



信毅学术文库

泛在商务环境下的 信息聚合与推荐

刘启华 著

国家自然科学基金项目（项目编号：71363022）研



信毅学术文库

泛在商务环境下的 信息聚合与推荐

刘启华 著

 复旦大学出版社

图书在版编目(CIP)数据

泛在商务环境下的信息聚合与推荐/刘启华著. —上海:复旦大学出版社, 2014. 11
(信数学术文库)

ISBN 978-7-309-10860-6

I. 泛… II. 刘… III. 信息管理-研究 IV. G203

中国版本图书馆 CIP 数据核字(2014)第 170217 号



泛在商务环境下的信息聚合与推荐

刘启华 著

责任编辑/张志军

复旦大学出版社有限公司出版发行

上海市国权路 579 号 邮编:200433

网址: fupnet@fudanpress.com http://www.fudanpress.com

门市零售:86-21-65642857 团体订购:86-21-65118853

外埠邮购:86-21-65109143

江苏省句容市排印厂

开本 787×960 1/16 印张 14.5 字数 205 千

2014 年 11 月第 1 版第 1 次印刷

ISBN 978-7-309-10860-6/G · 1400

定价: 30.00 元

如有印装质量问题, 请向复旦大学出版社有限公司发行部调换。

版权所有 侵权必究

总序

Preface

“书籍是人类进步的阶梯。”通过书面记载,语言、文字、图画能够较为完好地保存下来,可以大量印制和快速传播,大大地方便人类的阅读和学习。当下,国家和社会对创新性知识的巨大需求促成了中国学术出版的新一轮繁荣。学术能力已成为当前高校综合服务水平的重要体现,是学校价值追求和人才培养的关键影响因素。

科学合理的学科专业结构、能够引领学科前沿的师资队伍、作为知识载体和传播媒介的优秀作品,是高校作为学术创新主体必备的三大要素。江西财经大学较为科学合理的学科结构和相对优秀的师资队伍,为我校的学术发展与繁荣奠定了坚实的基础。学校教师中教材、学术专著编撰出版活动活跃。

为加强我校学术专著出版管理,淬炼教师学术科研能力,江西财经大学与复旦大学出版社经过充分磋商,达成共识,分批次推出高品质专著系列。为此,我们根据江西财经大学“信敏廉毅”校训精神,将之命名为“信毅学术文库”。前期我们已分批出版了“江西财经大学学术文库”和“江西财经大学博士论文文库”,为打造学术精品,突出江财特色,现将上述两个系列整合汇编为“信毅学术文库”。

本期“信毅学术文库”共选取了9部学术专著予以资助出版。这些学术专著囊括对经济、管理、法律、社会等各方面内容的研究,关注了社会热点论题和有重要研究和参考价值的选题,有一定的学术价值和现实指导意义。专著的作者既有学术领域的资深学者,也有初出茅庐的新聘博士。资深学者学术涵养深厚,且精于写作,他们专著的出版必定能够带来较好的学术影响和社会效益。作为青年学者,优秀博士学术思维活跃,容易提出新的甚至是有突破性的学术观点,从而成为学术研究或学术讨论的焦

点,把他们的学术研究成果编撰成书,其社会效益也不言自明。一般而言,国家级课题项目的研究在专业领域具有较强的创新性,必须达到国内甚至国际的领先水准,基于此,我们也吸纳了部分国家级科研课题项目研究成果。

“信毅学术文库”将分期分批出版问世,我们将严格质量管理,努力提升学术专著水平,力争将“信毅学术文库”打造成业内有影响的高端品牌。“信毅学术文库”的出版,得到了复旦大学出版社的大力支持,对他们的敬业精神和远见卓识,我感到由衷的钦佩。

王 乔

二〇一四年十一月六日

前言

Preface

信息技术和网络的发展不仅使信息数量呈几何级数增加,表现出“极端数据”或“海量数据”的特征,进一步导致“Web 数据危机”,而且也使信息资源的服务与利用日益社会化和个性化。新兴网络(社会化网络、微博、微信、Twitter 等)、新兴商务模式(团购、社会化商务等)、智能终端(智能手机等)以及新技术(泛在计算、物联网等)构建出了一个网络和应用无所不在的泛在商务环境。

泛在商务环境下,网络信息资源具有多源性、多样性、海量性和异构性等自然特性,同时用户信息服务需求存在着泛在性、情境敏感性和高度个性化等特征。现有的信息处理模式已无法适应新兴商务环境下人们研究、学习、创新与决策的要求,主要体现在:(1) 传统的信息组织和检索方式不适应泛在商务环境下多样化的终端设备的要求;(2) 用户信息需求的个性化和多元化特征给用户需求建模方法提出了更高的要求;(3) 传统的个性化信息推荐系统和相应算法大多千方百计地关注如何提高推荐算法的精确性,而忽略了推荐结果的多样性。

因此,泛在商务环境下的信息资源管理需要有新的思路和方法。在泛在商务环境下,用户的信息需求与用户所处的特定情境是密切相关的,建立基于情境计算的用户需求模型,进而实现用户需求与信息资源的智能匹配,并主动为用户推荐准确多样的信息资源,是泛在商务环境下智能信息处理领域值得关注的研究思路。

本书按照“按需聚合、智能推荐”的原则,以提供精准的信息聚合服务和推荐个性化的信息资源为目标,构建了适应泛在商务环境的智能信息处理框架,并以旅游信息检索为例,开发泛在旅游信息聚合与推荐原型系统(U-Travel)。

本书的主要研究内容如下:

1. 构建了适应泛在商务环境的智能信息处理框架

分析泛在商务环境下信息资源特性与用户信息服务需求特点,提出了涵盖用户需求建模、信息采集、信息检索、信息分类和信息推荐的基于用户情境和主题模型的信息聚合与推荐框架。

2. 提出了基于情境历史挖掘的用户需求建模方法

针对当前用户需求建模领域存在的“自动化程度不高,智能性不强,以及新用户需求预测难度大”的缺点,将用户档案、用户情境和用户兴趣结合起来,组成用户情境历史,利用多关系决策树的方法对其进行挖掘,在领域本体的基础上构建个性化的用户需求模型,从而提出了一种自动获取用户需求的方法。

3. 研究了基于主题模型的 Web 文本信息按需聚合的思路

以向量空间模型为代表的传统文本特征表示方法存在着语义缺失问题,本书引入主题模型(Topic Model)对 Web 文本建模,发现文本信息中隐藏的语义信息,进而实现 Web 文本与用户需求本体的智能匹配。主要包含以下三个方面的内容:基于 LDA(Latent Dirichlet Allocation)的主题信息采集、基于 LDA 的文本层次分类、基于 LDA 的文本语义检索。

4. 提供了基于兴趣社区和信任邻居的混合信息推荐模型

采用基于语义本体的层次聚类方法,构建多层语义兴趣社区(Communities of Interests, COI),集成基于兴趣社区的用户偏好匹配算法和基于信任邻居的多样性信息推荐算法,构建泛在环境下融合精确性和多样性的混合信息推荐模型,向用户主动推送准确而多样的信息资源。

5. 开发了涵盖武汉地区优势旅游资源的信息聚合与推荐系统(U-Travel)

应用情境历史挖掘,基于主题模型的信息按需聚合,混合推荐模型的理论和方法,构建了涵盖武汉地区优势旅游资源的信息聚合与推荐系统——U-Travel,并对 U-Travel 主要模块及其性能进行了一系列的测试。测试结果显示,基于用户情境和主题模型的信息聚合与推荐框架是有效的、可行的,比较符合泛在商务环境下用户信息服务的需求。

泛在商务是电子商务和移动商务发展的高级阶段,是一种新兴的电



子商务模式,是信息系统、管理科学、信息科学、电子商务和计算机科学等诸多领域的一个重要而崭新的研究课题。目前泛在商务理论体系与应用体系都不完备,国内外许多研究人员正在从不同视角进行研究。本书从智能信息处理的角度出发,研究泛在商务环境下的信息聚合与推荐技术,取得了一些成果。但是,由于水平有限,书中肯定有些观点是不成熟的,需要进一步探索和研究。书中还可能存在疏漏甚至错误,恳请读者批评指正。

本书的撰写参阅和应用了许多作者的研究成果,我们都以脚注的形式标注或以参考文献和引用文献的形式列于书末(也可能有因疏忽而遗漏的)。他们的研究成果为本书提供了丰富的素材,我们表示由衷的感谢!

本书可供电子商务、信息管理、信息系统、数据库、数据挖掘等相关领域研究与开发人员参考,还可供高等院校的电子商务、信息管理与信息系统、计算机应用、管理科学与工程、系统工程、情报学和信息资源管理等专业的研究生和高年级本科生作为参考书。

目 录

Contents

1 绪论	1
1.1 研究背景	1
1.1.1 泛在计算	1
1.1.2 物联网	3
1.1.3 商务模式的演变	5
1.2 研究问题及意义	8
1.2.1 研究问题的提出	8
1.2.2 研究意义	9
1.3 研究内容和组织	10
2 研究综述	12
2.1 泛在商务研究综述	12
2.1.1 泛在商务的基本概念	12
2.1.2 泛在商务的特点	13
2.1.3 泛在商务文献统计分析(2000—2009)	14
2.2 信息聚合研究综述	28
2.2.1 聚合、融合与集成	28
2.2.2 信息聚合及其应用	32
2.3 智能推荐研究综述	34
2.3.1 传统的信息推荐方法	35
2.3.2 新兴的信息推荐方法	40
2.4 本章小结	43

3 泛在商务环境下信息聚合与推荐框架	44
3.1 泛在商务环境下信息资源的特性	44
3.2 泛在商务环境下信息服务的特征	46
3.3 泛在商务环境下信息聚合与推荐框架	49
3.3.1 基于情境计算的用户需求模型	50
3.3.2 基于主题模型的信息按需聚合	51
3.3.3 基于兴趣社区和信任邻居的个性化推荐	52
3.4 本章小结	53
4 基于情境计算的用户需求模型	54
4.1 理论基础	54
4.1.1 情境计算	54
4.1.2 语义本体	57
4.1.3 Web 日志挖掘	61
4.1.4 多关系决策树	64
4.2 基于情境历史挖掘的用户需求获取	67
4.2.1 情境历史获取	69
4.2.2 情境历史推理	77
4.2.3 偏好预测和用户建模	79
4.3 与其他方法的比较	80
4.4 本章小结	82
5 基于主题模型的 Web 文本按需聚合	83
5.1 主题模型	83
5.1.1 主题模型的提出	83
5.1.2 主题模型的发展历史	86
5.1.3 LDA 模型	89
5.2 基于 LDA 的 Web 文本信息按需聚合	96
5.2.1 基本思路	96



5.2.2 基于 LDA 的主题信息采集	98
5.2.3 基于 LDA 的文本层次分类	108
5.2.4 基于 LDA 的文本检索	111
5.3 本章小结	120
6 基于兴趣社区和信任邻居的个性化推荐	121
6.1 个性化推荐系统的多样性研究进展	122
6.1.1 推荐系统多样性的类型	122
6.1.2 提高推荐系统多样性的方法	124
6.1.3 推荐系统多样性的主要度量指标	132
6.2 多层语义兴趣社区的构建	135
6.2.1 兴趣社区概述	135
6.2.2 基于用户需求本体的多层语义兴趣社区	138
6.3 基于兴趣社区和信任邻居的混合推荐模型	140
6.3.1 基于兴趣社区的用户偏好匹配算法	141
6.3.2 基于信任邻居的多样性信息推荐算法	141
6.3.3 基于兴趣社区和信任邻居的混合推荐算法	142
6.4 本章小结	142
7 实验研究：武汉泛在旅游信息聚合与推荐系统——U-Travel	144
7.1 U-Travel 应用背景	144
7.2 U-Travel 系统架构	146
7.2.1 系统框架	147
7.2.2 用户需求建模模块	149
7.2.3 旅游文本建模模块	155
7.2.4 旅游信息资源按需聚合模块	163
7.2.5 旅游信息推荐模块	164
7.3 系统运行结果展示	164
7.4 U-Travel 主要模块测试	168

7.4.1 基于情境历史的移动用户偏好挖掘效果测试	168
7.4.2 基于 LDA 的中文主题抽取效果分析	175
7.4.3 基于 LDA 的文本层次分类效果分析	178
7.5 U-Travel 性能测试	180
7.5.1 U-Travel 的检索效果测试	180
7.5.2 U-Travel 的个性化推荐效果测试	182
7.6 本章小结	185
8 总结与展望	187
8.1 总结	187
8.2 研究展望	188
参考文献	190

绪 论

1.1 研究背景

当今社会正经历着信息化所带来的巨大变革。泛在计算和物联网等技术推动了信息与通信技术(Information and Communication Technology, ICT)的快速发展,并促进了商务模式从电子商务向泛在商务(Ubiquitous Commerce, U-Commerce)的转变。“泛在的”(Ubiquitous)源自拉丁语,意为“普遍存在的、无所不在的”,即存在于任何地方。

1.1.1 泛在计算

泛在计算(Ubiquitous Computing)也称为普适计算(Pervasive Computing),是由 Weiser 提出的一种超越了桌面计算的全新计算模式。它强调把计算机嵌入到环境或人们日常生活用品中去,使我们生活的物理世界与信息空间中的虚拟世界融合为一个整体。在泛在计算环境下,人们可以随时随地、透明地获得数字化的服务。这一新型计算模式从根本上改变了人们对信息技术的思考,也改变了人们的整个生活和工作方式。

泛在计算是一种计算模式上的社会技术变革^①。从技术角度看,

^① Lyytinen, K., Varshney, U., Ackerman, M. S., Davis, G., Avital, M., Robey, D., Sawyer, S., and Sorensen, C. Surfing the next wave: design and implementation challenges of ubiquitous computing environments [J]. Communications of the Association for Information Systems, 2004, 31: 697-716.

泛在计算环境下,一系列计算装置能够在移动和泛在网络的支持下完成一系列自动化或半自动化的任务或服务。从社会角度看,泛在计算提供了一种不受时间和空间限制、让人们可以随时随地与其他人交流的模式。清华大学徐光佑教授指出,泛在计算具有两大本质:一是泛在计算环境下,物理空间与信息空间将融为一体;二是计算对人的透明化^①。

由于泛在计算巨大的吸引力和影响力,国内外学术界对其进行了广泛的研究,主要体现在:

(1) 纷纷召开与泛在计算相关的学术会议

泛在计算领域最重要的两个国际会议是从 1999 年开始的 UbiComp 会议和从 2000 年开始的 Pervasive Computing 会议。此外,一些以 Workshop 的形式所开展的泛在计算研讨会也纷纷召开。

(2) 广泛发行与泛在计算相关的学术刊物

目前,泛在计算领域比较著名的国际期刊有《Personal and Ubiquitous Computing》、《IEEE Pervasive Computing》、《Pervasive and Mobile Computing》、《IEEE Transactions on Mobile Computing》、《IEEE Internet Computing》以及 ACM(Association for Computing Machinery)等。

(3) 大量资助与泛在计算相关的科研项目

从 20 世纪 90 年代开始,美国国防部高级计划发展局(Defense Advanced Research Projects Agency, DARPA)、美国国家自然科学基金(National Science Foundation, NSF)、美国国家标准技术局(National Institute of Standards and Technology, NIST)、欧盟、中国国家自然科学基金、IBM 公司和微软公司等纷纷开始资助与泛在计算相关的科研项目。例如,麻省理工学院(Massachusetts Institute of Technology, MIT)的 Oxygen 项目(<http://oxygen.lcs.mit.edu>)、卡内基梅隆大学(Carnegie Mellon University, CMU)的 Aura 项目(<http://www-2.cs.cmu.edu/~aura>)、伊利诺伊州立大学(Illinois State University, ISU)的 Gaia 项目(<http://gaia.cs.uiuc.edu>)、加州大学伯克利分校(University

① 徐光佑,史元春,谢伟凯. 普适计算[J]. 计算机学报,2003,26(9): 57-69.

of California, Berkeley) 的 Endeavour 项目 (<http://endeavour.cs.berkeley.edu>)、华盛顿大学 (University of Washington, UW) 的 Portalano 项目 (<http://Portolano.cs.washington.edu>)、NIST 的 SmartSpace 项目 (<http://www.nist.gov/smartspace>)、IBM 的 DreamSpace 项目 (<http://www.research.ibm.com/natural/dreamspace>)、微软的 EasyLiving 项目 (<http://research.microsoft.com/easyliving/>)、欧盟资助的 DisappearingComputer 项目 (<http://www.disappearing-computer.net/>) 以及中国清华大学的 SmartClassroom 项目 (<http://media.cs.tsinghua.edu.cn/~pervasive/>) 等。

随着泛在计算研究的深入、泛在网络和分布式技术的飞速发展、电子器件的微型化及其成本的大幅度下降,泛在计算已经在工业、军事、医疗、农业和日常生活等领域得到广泛的应用。而随之产生的泛在商务、智能住宅、智能家庭和智能交通等将给人类的的生活和工作方式带来深远的影响。

1.1.2 物联网

物联网(Internet of Things, IOT),也称为传感网,是在互联网、射频识别技术(Radio Frequency Identification, RFID)、红外感应器、全球定位系统、激光扫描器和传感器等技术的基础上,利用约定的协议,将任何物品与互联网连接,进行信息交换和通信,以实现智能化识别、定位、追踪、监控和管理的一种网络技术。它可以通过感知、捕获、测量等技术手段,利用识别设备,随时随地对物体进行信息采集,获取各种前端传来的信息,然后依赖各种通信网络与互联网的融合进行可靠的信息传递与共享,最后利用各种智能信息处理技术对海量的跨地域、跨行业、跨部门的数据和信息进行分析处理,提升对物理世界、经济社会各种活动和变化的洞察力。

物联网是 2009 年和 2010 年中国科技界最热的词汇之一,受到了政府和企业的广泛关注。其实物联网的思想早在 20 世纪 90 年代末就提出来了。1995 年,微软前总裁比尔·盖茨在其《未来之路》一书中提出,可以将标签和传感器嵌入到人们平常用到的物品中,使这些物品连在一起,

形成一个网络^①。由于当时受到自动识别和传感设备的技术限制,这一理念并未引起重视。随着技术的不断进步,很多组织和机构开始研究RFID和传感器网络。2005年,在突尼斯举行的信息社会世界峰会(World Summit of Information Society, WSIS)上,国际电信联盟(International Telecommunication Union, ITU)在报告《ITU Internet Reports 2005: The Internet of Things》中正式提出了物联网的说法^②。

物联网的体系构架可以概括为三层:一是感知层,主要负责信息搜集,感知物质世界,即对物体属性进行标识;二是网络层,负责信息传输、信息初步处理、分类和聚合等,即识别设备对物体属性的读取和信息数据的转换,并将物体的信息通过网络传输到信息处理中心;三是应用层,例如智能农业、智能医疗和智能交通等,如图1-1所示。

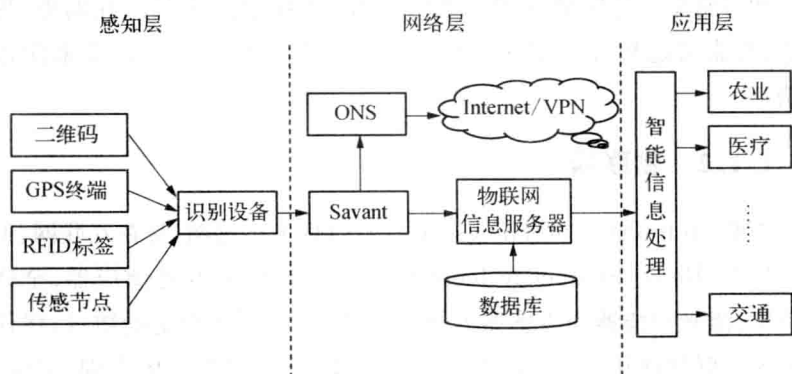


图 1-1 物联网体系结构示意图

为在即将来临的物联网社会中占据主导地位,各国纷纷将RFID和传感器等与物联网相关技术的发展提升为国家战略。2009年1月28日,奥巴马政府与美国工商业领袖举行圆桌会议,IBM首席执行官彭明盛提出新概念“智慧的地球”,得到美国总统奥巴马的大力推崇。欧盟下一代互联网研究小组在其研究报告“Future Internet: The Cross-ETP

① [美] 比尔·盖茨(著), [中] 辜正坤(译). 未来之路[M], 北京大学出版, 1996.

② ITU. ITU Internet Reports 2005: The Internet of Things, (2005-12-08)[2010-06-08], <http://www.itu.int/internetofthings>.

Vision Document”中指出：将来的互联网应该是人际互联网(Internet of People, IOP)、内容和知识互联网(Internet of Contents and Knowledgs, IOC)、物联网(Internet of Things, IOT)和服务互联网(Internet of Services, IOS)的有机统一^①。在此基础上,2009年6月18日,欧盟提出希望确保欧洲在下一代互联网中占据主导地位的14点计划。2009年韩国通信委员会通过的《物联网基础设施构建基本规划》,确定到2012年“通过构建世界最先进的物联网基础实施,打造未来广播通信融合领域超一流ICT强国”的目标。2010年,在十一届全国人大三次会议上,物联网作为中国国家战略被写进了政府工作报告。

1.1.3 商务模式的演变

随着泛在计算和物联网相关技术的飞速发展,商务模式将发生巨大的变革。继传统商务、电子商务和移动商务后,Watson提出了泛在商务(Ubiquitous Commerce、U-Commerce)的概念。他将泛在商务定义为使用泛在网络去支持企业和它的各个利益相关方进行超越传统商务模式的个性化和永不间断的交流和交易,以期达到超越传统商务价值的商务活动。Watson的这一说法得到了很多学者的推崇。Sheng进一步指出,这些利益相关方主要包括客户、供应商、政府、金融机构、企业经理、企业员工和普通大众等。

泛在商务是电子商务和移动商务发展的高级阶段,是一种新兴的电子商务模式,如图1-2所示。

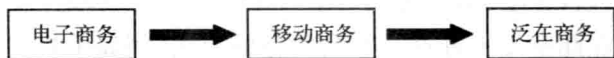


图 1-2 商务模式的演变

1.1.3.1 从电子商务到移动商务

电子商务指使用因特网在组织及其利益相关方之间所开展的一系列

^① Dimitri Papadimitriou. Future Internet: The Cross-ETP Vision Document, (2009-01-03)[2010-03-08], <http://www.future-internet.eu/news/view/article/the-cross-etp-vision-document.html>.