实践应用的热点。个性化信息推荐可以根据Web用户的喜好、访问留下的历史信息以及其他Web用户的相关信息，在Web用户访问Web站点的过程中为其提供个性化信息推荐服务，进而提高Web用户的满意度与忠诚度。

然而，基于Web挖掘的个性化信息推荐问题涉及的内容非常广泛，是一项复杂的系统工程。本书在总结国内外个性化信息推荐相关理论与方法的基础上，将Web挖掘理论与方法应用到个性化信息推荐中，并利用全信息理论和信息运动过程模型，对基于Web挖掘的个性化信息推荐机理与方法展开研究。

首先，研究了基于Web挖掘的个性化信息推荐机理与方法体系，从“点击流”信息运动视角建立了个性化信息推荐模型——基于Web挖掘的个性化信息推荐机理，并以此为依据提出了语法层次、语义层次和语用层次的个性化信息推荐方法体系。其次，研究了语法层次的Web用户偏好分析与推荐问题，提出了

基于 Web 交易事务聚类的 Web 用户偏好分析方法和基于频繁 Web 页面集的 Web 用户偏好视图生成方法。第三，研究了基于网络书签的个性化信息推荐方法，主要依据 Web 用户收藏的 Web 资源之间的关系建立 Web 用户关系网络，并利用派系过滤算法对其进行社团结构划分，以实现社团内基于协作过滤的个性化信息推荐和社团间基于“信息桥”的个性化信息推荐。第四，研究了基于 Web 文本挖掘的推荐规则获取与匹配问题，提出了基于 Web 特征词条聚类的推荐规则获取与匹配方法，讨论了基于关联规则的频繁 Web 特征词条集的生成方法。第五，研究了整合 Web 语义知识的个性化信息推荐方法，主要利用本体理论构建 Web 站点领域本体，进而实现基于 Web 领域本体的个性化信息推荐。第六，研究了基于社会化标签的 Web 用户兴趣建模问题，提出了基于密度聚类的 Web 用户兴趣建模方法和基于社会化标签网络的 Web 用户兴趣建模方法。最后，研究了语用层次的 Web 用户效用函数构建问题，阐述了利用 Web 用户点击行为所体现的效用权重来构建效用函数的具体方法，并尝试利用贝叶斯网络的学习机制来构建针对特定 Web 用户的面向此次 Web 站点访问的效用函数。

目　录

第1章 绪 论

1.1 本书研究背景

随着网络信息技术的发展和普及，互联网渗透到人们的生活、学习和工作的各个领域，将人类真正带入信息时代。互联网已经发展为当今世界上资料最多、门类最全、规模最大的信息资源库，同时 WWW 以超文本的形式呈现给 Web 用户各种各样的信息，构成了一个异常庞大的具有异构性、动态性和开放性等特点的分布式资源库。然而，在当前互联网环境下，一方面，互联网上的信息量大，使得 Web 用户不容易找到自己需要的产品和服务，妨碍了对 Web 站点的访问与利用；另一方面，Web 站点通常以 Web 的形式展现产品和服务信息以供 Web 用户浏览，是一种典型的“one-size-fits-all”的方法（谢中，2002），这样使得 Web 站点提供的信息根本没有考虑 Web 用户的个性化信息需求，而是以同一种方式对待所有的 Web 用户。

由此，Web 站点的这种模式必然会产生两个问题。第一，Web 站点没有针对性地提供产品和服务信息，Web 用户不能快速获得所需信息。根据中国互联网信息中心（CNNIC）发布的第 25 次中国互联网发展状况统计报告显示：截至 2009 年 12 月 31 日，中国网民达到 3.84 亿人，域名总量达到 1682 万个，Web 站点数量约为 323 万个。面对如此庞大的信息源，Web 用户想要快速找到自己所需要的信息，难度非常大。第二，Web 站点不能快捷地帮助 Web 用户找到感兴趣的产品和服务，Web 用户很容易产生转向其他 Web 站点的动机。事实上，随着互联网经济的到来，市场的主动权正逐渐转移到 Web 用户手中。Web 用户一旦对 Web 站点提供的产品或服务不满意，只要轻松点击一下鼠标就可以进行新的选择。吉顿·塞森（嘉信理财公司的执行副总裁）曾经说过这样一段话：“过去公司总是习惯于谈论如何锁住客户。他们想如何去拥有客户。他们诱导客户，待他们进来，然后就处心积虑地想：‘怎样从他们身上获取利润，以及怎样劝他们买这种产品’。但现在，公司已经不能再这样思考问题了。在互联网时代之前，公司可以有客户意识，但不必以客户为中心。现在，他们别无选择。互联网将交易主动权转移到了客户手中，你的客