Web 挖掘技术综述

王欣如

(重庆大学计算机学院　重庆 400044)

摘　要　随着 Internet 的迅速发展与广泛应用，Web 成为一个巨大的、蕴涵着具有潜在价值知识的分布式信息空间，为数据挖掘研究提供了丰富的数据源，同时也提出了新的挑战。该文首先概述了数据挖掘的概念、挖掘算法及其主要应用领域，然后详细阐述了 Web 内容挖掘、Web 结构挖掘和 Web 日志挖掘的概念和主要的挖掘算法及最新研究进展，最后介绍了 Web 挖掘的研究方向和发展趋势。

关键词　Web，数据挖掘，Web 挖掘

1　引言

Web 作为一个巨大的信息源，不仅内容复杂、而且形式各异。随着 Web 站点自身信息越来越丰富和拓扑结构越来越复杂，目前信息服务中普遍存在着“信息过载”和“资源迷向”的状况。近年来，因特网的飞速发展与广泛应用，使得 Web 上的信息量以惊人的速度增长，未来 Web 将包含人类信息的主要部分，因此，如何从 Web 中找到感兴趣的内容变得越来越重要。为数据挖掘提供了丰富的数据源和新的研究课题。面对 Web 丰富的信息内容，巨大的数据量，加之万维网分布、动态、海量、异质、复杂、开放性的特点，人们如何从这海量的数据中，查找自己想要的数据和有用信息，迫切需要一种新的技术能自动地从 Web 资源上发现、抽取和过滤信息，随之 Web 挖掘技术应运而生。

Web 挖掘就是从与 WWW 相关的资源和用户浏览行为中发现、抽取感兴趣的潜在的有用模式和隐藏的信息。它以从 Web 上挖掘有用知识为目标，以数据挖掘、内容挖掘、多媒体挖掘为基础，并综合运用计算机网络、数据库、人工智能、信息检索、可视化等技术，将传统的数据挖掘技术与 Web 结合起来。但是，Web 挖掘与传统的数据挖掘相比又有很多独特之处。首先，Web 挖掘的对象是大量、异质、分布的 Web 文档；其次，Web 在逻辑上是一个由文档节点和超链接构成的图，因此 Web 挖掘所得到的模式可能是关于 Web 内容的，也可能是关于 Web 结构的；此外，由于 Web 文档本身是半结构化或无结构的，且缺乏机器可理解的语义，而数据挖掘的对象局限于数据库中的结构化数据，并利用关系表格等存储结构来发现知识，因此数据挖掘技术要应用于 Web 挖掘，应当对 Web 文档进行预处理。这样，开发新的 Web 挖掘技术，以及对 Web 文档进行预处理以得到关于文档的特征表示，便成为 Web 挖掘研究的重点。

Web 挖掘可在多方面发挥作用，如电子商务中销售搭配、营销策略，搜索引擎结构的挖掘，搜索引擎的开发，改进网站结构，确定权威页面，Web 文档分类，智能查询，个性化信息服务等。

2　数据挖掘概述

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取潜在的、不为人知的有用信息、模式和趋势，是一种新兴的数据处理技术。

2.1　数据挖掘分类

从挖掘对象、挖掘任务、挖掘方法等几个方面，将数据挖掘划分为以下类型。

根据挖掘任务分：分类或预测模型知识发现、数据总结、数据聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等等。

根据挖掘对象分，有如下若干种数据库或数据源：关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据库、异构数据库、遗产数据库(Legacy)、Web 等。

根据挖掘方法可粗分为：统计方法、机器学习方法、神经网络方法和数据库方法。统计方法中可细分为：回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别等)、聚类分析(系统聚类、动态聚类等)、探索性分析(主元分析法、相关分析法等)等。机器学习中可细分为：归纳学习方法(决策树、规则归纳等)、基于范例学习、遗传算法等。神经网络方法中可细分为：前向神经网络(BP 算法等)、自组织神经网络(自组织特征映射、竞争学习等)等。

2.2　数据挖掘的技术方法

数据挖掘的方法通常可以分为两大类：一类是统计型，常用的技术有概率分析、相关性、聚类分析

和判别分析等;另一类是人工智能中的机器学习型,通过训练和学习大量的样品集得出需要的模式或参数。数据挖掘的应用中,最终的目标都是发现有价值的知识和信息,有共同的思路和步骤,但也存在很大的差异和区别。由于各种方法都有自身的功能特点以及应用领域,数据挖掘技术的选择将影响最后结果的质量和效果。下面对数据挖掘中常用的关联分析、决策树和神经网络等几种技术方法进行讨论,包括技术的基本思想、优势与缺点和主要应用领域(见表 1)。

表 1 数据挖掘的主要技术方法

技术方法	主要功能和特点	主要应用领域
关联分析	分类、聚类	零售业、保险业和通讯业
决策树	归纳分类、直观	制造业、医学和零售业等
遗传算法	聚类、优化、高效性	金融业、保险业和农业等
贝叶斯网络	分类、聚类和预测: 易理解	医学、制造业和电信等
粗糙集方法	不确定性分类	零售业、金融业和制造业等
神经网络	预测、分类和聚类: 解释性差	金融业、保险业和制造业等
统计分析	聚类: 结果精确、易理解	金融业、制造业和医学等

3 Web 挖掘

3.1 Web 挖掘的步骤

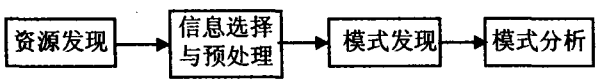

图 1 Web 挖掘的步骤

(1)资源发现,即收集所需的网络文档;(2)信息选择和预处理,即从检索到的网络资源中自动挑选和预先处理得到专门的信息;(3) 模式发现,即从单个的 Web 站点以及多个站点之间发现普遍的模式;(4)分析,对挖掘出的模式进行确认或者解释。

3.2 Web 挖掘的分类

Web 数据有三种类型:Web 数据,即人们通常所说的 Web 文档(主要是 HTML 或 XML 格式的)、Web 结构数据(如 Web 文档中的超链接)、用户访问数据(如服务器上的 Web 日志信息)。相应地,Web 挖掘也分为三类:Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘,如图 2 所示。

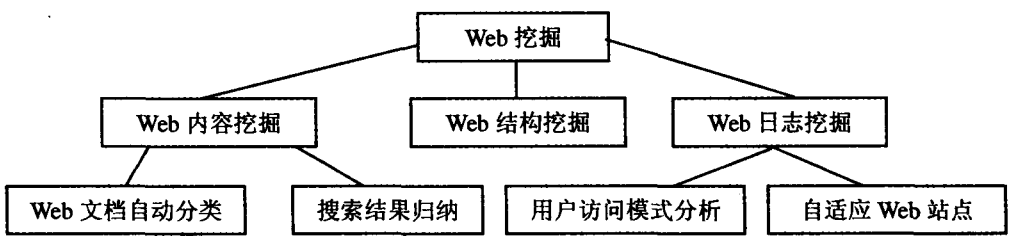

图 2 Web 挖掘的分类

3.2.1 Web 内容挖掘

Web 内容挖掘是从文档内容或其描述中抽取知识的过程。Web 文档文本内容的挖掘,是基于概念索引的资源发现。Web 内容挖掘有两种策略:直接挖掘文档的内容,或在其它工具搜索的基础上进行改进。采用第 1 种策略的有针对 Web 的查询语言 Web Log,WebOQL 等,利用启发式规则来寻找个人主页信息等等。采用第 2 种策略的方法主要是对搜索引擎的查询结果进行进一步的处理,得到更为精确和有用的信息。属于该类的有 WebSQL ,及对搜索引擎的返回结果进行聚类的技术等。

Web 内容挖掘的数据对象既可以是结构化的,也可以是非结构化的、半结构化的。Web 内容挖掘的结果既可以是对某个文本内容的概括,也可以是对整个文本集合的分类结果或聚类结果等。

目前 Web 内容挖掘的主要研究内容是对 Web 上大量文档集合的内容进行总结、分类、聚类、关联分析、科学文献资料浏览导航,以及利用 Web 文档进行趋势预测。

3.2.2 Web 结构挖掘

Web 结构包括不同网页之间的超链接结构和一个网页内部的可以用 HTML 或 XML 表示成的树形结构,以及文档 URL 中的目录路径结构等。Web 结构挖掘是从 WWW 的组织结构和链接关系中推导知识。主要是通过对 Web 站点的结构进行分析、变形和归纳,将 Web 页面进行分类,以利于信息的搜索。由于文档之间的互连,WWW 能够提供除文档内容之外的有用信息。利用这些信息,可以对页面进行排序,发现重要的页面。这方面工作的代表有 PageRank 和 CLEVER。此外,在多层次 Web 数据仓库(MLDB)中也利用了页面的链接结构。

Hits、Pagerank 以及在链接结构中增加了 Web 内容信息的 hits 改进算法等,主要用于模拟 Web 站点的拓扑结构,计算 Web 页面的等级和 Web 页面之间的关联度,典型的例子是 CLEAVER system

和 Google。

Web 结构挖掘所得到的模式，可以揭示许多蕴涵在 Web 内容之外的隐含着的有用信息。如通过文档之间的超链接，可以挖掘出文档之间的引用关系，从而有助于找到与用户请求相关的权威页面；通过分析 Web 网页内部树形结构，可以发现与给定页面集合相关的其它页面；Web 页面的 URL 同样可以反映页面的类型以及页面之间的从属关系，通过分析页面的 URL 信息，可以找到改变了位置的 Web 页面的新位置：

3.2.3 Web 日志挖掘

除了 Web 内容和 Web 链接结构，Web 挖掘的另一个重要任务是 Web 日志挖掘，它通过挖掘 Web 日志记录来发现用户访问 Web 页面的模式。通过分析和探究 Web 日志记录中的规律，可以识别电子商务的潜在客户，增强对最终用户的因特网信息服务的质量和交付，并改进 Web 服务器系统的性能和结构。目前研究较多的 Web 日志挖掘技术和工具可分为两大类：模式发现和模式分析。

在模式发现中，首先要解决的问题就是数据的预处理，它主要包括如下两个部分：数据清洗(data cleaning) 包括无关记录的剔除、判断是否有重要的访问没有被记录、用户的识别等问题；事务识别(transaction identification) 是指将页面访问序列划分为代表 Web 事务或用户会话的逻辑单元。在对事务进行了划分后，就可以根据具体的分析需求选择访问模式发现的技术，如路径分析、关联规则挖掘、时序模式以及聚类和分类技术。路径分析(path analysis) 可以用来发现 Web 站点中最经常被访问的路径，从而可以调整站点的结构。模式分析利用领域专家的知识以及其它一些可用的标准来分析这些模式，并过滤掉那些没有利用价值以及有偏差的模式，将发现的有价值的用户浏览模式以表格、饼图、曲线图、趋势图、直方图或者其它特殊表现形式显示出来。如果没有合适的技术和工具来辅助分析人员的理解，采用各种技术挖掘出来的模式将不能得到很好的利用。

3.3 Web 挖掘相关技术

模式发现是 Web 挖掘的主要阶段，采用的算法有统计分析、关联规则挖掘、路径分析、时序模式发现、聚类和分类算法等。

1)统计分析方法

它是抽取有关网站访问者知识的最常用方法。通过分析会话文件或事务数据库，可对诸如网页视图、浏览时间、导航路径长度等做出不同种类的描述性统计分析。很多 WebTraffic 分析工具还提供定期的报告，其中包含最大频繁访问页面、平均浏览时间、通过站点的路径的平均长度等统计信息。此类报告还能提供有限的低层次的错误分析，比如检测未授权入口点、找出最常见不变的 URL 等。尽管这种分析缺乏深度，但这类知识有助于改进系统性能、提高系统的安全性、便于站点修改，并能提供决策支持。

2)关联规则挖掘技术

关联规则主要关注事务内的关系。在网络用法挖掘中，关联规则挖掘就是挖掘出用户在一个访问期间(Session) 从服务器上访问的页面/ 文件之间的关系，找出在某次服务器会话中最经常一起出现的相关页面。挖掘发现的关联规则往往是指支持度超过预设阈值的一组访问网页，这些网页之间可能并不存在直接的引用(Reference) 关系。例如，用 Apriori 算法发现关联规则有可能发现访问包含网络搜索引擎网页的用户和访问有关 NASDAQ(纳斯达克) 市场网页的用户之间存在一定的联系。Apriori 算法是挖掘关联规则的常用技术，可从事务数据库中挖掘出最大频繁访问项集，该项集就是关联规则挖掘出来的用户访问模式。

3)序列模式挖掘技术

时序模式主要关注事务之间的关系。序列模式挖掘就是挖掘出交集之间有时间序列关系的模式，在 Web Log 中发现所有满足用户规定的最小支持度的大序列模式。在网站服务器日志中，用户的访问是以一段时间为单位记载的，经过数据精简和事件交易确认以后是一个间断的时间序列。利用对 Web 日志进行序列模式挖掘获得的知识，有助于网站管理人员：a. 改善网站的组织；b. 根据具有相同浏览模式的访问者所访问的内容来裁减用户与 Web 信息空间的交互，减少用户过滤信息的负担；c. 预测未来的访问模式，了解 Web 正在发生的变化。相关序列模式的存取分析，可对服务器的缓存、预取和交换参数进行调整。

4)分类技术

分类技术主要是根据用户群的特征挖掘用户群的访问特征(某些共同的特性)，这些特征可用于把数据项映射到预先定义好的类中去，即对新添加到数据库里的数据进行分类。在网络数据挖掘中，分类技术可以根据访问这些用户而得到的个人信息或共同访问模式得出访问某一服务器文件的用户特征。分类方法有很多种，常使用归纳学习算法，如决策树技术、贝叶斯分类法、K-邻近分类法等。

5)聚类技术

聚类技术是对符合某一访问规律特征的用户进行用户特征挖掘。在网络用法挖掘中，存在两种类型的聚类：使用聚类(用户聚类) 和网页聚类。用户聚类主要是把所有用户划分为若干组，具有相似特

(下转第 168 页)

(Transfer)可用:文件传输器(File transfer)和流传输器(Streaming transfer)。

文件传输器通过文件作为传输图像数据的媒介,而流传输器是指客户端使用 CD-SDK 提供的函数如 Open,Read,Write 来读取完成数据传输,最后使用 Close 来关闭流,这种方式不需要采集 Image Item。通过建立流的方式来传输数据对于需实时图像的场合犹为有用,还可以创建 Windows Device Independent Bitmap(DIB:Windows 设备无关的位图),并在应用程序中处理。

5 使用 CD-SDK 编程

CD-SDK 提供了丰富的 API 供开发人员使用,API 函数主要分为以下几类:基本函数,源设备控制函数,源设备选择和连接函数,设备管理函数,图像采集函数,图像管理函数,图像属性函数,拍摄控制函数。这些函数的原型都在 CDAPI. h,CDType. h,CDEvent. h,CDError. h,CDFncTyp. h 五个头文件中给予了定义。

使用 CD-SDK 的 API 开发程序的基本步骤可以从下面的典型例程中看出来(以建立流的方式传输数据为例):

```
#include "cdAPI. h" //cdAPI. h 包含了另外四个头文件
  CDStartSDK(&Ver, 0);// 调用 CDStartSDK,开始 CD-SDK 调用,分配系统资源
  CDEnumDeviceReset(1, &hEnumD);
  CDEnumDeviceNext(hEnumD, &SourceInfo);
CDEnumDeviceRelease(hEnumD);
  CDOpenSource( &SourceInfo, &hCam );// 选择并打开源设备,并获取源设备句柄,可以通过它设置相机的各种参数注册回调函数 */
    CDRegisterEventCallbackFunction ( hCam,
      &MyEventCallbackFunction,0,&hCallBack);
/ * 按下相机快门拍照 * /
CDRelease(m_hSource, FALSE, NULL, NULL, cdPROG_NO_REPORT, &NumData );
  CDCloseSource(hCam);//关闭源设备
  CDFinishSDK();//完成 CD-SDK 调用,释放系统资源
```

当按下快门的事件发生后,回调函数 MyEventCallbackFunction 启动一个线程来获取流数据,因此下面的关键代码应放在一个线程中执行:

```
cdStream * pStream;
for( NumData = m_NumData; NumData > 0; NumData--, i++ )
  {
  cdStgMedium MyMedium;
  CDCreateMemStream(iStartSize, iMinAllocSize, pStream);
  CDGetStreamInfo(pStream, pSize, ppMem);
  MyMedium. Type = cdMEMTYPE_STREAM;//表示内存流
  MyMedium. u. pStream = pStream;
  CDGetReleasedData ( m_hSource, ReleaseProgressFunc, NULL), cdPROG_REPORT_PERIODICALLY, &RelImgInfo, &MyMedium );
    //获取数据
  }
```

总结 通过以上对 CD-SDK 的研究,使用数码相机 SDK 能够实现对数码相机更精细的控制,可以开发出更专业的应用,本文给出的编程模型对开发其它种类的数码相机也极具参考价值。

参 考 文 献

1 邢萱. 数码相机高清晰图像采集的原理及软设计方法. 微处理机,2002(4)
2 唐朝京,鲜明,等. 平台上实现多媒体信息实时捕获的几种主要技术研究. 计算机应用研究,2003(5)
3 谢亚光,章琦,等. 基于 Microsoft DirectShow 的多媒体应用程序开发. 计算机应用研究,(4)
4 TWAIN Working Group. TWAIN Specification (Version 1. 9)
5 Canon Digital Camera Software Development Kit Software Developer's Guide

(上接第 129 页)

性(或浏览模式)的用户分在一组,这类知识对为用户提供个性化的服务特别有用。网页聚类可以找出具有相关内容的网页组,这对网上搜索引擎及提供上网帮助的应用特别有用。上述两类应用都能根据用户的询问或过去所需信息的历史生成静态或动态 HTML ,从而向用户推荐相关的超链接。目前,许多知名的门户网站如搜狐、新浪等均在用户浏览网页后给出相关链接服务,就是运用了这类技术。

4 Web 挖掘的发展方向

目前,在国内外 Web 挖掘的研究处于初级阶段,是前沿性的研究领域。在 Web 挖掘领域中面临下列诸多方面的挑战:

(1)在数据预处理方面,数据的收集机制与技术开发;

(2)研究和开发多种数据的智能集成系统,以期能提供完善的查询、优化和维护机制;

(3)高效、多能、自动导航的搜索引擎的研究;

(4)基于半结构化的 Web 数据的查询语言及查询系统的研究;

(5)现有挖掘方法与技术的改进;

(6)模式发现与分析智能化工具的研究与开发;

(7)新的数据模型与算法研究等。

结束语 Web 挖掘是当今世界上的热门研究领域,其研究有助于网络资源的开发利用,具有广阔的应用前景和巨大的现实意义。目前国内的 Web 挖掘尚处于学习、跟踪和探索阶段,许多问题有待于进一步的研究和深化。随着 XML 技术的发展,页面会蕴含更多的结构化和语义信息,这会使 Web 挖掘工作变得更有效,也更容易。

参 考 文 献

1 Facca F M. Mining Interesting Knowledge from Weblogs. Data & Knowledge Engineering. 2004,(11)
2 Kosala R Blockeel H. Web Mining Research: A Survey. ACM SIGKDD, July 2000
3 吉根林. Web 挖掘技术研究,计算机工程,2002(10)
4 黄晓斌. 网络信息挖掘. 电子工业出版社,2005