基于语义的分布式服务与资源发现

张莹　张昕　何慧　著

科学出版社

基于语义的分布式服务与资源发现

张 莹 张 昕 何 慧 著

科学出版社
北京

内 容 简 介

本书围绕互联网中服务与资源之间的联系，构建了服务与资源的统一描述方法，提出了基于语义的分布式服务与资源一体化发现原型系统及相应的一体化注册与发现方法，在其中基于服务描述文件将 QoS 等级信息的服务查询引入服务查询过程中，配合带有 QoS 信息的服务匹配算法，从而为互联网用户提供了更为便利、一体的服务与资源处理机制。本书阐述的内容为一体化可信网络与普适服务体系提供了服务与资源的统一描述与命名解决方案，以及从用户描述到服务表示的映射过程，以此为基础，为实现一体化搜索引擎提供了可行性。

本书所述内容线索明确，结构合理，主要面向从事互联网技术、知识工程、服务工程、语义网、知识图谱等方向的科研人员，为其在相关研究工作中提供研究思路、补充研究内容、启发研究方法，同时也可作为计算机相关专业硕博研究生的参考资料。

图书在版编目(CIP)数据

基于语义的分布式服务与资源发现/张莹，张昕，何慧著. —北京：科学出版社，2019.3

ISBN 978-7-03-060623-5

Ⅰ. ①基… Ⅱ. ①张… ②张… ③何… Ⅲ. ①网络服务-分布式数据处理-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 034992 号

责任编辑：闫 悦 / 责任校对：张凤琴
责任印制：吴兆东 / 封面设计：迷底书装

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

北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*

2019 年 3 月第 一 版 开本：720×1000 1/16
2019 年 3 月第一次印刷 印张：7 1/4

字数：136 000

定价：58.00 元

(如有印装质量问题，我社负责调换)

前　言

近年来，随着互联网的高速发展和广泛应用，其已逐渐成为人们在工作和生活中获得服务和资源的重要来源。面对互联网中的海量信息，从中选择能够真正转化为生产力的知识变得愈发困难。相应地，用户使用互联网时的需求焦点由早期对获得信息和资源的渴求，逐渐转变为对大量服务和资源的筛选，并以相对较低的行动成本实现对服务和资源的转化应用。鉴于此，本书面向新一代一体化网络中服务与资源并存的现实条件，提出服务与资源的一体化发现方法，着力解决数据移动性、多样性、融合性等问题，实现高效、准确、适用的服务与资源发现。本书主要研究内容与创新点如下。

(1)针对互联网的服务与资源查找，提出以本体为基础，将资源和服务通过属性关系联系起来的统一描述方法。

(2)基于 OWL-S 提出了带有 QoS 的语义服务描述方式，给出了带有 QoS 的服务匹配算法及其功能与性能测试，测试结果验证了带有 QoS 的服务描述与查询方法具有可行性和有效性。

(3)研究了现有服务或资源发现系统中基于关键词进行查询匹配所存在的问题，基于概念间的距离及概念的粒度提出了一种概念间语义相似度的计算方法，设计并实现了基于语义的服务与资源一体化发现原型系统。基于原型系统，给出了服务与资源一体化匹配算法，实现了服务与资源的统一注册与查询。原型系统的性能分析及测试结果表明该方案在查询效率、查准率及用户满意度方面均表现出明显的改进。

(4)提出将服务与资源的一体化描述方法应用于新一代网络——一体化网络中，实现了服务标识 SID 的生成。提出将服务与资源一体化发现方法融合到一体化网络的服务标识映射机制中，从而建立了一体化网络中从用户描述到服务标识的映射，实现了服务与资源的统一标识、注册与查找。

本书共 7 章。第 1 章介绍了基于互联网开展服务与资源发现的基本概况和发展背景，并集中阐述了全书的主要研究工作和创新点，为读者阅读本书提供线索建议。

第 2 章介绍了服务与资源发现研究的总体研究进展，并对现有的信息发现系统方案进行了对比分析，为读者了解相关研究重点和进展提供了理论基础。

第 3 章和第 4 章阐述了基于本体、语义网、服务质量等概念和理论构建形成的服务与资源一体化描述方法以及发现方法，引导读者深入了解和掌握具体的方法和技术。

第 5 章阐述了服务与资源一体化发现的技术架构和系统方案，向读者全面阐释了具体可执行的方案和流程。

第 6 章基于构建的研究基础阐述了服务与资源统一发现方法在一体化网络中的具体应用场景和流程，向读者展示了服务标识生成和一体化搜索引擎实现的技术方案。

第 7 章对全书内容进行了总结，为读者开展后续研究提供建议和引导。

本书在撰写过程中得到了北京交通大学黄厚宽教授的支持鼓励和宝贵意见，在此谨向黄教授致以衷心的感谢。另外，感谢华北电力大学(北京)控制与计算机工程学院信息安全教研室师生在本书成稿过程中给予的帮助支持。本书得到国家自然科学基金青年项目“网络资源的语义标识与分布式定位方法研究”(项目编号：61305056)、中央高校基金面上项目“地理信息集成方法及在智能电网中的应用研究”(项目编号：2018MS024)、吉林省科技发展计划项目“面向城市精细化治理的智能路径规划服务”(项目编号：20190303133SF)等的支持。在此对国家自然科学基金委员会、北京市教育委员会和吉林省科技厅的支持表示感谢。本书的出版得到了科学出版社的大力支持和帮助，在此表示诚挚的感谢。

由于作者水平有限，书中难免会有疏漏之处，敬请广大读者批评指正。

张　莹

2018 年 10 月于华北电力大学(北京)

目　录

第 1 章　服务与资源发现介绍 …… 1
1.1　概述 …… 1
1.2　本书的目标 …… 3
1.3　研究问题 …… 5
1.4　本书的内容结构安排 …… 5

第 2 章　研究进展综述 …… 7
2.1　资源与服务发现的目标 …… 7
2.1.1　目录式搜索 …… 7
2.1.2　基于机器人的搜索 …… 8
2.1.3　元搜索 …… 9
2.1.4　语义搜索 …… 10
2.2　服务发现的研究现状 …… 13
2.3　本章小结 …… 19

第 3 章　服务与资源一体化描述 …… 21
3.1　现有资源的描述、注册和查找 …… 21
3.2　服务描述、注册与查找的原理过程 …… 23
3.2.1　服务的描述 …… 23
3.2.2　服务的注册与查找 …… 26
3.3　基于语义网的服务与资源一体化描述 …… 27
3.4　基于本体的服务与资源一体化描述 …… 29
3.5　本章小结 …… 33

第 4 章　带有 QoS 的语义服务描述及发现方法 …… 34
4.1　带有 QoS 的语义服务描述——OWL-QoS …… 34
4.2　带有 QoS 的语义服务发现 …… 40
4.3　验证实验 …… 43

4.4 本章小结……49

第 5 章 服务与资源一体化发现方法……50
5.1 本体介绍……50
5.1.1 本体语言概述……50
5.1.2 领域本体……52
5.1.3 词典本体……54
5.2 基于本体论的语义相似度度量……56
5.2.1 语义相似度……56
5.2.2 语义相似度与概念间的距离……57
5.2.3 语义相似度与概念的粒度……59
5.2.4 语义相似度的计算……59
5.3 JXTA 技术……60
5.4 原型系统的结构及原理……62
5.5 服务与资源一体化注册与发现方法……64
5.5.1 服务与资源一体化注册……66
5.5.2 服务与资源一体化查询……68
5.5.3 相关算法……69
5.6 一体化发现系统性能分析……72
5.6.1 存储开销分析……72
5.6.2 查询响应时间分析……72
5.7 实验结果与分析……73
5.7.1 实验环境……73
5.7.2 正确性评估……74
5.7.3 效率评估……77
5.7.4 用户满意度比较……77
5.7.5 节点的加入和离开对原型系统影响的评估……78
5.8 本章小结……79

第 6 章 服务与资源统一发现方法在一体化网络中的应用……80
6.1 普适服务……80
6.2 现有网络的缺点与不足……81
6.3 一体化网络结构……82
6.4 统一命名与定位……84

6.4.1　服务与资源的统一描述与命名——SID ……84
6.4.2　用户描述到服务标识映射——一体化搜索引擎……89
6.5　本章小结……91
第 7 章　研究贡献……93
参考文献……95

第 1 章　服务与资源发现介绍

1.1　概　　述

随着信息社会的发展，互联网已经成为人们日常生活的一部分，取得了巨大的成功，其蕴含着具有巨大潜在价值知识的分布式信息空间，人们可以从中轻易获取大量信息[①]。然而，从信息中取得能够真正转化为生产力的知识却仍非易事，其主要困难不是信息的匮乏，反而是信息“过剩”，即难以快速有效地从信息中识别、拾取有价值的知识。

互联网是一个“网络的网络”，它把世界上的各种大大小小的网络联结汇聚在一起，推动了网络技术的迅速发展。在互联网发展初期，网站相对较少，网页数量也较少，信息的查找对于互联网用户来说是一件相对轻松的事。然而，伴随着互联网爆炸式的发展，普通互联网用户已无法掌控互联网这个信息海洋。1994 年，“蜘蛛程序(web spider)”的出现填补了这个空白；同年，超级目录索引 Yahoo!拉开了第一代搜索引擎的序幕。搜索引擎是一个面向互联网的信息搜集、整理和检索服务系统平台。当今搜索引擎的主流是基于网络爬虫的第二代网络搜索引擎，其主要采用了网络爬虫技术、索引技术、相关度及排序技术。目前以 Google 为代表的多家搜索引擎正在不断扩充和完善各自功能，提高其搜索服务的高效性和准确性。

从搜索引擎的迅速发展可以看出，用户对从海量信息中准确查找资源的需求非常迫切。在如今的互联网时代，各类网站分布式地支撑起了互联网的骨架，搜索引擎则是解决网络信息查找的理想解决方案之一。但是，互联网上还有很多信息游离于分布式的骨架之外，例如，按其他网络协议开发的网络服务器和客户端，即互联网上仍有很多资源和服务没有纳入搜索引擎的考察范围，若要找到此类资源和服务，则需要通过多次统一资源定位符(uniform resource locator，URL)的链接跳转才能实现。

为了解决此类问题，下一代网络中的一体化网络方案引入了分布式注册中心的概念。网络客户端将其可以提供的资源和服务在分布式注册中心进行注册；其他网络客户想要查找特定资源和服务时，直接向注册中心查找自己需要的资源

① 本书中，如无特别说明，信息特指资源与服务的集合。

或服务。在上述资源及服务的注册和查找过程中，各分布式注册中心处于对等的地位。

资源的获取和服务的接入是互联网的两种重要基本应用，几乎所有的互联网活动都需要两者的支持。互联网中的资源和服务不是孤立存在的，它们之间存在着很多潜在联系，但是在当前的互联网架构中，对资源和服务的描述与处理机制尚未完善，许多研究只是讨论了资源的查找或服务的查找，对两者之间联系的探讨和利用仍有较大的局限性。鉴于此，如果能充分发挥资源和服务之间的联系，构建服务和资源的统一描述与发现机制，将会为用户提供更为便利、高效的服务与资源统一处理机制，同时也能够减少分别开展服务和资源处理造成的网络资源浪费。

目前，很多信息检索的过程是基于关键词匹配技术达成的。此种方式的检索通常会向用户返回很多不相关的条目并需要用户从中手动挑选意向结果，且无法满足用户的更高需求。在信息检索中加入语义层面考虑有着重大的进步意义，其使得互联网中原本流动的单纯数据流被扩展为计算机可理解的语义信息。基于此，信息交换则可以建立在语义层面而非文字层面，进而可以使计算机精确地理解、采集和组合信息，即演进形成了语义网(semantic web)。

语义网是由万维网创始人蒂姆·伯纳斯·李在2000年的世界可扩展标记语言(extensible markup language，XML)大会上提出来的，他对语义网的概念进行了解释，并提出了语义网的体系结构。2001年5月，*Scientific American* 封面文章发表了 Berners-Lee 等[1]的文章，文中描绘了语义网应用的美好前景，并对其中的主要技术进行了简明介绍。语义网思想勾勒了一个计算机根据语义智能化地进行信息处理的下一代网络构想。其将“语义”的概念引入互联网，所涉及的“让机器理解信息的含义”已经成为近年来的研究热点。

为了实现对服务的语义查询，许多研究者采用网络服务的本体语言(ontology web language for services，OWL-S)[2]来描述服务。OWL-S 是一种用来描述服务的、计算机可以理解的本体语言。OWL-S 提供了三种类型的知识：service profile(描述服务是做什么的)，service model(描述服务是如何工作的)，service grounding(描述如何调用该服务)。其中，service profile 表达了服务的基本功能，可以用来实现服务的基本能力匹配，进而实现服务的发布、查找和匹配。

近年来，用户对互联网服务的要求越来越高，不同用户对同一服务的质量要求是不一样的，服务质量(quality of service，QoS)分级成了新一代网络的主要特点之一。不同的服务请求对应了不同的服务质量等级。因此，为用户提供带有 QoS 等级的服务查询是网络中服务查询的关键。

此外，现有的服务或资源发现多采用集中式的查询机制，描述信息(即元数据)

被存储在统一的网络节点。此种查询机制容易受制于单点故障，不适合大规模的服务和资源查询。这一缺陷与新一代网络中提出的普适服务是相冲突的，因为在普适服务中，越来越多的设备和实体以服务的方式加入网络，并且动态地保持更新，而集中式的查询机制则不能很好地适应服务与资源的动态性和可扩展性。

互联网中信息量的爆炸式增长使得集中式的服务和资源发现方案难以有效地满足实际需求，应运而生的点对点(peer to peer，P2P)，也称对等式技术(如 Napster、Gnutella、Aim 等)[3,4]变得越来越流行。然而，大多数 P2P 应用程序系统只适用于某一种特定的平台，相互之间不能进行通信和数据共享。例如，Napster 提供音乐文件的查找，Gnutella 提供普通文件共享，Aim 提供短消息发送。由于缺乏通用的基础机制，这些 P2P 系统互不兼容，难以实现互操作，而克服这些 P2P 系统的缺点正是 JXTA[5]的设计目标。JXTA 提供了一个跨平台、跨操作系统和跨编程语言的 P2P 网络应用程序平台，它构建的 P2P 应用程序具有三个特性：互操作性、平台无关性和广泛性。

基于当前网络的发展趋势和未来网络的新特性，研究如何基于现有技术(如语义网、P2P 等)[6,7]从海量信息中获取有用信息，具有较高的学术意义和应用价值。因此，本书针对现有网络中服务与资源发现系统存在的种种弊端，依托国家重点基础研究发展计划(973 计划)“一体化可信网络与普适服务体系基础研究”课题背景，研究服务与资源一体化描述、标识、查询的基本方法；结合未来网络的发展趋势及一体化网络的特点，研究带有服务质量等级信息的服务发现方法，并将其应用于一体化网络中生成服务标识(service identifier，SID)，实现从用户描述到服务标识的映射，以及服务与资源的统一标识、注册与查找。

1.2　本书的目标

本书依托国家重点基础研究发展计划(973 计划)“一体化可信网络与普适服务体系基础研究(2007CB307100)”课题背景，基于未来网络的特点，从现有的各种资源或服务发现原型系统出发，结合不同系统的优缺点，研究如何实现服务与资源的一体化发现，使其能为用户提供更加便利、一体的服务与资源处理机制；在此基础上将其应用于一体化可信网络与普适服务体系中，使其能够更好地服务于新一代网络。

本书所陈述的具体工作如下。

(1)分析现有服务与资源发现系统的研究进展，主要涉及服务与资源的描述方法、分布方式、匹配算法等。针对目前服务与资源发现系统中存在的一

些问题，本书提出一种新的服务与资源一体化发现系统，以更好地满足用户的需求。

(2)在传统互联网上进行资源、服务的注册与查询时，由于资源和服务的关系没有明确的定义划分，通常需要通过人工辨识进行判断，用户很难也没有必要自行区分所需要的是一个资源还是一次服务，或者是两者的混合。因此，对互联网资源和服务进行统一描述是一体化信息发现的首要任务。本书通过分析现有互联网中服务与资源描述、注册及查找所存在的不足，阐述一种以本体描述为基础，将服务与资源通过实体属性联系起来的统一描述方法。

(3)用户对网络服务的要求越来越高，不同用户对同一服务的质量有不同的需求，服务质量分等级成了新一代网络服务的主要特点之一，不同的服务请求要通过不同的 QoS 等级来应答。针对这一特点，本书提出一种向 OWL-S 中添加 QoS 信息的新方法用于描述服务(即 OWL-QoS)，并基于 OWL-QoS 提出带有 QoS 信息的语义服务匹配算法。

(4)现有互联网技术是围绕超文本系统展开的，其主要思想是通过统一资源标识符(uniform resource identifier，URI)对互联网上的信息进行标记，使人们可以迅速地对互联网上的信息进行定位。然而，现有互联网技术并没有描述信息的语义，计算机在处理信息时只是通过对照 URI 来定位信息，对信息的内容并不关心。由于现有互联网技术的局限，互联网上信息处理的自动化、智能化程度很低，计算机处理器的强大性能没有得到有效发挥。语义网的出发点正是改变现有互联网依靠文本对比实现资源共享的模式，通过本体来描述资源的语义信息，达成语义级的资源共享。通过语义网的引入，各类资源不再只是保持彼此间各种相连的信息，还包括信息的语义表达，支持用计算机可以理解的内容来描述资源与服务，使得网络中流动的不再是单纯的数据流，提高了计算机处理信息的自动化和智能化程度。因此，本书采用语义网的相关技术来描述服务与资源，并基于概念间的距离及概念的粒度提出一种语义相似度计算方法来评估概念间的语义关系，从而对信息实现快速、智能定位。

(5)基于本书提出的服务与资源一体化发现原型系统，研究相应的服务与资源一体化匹配算法，实现服务与资源的统一注册与查询。

(6)现有互联网体系架构是由美国设计的，互联网的核心设施则由美国主导控制，全世界现有的 13 个顶级域名服务器中，有 10 个坐落在美国。因此，在下一代的一体化网络中，研究新型的服务与资源统一描述、命名及名字解析映射具有重要的战略意义。本书将服务与资源的一体化描述方法应用于新一代网络——一体化网络的研究中，实现服务标识的生成。

(7)本书介绍基于语义的服务与资源一体化注册与发现方法及其在一体化网

络中的应用，实现一体化网络中从用户描述到服务标识的映射，以及服务与资源的统一标识、注册与查找。

1.3 研 究 问 题

本书重点探讨和阐述基于语义的分布式服务与资源一体化发现方法及其在一体化网络中的应用，主要研究问题及创新工作如下。

(1)研究了网络中服务与资源的相互关系，创新性地提出了一种以本体描述为基础，将服务与资源通过属性联系起来的统一描述方法。服务与资源的统一描述机制将根据使用频度向用户提供与其查询内容相关的服务和资源信息，即优先提供使用频度较高的结果，以此实现用户查询的方便化、智能化和语义化，以及实现服务与资源一体化发现的前提和基础。

(2)针对未来网络的发展趋势——服务质量分等级，本书提出了将 QoS 信息加入服务描述本体 OWL-S 中的服务描述方式，即 OWL-QoS。基于 OWL-QoS，本书作者提出了带有 QoS 信息的服务匹配算法。

(3)本书采用语义网的相关技术来描述服务与资源，基于概念间的距离及概念的粒度提出一种语义相似度计算方法。通过该方法，可以计算获得不同资源或服务间的语义相似度关系，其是基于语义的服务与资源一体化匹配算法的基础。为了向用户提供便利、一体的服务与资源查询处理机制，减少分别进行服务和资源处理产生的网络资源浪费，本书分析服务与资源的相互关联，发掘两者之间的联系，提出并设计基于语义的分布式服务与资源一体化发现原型系统。基于该原型系统，本书提出基于语义的服务与资源一体化匹配算法，详细地阐述服务与资源的统一注册、查询、更新等操作流程。性能分析实验表明本书中提出的服务与资源一体化发现方法具备合理性与有效性。

(4)通过将本书中提出的服务与资源一体化描述方法应用于一体化网络研究领域，提出服务与资源统一命名机制，实现服务与资源统一标识的生成；通过将服务与资源一体化发现原型系统整合至一体化网络中，实现一体化网络中从用户描述到服务标识的映射，以及基于语义的服务和资源的统一注册与查找。

1.4 本书的内容结构安排

本书后续的章节关系如图 1.1 所示，具体的组织结构如下。

第 2 章：对服务与资源发现系统的研究现状进行综述，重点介绍几种典型的信息发现系统的优缺点，讨论基于语义的分布式服务与资源一体化发现方法研究的重要性。

第 3 章：基于服务与资源的属性建立两者之间的紧密联系，统一描述网络服务与资源。

第 4 章：基于新一代网络中区分服务质量等级的重要特点，提出向基于语义的服务描述文件中添加 QoS 信息的新方法，并相应地提供面向 QoS 的服务匹配算法。

第 5 章：介绍基于语义的分布式服务与资源一体化发现原型系统的设计与实现，给出原型系统中基于语义的相似度计算方法，用于实现服务与资源的模糊匹配。重点研究并提出基于语义的服务与资源一体化匹配算法，分析服务与资源的统一注册、查询、更新等流程。

第 6 章：将本书提出的服务与资源统一发现方法应用在一体化网络中，提出服务与资源的统一描述和命名方法，实现从用户描述到服务标识的映射，以及基于语义的服务与资源统一注册和查找。这两部分的工作主要集中在一体化网络结构的“服务层”。

第 7 章：对本书的工作进行总结。

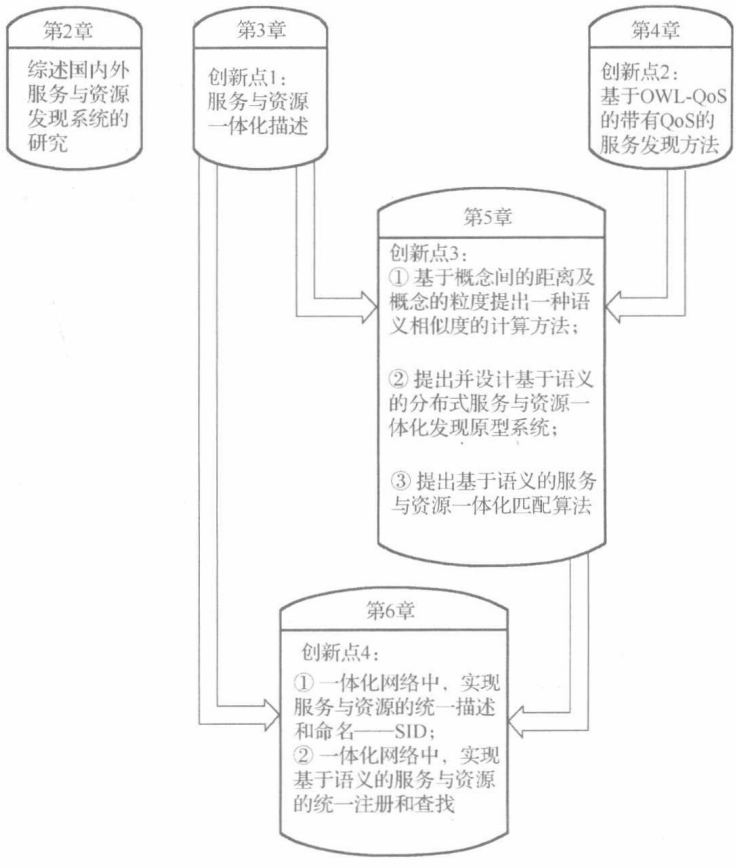

图 1.1　本书的内容结构及章节关系

第 2 章　研究进展综述

21 世纪是信息化、网络化时代，信息资源的广泛共享和高效传递是信息化、网络化发展的重要标志。随着互联网的快速发展，其已经成为世界上覆盖面最广、规模最大、信息资源最丰富的网络基础设施，成为全球范围内传播和交流社会信息、科研信息、教育信息以及商业信息的最主要渠道。互联网信息的特点主要包括三个方面：①数据量大，内容丰富，涵盖的内容种类繁多；②信息类型多样化，除文本、声音、图像等资源外，还包括多种类型的服务，如电子邮件 E-mail 服务、文件传输 FTP 服务、远程登录 Telnet 服务、网络新闻 Usenet 服务及 Web 服务等；③信息来源分散，信息无统一管理和统一发布的标准。

互联网信息主要呈现为资源与服务两种类型。资源获取与服务获取是目前互联网中最重要的两种应用，几乎所有的网络活动都需要两者的支持。本章针对资源与服务的发现问题，对国内外相关工作进行综述，以其构成本书后续研究工作的基础。

2.1　资源与服务发现的目标

越来越多的人习惯从互联网获取信息，其主要通过信息搜索达成[8,9]。如果说互联网是信息传递的通道，那么互联网信息检索则是寻求构建信息传递通道的关键工具。随着网络信息量的剧增，人们想在海量复杂的信息中找到真正需要的信息，无异于大海捞针。因此，为了更充分地利用互联网资源，满足人们不同的查询需求，互联网信息检索技术迅速发展起来。根据技术原理的不同，互联网资源搜索主要分为四类：网络信息目录技术、基于机器人的搜索技术、元搜索技术及基于语义的搜索技术。对应上述四种搜索技术，互联网资源检索方法主要包括以下四种类型。

2.1.1　目录式搜索

借鉴传统的图书情报管理方法，网络信息目录依靠人工(专门的信息管理人员)建立网络信息数据库。信息管理人员跟踪和选择有用的网络站点或页面，并按规范方式对其进行分类标引，组建信息索引数据库。构建网络信息目录所采用的分类法有主题分类法、图书分类法、学科分类法和分面组配分类法[10]。用户可以仅

靠分类目录找到需要的信息，而不用进行关键词匹配查询。目录索引中最具代表性的应用为 Yahoo！（雅虎），其他知名的还有万维网开放内容目录 DMOZ、LookSmart、About 等。国内的搜狐、新浪、网易搜索也都属于此种类型[11]。目录式搜索因为基于人的智能进行构建，具有信息准确、导航质量高等优点；其缺点是依赖人工输入、维护工作量大、信息时效性难以保证。

2.1.2　基于机器人的搜索

基于机器人的搜索技术是在网络信息目录技术基础上发展起来的，其实现了从人工录入到计算机自动化处理的转变过程，即通过编写特定程序完成索引项的自动维护更新。上述提及的特定程序称为机器人(Robot)，也称为 Spider、Crawler 或 Worm。此类程序自动搜索文件并跟踪文件的超文本结构，能够沿着网络链接漫游 Web 文档集合。机器人一般以 URL 清单为基础，提取网页上有价值的文本信息，并能够利用网络标准协议(如超文本传输协议(hyper text transfer protocol，HTTP)等)读取相应的文档，然后以所读取文档中的新 URL 为起点，继续重复漫游过程，直到不再出现满足条件的新 URL[10-12]。

按照搜索顺序划分，基于机器人的搜索一般采用两种策略，分别是广度优先和深度优先。广度优先搜索策略是指机器人会先抓取起始网页链接的所有网页，从中选择一个链接网页，继续以同理方式抓取在此网页链接的所有网页。该策略是最常用的搜索方式，并且可以通过构建机器人并行处理架构，具备较高的抓取速度。深度优先搜索策略是指机器人会从起始页中的某个链接开始，跟踪链接关联持续跳转并抓取，直至不再有新的跳转链接出现，再转入起始页中的另外一个新的链接开展同样的跳转和抓取[10]。

基于机器人的搜索具有信息量大、更新及时、不需要人工干预等优点，但其返回结果内容过多，包含的无关信息比较多，需要用户从中进行二次筛选。采用该类搜索技术的搜索引擎代表有 AltaVista、Excite、Infoseek 等；国内代表为天网、Openfind 等[11]。

互联网网页按存在方式可分为表层网(surface web)和深层网(deep web，也称 hidden web 或 invisible web)[13]。表层网指传统网页搜索引擎可以索引的页面，该类页面可以通过超链接访问，且部分页面属于静态网页[14]。深层网是指那些由普通搜索引擎难以发现其信息内容的 Web 页面[15]，它们存储在网络数据库中，不能通过超链接直接访问，而需通过动态网页技术进行访问。

Bright Planet 公司在 2000 年对深层网做了详细调查[16]，其调查结果显示，深层网中可访问的信息容量是表层网的 400～500 倍；深层网站点月访问量是一般站点的 1.5 倍，并且经常被链接；深层网是互联网中规模最大、发展速度最快的新

型信息资源；深层网站点比一般站点涉及范围小，内容更为精深。

尽管深层网中含有海量的有价值信息，但现有的搜索引擎，如 Google、Yahoo! 等，一般只搜索表层网中的静态页面、文件等资源，很少索引深层网中的资源。这是因为对深层网的搜索可能会使机器人陷入海量动态页面的跳转抓取任务中，从而浪费了网络带宽和存储资源，可见发现潜藏在网络数据库中的信息对于网络资源发现是一项艰巨的任务。

目前，很多研究者正积极地投入深层网搜索技术的研究中。现有的深层网爬虫技术大部分采用表单填写的方法，按照表单填写方式的不同可分为两类。①基于领域知识的表单填写。这种方法首先构建一个本体库，通过语义分析选取合适的关键词组合填写到表单中，开展后续搜索操作[17-21]。②基于网页结构分析的表单填写。该方法一般无须领域知识或者仅需要一定量有限的领域知识，将网页表单构建成文档对象模型(document object model，DOM)树，在 DOM 树中提取表单各字段值开展搜索操作[22-24]。

2.1.3　元搜索

元(meta)搜索是指基于前导搜索开展的进一步搜索，其通过将其他搜索引擎搜索到的信息进行融合，开展后续搜索。这类搜索引擎没有自己的数据，而是将用户的查询请求同时向多个搜索引擎递交，将返回的结果进行重复排除、重新排序等处理后，作为结果返回给用户。在搜索结果排序方面，有的元搜索直接按照来源引擎的排列展示搜索结果，如 Dogpile，有的则按自定的规则将结果重新排列组合。元搜索引擎的优点是返回结果的信息量更大、更全，缺点是不能够充分发挥所使用搜索引擎的功能，需要用户做较多的人工筛选。这类搜索引擎的代表是 WebCrawler、InfoMarket 等[11]。

元搜索中最重要的技术环节是融合算法的选择，即如何将检索结果融合到一起。在根据不同先验知识提出的各类融合算法中，比较经典的元搜索融合算法有以下四种。

(1)原始分值合成法。当已经知道文档的原始相关性分值，并且这些分值可以直接比较时，可以采用原始分值合成法。该方法直接依据每个文档的原始相关性分值决定其合成排列次序。

(2)规范分值法。若文档的原始分值不能直接比较，则可以通过对倒排文档频率等进行标准化来得到规范的相关性分值，并以之为依据确定文档的合成排列次序。

(3)加权分值法。若能得到文档的原始相关性分值，则可以计算出各个信息源相对于查询条件的重要性，再以此为权重乘以文档的相关性分值作为决定其合成排列次序的依据。