

Jobtong: 面向领域的 Deep Web 数据集成系统

王仲远 (Web 组)

一、介绍

随着网络与通信技术的迅速发展, Web上的信息呈现爆炸性增长, 互联网已经成为一个巨大的海量信息空间。整个Web看似杂乱无章, 但如果按其所蕴涵信息的“深度”可以划分为Surface Web和Deep Web两大部分。Surface Web是指通过超链接可以被传统搜索引擎索引到的页面的集合。而Deep Web目前还没有比较明确的定义, 通常是指那些可以在线访问的Web数据库[1]集合。

2000年7月BrightPlanet.com针对Deep Web的数量做了一次比较全面的统计, 其发表的白皮书[2]称: 整个Web上大约有43,000-96,000个Web数据库, 以及7,500TB的数据(约为Surface Web的500倍)。而经过这些年的发展, 根据UIUC在今年5月所发表的一篇Deep Web综述[3]估计, 截至2004年, Deep Web的网站数量已经达到307,000个, 其背后的数据库数量已经达到366,000-535,000个。面对如此多的“隐藏数据”, 虽然随着搜索引擎索引能力的提高, 以及一些Deep Web网站为优化搜索而提供的“直接浏览”功能, 主流搜索引擎例如Google、Yahoo已经能够覆盖到其中32%的数据, 但是大部分数据仍然不能够通过搜索访问到。因此, 在Deep Web上进行大规模数据集成显得越来越急迫。

本文就将介绍WAMDM实验室Web组的一个Deep Web数据集成研究项目: Jobtong项目。这个项目最初是研究在工作信息领域上的数据集成, 我们期望通过这种研究, 形成一套面向领域的Deep Web数据集成的方法。目前, 此项目已经取得初步成功, 并已不再局限于工作信息领域(但我们习惯上仍称此项目为Jobtong项目)。我们已经成功在工作领域、政府信息领域快速地构造了这样的应用, 拥有了两个演示网站: Jobtong(工作通, <http://www.jobtong.cn>) 以及 Govtong(政务通, <http://www.govtong.cn>)。

二、相关工作

面向领域的数据集成主要针对的是 Deep Web 上的数据源。目前, 在 Deep Web 上进行数据集成主要有两种方法: 接口集成与查询结果集成。

所谓接口集成, 是指将各个 Deep Web 网站的搜索框(即查询接口)集成起来, 这样用户可以通过这样一个集成的接口, 将关键字发送到各个网站的搜索框上, 然后再将搜索结果取回。使用这种方式进行集成, 本地服务器上是不保存数据的。在接口集成的研究中[4,5]试图建立一个复杂的查询接口来集成多个数据源, [6,7,8]则尝试建立能够处理自然语言查询, 并将其转换为结构化数据上的查询。

在查询结果集成中, 如果按照用户参与程度的大小来区分的话, 已有的信息抽取工作可以区分为: 手工, 半自动和自动的数据抽取方法。其中, 手工的方法[9]是最早提出来的, 但是它需要用户有较为丰富的相关知识与编程经验。半自动的方法[10,11]正式为了解决这个问题而提出来的, 通常它会提供一种图形化的界面来辅助生成爬取程序。全自动的方法[12,13,14,15]希望能够将用户彻底解放出来, 但是这些方法通常有过多的假设, 离实际应用还有一段距离, 同时, 当面对较为复杂的页面时, 往往不能得到理想的结果。

Jobtong 项目的面向领域的数据集成方法属于查询结果集成的一种。虽然在这方面的研究上, 已经有很多自动的方法来生成 Wrapper, 构建集成系统, 但是我们认为已有的自动方法, 实际表明, 它们所得到的结果距离真正的 Web 应用还有很大的距离, 主要集中在以下几点:

1. 自动的方法要么存在许多假设, 要么无法达到很好的效果。
2. 自动的方法要在初始时设置许多参数, 这些参数带有较多的主观因素。同时一旦变更领域, 这些参数往往就会失效, 又不得不重新耗费大量经历设置参数。
3. 自动的方法很难做到在记录(record)层次的区分, 而在属性(item)层次的区分就更不理想。

因此, 我们考虑了一种在保证高精确性的前提下, 尽量减少人工参与程度的介于半自动与全自动之间的集成方法。此方法的核心在于将 Deep Web 上的查询进行抽象, 使用统一的集成程序与分散式的配置文件构成配

置文件系统。同时，作为对配置文件系统的有力补充，构建一个具有更大规模集成能力的但是精确度略低的分布式爬取集成系统。两个系统合并使用，构成我们的整套集成解决方案。

三、Deep Web 数据集成项目框架介绍

正如上文所提到的，WAMDM 实验室的 Jobtong 项目旨在通过对工作信息集成的研究，总结出一套在 Deep Web 上进行分类信息集成的解决方案，并建立从底层 Deep Web 数据集成系统、中间信息的全文检索技术到上层的面向用户的产品及服务体系。以此提供给用户一个统一的快速访问 Deep Web 上的数据的方式。

基于以上思想，我们构建了一个四层的 Deep Web 数据集成系统的框架图，如图 1 所示。

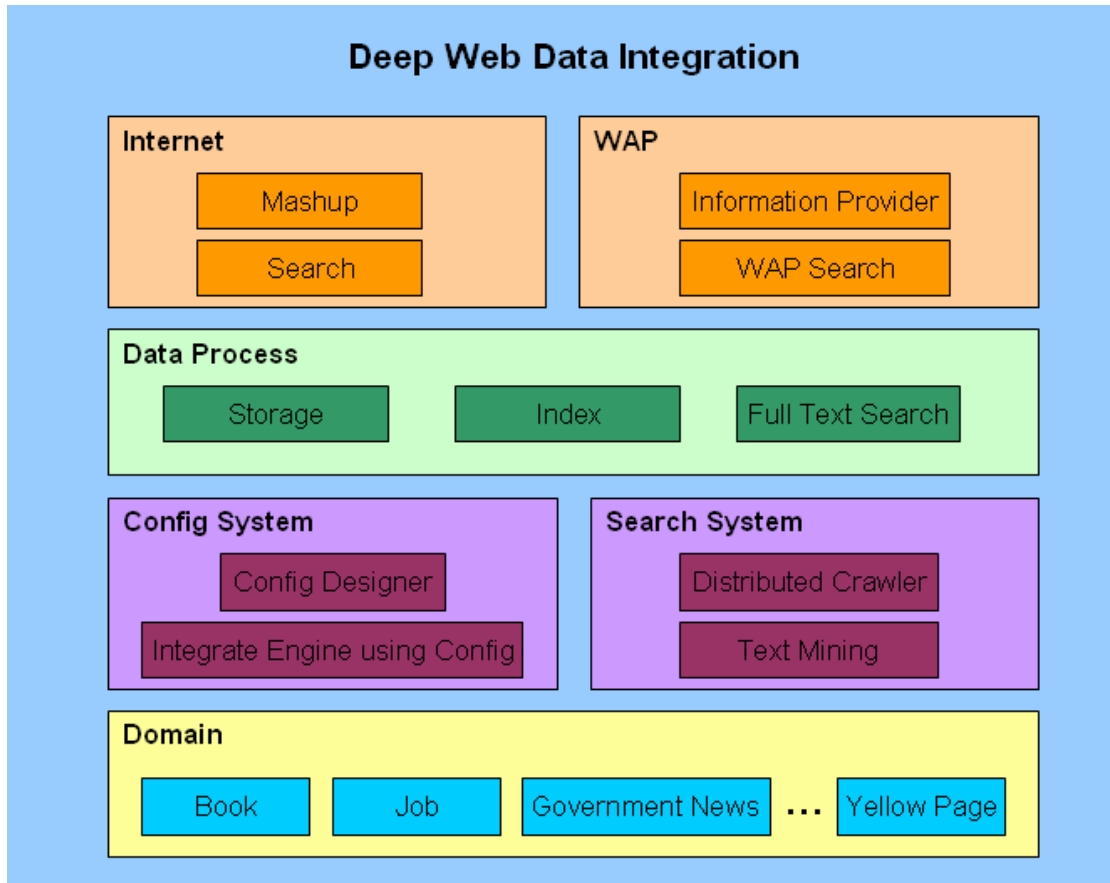


图 1 Deep Web 数据集成系统框架图

底层面向的领域是我们所要处理的 Deep Web 数据源，我们将其按照领域进行分类，可以分为图书信息领域、工作信息领域、政府新闻领域等等。由于我们的目标是将所有数据集成到本地再进行处理，但由于不同领域的数据结构大不相同，因此，我们在集成的时候，考虑的是面向领域的数据集成。

在领域之上，使我们的数据集成系统，主要分为两块，一块是基于配置文件的数据集成系统(Config System)，用于比较精确地对 Deep Web 上的数据进行抽取与集成；另一块是普通搜索引擎(Search System)，用于搭建大规模的、快速的、全自动的但准确率稍低的集成系统。基于配置文件的数据集成系统将在后面进行详细介绍。由于基于配置文件的数据集成系统的特殊性，它更关注的是获取精确的数据，以及 Deep Web 上那些普通搜索引擎无法爬取到的数据。但是规模也不可能过大。作为对基于配置文件的数据集成系统的有力补充，我们采用普通搜索引擎作为另一个数据集成途径。它可以对 Deep Web 网站中的 Surface Web 进行自动集成。同时也极大地增加我们的数据量。

在数据集成到本地之后，我们就需要考虑数据处理模块(Data Process)。该数据处理模块包括：数据存储单元，用于存储我们集成抽取出来的各个网站的数据；索引单元，用于对集成来的数据属性进行索引，以方便未来查询时的各项检索；以及全文索引单元，用于对集成过来的全部数据进行全文索引，以处理大规模数据集成后的关键字检索问题。

在我们框架的最顶层，就是基于我们集成系统之上的应用，包括因特网和 WAP 移动网上的应用。在因特

网上，我们可以利用集成出来的数据提供给用户进行垂直搜索，也可以整合多个数据源产生出数据混合(Mashup)应用。在 WAP 移动网上，我们可以将集成到的数据以简单的形式提供给用户，以便将因特网上的数据共享到 WAP 上，也可以将数据提供给手机用户进行移动检索。

这四层，构成了整个 Deep Web 上数据集成应用的一套解决方案。其中，整个项目最核心的部分是领域之上的基于配置文件的数据集成系统。这个系统以模拟用户在某个网站上实际点击行为的方式，通过专门针对这个网站的配置文件，实现对这个网站所有信息的集成以及对每条记录各个属性段的精确抽取，并实现数据的动态更新。以下我们将进行详细介绍。

四、基于配置文件的集成系统

Jobtong 基于配置文件的集成系统的核心思想是：用一个统一的集成程序，利用针对每一个网站的配置文件，对 Deep Web 上进行数据集成。它的基本工作机制类似于基于 Wrapper 的数据集成系统。但是有两点不同：第一，Jobtong 基于配置文件的集成系统将某一个具体网站的数据集成过程抽象出来，这样集成程序是统一的，而针对网站只需要写配置文件，并且这个配置文件与数据库属性挂钩的；第二，Jobtong 的这种工作机制比基于 Wrapper 的集成系统要节省内存，同时能够使用多线程进行爬取。

以下是 Jobtong 基于配置文件集成系统的整体框架图：

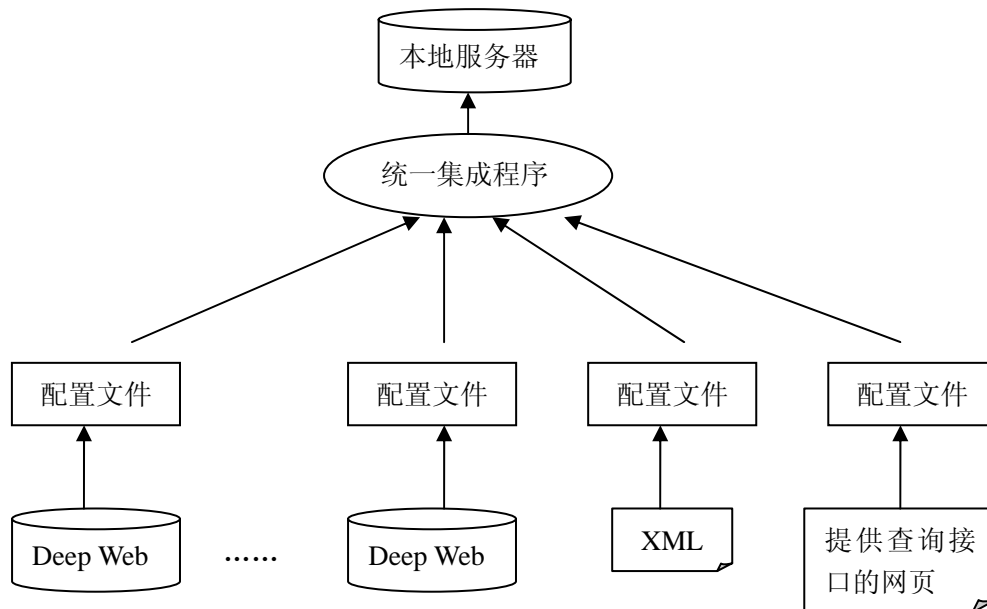


图 2 基于配置文件集成系统的整体框架图

如图 2 所示，该基于配置文件的集成系统主要包括：多个数据源，例如 Deep Web 数据、XML 数据、及其它提供接口查询的网页；多个配置文件单元，这多个配置文件单元的每一个与上述多个数据源的每一个相对应；统一的集成单元，用于集成底层的各个数据源，它利用数据源所对应的配置文件，采用统一的方式，对数据源中的数据进行抽取；以及本地服务器，用于保存集成起来的所有数据，这样用户检索时，便可以直接在本地服务器上进行搜索，提高效率。

我们这种方法的优点：

- 1、能够快速搭建起一个面向领域的集成系统；
- 2、非常容易商业化。目前，我们已经通过此系统在工作信息领域集成了超过 200 万条招聘信息，在政府新闻领域集成了近 70 万条数据，并通过两个网站 [JobTong](#)（工作通）和 [GovTong](#)（政务通）提供给外界进行搜

索。

3、抽取准确度高，抽取粒度细。这也是其它自动方法未能达到的目标。

4、面向大规模数据。我们的方法并不是做一个简单的应用程序，而是面对实际的海量的 Deep Web 上的数据。

5、模拟用户点击行为。使得这种方法既不获取冗余数据，也不会少获取数据。

五、工作通：面向领域的数据集成系统介绍

工作通(<http://www.jobtong.cn>)系统是一个面向工作领域的信息集成系统。其框架就是上文所提到的Deep Web上数据集成系统框架,其核心就是基于配置文件的Deep Web数据集成系统。如图4所示,系统先将Deep Web上工作信息领域的的数据爬取下来,放在本地数据库中。用户通过Jobtong网站提供的查询界面,可以查询自己想要申请的职位信息,而此查询关键字会被提交到本地服务器上进行处理,然后将查询到的结果返回给用户。由于对于用户的查询来说,直接再本地进行处理,无需再将查询分派到网络上的Deep Web相应查询接口查询,因此,能够大大提高查询速度,缩短用户等待时间。

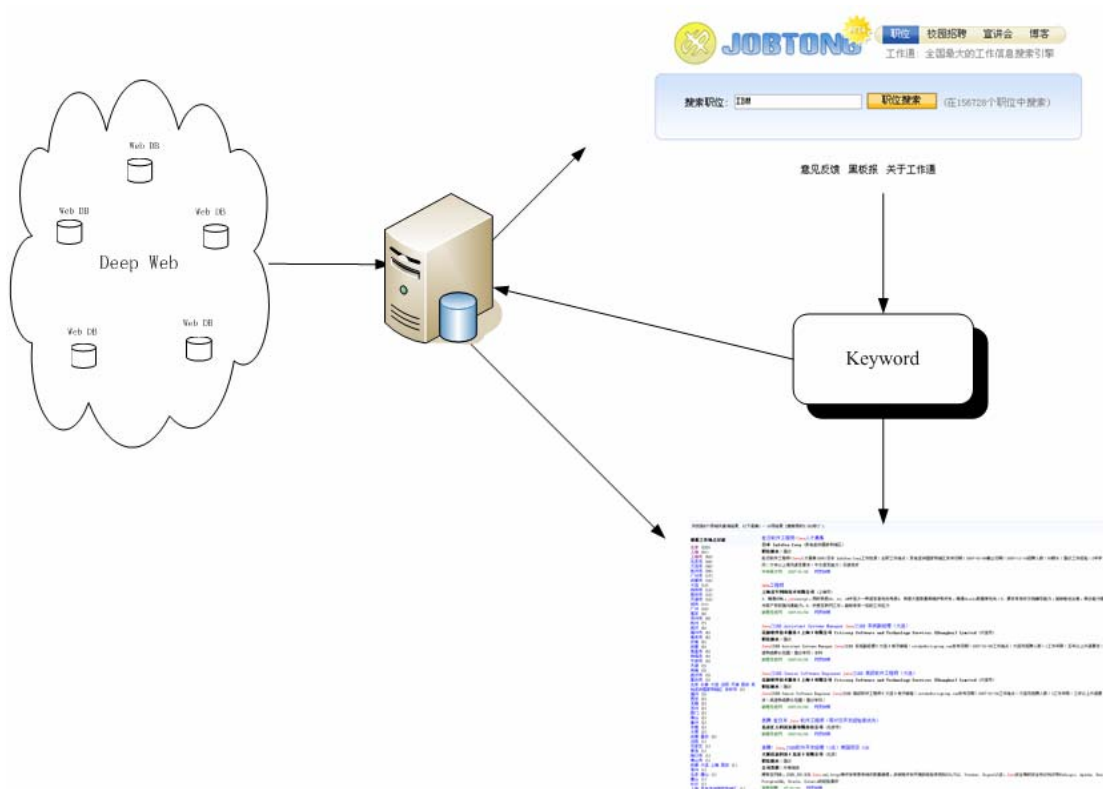


图4 工作通系统的数据流图

工作通系统于2007年1月1日向人大学生内部开放,受到了人大应届毕业生的热烈响应,成为毕业生了解职位招聘信息的一个重要渠道。3月28日,工作通系统向公网发布BETA版。

目前,工作通系统已经集成了超过200万条工作信息,并以每天数千条的速度在不断增长中。而Jobtong网站也每天为上千人提供近万次的招聘信息查询服务。

同时,我们还在政府新闻领域实现了一个集成系统的应用:Govtong(政务通,<http://www.govtong.cn>)。通过Govtong的实现,也证明了本文所提出的面向领域的数据集成系统是切实可行的。目前,Govtong已经集成了近70条数据,并以每天上千条的速度在持续更新中。

六、总结

Jobtong 这样一套面向领域的数据集解决方案，是我们实验室在 Deep Web 数据集研究方面多年研究成果的结晶。它既能够在某一个具体领域快速构造出一个大规模的数据集成系统，继而实现一个垂直搜索引擎；也能够将各个不同数据源的模式抽象出来，提供 API，构造 Mashup 应用。

在未来一段时间内，随着 Jobtong 项目的不断完善和改进[16]，以及在更多领域的集成实验，它一定会成为一个在 Deep Web 上进行数据集成的强大工具！

参考文献：

- [1] W. Liu, X. Meng, W. Meng: A Survey of Deep Web Data Integration. Chinese Journal of Computers, Vol 30, No. 9: 1475-1489, Sept. 2007
- [2] B. Michael K. The Deep Web: Surfacing Hidden Value[R]. The Journal of Electronic Publishing from the University of Michigan, July 2001.
- [3] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the Deep Web: A Survey. Communications of the ACM (CACM), 50(5):94-101, May 2007
- [4] Chang, K.C.C., He, B., Zhang, Z.: Toward large scale integration: Building a metaquerier over databases on the web. In: CIDR. (2005) 44–55
- [5] He, H., Meng, W., Yu, C.T., Wu, Z.: Wise-integrator: An automatic integrator of web search interfaces for e-commerce. In: VLDB. (2003) 357–368
- [6] Androutsopoulos, I., Ritchie, G.D., Thanisch, P.: Natural language interfaces to databases - an introduction. CoRR cmp-lg/9503016 (1995)
- [7] A. Popescu, O.E., Kautz, H.: Towards a theory of natural language interfaces to databases. International Conference on Intelligent User Interfaces. (2003)
- [8] X. Li, W. Meng, X. Meng: EasyQuerier: A Keyword Query Interface For Web Database Integration System. To appear in Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), Bangkok, Thailand, April 9-12, 2007.
- [9] G. O. Arocena, A. O. Mendelzon. WebOQL: Restructuring Documents, Databases, and Webs. In ICDE, pages 24-33,1998.
- [10] X. Meng, H. Lu, H. Wang. SG-WRAP: A Schema-Guided Wrapper Generation. In ICDE, pages 331-332, 2002.
- [11] R. Baumgartner, S. Flesca, G. Gottlob. Visual Web Information Extraction with Lixto. In VLDB , pages 119-128, 2001.
- [12] C. Chang, S. Lui. IEPAD: Information extraction based on pattern discovery. In WWW, pages 681-688, 2001.
- [13] V. Crescenzi, G. Mecca, P. Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In VLDB, pages 109-118, 2001.
- [14] Y. Zhai, B. Liu. Web data extraction based on partial tree alignment. In WWW, pages 76-85, 2005.
- [15] B. Liu, R. L. Grossman, Yanhong Zhai. Mining data records. In Web pages. In KDD, pages 601-606, 2003.
- [16] Zhongyuan Wang. Jobtong System Progress and Research Topics. In WAMDM Seminars: <http://idke.ruc.edu.cn/seminars/2007/12.09/Jobtong%20System%20Progress%20and%20Research%20Topics.ppt>, Dec 9, 2007.