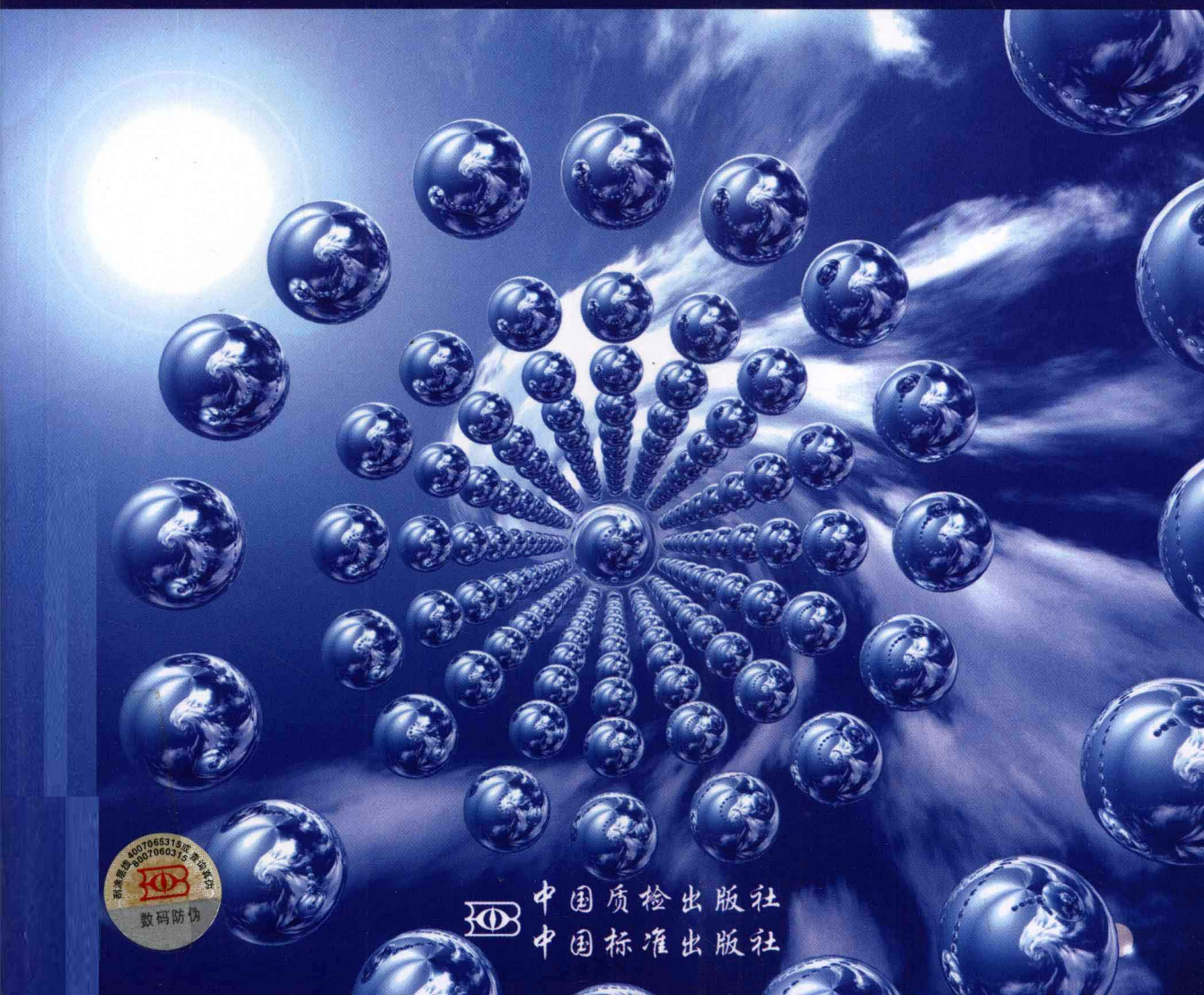
Standard Information Mining:
Theory, Method and Application

标准信息挖掘

——理论、方法与应用

刘 华◎著



中国质检出版社
中国标准出版社

标准信息挖掘

——理论、方法与应用

Standard Information Mining: Theory, Method and Application

◉ 刘华 著

中国质检出版社
中国标准出版社
北京

图书在版编目(CIP)数据

标准信息挖掘:理论、方法与应用/刘华著. —北京:中国标准出版社:中国质检出版社,2011
ISBN 978-7-5066-6360-1

Ⅰ.①标… Ⅱ.①刘… Ⅲ.①计算机应用-情报检索 Ⅳ.①G252.7

中国版本图书馆 CIP 数据核字(2011)第 131045 号

中国质检出版社
中国标准出版社 出版发行
北京市朝阳区和平里西街甲 2 号(100013)
北京市西城区复外三里河北街 16 号(100045)
网址:www.spc.net.cn
电话:(010)64275360 68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销
*
开本 787×1092 1/16 印张 17 字数 399 千字
2011 年 7 月第一版 2011 年 7 月第一次印刷
*
定价 40.00 元

如有印装差错 由本社发行中心调换

前言

Foreword

信息、物质和能源一起构成了客观世界的三个基本要素，信息的传递和交流是人类生存的基本需求。信息以纸张、光磁设备及网络等载体存储，借助声、光、电等媒介进行传递，通过人类的获取而起作用。

标准既是标准化活动的成果，也是人类技术积累的结晶。标准文献属于科技文献的一种，除了具有文献的基本属性以外，还具有规范性的显著特征。标准是现代化企业组织生产、提高产品质量、促进产品进出口的必备技术文献，也是质量技术监督部门、检验检疫部门进行产品检验的法律依据。在科技高速发展的今天，最新发布的标准往往是新技术的载体。

随着现代信息技术的发展，人类进入了信息超载和知识贫乏的信息时代，滚滚而来的各类信息大大超过了人们的处理能力和有效应用的需要。面对纷繁复杂的信息，如何精确、有效地获得成为了信息检索发展的目标。在此背景下，提出了广东省科技计划项目“数字化标准全文数据库及其应用建设”(项目编号:2006B80601002)。该项目采用分层设计思维模式，遵循在信息层进行规则挖掘、在数据层进行数据挖掘、在知识层进行知识发现的研发思想，建立了标准信息挖掘的理论和方法，并应用于标准信息产品的开发。

本书在前述研究工作成果的基础上编写而成，分三篇共九章：

第一篇为标准信息挖掘的理论部分，包括三章。第1章从经济社会发展、标准化发展、信息服务发展和技术环境发展方面阐述了标准信息挖掘理论建立的背景及意义，然后对标准信息检索的发展趋势进行归纳和总结；第2章从标准信息、信息组织、信息检索、数据挖掘和知识发现等主题阐述了标准信息挖掘理论建立的基础理论；第3章建立标准信息挖掘的概念模型和过程模型，然后应用集合与映射的方法，分别建立了信息加工、信息检索和信息抽取的数学模型。

第二篇为标准信息挖掘的方法部分，包括四章。第4章在前述标准信

息挖掘模型基础上，确立了标准题录信息著录的元素、著录元素描述模型以及著录细则，并给出了基于 XML Schema 信息描述；第 5 章在前述标准信息挖掘模型基础上，确立了技术法规题录信息著录的元素、著录元素描述模型以及著录细则，并给出了基于 XML Schema 信息描述；第 6 章在前述标准信息挖掘模型基础上，通过分析标准文献特征，建立了半结构化标准全文信息加工方法和全结构化标准全文信息加工方法两种标准全文信息加工方法；第 7 章分析了信息推送和 Rss 技术的特点，给出了基于 Rss 技术的标准题录信息推送和标准全文信息推送的数据格式和相应的代码。

第三篇为标准信息挖掘的应用部分，包括两章。第 8 章在前述标准信息挖掘理论和方法基础上，建立了由 Web Service、本体和知识挖掘等主要技术组成的标准信息挖掘应用技术架构，然后给出了标准信息挖掘应用模式；第 9 章为应用实例，介绍了数字化标准文献加工系统和全结构化标准文献存储与检索原型系统等软件工具，然后介绍了标准信息挖掘平台和标准信息挖掘光盘等服务产品。

在本书撰写过程中，有幸得到了各级领导和部门同事的大力支持，在此非常感谢熊勇院长、李木华副院长、郭龙祥副院长、黄克副院长、崔美莲副主席对课题研究的重视和关心，还有各有关部门的支持和配合，特别是部门的研究团队：伍文虹、陈莉、杨蕾、陈洪江、黎敬涛、冯宁霞、曹佳彦、覃耀青、徐丰华、黄禧贤，本书的出版是对你们支持的最好回报。

本书的出版完成，是本人在标准文献领域研究的一次深刻的思想总结和凝练，不妥之处在所难免，恳请广大读者批评指正，最后以《考工记》中的一段话与广大读者共勉：

“天有时，地有气，材有美，工有巧，合此四者，然后可以为良。材美工巧，然而不良，则不时，不得地气也”。

（联系方式：Email：chinaliuhua@hotmail.com；QQ：381508408）

著　　者

2011 年 5 月 25 日

目 录

Contents

第1篇 理 论

第2篇 方 法

附 录

第 1 篇

理　论

One

1 绪 论

1.1 经济社会发展

人类文明发展经历了农业化、工业化和信息化，现已进入了互联网知识经济时代。工业化延伸的是人类自然力中的“体力”，信息化延伸的是人类自然力中的“脑力”。信息化促使工业和现代农业走向数字化、智能化、网络化。

在农业社会，人类主要依赖物质资源。蒸汽机的发明推动了工业革命，能源资源的作用显现出来，人类进入了依赖物质和能源资源的工业社会。以微电子技术为代表的现代新兴技术的出现，使信息资源成为重要资源，人类开始进入依赖物质、能源和信息资源的信息社会。

工业化社会是有形产品创造新价值的社会，信息化社会则是无形的信息创造新价值的社会。没有物质，系统便无形体；没有能源，系统便无活力；没有信息，系统便无灵魂。信息是人类一切知识与经验的基础，是人类宝贵财富的源泉。

信息化社会中，信息的作用日益明显，而高速流动的信息所带来的效益与效率的提高使得物质与能源的作用相对降低。专有技术、诀窍、专利、品牌、知识型员工等信息知识资源取代资本成为组织最重要的战略资源，是组织发展核心竞争力的主要资源基础。在知识经济时代背景下，知识将取代土地、劳动力、资本和机器等成为生产力的最重要因素。“知识就是资本，知识就是财富”是这个时代的新理念。

知识分为隐性知识(tacit knowledge)和显性知识(explicit knowledge)。隐性知识是高度个人化的知识，难以表述清楚，隐含于过程和行动之中，由于难于规范化而不易传递，是一种主观的、基于长期经验积累的知识，包括信仰、隐喻、直觉、思维模式和所谓的“诀窍”(如特殊技艺)，主要指存在于人脑的经验、想法、判断、文化、习惯及员工潜能等。显性知识可以用规范化和系统化的语言进行传播，可用语言、文字、数字、图表等清楚地表达。以文件、数据、档案、图表、影像、程序等显示的结构化或半结构化信息，也叫编码知识，主要以专利、技术标准、科学发明和特殊技术等的文献形式存在。

构成现代文献的四要素是：知识信息、物质载体、符号系统和记录方式，四者缺一不可。知识信息是文献的内容，也是文献的主体；符号系统是用以揭示和表达知识信息的标识符号，如：文字、图形、数字、代码、音频、视频等，文献内容借助符号来表达；物质载体是文献信息赖以依存的物质基础，如：纸张、胶片胶卷、磁带磁盘、光盘、穿孔纸带等，文献的内容借助载体来展现；记录方式是指将表达知识信息的符号系统通过特定的人工记录方式使其附着于一定的文献载体上，如印刷、刻录、数字化等，将知识信息的内容与载体统一成为文献。

1.2 标准化发展

1.2.1 标准文献

标准是为了在一定的范围内获得最佳秩序，经协商一致制定并由公认机构批准，共同使用的和重复使用的一种规范性文件(GB/T 20000.1—2002《标准化工作指南 第1部分：标准化和相关活动的通用词汇》)。标准是一种特殊技术产品，表现为一种纸质或电子载体的文件。

标准文献是标准化活动的产物，狭义的标准文献主要指由公认的标准化组织发布的技术标准、管理标准、工作标准及其他规范性文件所组成的一种特种文献。

广义的标准文献，除了各类标准外，还包括标准分类资料、标准检索工具、标准化期刊、标准化专著、标准化管理文件、标准化会议文件、标准化手册、定型图册等有关标准化方面的一般文献。

标准是技术积累的结晶，标准文献是属于科技文献的一种，是现代化企业组织生产、提高产品质量、促进产品进出口的必备技术文献，也是质量技术监督部门、检验检疫部门进行产品检验的法律依据。特别是在当今科技高速发展的情况下，最新发布的标准往往是新技术的载体。标准与其他文献相比，有着规范的产生过程，明确的主题内容和适用范围，编排格式、叙述方式严格划一。另外，随着科学技术的迅猛发展，许多国家都对标准使用周期以及复审周期做了严格规定，以保证标准的时效性。

1.2.2 标准化战略

标准化是为了在一定范围内获得最佳秩序，对现实问题或潜在问题制定共同使用和重复使用的条款的活动(GB/T 20000.1—2002《标准化工作指南 第1部分：标准化和相关活动的通用词汇》)。

标准化是一门科学也是一门重要的应用技术，为工业化发展提供了技术保障，标准来源于社会实践且一直服务于这种社会实践。标准化是现代化大生产的必需条件，是科学管理的基础，是促进科学技术转化成生产力的平台，是推动贸易发展的纽带。

当今时代已经进入了网络经济时代，标准和标准化远远超出了原来指导和促进工业化生产的狭窄领域。农业、服务业、IT产业乃至社会生活的各个方面，无不渗透着标准和标准化的深刻影响，反映着对标准和标准化的强烈需求。从另一个方面说，标准化对生产、贸易，对各个国家、各个地区乃至全球经济发展的影响也越来越大。20世纪90年代后期，特别是进入21世纪以后，发达国家纷纷制定各自标准化战略，以应对经济全球化对自身带来的影响。

随着信息时代的到来，标准化在保证产品质量、沟通技术发展、促进国际贸易方面的作用越来越重要。世界科技竞争的目的在于产业竞争力和产业利益，而产业竞争的关键在于标准竞争。知识产权是政府(或政府间协议组织)管制授予的合法垄断权，属于私有领域；而主流标准是多方面博弈形成的法定或者事实垄断权，属于公有领域，两者结合是双重的产业垄断。

如图1-1所示，在某一领域，各科研机构研发出拥有自主知识产权的技术，通过标准组织，将某项技术转化或融入标准；依据标准生产企业生产该技术产品并进入市场，实现了该

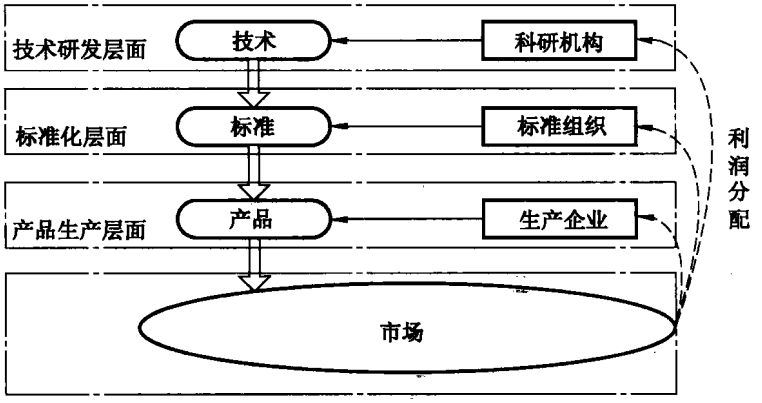


图 1-1 技术、标准和产品的相互关系

技术的商业利润。

在市场层面，国际标准已经成为国际贸易和市场准入的必要条件，越来越多的国家，特别是发展中国家，把进口产品或技术符合有关国际标准作为优先进口或引进的条件之一。

在技术研发层面，在某一领域，各研发机构依据自身的研究基础和对市场的判断，开发出拥有自主知识产权的技术。通过标准组织，将某项技术转化或融入标准，并通过标准掌握该产业相应领域的话语权，从而占据市场，获得垄断优势，攫取更多的商业利益，进一步掌握产业发展方向。

在产品生产层面，依据市场上的技术供给，生产企业依据标准生产相应的产品。在当今信息社会，敏捷制造、虚拟制造、网络制造等先进制造技术的出现，产品生产企业往往处于产品价值链的低端。

在标准化层面，国际标准是市场竞争的有力武器。国际标准的竞争不仅仅是技术问题，也不是战术问题，而是战略问题。发达国家纷纷出台标准化战略，具有很强的时代性和挑战性，标志着各国国际标准化工作由工业化时代向经济全球化时代的重大转变。

拥有知识产权就拥有影响产业的技术实力，拥有某一标准就占据这个产业相应领域的话语权，而主导某一行业的国际标准组织就占据控制这个行业的制高点。一个国家在更多的领域主导国际标准组织，从而成为世界科技中心，掌握世界前沿科技，吸引世界科技人才，控制产业的发展。

如图 1-2 所示，在某一领域，研发机构研制出各自的具有自主知识产权的科研成果，向标准组织提交该成果的技术方案。通过多方博弈，标准组织最后选择成果 i 的技术方案转化为标准，成果 i 获得了进入市场的准入条件实现了产业化生产。拥有研究成果 i 的科研机构从而能获得垄断利润，形成了创新的正反馈机制，使自己始终保持在产业的前沿。

在以上过程中，标准组织是技术从私有领域进入公有领域的关键环节。标准组织在标准运作中获得了产业控制权和主导权。全球的研发机构和专家源源不断地向标准组织提交技术论文和标准草案，创新成果源源不断地汇集到标准组织中、不断强化，主导该标准组织的国家和企业能够源源不断地获取新的技术和思想创新，形成正反馈循环的全球技术中心。

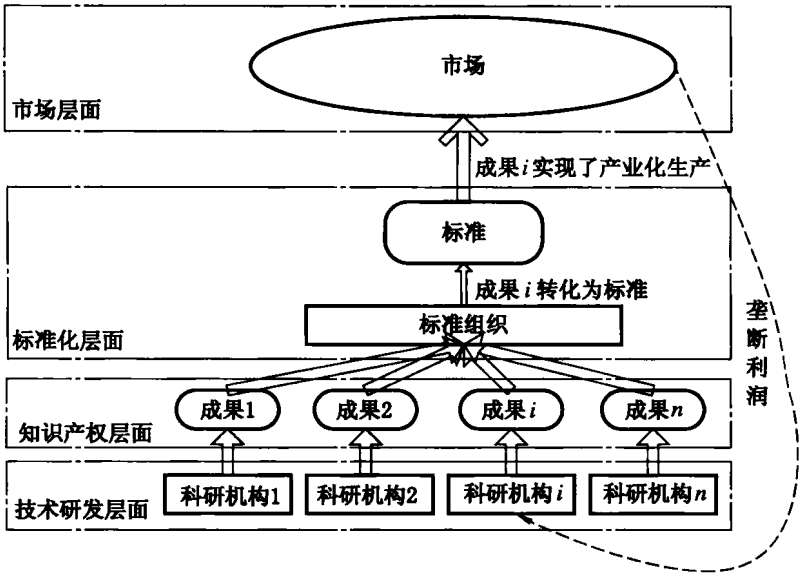


图 1-2 标准组织和科研机构在创新过程中的相互作用

1.3 信息服务发展

1.3.1 信息服务特征

21 世纪是以知识为基础的信息社会，如同技术是制造经济的生命一样，服务是知识经济的灵魂。所谓服务，就是一方能够向另一方提供的任何一项活动或利益行为。服务可以划分为劳动密集型服务和知识(智力)密集型服务，现在服务业的发展越来越强调服务活动中的知识附加值。每项服务活动或服务业务一般要包括三个基本要素：服务内容(产品)、服务对象和服务提供者。服务提供者是连接服务产品和服务对象的桥梁。

广义的信息服务泛指产品或劳务形式向用户提供和传播信息的各种信息劳动，包括信息的采集、整理、加工、存储、传递以及信息技术服务和信息提供服务等。

狭义的信息服务(或称信息提供服务)则是指专职信息服务机构针对用户的信息需要，将开发好的信息产品以用户方便的形式准确传递给特定用户的活动。

信息服务有五个基本要素：信息用户、信息服务者、信息产品、信息服务设施以及信息服务方法和策略。

(1) 信息用户，是信息接收者，是信息服务的对象，是信息产品的利用者，是信息服务业发展的需求动力；

(2) 信息服务者，是从事信息服务的各种机构及机构中的有关人员，是信息服务的主体，他通过提供信息产品来满足用户的信息需求；

(3) 信息产品，是指信息服务者采集、整理、加工的各种已知的或潜在的社会信息、科学知识及研究成果，它构成了信息服务区别于其他服务的本质特征；

(4) 信息服务设施，是信息服务的物质基础和必要手段，包括计算机、网络设备、通信设备等技术设备以及阅览室、信息咨询室等服务场所；

(5) 信息服务方法和策略，是指开展信息服务中的各类操作技巧、方式、程序和制度等，

如索引技术、软件技术、视频技术、服务岗位设立、服务评价等，它是实现信息服务效能的必备“软件”。

1.3.2 信息服务类型

按照信息服务的发展历史及信息内容的角度，可分为：文献服务、报道服务、检索服务、咨询服务和网络服务。

按照信息服务作用的信息客体类型，可分为：实物信息服务（包括样机、样品、样机信息服务）、交往信息服务（包括信息发布服务等）、文献信息服务（包括传统文献服务和电子文献服务）和数据服务。

按信息服务中的服务层次和信息加工深度，可分为：一次服务（以原始信息为内容的服务）、二次服务（包括目录、题录、文摘、索引服务）、三次服务（在原始信息基础上的研究、综述与评价服务等）。

按信息服务指向范围，可分为：单向信息服务（指向单一用户的服务）和多向信息服务（面向众多用户的服务）。

按信息服务的行业领域或主题内容，可分为：科技信息服务、经济信息服务、技术经济服务、法律信息服务、流通信息服务、军事信息服务等。

1.3.3 个性化信息服务

用户体验（user experience）指的是用户在操作或使用一件产品或一项服务时的所做、所想、所感，涉及通过产品或服务提供给用户的理性价值和感性体验。对信息服务来说，由于信息接受过程是一个复杂的心理过程，除了知识匹配等理性因素的作用外，感性因素如情绪、感受等在信息接受过程中同样起重要作用。呈现信息或传送信息的方式会影响用户接受和解释信息的方式，也会影响信息内容的传递效果。信息用户体验是信息用户与信息服务互动的客观反映，要求以用户为中心进行组织设计和提供服务。

个性化信息服务按照服务的层次可分为三类：(1)个性化定制服务模式。这是依据用户的主动需求以定制的方式来予以满足。按照定制的形式和要求，又可划分为个性化界面定制和个性化内容主题定制。个性化界面定制是指对用户访问、浏览界面的颜色、字体、栏目设置、栏目位置等方面进行的定制服务，其主要方法是通过开发各种界面模板提供给用户进行选择。个性化内容主题定制则是根据用户对信息内容主题的明确要求，通过相关内容的收集和整合，以电子邮件或频道推送的方式实时或定期提供给需要的用户。(2)个性化推荐服务模式。个性化推荐服务就是指信息检索或网站系统根据发现的用户喜好，以推荐方式动态地为用户提供观看的内容或浏览建议，简单地说就是为用户提供一对一的服务和指导，是个性化服务的高级阶段。(3)特色增值服务模式。依据用户需求，开发一些附属于信息产品的特色功能，即除检索、下载等基本功能外，还以用户为中心，提供信息通告、个人存储等服务。

个性化信息服务系统的一般体系结构如图 1-3 所示。

1.3.4 标准信息服务发展方向

网络信息环境下的信息服务走向为：服务理念从信息本位走向用户本位，服务目标从信息资源走向问题求解，服务对象从大众服务走向细分市场，服务内容从信息服务走向知识服