WAMDM 网络与移动数据管理实验室

Innovative Data Systems Research

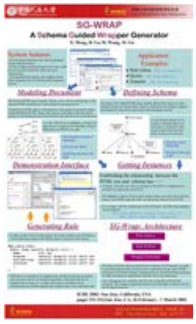
Lab of Web and Mobile Data Management

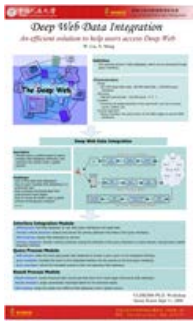
[English]

网络数据管理介绍

随着互联网的迅速发展，使得信息管理与访问的方式发生了巨大的变化。Web中蕴含的信息在以惊人地速度增长。目前Web中的信息量已经超过了7500千兆字节（或40亿个网页）。Web中的信息包括了人类现实世界中所有领域，这就为使用者从中受益打开了方便之门。因此，Web受到了越来越多的关注。

整个Web可以分为Surface Web和Deep Web两大部分。Surface Web是指静态的、链向其它网页的网页。Deep Web是指作为搜索结果而动态生成的网页。传统的搜索引擎爬取Surface Web中的网页，并为这些网页建立索引。因此传统的搜索引擎难以发现或获取Deep Web中的内容。平均说来，每个月对Deep Web的访问量要高于Surface Web—半以上。根据UIUC大学在2004年做的统计调查，目前Web中至少有30万个Web数据库和45万个查询接口可以访问，而且这两个数字仍然在迅速增加。Deep Web不仅规模巨大，而且覆盖了现实世界的所有主题。Deep Web正在成为互联网上一个新的信息源，并成为了一个热点研究问题。





[Top]

研究动机

尽管Web中有如此丰富的信息，但由于Web数据规模巨大，而且具有异质性和无结构的特点，要想从中获取想要的信息并非易事。因此，为了解决这个挑战，一个可行的方案就是从Web中将数据抽取出来并存储到结构化的数据库中作进一步的处理。为了集成Web数据的过程自动化并进行有效地利用，越来越多的研究者致力于该研究领域。

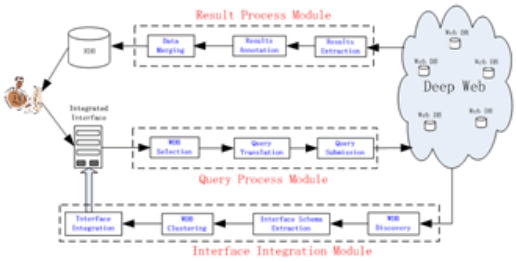
同样在Deep Web的环境下，由于Web数据库数量众多、规模巨大，用户准确地从Web中找到他们要查询的Web数据库并进行查询是一件十分困难的事情。为了能够有效地访问这些Web数据库，研究者们致力于寻求一个可行的解决方案对Web数据库进行集成，使得人们可以同一个统一的访问接口对Web数据库进行访问并自动地从中获取信息。

[Top]

现有研究工作

- **Deep Web数据集成**

Web中出现了越来越多的可访问的数据库。为了向人们提供一个对这些Web数据库统一的访问途经并从中自动地获取信息，我们针对Web数据库集成提出了一个综合的解决方案。下图是该方案的体系结构。该方案包括三个主要的模块：集成接口生成模块（integrated interface generation module）、查询处理模块（query processing module）、查询结果处理模块（results processing module）。



集成接口生成模块:在被集成的各个Web数据库的查询接口上产生一个集成的查询接口。该模块包括4个组件，其功能描述如下:

- *Web*数据库的发现（*Web database discovery*）：搜索具有Web数据库的网站，并从其中的网页中识别出可对该Web数据库查询的接口。
- 查询接口模式抽取（*Query interface schema extraction*）：从查询接口抽取出所包含的属性，以及这些属性相关的元信息。
- *Web*数据库分类（*Web database clustering by topic*）:把所有发现的Web数据库分为若干组，使得每一组属于现实世界中一个相同的主题领域。
- 接口集成（*Interface integration*）:给定若干同一主题下Web数据库的查询接口，把不同查询接口中表示同一语义的属性合并为一个全局的属性，并最终形成一个集成的查询接口。

查询处理模块: 处理用户在集成查询接口上填写的查询，并提交给每一个Web数据库。该模块有三个组件，其功能描述如下:

- *Web*数据库的选择（*Web database selection*）: 为给定的用户查询选择合适的Web数据库，使得用户可以在最小的代价下得到最合适的查询结果。
- 查询转换（*Query translation*）: 将集成接口上的查询尽量等价地转化为在各个Web数据库查询接口上的查询。
- 查询提交（*Query submission*）: 分析各个Web数据库的查询提交方式，并自动地完成查询的提交。

查询结果处理模块: 抽取Web数据库返回的查询结果，并把抽取到的结果合并到一个统一的模式下。该模块共有三个组件，其功能描述如下:

- 结果抽取（*Result extraction*）: 从Web数据库返回的结果页面中识别并抽取出纯粹的查询结果。
- 结果注释（*Result Annotation*）: 为抽取到的结果添加正确的语义。
- 结果合并（*Result merging*）: 把从各个不同Web数据库抽取的结果合并到一个统一的模式下。

- **Web**数据抽取

在Web中，信息主要以网页的形式发布。因此Web数据抽取是指从网页中把感兴趣的数据识别并抽取出来。众所周知，网页是半结构化的文档，Web数据抽取要达到较高的准确性具有很大的挑战性。所有的Web数据抽取工具可以分为三类：手工的、半自动的和自动的。下面对我们实现的抽取工具作简要的介绍。

ViDRE: 基于视觉的**Web**数据记录抽取
该工具主要针对Web数据库或搜索引擎返回结果页面中记录的抽取。已有的方法主要是通过分析网页的DOM树结构和Html标签来实现的。尽管这些方法可以达到较好的抽取效果，但过多的依赖与Html语言的规范与标准，因此当网页用其他标记语言编写时就会使得方法完全失效。为了避免这个缺陷，我们提出了一种新奇的独立于网页编写语言的抽取技术。我们通过分析结果页面中一般的视觉特征作为该技术的实现基础，包括位置特征、布局特征、外观特征以及内容特征。基于这些视觉特征，我们实现了抽取工具ViDR ，该工具可以只利用结果页面中的视觉信息完成对页面中数据记录的抽取。

TSReC: 一种自动抽取网络新闻内容的混合式方法
网络新闻的搜索引发了一系列传统信息检索技术无法解决的新问题。如何区分网页中的新闻与其他内容就是其中之一。对于目前的搜索引擎，尤其是新闻搜索服务，网络新闻内容的抽取对于提高新闻的索引和搜索效果有着及其重要的作用。在这篇论文中我们研究了这一问题，并且提出了一种混合式自动抽取方法，能够同时利用序列匹配与树匹配技术的优点。我们还提出了一种适于同时利用上述两种技术对网络新闻内容进行自动抽取的标签序列描述方法，即TSReC，以及相应的算法。

RecipeCrawler: 从互联中搜集菜谱数据
互联网是一个巨大的数据仓库。人们把互联网作为数据源是一个自然而然的选择，并且已经做出了许多的努力与尝试。我们关注于建立一个健壮的系统来从Web中搜集结构化的菜谱数据。我们相信，这是实用的、持续的、可靠的Web数据抽取系统的关键一步。我们采用增量方式的原因包括两个方面：一、由于Web的高度动态性，一次爬取所有的菜谱页面是不现实的；几乎不可能从初始的菜谱页面中推导出一个一般的适用于将来的Wrapper。我们构建了系统RecipeCrawler，能够以增量的方式收集菜谱数据。

SG-WRAP: 模式导航方式的**Wrapper**生成器
随着互联网的发展，wrapper技术逐渐兴起并致力于将无结构/半结构化数据向半结构化/结构化数据转化，从而使得利用现有成熟的数据库等技术对这些数据进行分析和查询成为了可能。SG-Wrap采用了一种新型的以模式（shema）为导向的方法来指导wrapper的生成。基于这种以模式为导向的方法，用户可以以XML中提供的数据类型描述（DTD）的形式来定义想要从HTML页面上抽取出来的数据的具体格式。同时，用户还需要以示例的方式提供关于HTML页面中数据和DTD中元素之间的匹配关系。从而系统根据学习到的匹配规则生成适应于该HTML页面的wrapper，以便从该HTML页面中抽取数据，并根据指定的DTD格式将抽取出来的数据以XML文档的形式展现出来。

SG-WRAM: 模式导航方式下的Wrapper维护

SG-WRAM是一种新颖的模式（schema）导向的wrapper维护机制。它是基于我们之前的工作——**SGwrap**（模式导向的Wrapper生成机制）开发而成的。观察发现，尽管网页是经常变化的，但是网页中的一些重要特征是保持不变的，比如句法模式，释义信息以及数据项的链接信息等。利用这些特征，我们就有可能在不断变化的网页中重新定位我们所要找的数据项，而且使得所要抽取的数据的模式保持不变。具体说来，wrapper的维护主要基于以下4个步骤：首先，我们从用户自定义的模式信息、数据抽取的规则和抽取结果中提取特征。其次，根据这些特征从变化了的网页中识别数据项。接下来我们根据模式信息对这些数据项进行分组。每一组都是关于某个特定模式的实例，被称作语义块（semantic block）。最后，我们选择一些典型的实例为该变化之后的新网页生成新的抽取规则。

PIM (personal information management)

随着个人信息量的迅速增加，个人信息管理问题日益突出，在现实生活中，人们往往不能及时发现所需信息。有些重要的信息，由于时间或空间问题不能及时被保存，以至被完全忘记，从而错过信息所可能带给我们的机遇。

PIM对信息获取、组织、存储、访问等技术进行研究，并为人们提供实际的应用工具。理想的PIM可以使我们在正确的时间、以正确的方式获得所需要的信息，以满足需要。

PIM设一个跨学科的研究领域，涉及许多关键技术：数据空间、场合感应、信息搜索、人机接口等等，对于数据管理研究领域的研究者来说，PIM研究即使一个重要的机遇，也是一个巨大的挑战。

[Top]

未来研究方向

- 数据空间

传统数据库的各种数据存储方式，关系也好，XML也好，无不强调一个格式，总是先有一个格式，然后使数据服从于这个格式，如此才能存储数据，进而提供查询等服务。但是任何形式的数据，其核心都是数据本身，形式只是一种载体，如果将数据限制于某种形式之中，多少显得有些许被动，所以就是一种“被动”的方式，也就是说如果你有一份不同格式的数据要想存储于数据库中，必须将其转化为数据库中数据的存储格式。因此，对于这种格式性很强的存储，可以称之为“先有格式，后有数据”。

数据空间不同，从它的名字可以看出，它与数据库不同并且强调的是一个Space，Space是什么？是空间，广阔的宇宙是一个Space，是个ObjectSpace，不管这些Object在其中如何排列，如何组织，只要是属于这个Space的就是符合要求的。同样，数据空间是一个满是数据的空间，数据在其中如何组织都可以，表也罢，XML也罢，文本也罢，只要你是数据的一种载体，你就可以存在于这个Space中，对数据的组织排放不做任何要求，“一个数据空间应该包含与某个组织或个体相关的一切信息，无论这些信息是以何种形式存储、存放于何处”。这样一来，无论你有一份怎样格式的数据，XML文档也好，文本文档也好，都可以存储于数据空间中，并且通过数据空间来对其进行掌控，这可以称之为“淡化形式，凸现数据”。

[Top]

获资助课题

- 2002-2005 国家自然科学基金项目“Web数据抽取和集成技术研究”，项目负责人，编号：60273018
- 2002-2004 国家863项目“基于Web Service的Web数据集成技术”，项目负责人， 编号：2002AA11304

[Top]

论文著作

- 李先，刘伟，孟小峰，EasyQueries：一种基于关键词的Web集成查询接口. 计算机研究与发展, 卷43(增刊)：54 - 60，2006. (第23届中国数据库学术会议，广州.)
- 凌妍妍，刘伟，王仲远，艾静，孟小峰：Deep Web数据集成中的实体识别方法. 计算机研究与发展, 卷43(增刊)：46 - 53，2006 . (第23届中国数据库学术会议，广州.)
- W. Liu, X. Li, X. Meng, et al: A Deep Web Integration System for Job Search. Wuhan University Journal of Natural Sciences, 11(5):1197-1201, Nov., 2006. (The Third Web Information System and

Application(WISA2006), Nanjing, Nov 3-5, 2006.)

- W. Liu, C. Lin and X. Meng: Web Database Query Interface Annotation Based on User Collaboration. Wuhan University Journal of Natural Sciences, 11(5):1403-1406, Nov., 2006. (The Third Web Information System and Application(WISA2006), Nanjing, Nov 3-5, 2006.)

- Y. Li, X. Meng, Q. Li, L. Wang: Hybrid Method for Automated News Content Extraction from the Web. In proceeding of 7th International Conference on Web Information Systems Engineering(WISE2006),pages 327-338, Wuhan, China, October 2006

- W, Liu, X. Meng: Web Database Integration. In Proceedings of the Ph.D Workshop in conjunction with VLDB 06 (VLDB-PhD2006), Seoul, Korea, September 11, 2006.

- W. liu, X. Meng, W. Meng: Vision-based Web Data Records Extraction. In Proceedings of the 9th SIGMOD International Workshop on Web and Databases (SIGMOD-WebDB2006), Chicago, Illinois, June 30, 2006. (12/48=25%) [PDF]

- Y. Ling, X. Meng, and W. Meng, Automated Extraction of Hit Numbers From Search Result Pages. In Proceedings of the Seventh International Conference on Web-Age Information Management(WAIM2006), pages 73-84, Hong Kong, China,17-19 June, 2006. Lecture Notes in Computer Science 4016, Springer 2006.

- Y. Li, X. Meng, L. Wang, Q. Li, RecipeCrawler: Collecting Recipe Data from WWW Incrementally. In Proceedings of the Seventh International Conference on Web-Age Information Management(WAIM2006), pages 263-274, Hong Kong, China, 17-19 June, 2006. Lecture Notes in Computer Science 4016, Springer 2006.

- D. Hu and X. Meng: Automatically extracting data from data-rich web pages. In proceedings of the 10th International Conference on Database Systems for Advanced Applications (DASFAA 2005), pages828-839, Beijing, China, April 17-20, 2005. Lecture Notes in Computer Science 3453, Springer. (Full paper)

- C. Lin, Q. Zhang, X. Meng, W. Liu: Postal Address Detection from Web Documents. In Proceedings of the ICDE International Workshop on Challenges in Web Information Retrieval and Integration (ICDE-WIRI2005), pages 40-45 , Tokyo, Japan, April 8-9 2005.

- X. Meng, D. Hu, C. Li: Schema-Guided Wrapper Maintenance for Web-Data Extraction. In Proceedings of ACM Fifth International Workshop on Web Information and Data Management (WIDM 2003), pages 1-8, New Orleans, Lousiana, USA, November 7-8, 2003

- X. Meng, H. Wang,D. Hu, M. Gu: SG-WRAM Schema Guided Wrapper Maintenance: A Demonstration. In Proceedings of the 19th International Conference on Data Engineering (ICDE2003), pages 750-752, Bangalore, India, March 5-8, 2003

- X. Meng, H. Lu , et al.: Data Extraction from the Web based on Pre-defined Schema. JCST, Vol.17(4):377-388, 2002,7

- X. Meng, H. Lu, et al.: SG-WRAP: A Schema-Guided Wrapper Generator. In Proceedings of the 18th International Conference on Data Engineering (ICDE2002), pages 331-332,San Jose, CA., 26 February - 1 March 2002

[Top]

Maintained by Zhongyuan Wang(zhywangchina@163.com)