长沙通信职业技术学院学报
第5卷第4期 Journal of Changsha Telecommunications Vol.5 No.4
2006年12月 and Technology Vocational College Dec.2006

基于XML的WEB数据挖掘

蒋社想

（安徽理工大学计算机科学与技术系,安徽淮南 232001）

【摘 要】文章叙述了Web数据挖掘的概念、分类、技术等，重点讨论了基于XML语言的Web数据挖掘技术，解决了Internet上绝大多数非结构化甚至是无结构的、Web信息的组织结构性差而导致的Web数据挖掘困难的问题。

【关键词】数据挖掘；Web挖掘；XML

【中图分类号】TP311.13 【文献标识码】A 【文章编号】1671-9581(2006)-04-0030-04

WEB data mining based on XML

JIANG She-xiang

(Dept of Computer Science and Technology,Anhui University of Science and Technology,Huainan, Anhui, China 232001)

Abstract: This paper describes the concept, classification, technology of web-based data mining, and then discusses the web data mining based on XML. It solves the web data mining problem which is caused by the non-structure of the much Internet data and the poor structure of the information on the Web.

KeyWords:data mining; Web mining; XML

1 数据挖掘、Web数据挖掘的基本概念

1.1 数据挖掘 (Data mining)

根据W. J. Frawley和G. P. Shapiro等人的定义，数据挖掘 (DataMining，DM) 是指从大型数据库的数据中提取出人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在的有用信息[1]。数据挖掘的主要目的是提高市场决策能力，检测异常模式，在过去的经验基础上预言未来趋势等。

1.2 Web数据挖掘

Web数据挖掘是从数据挖掘发展而来的，是数据挖掘技术应用于Web信息的一个崭新领域。Web上的数据与传统数据库中的数据不同之处在于传统数据库都有一定的模型，可以根据数据模型来对具体的数据进行描述，而Web站点中的数据不存在统一的模型，各站点都是独自设计，并且站点中的数据是处于不停变化之中的，因此传统的数据挖掘技术并不适应Web挖掘。但因为Web有自身的结构，大体上站点的结构差异并不是特别大，所以可以认为Web数据是一种半结构化的数据，这是Web数据的另一个重要的特点[2]。

2 Web数据挖掘的分类

Web数据有三种类型：它们分别是HTML标

[收稿日期] 2006-09-16

[作者简介]蒋社想(1981-),男,安徽砀山人,安徽理工大学在读硕士，研究方向：计算机网络。

记的 Web 文档数据、Web 文档内的结构数据和用户访问日志数据，相应地，Web 数据挖掘可分为三类：内容挖掘（Web content mining）、结构挖掘（Web structure mining）和用户使用挖掘（Web usage mining）[3]。如图 1 所示。

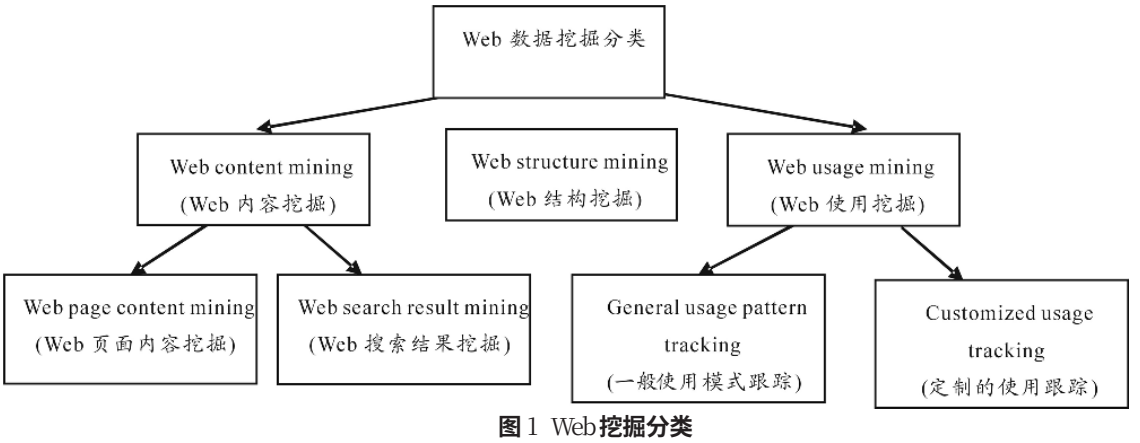

图 1　Web 挖掘分类

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有趣知识的一种过程，是一种基于网页内容元素对象的 Web 挖掘。这些元素对象既有文本和超过文本数据，也有图形、图像等多媒体数据；既有来自于数据库的结构化数据，也有用 HTML 标记或 XML 标记的半结构化数据和无结构的自由文本。Web 内容挖掘可以协助用户搜索信息，可以根据用户搜索条件过滤无用的信息。

2.2　Web 结构挖掘

Web 结构挖掘是从 Web 的组织结构和链接关系中推导有趣知识的过程。挖掘页面的结构和 Web 结构，可以用来指导对页面进行分类和聚类，找到权威页面，从而提高检索的性能。同时还可以用来指导页面采集工作，提高采集效率。Web 结构挖掘的目的是发现页面之间内在的有趣的联系，用户的访问模式与访问习惯，以便更好的组织页面和使用页面。

2.3　Web 使用挖掘

Web 使用挖掘是从服务器端记录的用户访问日志或从用户的浏览信息中抽取有趣知识的模式，通过分析这些数据可以帮助理解用户隐藏在数据中的行为模式，做出预测性分析，从而改进站点的结构或为用户提供个性化的服务。

3　Web 挖掘技术

目前应用在数据挖掘上的技术有很多，比较流行的有人工神经网络、遗传算法、决策树、近邻算法、规则推导等。

1）人工神经网络（Artificial Noural Network）：它是仿照生理神经网络结构的非线性预测模型，通过学习进行模式识别，可以完成分类、聚类、特征挖掘等数据挖掘任务。

2）遗传算法（Genetic Algorithm）：基于进化理论，并采用遗传变异、遗传组合和自然选择等设计方法的优化技术。将数据挖掘任务表达为一种搜索问题从而可以发挥遗传算法的优化搜索能力。

3）决策树（Decision Tree）：用树形结构来表示决策，这些决策通过对数据集的分类产生规则。这种方法一般用于分类规则的挖掘，典型的决策方法有分类回归树（CART）。

4）近邻算法（Neighbor Algorithm）：将数据集合中的每一个记录进行分类的方法，这种方法可以用作聚类、偏差分析等挖掘任务。

5）规则推导（Rule Induction）：从统计意义上对数据中的“if-then”规则进行寻找和推导，是统计学在数据挖掘中的应用，这种方法可以用作关联规则的挖掘。

4　XML 与 Web 数据挖掘

4.1　XML 介绍

XML（Extensible Markup Language）即“可扩展的标置语言”，它是由万维网协会（W3C）设计，特别为 Web 应用服务的 SGML（Standard General Markup Language）的一个重要分支。它是一种中介标示语言（Meta-markup Language），可提供描述结构化资料的格式。XML 描述的是数据内容和语义，而不像 HTML 那样描述显示样式和布局，XML 文档除了可以用文本编辑器浏览外，由于它有天然的层次结构，更为复杂的输出样式需要用到过滤器，XML 文档的 Web 输出如图 2 所示。

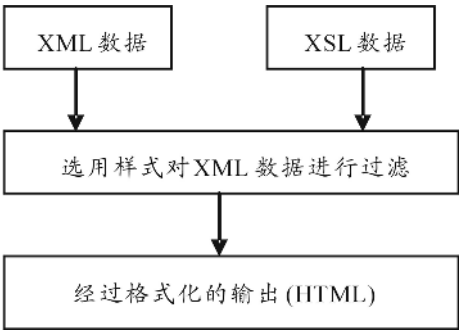

图2 XML格式化文档输出

4.2 XML的主要特点

1）简洁有效：XML是一个精简的SGML，它将SGML的丰富功能与HTML的易用性结合到Web应用中，它保留了SGML的可扩展功能，这使得XML从根本上区别于HTML。并且XML种还包括可扩展格式语言XSL（Extensible Style Language）和可扩展链接语言XLL（Extensible Linking Language）使得XML的显示和解析更加方便快捷。

2）开放的国际化标准：XML是W3C正式批准的，它完全可用于Web和工具的开发。XML具有标准的名域说明方法，支持文档对象模型标准、可扩展类型语言标准、可扩展链接语言标准和XML指针语言标准。使用XML可以在不同的计算机系统间交换信息，而且还可以跨越国界和超越不同文化疆界交换信息。

3）高效可扩充：XML支持复用文档片断，使用者可以发明和使用自己的标签，也可以与他人共享，可延伸性大。在XML中，可定义一组无限量的标准，可以有效地进行XML文件的扩充。

4.3 XML在Web数据挖掘中的应用

XML已经成为正式的规范，开发人员能够用XML的格式标记和交换数据。XML在三层架构上为数据处理提供了很好的方法。使用可升级的三层模型，XML可以从存在的数据中产生出来，使用XML结构化的数据可以从商业规范和表现形式中分离出来。XML在三层模型中的位置如图3所示。

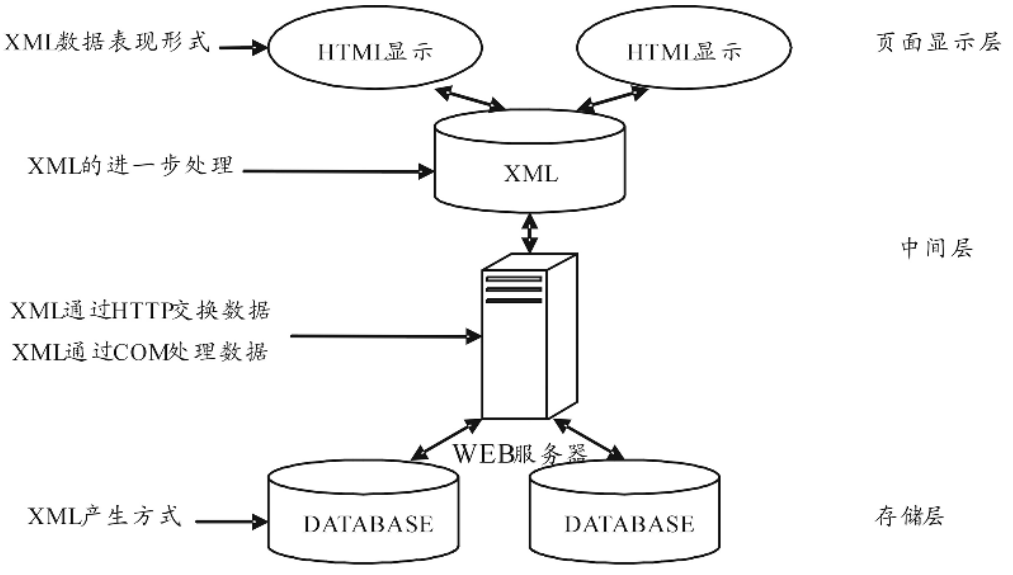

图3 XML在三层模型中的位置

XML给基于Web的应用软件赋予了强大的功能和灵活性。因此它给开发者和用户带来了许多好处。利用XML，Web设计人员不仅能创建文字和图形，而且还能构建文档类型定义的多层次、相互依存的系统、数据树、元数据、超链接结构和样式表。由于Web数据可被XML唯一地标识，这样可以给开发者和用户带来了更有意义的搜索。如果没有XML，搜索多样的不兼容的数据库实际上是不可能的，这是因为每个数据库描述数据的格式几乎都不相同，但XML能够使不同来源的、结构化的数据结合在一起，软件代理商可以在中间层的服务器上对从后端数据库和其它应用处来的数据进行集成。然后，数据就能被发送到客户或其他服务器做进一步的集合、处理和分发。

5 结论

面向Web的数据挖掘是一项复杂的技术，XML的出现为解决Web数据挖掘的难题带来了机会。由于XML能够使不同来源的结构化的数据很容易地结合在一起，因而使搜索多样的不兼容的数据库成为可能，从而为解决Web数据挖掘难题带来了希望。XML的扩展性和灵活性允许XML描述不同种类应用软件中的数据，从而能描述搜集的

Web 页中的数据记录。同时，由于基于 XML 的数据是自我描述的，数据不需要有内部描述就能被交换和处理。相信随着 XML 作为在 Web 上交换数据的一种标准方式的出现，面向 Web 的数据挖掘将会变得非常轻松。

[参考文献]

[1] Usama M. Fayyad，Gregory Piatetsky Shapiro et al. Advances in Knowledge Discovery and Data Mining [M] . California. AAAI / MITPress，1996.

[2] 马保国，侯存军，王文丰等. Web数据挖掘技术及应用 [J] .计算机与数字工程，2006, (6) :20- 22.

[3] 陈文伟，黄金才. 数据仓库与数据挖掘技术 [M] . 北京: 北京大学出版社，2002.

[4] [美] MarkGraves，尹志军等译. XML数据库设计 [M] .北京:机械工业出版社，2002.

[5] 陈京民. 数据仓库与数据挖掘技术 [M] . 北京：电子工业出版社，2002.

[6] 刘　云，刘东苏. 基于Web的数据仓库与数据挖掘技术 [J] . 情报理论与实践，2001，24 (4) :289- 290，320.