Deep Web 研究现状与展望

高 明 黄哲学
（中国科学院深圳先进技术研究院深圳市高性能数据挖掘重点实验室 深圳 518055）

摘 要 随着Deep Web数量和规模的快速增长，通过对其发起查询请求以得到存储在后台数据库中的相关信息，日渐成为用户获取信息的主要方式。为了方便用户有效地利用Deep Web中的信息，越来越多的研究者致力于这一领域的研究，重点之一是Deep Web后台数据库的数据集成。由于Deep Web后台数据库存储的主要是文本信息，使得从文本处理角度出发，针对Deep Web中存储的内容进行查询与检索的研究具有十分广阔的应用前景。本文对Deep Web的研究现状进行了较为详细的分析，同时对研究的发展方向进行了展望。

关键词 Deep Web；Web数据库；查询接口；Web数据集成

A Review on Deep Web Research and Prospects

GAO Ming HUANG Zhe-xue
(*Shenzhen Key Laboratory of High Performance Data Mining, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract With the rapid increase in numbers and scales of deep web sites on the Internet, search for data or information from deep web sites by submiting queries to and obtaining results from the backend databases has become a major means in information retrieval from the Web. This area has attracted many researchers to devote their efforts on development of technologies to make better use of information in th deep web. One challenge is searching for and integration of data from various databases in deep web. Since deep web is dominated by text data, research and development of technologies for text information retrieval from deep web have a broad application potential. In this paper, we review the state-of-the-art of deep web research in details and propose some future research directions.

Keywords deep web; web database; query interface; web data integration

1 引 言

随着万维网技术的发展，网络中出现了越来越多的在线数据库，这些数据库的共同特点是可通过填充特定的查询表单来使用其中的数据。这些数据一般不被搜索引擎通过静态链接而得到[1]，而是需要通过HTML表单提交查询，由服务器根据请求动态生成页面，并把相应的结果返回给客户端。我们通常把这些隐藏在后台的在线数据库称为是Deep Web[2, 3]，也称为Hidden Web。页面内容根据用户

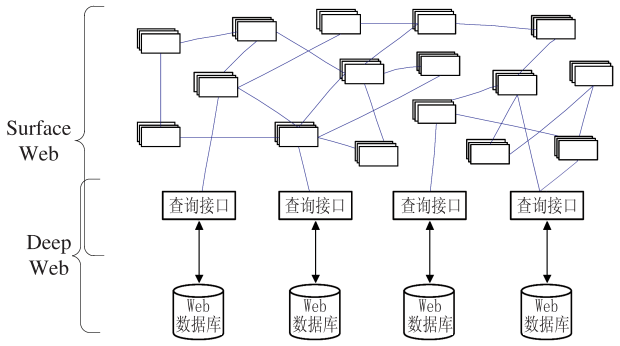

图1 传统Surface Web与Deep Web的关系示意图

作者简介：高明，博士后，研究方向为数据挖掘与推荐系统，E-mail: ming.gao@siat.ac.cn；黄哲学，研究员，博士生导师，研究方向为数据挖掘与商务智能。

的请求动态生成，且没有直接指向这些信息的超链接，这是Deep Web中数据和传统万维网静态页面的本质区别，如图1所示。

由于搜索技术的限制，传统的搜索引擎一般只抓取并索引网络中那些可以通过超链接直接指向的静态页面、文件等，而很少能够索引这些隐藏在后台数据库中的资源。相对于Surface Web，Deep Web蕴藏的内容更加丰富、更加专业，这些资源大多是高质量、权威的信息，检索查询Deep Web的信息成为搜索技术的研究热点。

早在2004年，Chang[1]等人对全球的Deep Web规模进行调查后得到的统计数据是：整个Web上大约有307,000个Deep Web站点使用了后台数据库，数据库总量在450,000个左右，其信息量大约为普通静态页面的500倍左右。这一数据是Bergman等人在2000年做的类似调查结果[2]的6~7倍。

Deep Web在中国的增长情况同样迅速，2005年7月CNNIC发布的第十六次中国互联网信息资源数量调查报告显示[4]：中国在线数据库的总量为3.06万个，中国网站中拥有在线数据库的网站个数为16.1万个，约占全部网站的24.1%。以拥有在线数据库的网站为基准，全国平均每个网站拥有1.9个数据库。在线数据库总数量年增长近13万个，增长率为80%。2010年，清华大学以搜索引擎抓取到的数据为基础，通过实验给出了一个最新的中文Deep Web规模的统计结果，“当前中文深度万维网上有60多万个查询接口，与2006年初人们统计所得的74000个项目相比增长了近7倍”[5]。

从Deep Web的系统架构来看，其页面的内容存储在后台的网络数据库中，这些页面是由用户通过特定的查询接口提交查询而生成的。虽然Deep Web数据库中的内容对用户来讲是十分有用的高质量信息，但是由于Web数据库数量众多，仅通过人工遍历所有的Web数据库并通过其接口提交查询来获取相应的数据对人力和时间的耗费无疑是巨大的。因此，国内外研究人员的重点越来越多的倾向于Deep Web信息的集成，以方便用户对其蕴含信息的使用。

从应用角度来讲，通过Deep Web数据搜索平台可以获得更为实用的数据和服务。对于互联网门户来说，可利用Deep Web数据搜索平台来收集、存储、分类和索引各类信息，为用户提供更方便和个性化的Web信息搜索服务。因而对Deep Web数据处理的研究将会产生可观的经济和社会效益。

2 Deep Web主要研究内容概述

面对大规模的Deep Web信息，由于针对某个具体领域的信息有许多共性的特征，因此一个折中的方案就是基于领域的集成。目前来讲，针对大规模的数据集成需要解决以下几个问题：

2.1 数据源的发现与识别

Deep Web虽然蕴含了丰富的信息，但是它没有实际的物理页面存在，而是需要用户填写相应的查询接口，然后动态的从数据库中提取相关内容并格式化后以虚拟结果页面的形式返回。这就使得查询接口的发现与识别在Deep Web集成系统中占有十分重要的地位。

从形式上来说，Deep Web查询接口均以表单的形式出现在页面中，因此利用表单的结构特征来作为Deep Web查询接口的判断依据是一种合理的解决方式。但是，并不是所有的表单都是查询接口，比如：用户注册、bbs留言板、搜索引擎等也都以表单的形式出现，却均不属于查询接口的范畴。因此，在实际应用中需要对Web中的表单进行提取和识别，识别出那些真正的查询接口。

同时，互联网中为数众多的查询接口又时刻处于变化之中，不时有新的查询接口出现，旧的查询接口可能会被改动甚至消失，这无形中增加了识别查询接口的难度。

2.2 数据源的分类与集成

Web上网络数据库的数量众多，对这些网络数据库的使用只能通过填充其各自的查询接口来进行。对于某一领域的问题，可能存在数量众多的网络数据库，这些数据库的内容不同而且每个数据库都有自己的查询界面，使用时需要用户消耗大量时间去访问多个数据源去查找所需的内容。

为了解决这一方面的问题，需要在众多网络数据库的基础上构建一个虚拟的统一的集成系统。即通过检索界面集成来给用户提供一个针对某一特定领域网络数据库的统一的访问接口。这一过程的本质就是根据各自独立的Deep Web数据库本地视图生成一个建立在这些视图之上的全局视图。

Deep Web查询集成过程如图2[6]。

集成后的检索系统将用户的查询请求转换成对多个Web数据库的查询，并接收查询结果。进而由集成系统将查询结果进行归并和排序等处理后将内容进行归一化处理返回给用户，由此来简化用户的检

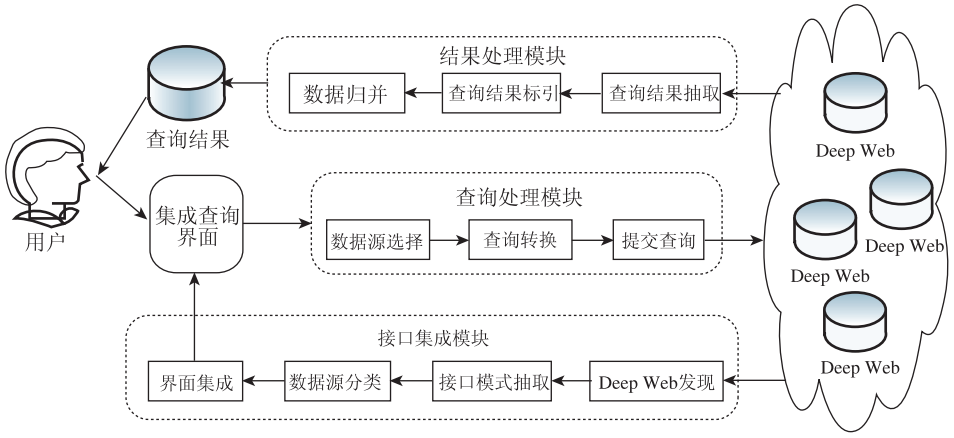

图2　Deep Web的查询流程

索操作。

在完成查询接口的集成后，通过集成查询接口对Deep Web的访问过程如下：

（1）用户通过查询界面填写查询表单，并提交查询请求；

（2）查询集成系统根据集成查询接口和各独立数据库系统之间的映射关系将用户的查询转换成针对各个独立Web数据库的查询并提交转换后的查询请求；

（3）查询处理的后台服务器程序处理该查询请求；

（4）集成查询系统接收各独立Web数据库返回的数据，并合成结果；

（5）将查询结果返回给用户。

2.3　数据源的选择与查询转换

由于Deep Web的自治性，造成了Deep Web可接受请求受限于其查询接口的表达能力，其后果是集成后的查询接口与各目标查询接口的查询能力不等价，进而造成返回的查询结果不满足用户查询的需求。因此，需要对通过集成查询接口提交的请求针对各Deep Web目标查询接口进行查询转换，即，把集成查询接口S上的用户查询“翻译”到多个目标查询接口T上，从而在多个目标查询接口上产生相应的查询。

在将用户查询从集成查询接口转换到某一目标查询接口时，应找出从集成查询接口映射到目标查询接口所有查询所构成的集合，称之为联合查询Q_t。但为了尽可能降低在数据传输和查询结果处理阶段的开销，在实际操作过程中，大多采用基于最小超集的查询转换方法。一般是使用过滤器对联合查询Q_t进行过滤，找到其最小包含，以减少不相关的检索结果，并尽量使过滤后的查询与原查询保持一致。

对于某些相同类型的数据，在各目标查询接口之间可能存在数据范围或者数据格式之间的不一致问题，如何在集成查询接口向目标查询接口映射时对这种不一致进行统一，是目前查询转换亟待解决的问题。另外，目前从集成查询接口到目标查询接口之间的映射大多是通过事先给定的规则来进行的，当某一目标查询接口结构发生变化时，如何即时更新规则也是查询转换方面需要进一步研究的课题。

2.4　Deep Web数据的抽取

Deep Web的返回结果是HTML标注的结果页面，其中HTML标注只是用于格式化页面的显示。Deep Web数据的抽取主要是指基于HTML结构的信息抽取，也就是从结果页面中提取用HTML标签标注的无结构或者半结构化的内容，并以计算机可自动处理的结构化的格式进行保存。

目前针对基于HTML结构信息抽取的主要途径是：通过解析器将文档解析成相应的语法树[7]，借助于自动或半自动方式产生的抽取规则，把信息抽取转化为对语法树的操作来实现信息抽取。将此方法应用于抽取实践中时，大多将一个HTML页面转化为一个对应的DOM树，将页面中的标签转化为内部节点，而将具体的文本内容、图像等转化为叶节点的方式来进行。在此基础上，又有研究人员提出了“风格树”的概念，将一系列的标签序列识别为一个风格节点，以简单DOM树方式很难通过单个或少量HTML页面学习到一组页面所有布局的问题。

Deep Web数据抽取的另一个思路是基于“网页去噪”来进行。对于网页中出现的类似版权信息、公司信息等内容可被看作是页面噪声，此方法主要是通过去除“网页噪声”来获得所需的内容。目前常用多个网页的正文信息对比的方法来检测页面中的公共字串，并将长度超过某一阈值的字串看作是网页噪声予以去除。

目前Deep Web数据抽取面临的问题主要集中在：无论是通过语法树生成抽取规则方面，还是在网页去噪学习阶段，都需要较多的人为干涉，这不利于大规模Deep Web数据抽取的进行。如何通过对页面内容自动聚类，并通过剔除网页中的噪声来实现对Deep Web数据的自动抽取，将是Deep Web数据抽取的一个重要研究方向。

3 Deep Web相关问题的研究现状

随着越来越多的网络站点开始使用动态页面发布信息，Deep Web也引起了研究人员的广泛关注，不仅期刊论文数量逐年递增，而且在很多国际顶级会议如WWW、VLDB等也刊载了与之相关的研究成果。

3.1 国外研究现状

自Deep Web概念被提出以来，在国外的高校、科研及商业机构中逐渐掀起了针对Deep Web资源研究的热潮。为了能够使用户可以更加方便的使用这些隐藏在后台数据库中的资源，这些研究遍及查询接口的自动发现与集成、数据抽取、数据库分类等方面。

3.1.1 数据源发现

数据源发现是对Deep Web进行集成的前提与基础。由于互联网上的Deep Web数据源不断变化，不时有新的数据源产生，同时也有大量的数据源会发生变动甚至消失，因此要求Deep Web集成系统应能够主动的去发现互联网上的数据源。由于访问数据源后台数据的主要入口就是查询接口，而查询接口又主要以表单的形式体现，因此数据源发现的主要思路就是对包含表单的页面进行特征提取，进而对表单进行判断是否为数据源的入口。

基于此思路，在数据源的发现方面主要的解决方案大致有如下几种：

首先是利用互联网上已经存在的很多专业针对Deep Web数据源进行统计的站点，如completeplanet.com和invisible-web.net等。虽然这些站点不可能统计所有的Web数据库，但是这些Web数据库都已按照领域做了分类，对于小规模的集成来讲仍然是一个有效的方案。

与此类似的一个方案就是利用搜索引擎进行搜索[8]，虽然搜索引擎不能获取Web数据库中的内容，但可以用来找到Web数据库所在的网站。其关键在于如何向搜索引擎提交有效的查询，使得含有Web数据库的网站尽可能多地出现在查询结果中，并使其排名尽量靠前。由于必须向搜索引擎提交查询，因此这种方案是基于某个领域的Web数据库的发现，也使得此方案更加具有实际应用意义。

同时，部分研究人员在传统搜索引擎爬虫的基础上开发出了Deep Web爬虫，并通过对网页的爬取、分析来定位Deep Web数据源[9, 10]，在此基础上文献[11][12][13]增加了对数据源所属主题的分类，使得通过Deep Web爬虫发现的数据源更加具有针对性，同时也减轻了后续在数据源分类阶段的工作量。

数据挖掘和机器学习方法在查询接口发现方面也有比较好的应用，如Lage等人提出了两个根据实际经验总结出的启发式规则来判断网页表单是否为查询接口[14]；Jared等人则首先提取查询接口的特征，在这些特征之上利用C4.5算法得到一棵决策树，通过这棵决策树来找出真正的查询接口，从而实现了查询接口的识别[10]。

为了尽可能多的发现存在于网络中的数据源，有科研人员提出了通过遍历互联网的方式，首先收集WWW站点的IP地址，进而按照一定的策略遍历所有[15]。虽然这种方案在理论上可以把所有的Web数据库完整地找出来，但如果遍历所有几亿个IP显然代价太高，因此大多仅作为一种研究统计的手段。

3.1.2 查询接口的识别

目前，对查询接口的识别主要是通过两种方式来实现：

一类是提交查询法：通过提交试探性查询，根据返回的结果进行判断是否为Deep Web查询页面。该方法一般通过表单的结构特征，按照一定的策略自动填写表单，并根据返回结果的情况来对其是否为Deep Web查询接口进行判断[16]，此方法适用于结构简单的搜索页面。采用这种方法时，虽然加大了网络的开销，但是可以根据返回结果的情况对数据库内容进行分析，从而可以对Deep Web接口进行较高精度的识别。

另一类是非提交查询法：直接利用网页表单的结构信息，对控件的类型、其内部属性以及描述控件的标签进行特征提取，从而实现对查询接口的判断。当数据库中表结构可以完全由页面表单的特征来表示的时候常采用此方法。

查询接口识别也可被看作是查询接口理解的问题[17]，可以将查询接口理解分为4个阶段：建模、（模型的）语法分析、分割和局部处理。建模阶段将一个查询接口视为后台关联数据库的一系列查询序

列。此序列在语法分析阶段被分割为一个可处理的结构，例如文[18]将一个建模后的查询接口视为一个包含3个组件的字符串，其中：‘t’为任意文本，‘e’为任意表单组件，‘|’为行分隔符；查询接口完成建模和语法分析后，被分割成为针对关联数据库的查询片段；然后在局部处理阶段重点关注与后台数据库内容抽取相关的完整性约束，如文[19]使用机器学习分类提取诸如领域相关的数据类型和个体元素。

针对现有研究主要通过复杂查询接口（SQI）对Web后台数据库（WDB）作探测查询和模式识别的方式。文[18]提出了一种基于SQI的WDB探测查询和模式识别方法：在查询接口模式识别阶段，提出了基于SQI的满条件查询定义及其生成策略。在结果模式识别中，通过在服务器返回的结果页中寻求可作为扩展的非查询接口关键词的方式，大大提高了结果模式识别的属性召回率。

由于网页表单很容易获取，并且非提交查询法比较适合对内容和结构多变的表单进行判断，因而大部分研究者倾向采用此方法来识别Deep Web查询接口。但是到目前为止，尚无一种高效的方式来实现查询接口的自动发现与识别。

3.1.3　数据源的分类与查询接口集成

由于Deep Web数据源广泛分布于互联网中，虽然其蕴含了丰富的信息与资源，但是用户在使用时必须从找到的众多的数据源中检索出自己需要的数据源，以便进一步的信息检索。为了方便用户的使用，必须对数据源进行分类，并根据类别对查询接口进行集成显的尤为重要，这也是Deep Web信息集成系统的核心工作之一。

纽约州立大学和伊利诺伊大学合作开发Deep Web数据库集成项目WISE[20-22]，以提供集成访问Web数据库技术为目标，针对Deep Web信息集成的整个过程展开了研究，最终开发了用于查询接口聚类的WISE-Integrator软件，和用于查询接口抽取的工具WISE-iExtractor。

哥伦比亚大学的Qprober[23]研究小组主要关注于Deep Web后台数据库的自动分类。

通过机器学习生成一套基于规则的分类器，然后通过分类规则来重写查询url以实现数据库内容的探查，并根据返回的结果来对数据库进行分类。在此基础上，针对非结构化的文本数据库，进一步提出了对其内容摘要选择的算法[24,25]。

亚利桑那州立大学的Deep Web信息源集成系统Factal[26]使用了SourceRank[27]的方法，它基于不同信息源之间信息的可信和相关性，完成了查询接口的集成与数据抽取，主要包括信息源爬取、一致性评估、SourceRank计算、匹配评估和在线查询处理几个步骤。该系统建立在电影和图书领域的40个在线数据库之上，使用了以“集成接口与数据源直接映射”的方式为用户提供查询服务。

3.1.4　Deep Web数据的抽取

早在1997年，华盛顿大学就开发了通过自动填充表单来获取在线商品详细信息的系统ShopBot[28]，并能自动的匹配出最符合用户需求的商品。虽然此时Deep Web的概念刚刚提出，而且ShopBot针对的对象范围也很狭窄，但是ShopBot系统已经初步实现了后来Deep Web系统的最基本功能。

斯坦福大学的Deep Web爬虫HiWE[9]通过使用url重写技术来对网络爬虫进行改进，使其能够自动识别出页面表单及其各控件，然后自动从预先准备好的数据集合中匹配出合适的数据来完成对各表单控件属性值的填充，即自动构造请求字符串，然后，以经过重写的url作为爬取的位置来抓取数据。这也是后来Deep Web数据抽取技术中的主要技术之一。

伊利诺斯大学的MetaQuerier[29]研究小组建立一个元查询系统，目标是有效获取Web上结构化的信息。首先，MetaExplore项目聚焦于发现、模型化和重构Web数据库来建立一个可搜索的数据源知识库。特别是，MetaQuery系统开发了一个Web数据库搜索引擎，它可以发现Web上含有数据库的站点，设计模型来描述这些数据库，设计包装器自动抽取这些模型中的参数，重组和索引可搜索的Web数据库。其次，MetaExplorer项目聚焦于集成在线数据库，同时还研究了数据源选择，查询转换和模式集成问题。在研究大规模信息集成的过程中将依赖于前面所建立的数据源知识库，其研究重点是动态信息集成技术。与传统的信息检索不同，MetaExplorer项目的MetaQuery系统是动态的，即可以将实时发现的新数据源加入系统中，同时可以动态选择数据源并将用户查询进行相应转换，从而获取用户查询结果。

3.2　国内对Deep Web的研究与探索

国内的各大高校和科研院所对Deep Web也展开了广泛而有效的研究：

中国人民大学的孟小峰、刘伟等在国内较早介绍了Deep Web数据集成方面的内容[8]。该团队在Deep Web结果页面内容抽取方面，创新性的分析了几种结

果页面的视觉特征，并使用结果页的视觉特征来完成抽取工作[30]，此方法最大的特点是抽取过程与页面语言种类无关，适合在多语种环境中的使用。同时他们在查询接口集成和数据抽取方面，首先通过将关键词匹配到相对合适的领域，进而将集成接口查询根据查询领域转换到目标接口查询的方式[31]，解决了Deep Web集成中领域众多和集成后查询接口结构过于复杂的问题。

东北大学申德容团队在近几年内也取得了丰硕的成果。他们将从Deep Web中抽取到的数据转化为实体记录的形式，进而在异构记录处理模块中，利用在同构记录处理模块所得到的权值、计算各实体记录的相似度，最终得到从不同数据源中抽取到的重复记录[32]。文[33]将数据抽取工作分为结果模式生成和数据抽取两个阶段，属性语义标注放在结果模式生成阶段来完成，有效解决了重复语义标注问题，同时针对嵌套属性问题提出一种有效的解决方法。针对Deep Web环境中，特别是在从集成查询接口到本地具体查询接口的映射过程中，存在的查询转换失败问题，他们提出了一种应用于Deep Web数据集成系统中的查询松弛策略[34]，将Deep Web资源按查询接口属性分组，合并成全局数据源关系图(DRG)，再针对特定查询将DRG转换为对应该查询请求的数据源关系图，然后按特定的规则进行查询松弛和执行处理。针对如何有效地抽取Deep Web中结果页面所包含的实体信息问题，文[35]通过分析Deep Web结果页面的特点，提出了一种基于DOM树的抽取机制，有效的解决了Deep Web环境中的实体抽取问题。

苏州大学的崔志明、赵朋朋等对Deep Web的研究集中在数据源发现、分类以及查询接口模式抽取等方面，提出了基于视觉特征的查询接口自动抽取方法[36]，总结了Deep Web查询接口和结果页面的视觉特征，运用网页分割算法对接口和结果页面进行分割和有效的模式抽取。在查询接口分类方面提出一种基于查询接口特征的Deep Web分类方法和基于查询接口连接图的Deep Web聚类方法[37-39]，从而达到了对Deep Web数据源按领域分类的目的。

复旦大学的胡运发、徐和祥等研究人员在Deep Web查询接口分类的方面做了大量细致而且有效的工作，在提出了Deep Web查询接口本体模型和建立本体推理规则的基础上，建立了基于本体模型的查询接口的向量空间模型[40]，通过实验验证得到了较好的分类效果。针对数据库分类问题，他们提出了一种Web数据库表示模型和基于机器学习的分类方法，并给出了一种基于语义的权重计算方法[41]，使用这些方法，仅需要少量样本的训练，就能达到较好的分类效果。该团队还就Deep Web集成环境变化处理问题进行了研究[42]，提出了一种基于知识的环境变化处理方法，包括Deep Web集成环境变化处理模型以及适应Deep Web环境变化的动态体系结构和处理算法，这些成果可对大规模Deep Web集成的进一步探索和应用走向提供参考。

清华大学在Deep Web的规模估计[5]、查询接口及查询模式识别[20]等方面也做了很多有意义的探索。

为了推动Deep Web数据集成在国内的发展，《软件学报》在2008年推出了Deep Web数据集成专刊，刊登了国内在数据集成方面的最新进展，展现了国内在此方面的较高水平的研究成果与项目。

3.3 语义层面的Deep Web研究[6]

总体来说，语义层面的Deep Web研究还在起步阶段。

很多机构在语义Web层面研究Deep Web资源的一致访问规范[43,44]，对Deep Web资源内容语义标注和Deep Web资源进行分类索引。目前利用元数据或者本体概念来标注Deep Web资源的研究正在起步[45]。元数据是用于网络信息资源的识别、描述和定位的数据，经过整理加工后的元数据通常存放在数据库服务器上供用户使用。网络动态信息资源采用统一的元数据加工和标引后，形成基于RDF的可共享的XML数据格式，可以实现资源的长期保存和发布。此外，通过构造本体嵌入的语义搜索引擎，在搜索端增加知识本体及相应的推理功能。利用语义Web技术，本体对查询请求的语义分析及扩展技术，在语义词典或本体的协助下进行语义匹配计算完成检索。Deep Web和语义Web的研究将会相互渗透形成语义Deep Web。

4 Deep Web研究发展趋势

Deep Web的研究总体上来说将随着业界技术的发展而演变，并趋于与其他研究领域相结合，进一步提高分析、处理过程的自动化程度和精度。其中的主要结合方向是语义技术、文本挖掘和自然语言处理等领域。

4.1 研究领域将紧跟业界流行技术

Deep Web与互联网技术的发展紧密相关，随着互联网技术的发展，Deep Web研究的方向也将随之发生

演变。

例如，开放API是当今互联网领域的一个重要热点，与传统的通过查询接口获取Deep Web数据相比，通过开放API能获取数量更多、质量更高的后台数据库中的信息，但此领域的研究尚处于探索阶段[47]。

再者，最早的查询接口大多是静态的HTML表单，过去对Deep Web的研究大多基于此规范。随着互联网技术的进步，近年来Ajax技术在互联网领域快速发展，但关于在嵌入了Ajax技术的Web应用程序中，对用户交互的形式化描述一直欠缺。牛津大学的学者尝试了使用XPath的超集OXPath来虚拟用户与系统的交互，在使用了大量客户端脚本的旅行网站Kayak上进行了测试，取得了令人满意的结果[48，49]。此领域类似的研究已吸引了越来越多科研工作者的注意。

4.2　基于语义的Deep Web研究将进一步发展

如前所述，语义层面的Deep Web研究仍处于起步阶段，研究的重点在Deep Web语义资源以及元数据的标注等方面。语义技术在Deep Web集成技术的其他各阶段也能发挥应有的作用，例如：在查询接口的识别过程中加入语义的支持，使用“白名单”与“黑名单”可以大大提高对特殊查询接口的识别效果。在对查询接口的集成过程中，语义本体对接口集成也有望起到指导性的作用。本体在查询关键词的匹配以及后台查询的构造阶段也能起到一定的指导作用。

4.3　与文本挖掘、自然语言处理技术的融合将更加紧密

Deep Web的数据抽取分为两个步骤：

（1）通过查询接口获取Deep Web后台数据库中尽可能多的信息；

（2）以第一步获取到的信息为处理对象，对其进行处理，如：实体、概念等抽取。

在数据抽取的第一阶段，主要是通过自动填充Deep Web表单，由后台服务器通过填充内容自动构造数据库查询语句来获取相应的数据，以哪种方式实现表单的填充，以及哪些填充数据使获取尽可能多的返回结果概率更高的问题等，都可以部分的通过文本挖掘以及自然语言处理等技术来实现。对于数据抽取的第二个步骤，其本身就可以被看作是一个文本与自然语言处理问题。

4.4　对返回查询结果方面的研究仍需深入

当今Deep Web研究的重点大都集中于查询接口的集成、通过集成接口的数据抽取等专注于查询接口本身的问题，在通过集成接口完成对数据库内容抽取后的相关工作也是有待于深入研究的课题。

例如，在完成查询后，查询结果的排序问题：虽然Deep Web大都不生成静态页面，但大多数Deep Web应用系统在动态生成返回结果时往往同时生成与之关联的资源信息（如在文献数据库中检索到某一文献时可同时会显示与之相关的文献，如引用、被引等信息），因此，虽不能直接使用当前IT界广泛使用的PageRank以及HITS等算法来对查询结果进行集成排序，但可通过使用类似的技术来完成。

5　结束语

获取Deep Web中蕴含的海量信息，以便进一步的分析和使用具有重要的意义。但是由于Deep Web具有的异构性、自制性、分布式等特点，决定了传统的信息处理技术无法满足人们的需求。由此，Deep Web相关内容的研究得到了越来越多科研人员的关注。本文以此为切入点回顾总结了近期Deep Web文本数据集成领域的研究动态与最新进展，并展望了Deep Web可能的发展方向。虽然此领域的研究已经取得了一定的成果，但总体上来讲还是处于发展阶段，仍需对此领域的相关问题进行深入、广泛的研究。

参 考 文 献

[1] Chang K C, He B, Li C K, et al. Structured databases on the web: observations and implications [C] // IGMOD Record, 2004, 33(3): 61-67.

[2] Michaee B K. The deep web: surfacing hidden value [J]. Journal of Electronic Publishing, 2001, 7(1): 1-7.

[3] Priee G, Sherman G. Exploring the invisible web: seven essential strategies [J]. Online 2001, 25(4): 32-34

[4] Pang B, Lee L, Shivakumar V. Thumbs up sentiment classificationusing machine learning techniques [C] // Proceedings of the 2002 Conferenceon Emprical Methods in Natural Language Processing. NJ, USA: Association for Computational Linguistics Morristown, 2002: 79-86.

[5] CNNIC.第十六次中国互联网络发展状况统计报告 [EB/OL]. http://www.cnnic.net.cn/.

[6] 刘玉奎,周立柱,范举.中文深度万维网络数据库的现状研究 [J].计算机学报, 2011, 34(2): 360-370.

[7] 金海,袁平鹏.语义网络数据管理技术及其应用 [M].北京:科学出版社, 2010.

[8] 刘伟,孟小峰,孟卫一.Deep Web数据集成研究综述 [J]. 计算机学报. 2007, 30(9): 1475-1489.

[9] Sriram R, GMHector. Crawling the hidden web [C] // Proceedings ofthe 27th International Conference on Very Large

DataBases.Roma. 2001: 129-138.

[10] Cope J, Craswell N, Hawking D. Automated discovery of searchinterfaces on the web [C] // Proceedings of the 14th Australasian Database Conference (ADC2003). Adelaide, Australia, 2003.

[11] 王辉,刘艳威,左万利.使用分类器发现特定领域深度网入口 [J].软件学报, 2008(2): 246-256.

[12] Barbosa L, Freire J. Searching for hidden web databases [C] // Proceedings of the8th International Workshop on the Web and Databases. 2005.

[13] Barbosa L, Freire J. An adaptive crawler for locating hidden web entrypoints [C] // Proceedings of the 16th International Conference on World WideWeb. New York: ACM, 2007: 441-450.

[14] Lage J P, A S da S, Golgher P B, et al. Automatic generation of agents for collecting hidden web pagesfor data extraction [J]. Data and Knowledge Engineering, 2004, 2: 177-196.

[15] Lim C, McAleer M. Time series forecasts of international travel demandfor Australia [J]. Tourism Management, 2002, 23: 389-396.

[16] Khare R, An Y, Song I Y. Understanding deep web search interfaces : asurvey [C] // SIGMOD Record. 2010, 39(1): 33-40.

[17] He H, Meng W Y, Liu Y, et al. Towards deeper understanding of the searchinterfaces of deep web [C] //Proceeding of the WWW'07, 2007: 13-25.

[18] 林玲,周立柱.基于简单查询接口的Web数据库模式识别 [J]. 清华大学学报（自然科学版）, 2010, 4: 551-555.

[19] He H, Meng W Y, Yu C, et al. WISE-Integrator: an automatic integrator of web search interfaces for e-commerce [C] // Proceedings of the 29th International Conference on Very Large Data Bases(VLDB'03). Berlin, German, 2003: 357-368.

[20] He H, Meng W Y, Yu X, et al. Automatic integration of web search interfaces with wise-integrator [J]. VLDB Journal, 2004, 13(3): 256-273.

[21] He H, Meng W Y, Yu X, et al. Wu. WiSE-Integrator: a system for extracting and integrating complex web search interfaces of the deep web [C] // Proceedings of the 31st International Conference on Very Large DataBases(VLDB'05). Trondheim, Norway, 2005: l314-1317.

[22] Qprober Research Group[EB/OL]. [2005]. http://qprober.cs.columnbia.edu.

[23] Ipwirotis P G, Gravano L. When one sample is not enough: improving text database selectioin using shrinkage [C] // Proceeings of the 2004 ACMSIGMOD International Conference on Management of Data. 2004: 767-778.

[24] Ipeirotis P G, Gravano L. Classification-aware hidden-web text database selection [J]. ACM Transactions on Information Systems, 2008, 26(2): 1-66.

[25] Balakrishnan R, Kambhampati S. Factal : Integrating deep web based on trues and relevance [C] // Proceedings of WWW. 2011.

[26] Balakrishnan R, Kambhampati S. Sourcerank : relevance and trust assessment for deepweb sources based on inter-source agreement [C] // Proceedings of International World Wide Web Conferences. 2011.

[27] Doorenbos R B, Etzioni O, Weld D. A scalable comparisonshopping agent for the world-wide web [C] // Proceedings of the 1st International Conference on Autonomous Agents. 1997: 39-48.

[28] Tay F E H, Cao L. Application of support vector machines in financialtime series forecasting [J]. Omega, 2001, 29: 309-317.

[29] MetaQuerier Research Group [EB/OL]. [2005]. http://metaquerier.cs.uiuc.edu.

[30] Liu W, Meng X F, Meng W Y. Vision-based web data records extraction [C] // Proceedings of the 9th International Workshop in Web and Databases. NewYork: ACM, 2006: 20-25.

[31] 李先,刘伟，孟小峰. EasyQueries：一种基于关键词的Web查询接口集成 [J]. 计算机研究与发展. 2006增刊: 54-60.

[32] Liu L N, Kou Y, Sun G S, et al. Duplicate Identification Model for Deep Web [J]. Journal of Southeast University, 2008.24(3):315-317.

[33] 马安香,张斌,高克宁,等.基于结果模式的Deep Web数据抽取 [J].计算机研究与发展, 2009, 46(2): 280-288.

[34] 申德荣,马也,聂铁铮,等.一种应用于Deep Web数据集成系统中的查询松弛策略 [J]. 计算机研究与发展, 2010, 47(1): 88-95.

[35] 寇月, 李冬,申德荣,等.D-EEM: 一种基于DOM树的Deep Web实体抽取机制 [J]. 计算机研究与发展, 2010,47(5): 58-865.

[36] Zhao P P, Cui Z M, Gao L, et al. Vision-based deep web query interfaces automatic extraction [J]. Journal of Computational Information System, 2007, 3(4): 1441-1448.

[37] Zhao P P, Lin C, Fang W. Deep searcher: organiziing structurerd deep web classified search engine [J]. Journal of Computationl Information System, 2008, 4(1): 111-117.

[38] Zhao P P, Huang L, Fang W. Organizing structured deep web by clustering query interfaces link graph [C] // Proceedings of the 4th International Conference on Advanced DataMing and Applications. 2008.

[39] Cui Z M, Zhao P P, Zhong H, et al. Automatic hierarchical cluster of structed deep web by qury probing [J]. Journal of Computational Information System, 2007, 3(4): 1433-1440.

[40] Xu H X, Hao X L, Wang S Y, et al. A method of deep web classification [C] // Proceedings of International Conference on Machine Learning and Computing. 2007: 4009-4014.

[41] Xu H X, Zhang C H, Hao X L, et al. A machine learning approach classification of deep web sources [C] // Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery. 2007: 561-565.

[42] 徐和祥,王述云,胡运发.基于知识的Deep Web集成环境变化处理的研究 [J]. 软件学报, 2008, 19(2): 257-266..

[43] Geller J, Chun S, An Y. Towards the semantic deep web [J]. IEEE Computer, 2008, 41(9): 95-97.

[44] A Wright. Searching the deep web [J]. Communications of the ACM, 2008, 51(10): 14-15.

[45] 黄晓东.Invisible Web研究综述 [J].情报科学, 2004, 22(9): 1144-1147.

[46] 郭文宏.基于领域知识的Deep Web信息处理技术研究 [D].同济大学博士学位论文, 2009.

[47] Alba A, Bhagwan V, Grandison T. Accessing the deep web : when good ideasgo bad [C] // Proceedings of Object-Oriented Programming, Systems, Languages & Applications Conferences. 2008: 815-818.

[48] Selleres A. The OXPath to success in the deep web [C] // Proceedings of International World Wide Web Conferences. 2011: 409-413.

[49] Sellers A, Furche T, Gottlob G, et al. Taking the oxpathdown the deep web [C] // Proceedings of International Conference on Extending Database Technology. 2011: 542-545.