●张 亮 黄河燕 胡春玲

基于 Ontology 的中文问答系统问题分类研究*

摘 要 问题分类是问答系统处理的基础。现在绝大多数的问答系统把问题局限在 person, location, date, quantity, manner, works, organization 等类型。不利于对更多情况和更深语义的问题的处理。可以基于 Ontology 的思想建立完整的、全面的、多层次的问题分类模型。表 1。图 8。参考文献 6。

关键词 问答系统 问题分类 Ontology Hownet
分类号 TP391.1

ABSTRACT Question classification is the basis of a question-answering system. Most question-answering systems limit questions to person, location, date, quantity, manner, works, organization, etc., which are not sufficient for the processing of more diversified and semantic questions. The authors propose to use the conception of ontology to build a comprehensive and stratified question classification model. 1 tab. 8 figs. 6 refs.

KEY WORDS Question-answering system. Question classification. Ontology. HowNet.
CLASS NUMBER TP391.1

现有的计算机检索主要是基于关键字的检索，检索入口是关键字或关键字的逻辑组合，检索结果是与关键字相关的大量的文本或网页。问答系统是一种计算机信息检索的高级形式，它有别于关键字检索系统的特征表现为：检索入口是自然语言形式的问句；检索得到的结果是和问句直接相关的一个词或一句话，即答案简洁明了，与主题直接相关。目前的问答系统各有不同的技术处理，但总体流程都有相似的处理步骤[1]，即问题分析、问题分类、问题形式化、形式化扩展、文本检索、候选句检索、答案抽取。

问答系统总要涉及问题的分类，大多数的问答系统将问题分类作为整个系统处理的开始。所谓问题分类，是指预先根据一定的标准，定义一个问题类型集合，将用户随机的问题在集合中找到对应的类型。显然，问题归类有助于对问题分门别类地处理，因为不同的问题有不同的内涵和形式，答案也有相应的内涵与形式。如，询问时间的问题，在问句中必然含有特殊的信号词(为了表述方便，下面称为疑问词)，在汉语中如“几点”，“何时”，“哪一年”，“什么时间”，“公元前多少年”等，在英语中如 when, what day, which year, what time 等，而在答案中则必须含有相应的时间信息，因此通过命名实体的识别，就可以比较好地从检索到的文章或段落中抽取出答案[2-3]。

1 问答系统问题分类的现状

几乎所有的问答系统都有自己的问题类型集合和相应的分类算法，一般而言，问题的分类与答案的内容相对应。例如“某某是谁?”则答案中心内容为人，即可归类为“HUMAN”，又如“…什么时候发生的?”即可归类为“TIME”。在美国 TREC 评测中，涉及的问题类型主要为人、地方、日期、数量、样式、作品、组织等，这些类型还包含相应的子类型，如“地方”就包含国家、城市、山、河流、湖泊等。

一个典型的代表是新加坡国立大学的问答系统[4]，其分类基于问题焦点和答案内容，是一个目前较为复杂的分类模型。如表 1 所示。子类的划分有利于对问句类型更精确地把握，从而在答案抽取中定位得更精确。例如：How many chromosomes does a human zygote have? 对应 NUM_COUNT; How much does it cost to register a car in New Hampshire? 对应 NUM_PRICE，同样是 NUMBER 型，数目和价格还需区别。

* 本文系国家自然科学基金资助项目(60272088)研究成果。

表 1 一个有代表性的问题分类

大 类	子 类
HUMAN	Person, Man, Woman, Child, Younger
TIME	Day, Month, Year
LOCATION	City, Continent, County, County, Island, Lake, Mountain, Ocean, Planet, Province, River
NUMBER	Age, Area, Count, Degree, Distance, Frequency, Money, Percent, Period, Range, Size, Speed
CODE	URL, Telephone, Post code, email address, Product index
OBJECT	Animal, Breed, Color, Currency, Entertainment, Game, Language, Music, Plant, Profession, Religion, War, Works

问句中含有的问句词(信号词)对问句的正确归类起着非常重要的作用，一些 Q&A 系统建立了可以用于语义层次分析的词典来处理问句词，特别是对于如 what, which, who, when, where, why, how 这样的词。what 的定义位于顶层，询问某特定事物的信息，what person(who)则位于其下一层，询问某个或某些特定的人。图 1 是疑问词 what 和 how 可以引领的疑问内容的进一步细化，其中 how adj 为询问某事的程度。实际分类处理时，要看疑问词和其他词的搭配，如 what 后面跟 agency, company 和 university，则一般可以归类到 organization(机构)中。但有例外，如 Who is the largest producer of laptop computers in the world? 可以很容易地归类到机构，但问答系统仅凭词典和简单的规则很难作出正确判断[5]。

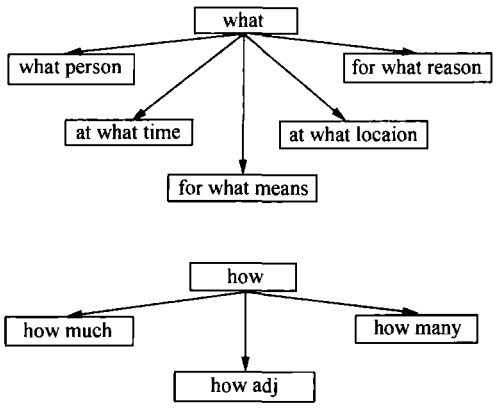

图 1 疑问词语义细化的示例

2 目前的问答系统问题分类的缺陷

以上的问题分类方法对于如 how many 这样的问题，有很好的效果，但还是有许多不足。这样的分类模型依赖于机械的规则，规则不能处理的，就得修改或增加规则，而语言的复杂性决定了要穷尽所有的情况几乎是不可能的。更为关键的问题是，现在的问题分类方法过于机械，可以对付大多数的浅层语义的常见问题，如 TREC 中问句涉及的答案往往与 person, location, date, quantity, manner, works, organization 有关，问题的性质也只是事实性问题、罗列性问题、定义性问题。现有的问句分类方法从根本上讲还是一种应急方法，或是问答系统的初级阶段，很难解决更深和更广泛的问题。例如对于常见的一个问题，“某某是谁”，通常情况下人们想知道的是关于某某的某些方面的信息。但基于现在的处理方法，就需要添加许多的规则。又如：9.11 事件是怎么回事？中国古代有什么重大发明？中国以什么身份加入 WTO？戴妃死于何因？等等，这些问题用现在的方法难以归类。主要原因是，问答系统对问题类型的预设数量太少，而且类型层次不够丰富，不能进行简单的推理判断。

理论上讲，几乎所有普通陈述句都可以转换成疑问句，陈述句中的每一个词都可以是疑问点，如这样一个简单的陈述句“搜索引擎每年为美国的网络商创造 20 亿美元的收入”，其中的每一个词可以作为疑问点，产生对应的疑问句，如图 2 所示，其中有一半问题的类型无法确定。

甚至可以从一个陈述句派生出多个疑问点，或者说对一个陈述句中多个语义信息进行组合疑问。如图 3 所示，可以发现，对同一个陈述句，被抽出的语义点越多，对应的问句就越短，可能的答案范围就越不确定，可以成为答案的内容就越多，疑问点与原句中的语义点的对应就越不严格。当然如果抽出多个语义点，分别作为疑问点，则答案的内容及范围又是明确的(当然这样的疑问方式不常见)。

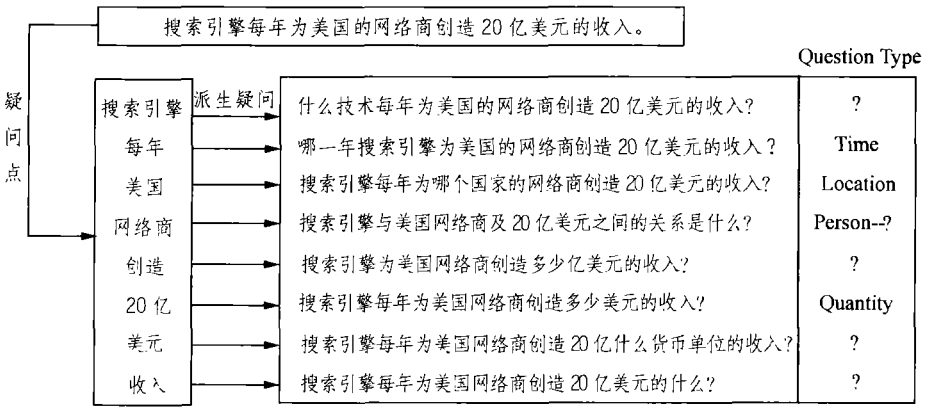

图2 一个简单陈述句可派生出的疑问

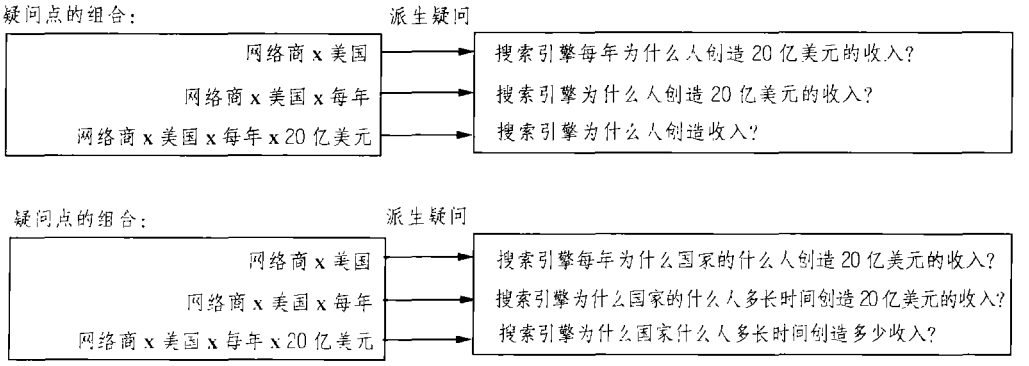

图3 一个简单陈述句中多疑问点的疑问

以上分析说明,疑问或者疑问句的产生方式是多变的,可以成为疑问点的内容是任意的,几乎所有的语义都可以在疑问句中被疑问,因此仅以 person, location, date, quantity, manner, works, organization 等作问句分类的标准显然不能涵盖许多疑问内容,更不要说涉及"为什么""怎么""如果…那么"之类的需要推理、阐述、分析的问题。本质上,问题的范围可以涵盖语言表述的一切范围,包括命名实体和非命名实体、主观和客观的内容、具体与抽象的事物以及各种事物的属性和关系等等,所以必须在更全面的范围和更丰富的层次解决问题分类,Ontology 理论思想及其概念模型为问题分类提供了很好的解决方案。

3 基于本体论的问题分类

3.1 本体论

本体论(Ontology)本是哲学上的一个概念,是对客观存在的一切事物的系统解释或说明,它关心的是客观事实的抽象本质及其相互之间的关联。将本体论引进计算机信息处理是近年来的研究热点,说明要更好地解决信息领域的问题,还是需要从更高的高度,更深更全面地把握事物(信息)的本质。它最基本的表现形式是,一个带有详尽信息和数据结构的便于计算机处理的语义词典或术语表,根据应用的需要可以将它分成不同的层次[6],如图4所示,其中 Top-lever ontology 定义最基本的概念类、属性及其语义关系,例如时间、空间等,领域 Ontology 和任务 Ontology 来细化定义不同的应用领域或具体的通用任务的专用概念类、属性及其语义关系,应用 Ontology 则利用领域和任务的概念集来进一步定义针对每个具体的应用的概念集(例如图书交易概念集)。

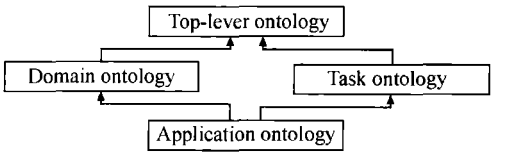

图4 面向应用的多层次 Ontology

3.2 知网

知网(HowNet)是一个以英汉双语所代表的概念以及概念的特征为基础的，以揭示概念与概念之间以及概念所具有的特性之间的关系为基本内容的常识知识库。HowNet 的设计基于在某一时空中运动变化发展着的物质和精神。它运算和描述的基本对象是万物，其中包括物质的和精神的两类，基本运算单元为部件、属性、时间、空间、属性值和事件。如图 5 所示，是 HowNet 中关于实体定义的层次结构的一部分。

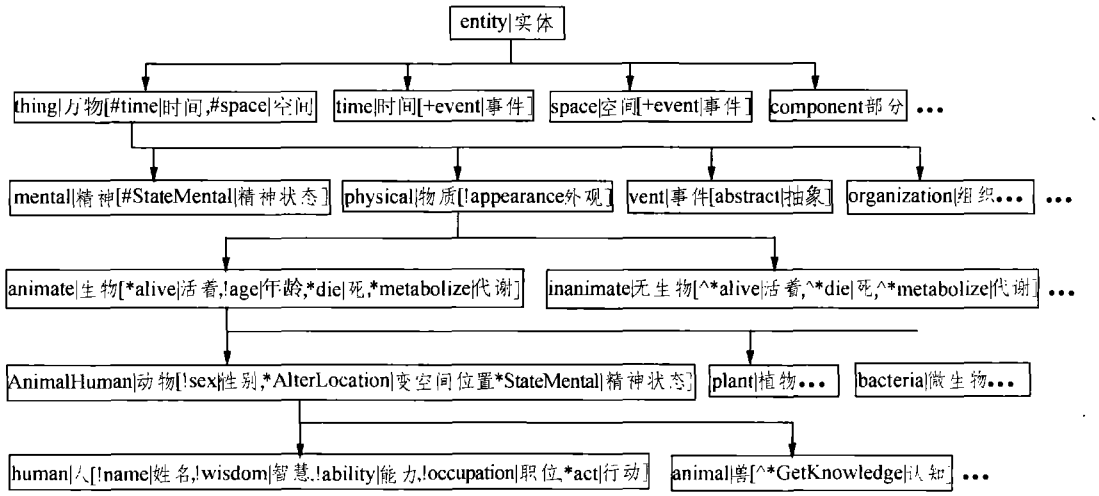

图 5 HowNet 中的实体

知网较全面地描述了概念之间和概念的属性之间的各种关系，不仅包含同类概念之间的关系，如上下位关系、同义关系、反义关系、对义关系、部件与整体关系、属性和宿主关系等，还包含非同类概念之间的关系，如属性值和属性的指向关系、事件和角色关系。如图 6 所示，由义原作为基本概念单位，由实体、事件、属性以及属性值构筑的语言描述体系的 HowNet 相当于 Top－lever ontology，对汉语言的基本概念及其表示方法和结构层次有较好的刻画，可以以此为基础构建面向问题分类甚至是整个问答系统任务的 Task ontology。

图 6 HowNet 对语言的解构

3.3 基于 Ontology 的问题分类

任何事实、任何知识以及它们的组合，都可以派生出问题。一般情况下，简单事实派生简单问句，复杂事实或事实的组合派生复杂问题，浅层知识派生浅层问题，深层知识派生深层问题。要全面深刻地解析问题，必须有全面和多层次的语义知识库做支撑。Ontology 理论及方法，正好适应了解决这个问题的内在需求。如图 7 所示，从计算机的角度看，一个事实或知识或它们之间的关系是一个黑体，其中的每一个部分(语义点或其组合)，都可以成为疑问点，而 Ontology 就如同一张网一样，固定住黑体中的语义点，解析各点之间的关系，判断已知和求解(未知)之间的约束条件。

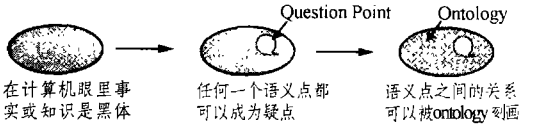

图 7 事实(知识)、疑问点与 Ontology 的关系

如对于“某某人是谁?”这样的问句，根据前面从陈述句派生疑问句机理分析，可以知道，疑问越短，疑问的句型越容易把握，疑问的辐射面越广，答案涉及的内容越多，疑问点与答案之间的语义约束越不严格，因此用以往的机械的类型分类的方法就不易处理，而通过 HowNet 的帮助，就可以找出一条解决这类问题新路。疑问点是“某某人”，除了知道是人外，没有别的约束，可以查找 HowNet 中有关人的定义，如图 5 所示，我们可以从智慧、能力、职位、行动等方

面去回答这个问题，还可以进而扩展到“人”上位词“动物”，和上上位词“生物”，这样，我们可以将“某某人是谁?”转化为图8所示的一些具体的、疑问点明确的问题。

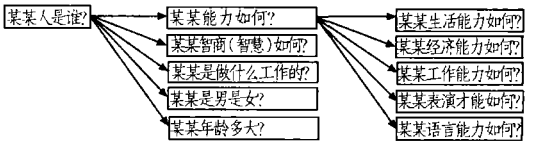

图8 复杂问题转化为具体问题的示例

当然，需要在HowNet的基础上，建立面向问题分类的或面向问答的task ontology。因为问题分类还有许多特有的性质，就比如“某某人是准?”，通过高层ontology可以将抽象问句具体化，但是任意具体下去，将会把一些无关紧要的属性也扩展进来，如“身高”、“体重”等等，这就需要建立task ontology，对扩展属性相关度或重要程度设置权重，进行排序，取大于设定阈值的属性。进而可以在此基础上建立一定的推理功能，如先对各项指标进行检索，对检索到的文献做定量分析，如收集到关于此人的文献中体育方面的资料多，则推断他可能是体育界人士或体育明星，最终抽取答案时，将与体育相关的属性的权重加大。

同样，对于“9.11事件是怎么回事?”这样的问题，也是借助ontology先将抽象问题具体化，事件=人物+时间+地点+工具+结果+原因+……，根据需要可以进行多次多层次的具体化。即大问题=小问题1+小问题2+小问题3+……，抽象问题=具体问题1+具体问题2+…

面向问题分类的task ontology和其他的task ontology一样，一方面要考虑问题类型的内在关系和答案抽取的最终需求，另一方面要考虑如何充分利用如HowNet这样的top-lever ontology提供的语义资源。建立the task ontology可以从以下几个方面深入展开：

(1)问句形式结构的种类。问句内容无限多样，但问句表现形式是有限的，能起发问作用的疑问词是有限的。疑句词如何与其他词搭配形成问句，搭配的规律是什么，这些应该是task ontology中的一个重要内容，因为它是分析处理的基础。疑问词中，“什么”的搭配能力最强，如“什么人”、“什么时间”、“干什么”、“是什么”、“为了什么”等等。

(2)同一问题多形式表示。汉语言中许多意思可以用不同的词和句子表示。问句也是一样，语义相同的问句往往可以有几种不同的表达形式，如询问时间就可以有“什么时间”、“几点”、“何时”等，询问人有“谁”、“什么人”、“哪个”、“哪一个”、“何人”等。如何向系统提供判断不同问句形式之间是否具有等价语义的能力，以及具有等价语义的不同问句形式之间的转换能力，是task ontology需要建立的另一个内容。

(3)问题分类的不同角度。不同的角度可以有不同的问题分类，如形式上分疑问、设问、反问，或特指问、选择问、是非问；目的上分查找信息、验证事实、收集资料；从性质上分开放型、封闭型等等。最重要的是从内容上分，可以直接利用top-lever ontology的概念分类体系，较全面多层次进行分类。从不同的角度分析问句，有利于更准确地把握问句的含义，从而保证最终抽取出正确答案。

(4)问题复杂性的谱系。问题有简单问题和复杂问题，但什么是复杂问题，什么是简单问题，它们的特性各是什么，这也是问答系统里有潜在价值的问题。将问题的复杂性加以区分有利于对不同问题采取不同策略和算法、分门别类地处理。分析问题的复杂性可以从多个角度出发，问题开放程度是一个考虑方向。一般来说，封闭的问题较容易回答，最封闭的问题是是非问题或多选一问题，如“中国位于亚洲还是美洲?”而开放的问题较难处理，往往没有直接的答案，如“9.11事件是怎么回事?”“台海局势将如何发展?”

(5)基于ontology的简单推理机制。概念和概念是相互关联和相互制约的，开放性问题和语境相关问题往往需要通过Ontology提供的概念关系和上下文而做出一定的判断和推理。如“克林顿是谁”与“姚明是谁”这两个问题在类型上是一样的，都是查询某人的资料，但回答的侧重是不同的。首先都是从Ontology中获取人的相关属性进行检索项的扩展与细化，再从检索项扩展后检索到的文献中判断特征属性，如前者和政治概念相关度高，后者与篮球概念相关度高，在答案抽取中做相应的取舍，突出重点。

4 总结

将问题的类型局限于person, location, date, quantity, manner, works, organization，可以解决大部分浅语义的简单问句。但从根本讲，问句涉及的内容范围是无所不包的，而且涉及的语义是多层次的，因此这样的问题分类和以此为基础的问答系统难以进一步拓

展和深化。Ontology 是对客观存在的一切事物的系统解释或说明，它关心的客观事实的抽象本质及其相互之间的关系，我们正可以基于 Ontology 的思想建立完整的全面的多层次的问题分类模型。

现在许多的问答系统也用到像 WordNet、HowNet 这样的语义资源，但仅仅是在处理的某一部分作为工具使用，并不是完整的基于 Ontology 思想的体系结构。

参考文献

1 Jimmy Lin and Boris Katz. 2003. Question Answering Techniques for the World Wide Web. The 11th Conference of the European Chapter of the Association of Computational Linguistics(EACL－2003)

2 Yi Chang, Hongbo Xu, Shuo Bai. 2003. TREC 2003 Question Answering Track at CAS－ICT. In the Twelfth Text REtrieval Conference(TREC2003).

3 Zhiping Zheng. 2002a. AnswerBus question answering system. In Proceeding of 2002 Human Language Technology-Conference(HLT 2002).

4 Yang, T. S. Chua, 2002. The Integration of Lexical Knowledge and External Resources for Question Answering. In the Proceedings of the Eleventh Text REtrieval Conference.

5 Kenneth C. Litkowski. 2003. Use of Metadata for Question Answering and Novelty Tasks In the Twelfth Text REtrieval Conference(TREC2003).

6 R. Studer, V. R. Benjamins, D. Fensel. 1998. Knowledge Engineering: Principles and Methods. Data & Knowledge Engineering. 25:161－197.

张　亮　南京理工大学计算机系博士研究生，江苏警官学院科技系教师。通信地址：北京北四环中路 257 号科群大厦西楼。邮编 100083。

黄河燕　中国科学院计算机语言信息工程研究中心博士生导师。通信地址同上。

胡春玲　中国科学院计算所博士。通信地址同上。

（来稿时间：2005－06－07）

深圳市公共图书馆图书“通借通还”全面开通

深圳市“图书馆之城”建设又有新举措，从 2005 年 12 月 18 日起，深圳市又有 4 个区图书馆加入深圳市公共图书馆“通借通还”系统。至此，深圳市、区公共图书馆全部实行联网，深圳图书馆与 6 个区图书馆之间实现图书“通借通还”。读者持有一张借书卡，便可在市图书馆和 6 个区图书馆之间通借通还图书，享受到一卡通行的便利。

图书“通借通还”服务是深圳市“图书馆之城”建设推进办公室 2004 年推出的方便市民借阅图书的一项新服务。2004 年 8 月 12 日，深圳市公共图书馆图书“通借通还”系统试运行。深圳图书馆、南山图书馆、宝安图书馆加入该系统，三馆联合开展图书“通借通还”服务。一年多来，开通“通借通还”功能的读者逐渐增多，图书借阅册次明显增加，“通借通还”系统运行平稳，图书物流运送正常，极大地方便了读者就近借还书，受到广大读者的好评和赞扬。一些读者来电话或来函，感谢图书馆为方便市民读书办了件大好事。为满足广大市民日益增长的阅读需求，深圳市“图书馆之城”建设推进办公室决定将图书“通借通还”范围扩大到福田、罗湖、盐田、龙岗等 4 个区图书馆。2005 年 11 月，深圳图书馆网络中心技术人员到 4 个区图书馆对各馆的网络系统进行了“通借通还”系统的安装调试，系统运行正常。12 月 18 日，深圳市 7 个市、区公共图书馆正式开通“通借通还”系统，为读者提供图书通借通还服务。

深圳市公共图书馆全面实行图书“通借通还”后，深圳图书馆、南山图书馆、宝安图书馆、罗湖图书馆、福田图书馆、盐田图书馆和龙岗图书馆七个馆的读者，通过申请开通“通借通还”功能，可享受七馆开展的中文图书“通借通还”服务。开通了此项服务的读者，除享受原借书证所规定的借阅服务外，新增了两项服务功能：一是读者可持七馆中任一图书馆的借书证到七馆中任一图书馆借阅中文图书，数倍地扩大了借阅范围；二是读者在七馆中任一图书馆所借的中文图书，可就近在七馆中任意一家图书馆归还，免除了读者往返奔波之辛苦，极大地方便了读者。同时，七馆推出的图书“通借通还”实行免费服务，读者申请开通“通借通还”服务功能，不另收取图书押金和手续费。为方便读者查询和借阅，深圳图书馆在“网上深图”网站主页上设置了“图书馆之城”、“通借通还”栏目（网址：http://www.szlib.gov.cn/tongjie/），设置了“七馆书目查询”、“读者借阅信息查询”等功能，方便读者在网上查询七个公共图书馆的书目数据和读者借阅信息。

深圳市公共图书馆实行联网，全市公共图书馆实现图书“通借通还”，标志着深圳“图书馆之城”建设迈上了一个新台阶。开展图书“通借通还”服务，一方面极大地方便了广大读者就近、便捷地使用图书馆文献资源，是实现市民文化权利的具体体现；另一方面，通过联网和合作，进一步加强深圳公共图书馆之间的联合，逐步构建深圳市公共图书馆网络共享平台，为全市公共图书馆文献资源与服务的整合，实现文献资源共享打下了良好基础。

（余胜）