ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.177−178
http://www.jos.org.cn/1000-9825/19/177.htm

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

Deep Web数据集成专刊前言*

孟小峰[1+], 于 戈[2]

[1](中国人民大学 信息学院,北京 100872)

[2](东北大学 信息科学与工程学院,辽宁 沈阳 110004)

+ Corresponding author: E-mail: xfmeng@ruc.edu.cn

孟小峰,于戈.Deep Web 数据集成专刊前言.软件学报,2008,19(2):177−178. http://www.jos.org.cn/1000-9825/19/177.htm

随着 World Wide Web 的飞速发展,出现了越来越多的可以在线访问的数据库,我们把这些数据库称作 Web 数据库.据统计,目前 Web 数据库的数量已经超过了 45 万个,在此基础上构成了 Deep Web.Deep Web 蕴含了大量有用的信息,其价值远远超过了仅由网页构成的 Surface Web.但由于对 Web 数据库的访问只能通过其提供的查询接口,因此很难被一般的搜索引擎获取到.由于 Deep Web 的大规模性、动态性以及异质性等特点,通过手工方式远远不能在效果和效率上满足用户对信息获取的需要.为了帮助人们快速、准确地利用 Deep Web 中的海量信息,研究者们已经在 Deep Web 数据集成方面展开了研究.这逐渐成为数据库领域的一个研究热点.研究者力图提出一种通用的集成方法,可以实现对现实世界各个领域的 Deep Web 数据的集成,并在查询接口集成和数据抽取等方面取得实质性的进展.近几年来,已有大量的研究成果在 SIGMOD、VLDB 等高级别的国际会议和期刊上发表.国内对 Deep Web 数据集成的研究也取得了一定的成果,但与国际水平相比还有一定的距离,主要表现在研究问题和解决方法上尚缺乏突破性的成果.

为了推动 Deep Web 数据集成在国内的进展,本专刊关注于当前国内在该研究领域最新的基础性、前瞻性、战略性的重大理论问题和关键技术的问题,目的在于为大家展示当前该领域的研究状况和最新的研究成果,为该领域的研究者们提供一个相互学习交流、借鉴指导的机会.

本专刊得到了国内同行的广泛响应与支持,收到稿件 60 余篇.本专刊严格按照《软件学报》审稿流程和评审要求对稿件进行了认真评审.审稿工作由本领域从事 Deep Web 数据集成的海内外专家组成的评审委员会来组织,每篇稿件均经过两位以上评审委员的认真评审.最后,经过《软件学报》编委会终审,遴选出具有代表性的研究工作 9 篇.这些论文涉及了 Deep Web 数据集成的若干关键问题,研究的内容注重理论创新与实际应用相结合,立足于国际上最新的研究和应用状况,真实反映了当前我国的 Deep Web 数据集成技术在重要科学领域的应用研究状况.这里,我们要再次感谢大家的关注和向本专刊投稿的各位作者.

论文“一种基于图模型的 Web 数据库采样方法”把 Web 数据库模型化为一种图结构,在这个图结构上实现对 Web 数据库的采样,可以增量的方式获取近似随机的样本.该方法的一个重要特点是不受查询接口中属性表现形式的局限,因此是一种通用的 Web 数据库采样方法.

论文“一种基于语义及统计分析的 Deep Web 实体识别机制”提出了一种基于语义及统计分析的实体识别机制(SS-EIM).SS-EIM 主要由文本匹配模型、语义分析模型和分组统计模型组成,采用文本粗略匹配、表象关联关系获取以及分组统计分析的三段式逐步求精策略,基于文本特征、语义信息及约束规则来不断精化识别结果.该方法可有效解决 Deep Web 数据集成中数据纠错、消重及整合等问题.

论文“针对模板生成网页的一种数据自动抽取方法”提出了一种新颖的模板检测方法,并利用检测出的模板自动地从实例网页中抽取数据.与其他已有方法相比,该方法能够适用于“列表页面”和“详细页面”两种类型

* Received 2007-12-28

的网页.

论文“基于属性相关度的 Web 数据库大小估算方法”提出了一种基于词频统计的解决方法,通过分析 Web 数据库查询接口中属性间的相关度来获取某个属性上一组随机样本,并以对该属性分别提交由前 k 位高频词形成的试探查询的方式,估算出 Web 数据库中记录的总数.

论文“基于本体的 Deep Web 数据标注”借鉴语义 Web 领域中深度标注的思想,将领域本体作为 Web 数据库遵循的全局模式,引入到查询结果语义标注过程中,并将本体与接口模式、结果模式相结合,辅以查询条件重置的策略,对查询结果进行统计及结构特征分析,确定查询结果数据的语义标记.

论文“使用分类器自动发现特定领域的深度网入口”提出了一种三分类器的框架,用于自动识别特定领域的深度网入口.查询接口得到以后,可以将它们进行集成,然后将一个统一的接口提交给用户以便于查询信息.

论文“基于知识的 Deep Web 集成环境变化处理的研究”研究了 Deep Web 集成环境中构件的依赖关系(执行偏序依赖和知识依赖),并在此基础上提出了一种基于知识的环境变化的处理方法,包括 Deep Web 集成环境变化处理模型、适应 Deep Web 环境变化的动态体系结构和处理算法,可以对大规模 Deep Web 集成的进一步探索和走向应用提供参考.

论文“基于网页上下文的 Deep Web 数据库分类”给出了采用分层模糊集合对给定学习实例所发现的领域和语言知识进行表示和基于这些知识对标记词归一化的算法.基于上述预处理,给出了计算 Deep Web 数据库的 K-NN 分类算法,其中对数据库之间的语义距离计算综合了数据库表之间和含有数据库表的网页的内容文本之间的语义距离.

论文“基于页面 Block 的 Web 档案采集和存储”提出了基于页面 Block 的采集和存储方式,并详细表述了该方法如何完成基于布局页面分区、Block 主题的抽取、版本和差异的比较以及增量存储的方式.本文还实现了一个 Web 归档原型系统,并对所提出的算法进行了详细的测试.

这些论文集中反映了国内研究者在 Deep Web 数据的分析、集成和检索等方面的最新研究成果,对于促进针对下一代信息系统的创新性研究,以及鼓励数据库技术与其他相关领域的交叉研究具有重要的意义.

孟小峰(1964－),男,博士,中国人民大学信息学院教授,博士生导师.现为中国计算机学会理事,普及工委主任,中国计算机学会数据库专委会委员、秘书长,办公自动化专委会副主任委员,《计算机研究与发展》等期刊编委,MDM、WAIM 等国际学术会议指导委员会委员,IEEE CS、ACM SIGMOD 会员.曾先后在香港中文大学、香港城市大学、新加坡国立大学、法国 Prism 实验室访问研究.主持或参加过 20 多项国家科技攻关项目、国家自然科学基金、国家高技术研究发展计划(863)、信息产业部发展基金项目以及国际合作交流项目等.先后获得国家科技进步二等奖 1 项、电子部科技进步特等奖 1 项、北京市科技进步二等奖 2 项,以及第 7 届“中创软件人才奖”、“新世纪优秀人才”、“第三届北京市高等学校名师奖”等奖项.研制开发的主要软件产品有国产数据库系统 COBASE、嵌入式移动数据库系统“小金灵”、中文自然语言查询系统 NChiql、并行数据库系统 PBASE/1 等.近 5 年先后在国内外学术期刊以及 VLDB、SIGMOD、ICDE 等重要国际会议发表论文 100 余篇.多次应邀担任国际会议程序主席或委员,如 SIGMOD、ICDE、ER、DASFAA、MDM 等.近期主要研究领域为 Web 数据集成、XML 数据库、移动数据管理.

于戈(1962－),男,博士,东北大学教授,博士生导师,中国计算机学会理事,数据库专业委员会副主任委员,电子政务与办公自动化专业委员会副主任委员,YOCSEF 学术委员会荣誉委员,美国 ACM 学会和 IEEE 学会会员.1982 年和 1986 年在东北大学分别获得计算机应用专业学士学位和硕士学位,1996 年于日本九州大学获得计算机工学博士学位.1986 年起在东北大学计算机科学与工程系任教.曾在日本九州大学、香港中文大学和香港科技大学做访问学者.研究方向涉及数据库系统、嵌入式软件、信息安全等相关领域.近年来,在国内外重要学术期刊和会议上发表论文 100 余篇,获得省部级自然科学奖 3 项、科技进步奖 2 项.