学校编码：10384　　分类号_____密级_____
学号：200428015　　UDC _____

厦门大学

硕 士 学 位 论 文

基于树形结构的 Web 信息抽取技术研究

Research of Web Information Extraction Based on Tree structure

任仲晟

指导教师姓名：薛 永 生　教授
专 业 名 称：计算机软件与理论
论文提交日期：2007 年 4 月
论文答辩时间：2007 年 5 月
学位授予日期：

答辩委员会主席：_____
评　　阅　　人：_____

2007 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘 要

随着 Internet 的快速发展，Web 已经发展成为一个巨大的、分布式的和共享的信息资源。目前 Web 数据大都以 HTML 页面的形式出现。由于 HTML 描述的数据是一种半结构化的数据，这使得由 HTML 描述的 Web 页面只适合人类的浏览，应用程序无法直接解析并利用 Web 上的丰富信息。为了增强 Web 数据的可用性，提供更多的增值服务，出现了 Web 信息抽取技术。它通过包装(wrapper)现有的 Web 信息源，将网页上的信息以结构化的方式抽取出来，为应用程序利用 Web 中的数据提供了可能，因此有着广阔的前景，是当今数据库领域的研究热点之一。

本文首先对 Web 信息抽取的一些基本概念做了简要介绍，并简述了 Web 信息抽取技术的产生和发展。在此基础上，给出了适用于本文算法的 Web 页面的定义。

其次详细介绍了当前 Web 信息抽取技术的一些常用方法，并对这些方法进行分类，进而对这些方法进行对比分析，指出各种方法的优缺点。在分析了多种方法的基础上，讨论了未来 Web 信息抽取技术研究发展的方向。

最后，提出了一种基于树形结构的 Web 结构化数据抽取算法。该算法基于 HTML 的树形层次结构，包括 HTML 树构造算法，数据区域挖掘算法，数据记录挖掘算法，以及数据记录模式生成算法。本算法引入了页面元素布局位置等信息用于清洗页面，采用层次划分思想实现页面数据区域的挖掘，并通过树匹配生成记录模式，实现最终数据项抽取。通过理论分析和实验表明，该方法可以有效地实现 Web 结构化数据抽取。

关键词：Web 数据抽取，Web 挖掘，信息抽取

Abstract

With the rapid development of Internet, Web is becoming a vast, distributed, and shared information resource. Most of Web data are in the form of HTML. Due to the semi-structured nature of HTML pages, Web pages are easy for exploring by human beings while it is difficult for applications to process and use the data in the Web pages. To strengthen the availability of Web data, providing more value-added services, Web information extraction technology comes out, which wraps the Web resources, extracts semi-structured data, and provides supports to applications using Web data. Therefore, the research of Web information extraction is one of the hottest research areas in database field and has a promising future.

In this paper, we first briefly introduce some basic concept of Web information extraction and also give a short introduction to the development of the technology of Web information extraction. Then we describe the definition of the web pages used by our algorithm.

Secondly, we describe, compare, and analyze several kinds of Web information extraction methods commonly used at present in detail, pointing out advantages and disadvantages of each method. Furthermore, we discuss the future direction of research and development of Web information extraction.

Finally, we propose tree structure based Web data extraction algorithm in view of the inadequacies of the existing methods. Our tree structure based algorithm includes: the algorithm of HTML tree construction, the algorithm of data region mining, the algorithm of data record mining, and the algorithm of record schema generation. Our algorithm cleans the Web pages using the position information of page elements, mines data region by hierarchical clustering, and generates record schema finishing data item extraction through tree matching. Theoretical analysis and experimental results show that our algorithm can improve the accuracy and efficiency of Web data extraction.

Key Words: Web data extraction, Web mining, information extraction

目 录

Contents

第一章 绪论

本章首先简述论文的研究背景，包括当前 Web 数据存在的问题，推动 Web 信息抽取技术发展的客观需要，以及 Web 信息抽取技术面临的挑战；然后介绍本课题研究的科学依据和意义；最后介绍了本文的主要内容和组织。

1.1 论文的研究背景

随着 Internet 的快速发展，互联网成为全球信息传播与共享的重要途径。互联网上的数据大多数出现在用 HTML 语言描述的页面中，而 HTML 语言最大的特点就是结构不完整，用 HTML 语言描述的 Web 数据是一种半结构化(semi-structured)的数据。半结构的数据缺乏语义信息，数据集成性差，给应用程序的解析带来了很大的困难，从而使得 Web 上的海量数据不能被充分利用。其次，由于人类的审美观，以及商业上的要求，现在的 Web 页面中除了与主题相关的数据外，还充斥着大量的与主题无关的信息，诸如一些图片，广告信息，脚本语言，超链接等等。这些无关内容的存在，也影响了对有用信息的判别。如何避免“数据爆炸，知识匮乏”的尴尬，从海量的半结构化信息中抽取出结构化的，符合主题的数据，推动了 Web 信息抽取技术的产生与发展。

Internet 上的信息，数量大，更新快，这些特性使得 Web 信息抽取不同于传统的信息抽取。为了增加 Web 数据的可用性， Web 信息抽取技术通过包装器，将网页上的信息以结构化的方式抽取出来，为应用程序直接利用互联网上的海量数据提供了可能。

Internet 所具有的海量、异构、动态等特性也给 Web 信息抽取技术的研究带来了挑战。首先，互联网是一个巨大的信息空间，Web 页面数以亿计，而且仍在以几何级数增长，如何自动高效地处理海量 Web 信息是一个难点；其次，Web 页面的异构性，例如同一主题的信息分散在多个形式各异的 Web 页面中，使得如何在这些异构的网页里准确识别所需要的信息变得复杂;最后，Internet 是一个动态的空间，Web 网站的页面格式和内容瞬息万变，如何保证 Web 信息抽取技术的适应性也是一个有待解决的问题。

1.2 本课题的研究价值

Web 信息抽取的直接应用，就是让人们在互联网中迅速准确地查找自己所需要的信息，加快信息获取的速度。例如，当网页上的数据以更加结构化的形式出现在人们面前时，人们就能够应用传统的数据库查询方式，直接定位到自己所需要的信息。

Web 信息抽取的另外一个重要应用，就是为高一级的应用程序提供增值服务。例如提取出的结构化信息，能够进一步地提供给信息检索，数据挖掘，机器翻译等其它 Web 信息处理系统。同时，通过将半结构化的数据转换成更加结构化的数据，为集成来自不同网站的数据提供了可能。数据的集成，方便了某个领域信息库的生成，给人们的检索提供了方便。例如，将众多不同商务网站上的价格信息进行抽取，以统一的界面呈现给用户，使得比较购物(comparative shopping)成为了可能。另外，通过对某一领域知识的抽取、汇总、集成，可以形成该领域的相关知识库。这一切都为人们更好地利用互联网上的海量信息提供了一种新的途径。

上述种种 Web 信息抽取技术带来的效益，充分说明了 Web 信息抽取技术是一个非常有前景的研究领域。另外一个方面，当前大量网站设计的一个特点，就是事先生成网页模版，系统根据用户的请求做出响应，从后台数据库中提取相关信息，完成模版的填充，最后再返回给用户。因此，对这一类型网页的信息抽取，更具有价值。这种类型的网页，在许多文献中，也被称作模版页面(template page)，deep page，hidden page 或者是数据导向型网页(data intensive page)。

本课题的研究重点就是如何从这种数据导向型网页中，有效地提取出结构化的数据信息。本文通过利用 HTML 页面的固有层次结构，引入叶子节点相似度划分，路径分裂，树匹配等技术，能够有效地识别抽取出这种类型网页中的数据，并转换成结构化的数据。

本课题另外一个方面的工作是试图对各种 Web 信息抽取技术做出一个对比归纳，并探讨未来 Web 信息抽取技术的研究发展方向。