

一种 Web 数据挖掘系统的设计和研究

李 健,徐 超,谭守标

(安徽大学 电子科学与技术学院,安徽 合肥 230039)

摘 要: Web数据挖掘是指从众多的 Web 网站、网页上挖掘出有用数据和知识的过程,因其具有广泛的应用前景而成为当前 IT 研究的热点之一,同时它也是一个具有挑战性的研究领域,存在很多问题亟待解决。针对一个案例,设计一个系统(或模型)实现 Web 数据的挖掘,是一次实践性研究。系统采用当前流行的软件工具(VS2005 和 SQL2000 数据库)和编程语言(C#)进行开发设计,主要由数据的下载、预处理、后处理和前台检索等模块组成,基本达到 Web 数据挖掘的目的。

关键词: Web 数据挖掘;下载;预处理;后处理;前台检索

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2009)02-0070-04

Design and Research of a Web Data Mining System

LI Jian, XU Chao, TAN Shou-biao

(School of Electronic Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Web data mining is the process of extracting useful data and knowledge from lots of websites and web pages. It becomes one of the hottest issues in IT at present, but it is also a challenging research field because there are many questions expected to be resolved. Introduces a system (or model) which based on a case to achieve web data mining. It is a practice. The system is designed by using the nowadays popular software (VS2005 and SQL2000 database) and programming language (C#). It consists of the blocks: download, pre-dealing, latter-dealing and foreground searching interface. The system basically gains the ends of web data mining.

Key words: web data mining; download; pre-dealing; latter-dealing; foreground searching interface

0 引 言

数据挖掘(Data Mining)是指从大量的数据(结构化和非结构化的)中提取有用的信息和知识的过程^[1]。最初,数据挖掘是研究从数据库(结构化的数据)中发现知识(KDD, Knowledge Discovery In Database)^[2],但在 Internet 技术迅猛发展的今天,Web 页面为人们提供了海量的数据信息,于是基于 Web 的数据挖掘应运而生,并成为当前数据挖掘的热点,称之为 KDW (Knowledge Discovery In Web)^[2]。

Web 数据挖掘是指从众多的 Web 网站、网页上挖掘出有用数据和知识的过程^[1]。Web 数据挖掘的任务是帮助人们从大量的 Web 页面上有效地收集、选择和存储所感兴趣的信息以及在日益增多的信息中发现新的概念和它们之间的关系,做到信息处理的自动化。它在实际中有着广泛的应用,尤其对企业获取有用可

靠的外界信息,商业运作过程中收集、分析数据从而做出正确决策有着十分重要的意义^[3]。

Web 数据挖掘比基于关系数据库或数据仓库的数据挖掘要复杂得多。如果把 Web 看作一个巨大的、复杂的分布式数据库,每一个站点都是一个独立的数据源,它们之间的数据组织形式与结构是不相同的。因此,Web 上的信息可视是为一个异构的数据库环境,对这些数据进行挖掘,首先要解决站点之间异构数据的集成问题,为用户提供一个统一的视角来看待 Web 资源。另外,除了不同站点的异构以外,Web 页面上大量的数据还是半结构、无结构的文本和多媒体信息。而且,对于某一特定应用或某一个人来说,Web 页面上大部分的数据是无用或“垃圾”信息,只有很小一部分是我们需要的^[4]。这些问题使得 Web 数据挖掘成为一个具有挑战性的研究领域。

当前 Web 数据挖掘的理论还不够成熟,Web 数据挖掘技术也期待研究,文中根据实际应用,针对一案例(交通信息的挖掘),设计一个系统(或称之为模型)达到 Web 数据的挖掘、处理和再利用。这是 Web 数据挖掘的一次实践性研究。由此还可以将之推广到其他的

收稿日期:2008-06-19

基金项目:安徽省自然科学基金项目(2005KJ004ZD)

作者简介:李 健(1985-),男,硕士研究生,研究方向为网络与智能系统;徐 超,教授,研究方向为网络与智能系统;谭守标,副教授,研究方向为网络与智能系统。

应用方面。

1 系统目标和方案设计

根据Web上数据的特点和数据挖掘的一般过程, Web数据挖掘的流程可分为以下5个功能模块,如图1所示^[5]。

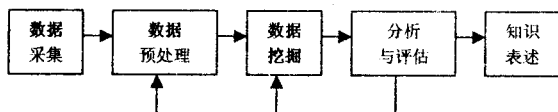


图1 Web数据挖掘的一般流程

此次设计的系统也是在此大思路的指导下,针对Web网页上大量的各种交通信息进行下载、处理,从而使用户轻松方便地获得需要的信息。例如,有的网页提供火车信息,有的提供飞机信息,有的提供具体某一城市的公交信息,而我们的目标在于提取这些网页中可靠的信息并进行加工综合,满足用户多方位的需求,不必烦琐地搜索各种网页来获取信息。系统方案如下所述。

整个系统分成以下几个部分,如图2所示。

1) 下载。

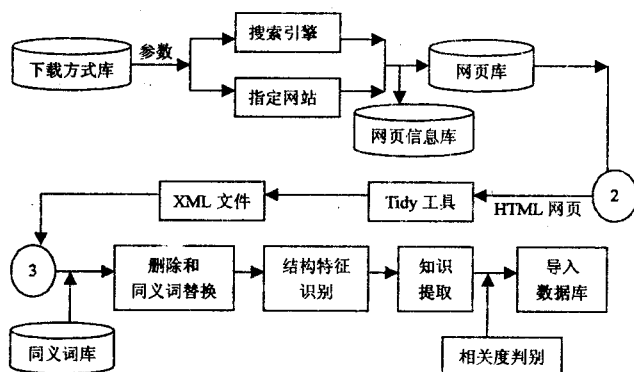


图2 系统后台框图

下载方式有多种,目前很多网站提供搜索服务,常用的如百度,google等,通过在这些网站键入关键字,然后搜到我们想要的网页;其二,通过指定网站,如铁道部官方网站,传入参数找到我们需要的信息。如上框图所示,需要传递的参数存放在数据库中,搜索得到的网页放入网页库。同时,下载的网页日志和数据的可靠度均放入网页信息库,用来作为判别网页有效性的根据,而不致得到过时甚至错误的数据。

2) 预处理。

这部分是利用Tidy工具,将下载的HTML网页转化成XML文件。这是因为下载得到的HTML网页中的数据很混乱,不具有结构化的特点,无法直接提取,而XML是严格的格式化定义数据的,所以将XML作为数据挖掘的对象可以获得和数据库挖掘一样的效

果。

3) 删除、替换、提取和导入。

对于转化后的XML文件还包含许多不需要或无用的信息,要将之删除;对于有用的信息,但由于不同网页词语表达方式的不统一,所以首先要用同一词语替换网页上的同义词,例如用始发站替换开始站、出发站等同义词,从而方便后面的提取工作;同样,不同网页存放数据的格式也不相同,有的采用横排,有的采用竖排,要识别其结构,然后针对不同的结构进行相应的知识提取(即我们需要的信息)。提取的信息还要和数据库中已有的数据进行比较,从而确定是否有价值,是否需要导入数据库中。

4) 前台检索。

从网页上下载提取出的信息导入数据库后,就是知识库。数据挖掘的目的是达到数据的重组织和再利用,再做一个前台检索界面,使用户可以从知识库中检索出他们需要的信息,从而完成数据挖掘的任务。如图3所示。

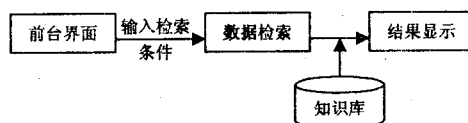


图3 前台检索框图

2 系统设计的实现

上面介绍了系统设计的方案,整个系统就是按此方案的框图来设计实现的。下面简介一下用到的软件及知识点,然后介绍各个部分的设计并演示结果。

2.1 软件及知识点简介

2.1.1 开发平台

Microsoft Visual Studio 2005 集成开发环境,SQL Server2000 数据库。

我们设计的系统是利用C#语言编程,通过数据库存取数据实现的。在Visual Studio中,C#语言包含大量丰富的类可供开发人员使用,在用C#语言进行各模块的编程时,需要经常调用相关类中的各种方法。

2.1.2 数据库说明

为方便下面各模块设计的介绍,先对本系统所需建立的数据库作一说明:

下载表,存有下载的网页地址URL,网页处理未处理,网页的有效性,以及下载时间等字段信息;参数表,提供下载网址所要的参数;同义词表,用来消除网页中的同义词,方便提取;信息表,提取之后有用数据的存放处。

上面说明的是数据库中主要的表和字段,另外,前

台检索是基于信息表,但由于检索算法的需要,还可能建立多张表。根据编程的需要,还要用 SQL 语言建立相关的存储过程。

2.2 各模块(部分)设计及结果演示

2.2.1 下载和预处理

下载是一个给定初始条件,具有自组织自学习的过程。下载的三种方式是:

(1)搜索引擎搜索:通过下载表和参数表向搜索引擎网址传递参数,搜索到相关链接,提取出这些链接网址;

(2)固定网址加参数下载:预知某些网页如铁道部网站首页,在其中键入地名(即参数),就会跳到我们需要的信息页面,这些网页的网址存放在数据库中;

(3)网页直接下载:这些网页提供的就是我们需要的信息,直接下载即可,这些网页的网址也存放在数据库中。

由上可以看出,这三种下载方式相辅相成。由 1 可以搜到 2,3 的网页,由 2 可以搜到 3 的网页,而这些网页的地址都存放在数据库中。通过下载表中的网址和参数表中的参数就可以搜到具体的网页,再送给处理。处理的结果可以得出下载网址的有效性,从而确定保留或删除网址。保留下可用的网址,根据当前时间与下载时间的距离确定是否再次实施下载动作,以更新数据库中可能过时的信息,同时也更新下载表中的网址。

预处理,即使用 Tidy 工具(一种转化工具,读者可上网下载)将下载下来的 HTML 页面转化成 XML 格式文件。现在的网页绝大多数是 HTML 格式,转化后的 XML 文件只能说是符合 XML 规范的 HTML 文件,也可称之为 XHTML。因为有的 HTML 网页本身结构很不规范,所以 Tidy 工具并不能转化所有的 HTML 网页,对于这些网页,我们将之删除。Tidy 工具的使用可以嵌入到下载模块中,下载的 HTML 文件直接送给 Tidy 工具处理,这一预处理后删除原 HTML 文件,存放转化后的 XML 文件到本地磁盘,留作后处理。图 4 是下载并经过预处理后的 XML 文件,图 5 是之后缀改成 .html 后的页面(点击右键查看源代码,与图 4 同)。

2.2.2 后处理

如图 4 所示,预处理后的文件还含有很多无用数据(有用部分在下面,因无法完全显示,见图 5),要将其删除。从图 5 可以看到它使用了“终到站”,为了统一,将之替换成“终点站”,其他表头字段不变。如此经过重新读写后的 XML 文件见图 6(对比图 4,可以看出前面无用的数据已被删除)。

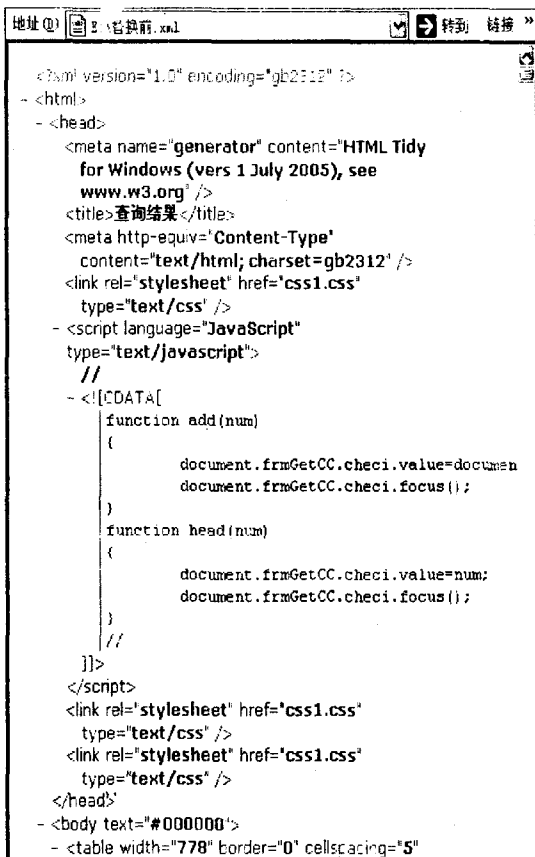


图 4 下载并经过预处理后 XML 文件

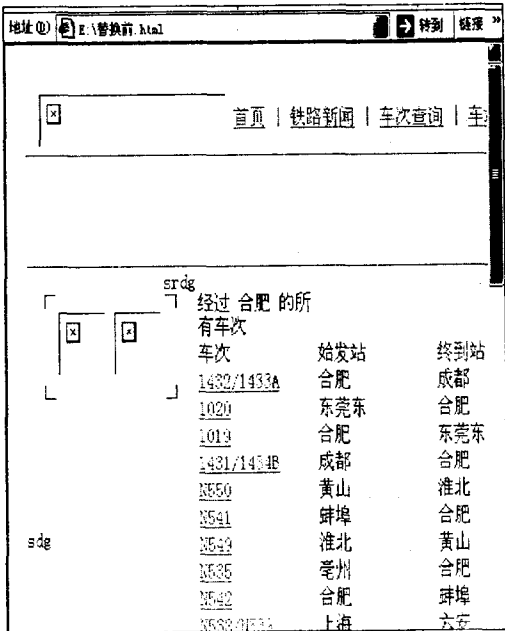


图 5 预处理后的 .html 文件

2.2.3 导入数据库

图 6 所示的文件具有规范的 XML 结构样式,是明显的树状结构,可采用扫描文件,节点提取的方法进行数据的提取,并导入到信息表中。同时下载表中的处理字段的值更新为已处理。网页数据的有效性是通过相关度判别来实现的,如果多次(或多个网页)下载

的数据基本一致,那么该数据的可靠性就越大。通过此法更新数据库中的信息,同时更新下载表中的有效性字段值。

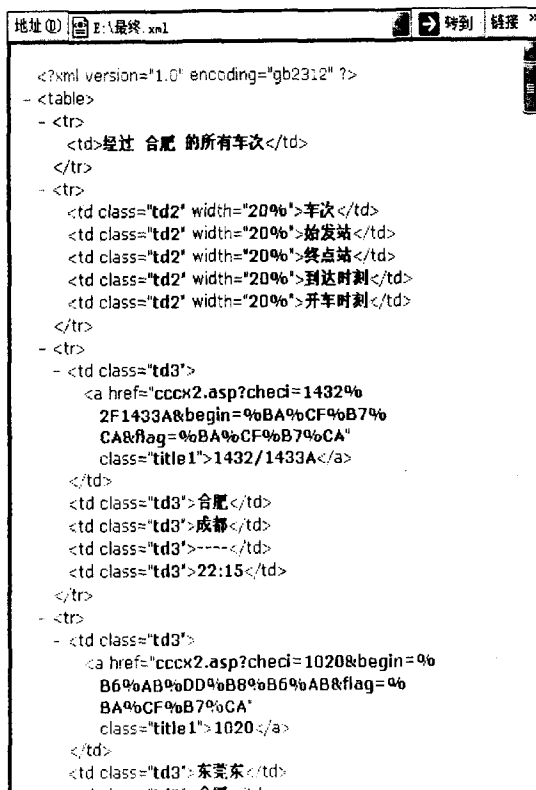


图6 后处理后的XML文件

2.2.4 前台检索

利用 Visual Studio 2005 集成开发环境和 C# 语言,可以很轻松地开发 Asp.net 网站。我们编写建立自己的网页,利用下载到数据库中的信息,可以提供给用户多方位和快捷的服务。需要指出的是,前台页面

(上接第 69 页)

与个性化分类。

Agent 的自主性使得 Agent 能够在没有管理员指导的情况下由自身的决策机制决定采取何种行动。

Agent 的反应性使得 Agent 能够学习用户的反馈,不断调整和改进分类算法,完善分类系统。Agent 的社会性使得系统的各组成 Agent 能够方便地相互交流,更好地为用户服务。

4 结束语

文中提出了一种基于 Agent 的邮件过滤与个性化分类系统,不但可以过滤垃圾邮件,还可以对正常邮件进行自动分类,有效地减少了用户手工归档的工作量。目前,该系统的具体实现还在实验之中,有关 Agent 与 Agent 之间的高效通信等问题还有待进一步研究。

的数据检索是基于信息表,但由于其中的数据不具有 consistency,需要经过重组织方能使用,方能提供给用户完善的信息。

3 结束语

作为 Web 数据挖掘的一次实践性研究,设计的系统还不够完善,在实际的设计中还存在很多问题,比如:因为是从网上下载数据,而提供网页的服务商会更新网页,从而引起网址格式的变动,这必然使大量下载成为难题;各个网页提供的数据结构不一,信息也不一致(有的信息不完整),这对提取工作提出挑战;同样,存入数据库的信息的不一致性,也给检索算法的设计带来麻烦。这些都是亟待解决的问题。

总的来说,Web 数据挖掘目前还处于发展时期,其技术还不成熟,但由于其广泛的应用价值,必然会让更多的人来参与研究,其路会越走越远。

参考文献:

- [1] 苏新宁,杨建林,邓三鸿,等.数据挖掘理论与技术[M].北京:科学技术文献出版社,2003.
- [2] 恽爽,韩立新,董浚,等.KDW 综述:基于 Web 的数据挖掘[J].计算机工程,2003(1):284-286.
- [3] Linoff G S, Berry M J A. Web 数据挖掘:将客户数据转化为客户价值[M].沈钧毅,等译.北京:电子工业出版社,2004.
- [4] 李雄飞,李军.数据挖掘与知识发现[M].北京:高等教育出版社,2003.
- [5] 周琪峰.基于 Web 的数据挖掘技术的研究[J].电脑知识与技术,2007(1):97-103.

参考文献:

- [1] 徐俊萍,翟玉庆.基于 Agent 的个性化信息服务技术的研究[J].计算机工程与科学,2002,24(3):74-76.
- [2] 张克,邵长胜,强文义.基于面向 Agent 技术的任务规划系统研究[J].高技术通讯,2002(5):82-86.
- [3] 王新梅,芦苇,尹朝庆,等.基于文本挖掘的邮件分类与过滤[J].计算机工程与应用,2006(2):135-137.
- [4] 李志博,余正红,尹朝庆,等.邮件服务智能代理的研究[J].计算机工程与设计,2007,28(3):683-686.
- [5] Zhang PHaiyi, Li PDi. Naive Bayes Text Classifier[C]//Proceedings of the 2007 IEEE International Conference on Granular Computing. USA: [s. n.], 2007.
- [6] 成宝国,冯宏伟.一个基于 Naive Bayesian 垃圾邮件过滤器的改进[J].计算机技术与发展,2006,16(2):98-99.
- [7] 王宁,张建忠,何云,等.基于改进贝叶斯模型的中文邮件分类算法[J].计算机工程与应用,2006(31):97-100.