计算机科学 2002Vol. 29№. 10

XML 和数据挖掘的关系探讨

Discussing the Relationship between XML and Data Mining

朱美琳　王崇骏　陈世福

（南京大学计算机软件新技术国家重点实验室　南京大学银河软件开发中心　南京 210093）

Abstract The information is exploding today. How to discover new knowledge from large information database is an important question in the information times. This paper introduces the survey of XML and data mining at first, and then discusses the combination of XML and data mining on the KD and PM domain. Otherwise, because XML is widely applied on Web site, this paper emphasizes the relation of XML and data mining in Web.

Keywords XML, Data mining, Predictive model, Web mining

1 引言

由于 Internet 的发展，在信息膨胀的今天，造成数据丰富而知识缺乏的现状，如何有效地、快速地从海量的数据里面提取有用的信息，如何更方便地传递、交流、获取信息，成为当前高科技领域的关注热点。XML 以及数据挖掘都是近几年兴起的新技术，在很多领域发挥了重大作用，而两者的结合能够迅速地应用到各个领域，例如：Web 服务、电子商务、图书馆、医学等。XML 促进了信息间的交流，为数据挖掘提供了更广阔的结构化的信息平台，数据挖掘从大量信息中提取有用的知识，从而提高学术上的可研究性，以及商业上的决策性。

本文首先介绍了 XML 和数据挖掘的概况，然后，从预言模型、知识发现等方面阐述了 XML 与数据挖掘的关系，另外，由于 XML 是 SGML 的一个子集，广泛应用于 Web 网站上，因此本文还着重强调了在 Web 上的数据挖掘。

2 XML 概述

XML[1]是 W3C 于 1998 年 2 月设计的，是 SGML 的一个子集，是针对 SGML 和 HTML 的局限性而创立的（尤其是在 Internet 上）。XML 是一种元标记语言，它提供了描述结构化资料的格式，作为一种在 Web 上使用数据和元数据的语言正在迅速发展。

与 HTML 相比，XML 增加了结构和语义信息，从而将杂乱的信息进行初步归类，另外，XML 中的标记 tag 可以根据用户要求来定义标记的含义，属性名也可以包含描述法，突破了 HTML 固有标记集合的约束，使文件的内容更丰富、更复杂、更灵活，组成一个完整的信息系统。

除了 XML 标记和标记的内容，XML 文档的另一主要内容是 DTD（Document Type Definition），用来定义 XML 标记及其相互关系，DTD 规定文档的逻辑结构和语法，也定义了页面的元素、元素属性以及两者的关系。DTD 使得 XML 走向国际标准化，通过使用共同的 DTD，不同的平台可以无障碍地进行交换。

此外，XSL（Extensible Style Language）定义了 DTD 的样式表，能使 Web 浏览器改变文档的表示法，给 Web 提供了高级的布局特性，消除 HTML 在 Internet 上的一些限制，展现更多的文档样式，提供了结构化的数据表示方式，从而使得数据挖掘工作更加深入。XML 文档间的超链接功能由独立的 XLL（Extensible Link Language）来支持，可以多方向链接，使得信息传输和查询更加灵活方便。

XML 比较简单，容易学习，采用 Unicode 编码系统，不同语言文本可以在同一个文档中混用，使得 XML 成为数据表示的一种开放标准，独立于机器平台、提供商和编程语言，从而为不同的系统、不同的数据库、不同的语言之间搭起沟通的桥梁。

XML 不仅仅应用于 Web 站点，它在电子商务、数据库、知识管理、数据交流与共享、自然语言转换等方面都得到广泛的应用，另外还应用在科技领域，MathXML（数学 XML）、CML（化学 XML）、AML（天文 XML）[2]等语言，使得专业语言的表示不再困难。

3 数据挖掘概述

20 世纪 80 年代末，数据挖掘作为一种新的知识获取技术从人工智能的机器学习中分支出来，在商业分析中引起很大的关注，从而得到飞速发展。数据挖掘是从大型的数据存储库（数据库、数据仓库或其他信息存储库）中发现有价值的知识的过程[3,4]，广泛应用于货篮数据（basket data）分析、金融风险预测、产品产量、质量分析、分子生物学、基因工程研究、信息搜索和分类、工程诊断等方面。

数据挖掘一般包括下面几个步骤：数据准备、数据选择、数据预处理、数据缩减、KDD 目标确定、数据挖掘算法确定、数据挖掘、模式解释以及知识评价等。对数据进行分析的时候，主要用到统计法、人工神经网络、决策树、遗传算法、近邻算法、模糊逻辑、粗糙集（rough set）、规则推导等方法和技术。现在，数据挖掘技术和其他技术相结合，在更广阔的领域里得到发展，例如可视化的数据挖掘、基于并行的数据挖掘、分布式数据库的挖掘等等，另外，随着 Internet 的发展，基于 Web 的数据挖掘研究逐渐成为新的热门话题。

尽管数据挖掘技术的研究与应用已取得了很大的成果，但是它也面临着许多问题，例如：

（1）由于数据挖掘的方法和模型多种多样，彼此又互相孤立，联系很少，没有统一的约定对模型进行描述和定义，因此造成各数据挖掘系统之间的封闭现象；

（2）数据挖掘系统仅提供孤立的知识发现功能，难于嵌入大型应用，也很难和大型数据库紧密结合，虽然现在数据挖掘原语（如 DMQL）的发展，将有助于解决这类问题，数据挖掘原语允许用户在挖掘过程中从不同的角度或深度与数据挖掘系统进行交互式的通信，但是如何更有效地进行知识发现还

需进一步的研究。

如今，数据挖掘行业是高度分散的，公司和研究机构独立开发各自的数据挖掘系统和平台，没有形成开放性的标准；同时数据挖掘技术本身也是综合多学科知识，跨度非常大，这两点是上述问题存在的根本原因。随着 XML 的出现，对于上述问题的解决，提供了新的思路，由于 XML 结构化、可扩展性的特点，为数据挖掘在内部环境以及外部沟通方面，创造了统一的平台。

从数据分析的角度，数据挖掘文化可以分为两种：知识发现文化和预言挖掘文化，知识发现(KD)文化是利用数据挖掘算法以简明概要的方式提取正确的、新的、有用的知识。预言挖掘(PM)文化是利用数据挖掘算法自动生成一个或一组预言模型，并试图预测新数据集。这两种文化的共同点在于输入的都是学习集，不同的是一个输出规则，一个输出预言模型，在应用上各有各的优点，服务于不同的模型。本文将分别对这两种文化阐述与 XML 的关系。

4 XML 与预言模型

XML 在预言模型中的主要应用是 PMML(Predictive Model Markup Language)[5]，它是对数据挖掘预言模型进行描述和定义的语言，使得数据挖掘系统在预言模型定义和描述方面有标准可以遵循，各系统之间可以共享模型，从而解决目前各数据挖掘系统之间封闭性的问题，又可以在其它应用系统中间嵌入数据挖掘模型，解决孤立的知识发现问题。

PMML 主要目的是允许应用程序和联机分析处理(OLAP)工具能从数据挖掘系统获得模型，而不用独自开发数据挖掘模块。另一个目的是能够收集使用大量潜在的模型，并且统一管理各种模型的集合。这些能力在商业应用领域是有效的配置分析模型的基础。

PMML 是一种基于 XML 的语言，它为各个公司定义预言模型和在不同的应用程序之间共享模型提供了一种快速并且简单的方式。通过使用标准的 XML 解析器对 PMML 进行解析，应用程序能够决定模型输入和输出的数据类型，模型详细的格式，并且按照标准的数据挖掘术语来解释模型的结果。

PMML 是基于 XML 的标记语言，非常适合部分学习、元学习、分布学习和其他相关领域[5]，PMML 可以看成是预言模型的互换格式。

使用 PMML 进行模型定义由以下几部分组成：

(1)头文件(a header)

(2)数据模式(a data schema)

(3)数据挖掘模式(a data mining schema)

(4)预言模型模式(a predictive model schema)

(5)预言模型定义(definitions for predictive models)

(6)全体模型定义(definitions for ensembles of models)

(7)选择和组合模型及全体模型的规则(rules for selecting and combining models and ensembles of models)

(8)异常处理的规则(rules for exception handling)

PMML 能够很好地处理预言模型和预言模型的集合，所谓预言模型的集合是指在对一个问题进行挖掘分析的时候，通常不仅仅只产生一个预言模型，还可能会产生一组预言模型。对于预言模型的集合，我们可以通过运用选择法则或平均法则，对模型集合中的模型进行组合，来产生唯一的预言结果。

模型集合在很多地方被广泛应用，主要原因有：

(1)模型集合能生成更精确的预言模型，若干预言模型的组合避免了一个预言模型的偏颇；

(2)处理海量数据时可以先把数据库进行分割，每个小的数据库可以分别被挖掘，并行地生成一个模型集合，然后再将结果组合；

(3)模型集合可以比较容易地处理分布数据，分布数据先被分别分析产生一个模型集合，再从这个模型集合里产生一个新的预言模型。

我们将第二种情况称为部分学习，第三种情况称为分布学习，PMML 的使用，使这两种情况中的模型变得统一，从而易于组合，得出唯一的结果。

在 PMML 中头文件可以用来描述学习集、算法、数据挖掘应用相关信息，使得预言模型显得更加清晰，另外，PMML 中的一些属性可以发生转变，以适应数据删改的需要，从这里我们可以看到 PMML 作为预言模型的互换格式的灵活性和通用性，而好的预言模型的互换格式以前并没有得到重视，在很多数据挖掘应用中，我们已经感觉到支持模型集合多种属性的机制的好处，开放的、弹性的互换格式的发展，正是数据挖掘应用的迫切需要，XML 在预言模型上的应用，无疑是一条便捷之路。

5 XML、知识发现和 Web 挖掘

上一节阐述的是在数据挖掘中应用 XML，将被发现的知识用 XML 表示，那么，这一节讨论的是在 XML 文档中进行数据挖掘，这部分内容主要集中在知识发现上，即 XML 文档的知识发现。

这方面的一个典型系统是 Scientio 公司开发的 XML Miner 系统[6]，该系统是专门针对 XML 代码而进行数据挖掘的，首先读取 XML 格式的数据，然后利用模糊逻辑法则进行数据挖掘，写出 XML 下的规则集，这些规则是基于“元规则”的，可以转换成“if…then”的形式，也支持用户自己定义的新规则和符号表示。

XML 文档应用在不同的领域，因此对其进行知识发现也有很多不同的方法，现在 XML 应用最广的是在 Web 和图书馆资料整理上，下面将介绍在 Web 上的数据挖掘(Web Mining)，并简单讨论与 XML 的关系，毋庸置疑，随着数据挖掘和 XML 在 Web 上的应用普及，它们之间的联系会越来越紧密。

现在，Web 已经变为全球最大的信息源，Web 拥有巨量的、多变的、动态的，而且大多是非结构化的数据，对于不同的用户、Web 服务商、商业分析者来说，所需要处理的信息又各不相同，但他们都希望利用某些工具和技术，从 Web 上找到需要的信息以解决实际问题，因此，Web Mining 在最近几年变得相当活跃而普遍了。

Web Mining 这个概念，是由 Oren Etzioni 于 1996 年提出的[7]，Web Mining 是应用数据挖掘技术自动地从 World Wide Web 的文档和服务器上发现和提取有用的信息的过程。尽管 Web Mining 用到很多数据挖掘的技术，但 Web Mining 又不能和普通的数据挖掘等同起来，因为 Web Mining 具有如下特点：

(1)算法必须有很高的效率：由于数据量非常庞大，而且每天都在迅速增长和更新，从如此巨量的数据中有效地提取有用的信息要求数据采掘速度必须很快；

(2)有强大的并行性：分布在网络上各个站点的资源通过 Internet 连成一个大型分布的数据库，数据的巨大规模和广

泛分布要求并行性很高；

(3)具有动态性：Internet 中数据更新非常迅速，有些信息可能很快过时，针对当前状态的信息能快速更新知识，提供准确的决策支持要求数据挖掘的动态性；

(4)有效地组织和管理数据：目前数据挖掘多应用于关系和面向对象数据库，它们有完美的结构，按照预先定义的模式进行组织、存储和存取，而 Internet 的信息往往具有半结构化或非结构化特性，比如大量的文本，难以映射到一个固定的模式，使传统数据模型和数据库系统难以支持 Internet 上的信息管理。XML 结构化的特点，使得 Web Mining 变得相对容易一些。

现在，公认的 Web Mining 分类是 Web 内容挖掘(Web Content Mining)、Web 结构挖掘(Web Structure Mining)、Web 使用挖掘(Web Usage Mining)，下面将对这三种挖掘类型进行详细描述。

5.1 Web 内容挖掘

Web 内容挖掘是自动对在线信息进行查询，对数据的内容进行挖掘[8]。从挖掘的内容来说，Web 内容挖掘和普通的数据挖掘有很多相似之处，普通的数据挖掘工作在 Web 内容挖掘上都可以应用。

Web 数据通常有多种类型，如文本、图片、声音、视频、元数据和超链接等，大多数都是非结构化数据，对这些文档的内容进行查询和分析，无疑比较复杂，如果 Web 文档是用 XML 来编写的，则不仅提高了文档的结构化程度，也为建立分析所需的内容数据库提供了方便。

由计算机科学公司(Computer Sciences Corporation)和国防科技组织(the Defence Science and Technology Organization)联合开发的智能健康系统(HINTS)就是基于 XML 的知识发现系统[9]。该系统首先自动对多种格式的 Web 数据进行元数据的提取，元数据包括关键字、摘要等内容，然后用 XML 格式来存放元数据。这里使用 XML 的 RDF(Resource Definition Framework)提供的一般存储机制，便于复杂的文档内容查询。RDF 着重于计算机间的自动化交流，能表达资源与资源间的相互关系，因此，可以方便地动态浏览文档和标题。HINYS 系统对 XML 格式的数据进行查询，在用户图形界面显示出来。

5.2 Web 结构挖掘

Web 结构挖掘用于总结 Web 站点和 Web 网页的结构特征，Web 内容挖掘主要针对的是内部文档，而 Web 结构挖掘主要针对的是外部文档的超链接结构。根据超链接的拓扑结构，Web 结构挖掘可以对网页进行分类，并且取得不同网站的相似信息和其他关系[10]。

另外，Web 结构挖掘也能发现 Web 文档自身的结构，这种结构挖掘能更有助于用户的浏览，也利于对网页进行比较和系统化。通过提供一个指导性的方案，从而达到更方便地访问信息的目的。

XML 本身可以使 Web 文档结构化，因此，对结构挖掘来说，用 XML 开发的网站，将节省很多数据的预处理工作。

对 Web 进行结构挖掘，可以得到以下信息：

(1)同一网站里不同网页链接的频率；

(2)同一网站里同一网页内部链接的频率；

(3)不同网站间链接的频率等。

通常，如果一个网页直接链接到另一个网页，或者网页是相邻的，我们就可以发现这些网页间的关系，这些网页可能具有相同的结构、类似的内容，或者位于相同的 Web 服务器，XML 可以使之很容易地形成网页之间的比较，找出相同点和差异。

Web 结构挖掘的另一个功能是发现专业网站中的层次关系和链接网络关系，获取信息流的流向，从而使查询更方便容易，XML 文档的拓扑结构比较明显，通过 Xlink 进行超链接，因此能迅速地发现文档间的关系。

因为 Web 结构挖掘和 Web 内容挖掘都需要处理有链接的 Web 文档，都需要使用 Web 上的原始数据来进行挖掘，所以这两种数据挖掘通常联系在一起应用。在一些文献中，就把它们通称为内容挖掘[11]。

5.3 Web 使用挖掘(WUM)

当用户在网站上浏览的时候，Web 可以得到诸如日志的一些二手数据，Web 使用挖掘就是在这些数据中发现有用的信息，从而预测用户的网上行为[12]，是现在三种 Web 挖掘中受关注最多的研究方向。

通过 WUM，Web 服务商可以根据实际用户的浏览情况，调整网站的网页的连接结构和内容，更好地服务用户，也可以从 proxy 的访问信息中分析用户的访问模式，从而预测用户的网页访问，提高 Web Caching 的性能；另外，对于从事电子商务的网站来说，可捕捉到大量的用户进行商务活动的细节，提供更加深入分析的可能；现在个性化网站逐渐兴起，通过 WUM，可以发现用户的喜好，动态地为用户定制观看的内容或提供浏览建议，使得网站更加生动而独特。

WUM 的过程包括：

(1)数据收集。收集由服务器、用户、代理商提供的包括页面的实际内容、结构、用户信息、使用事件在内的 Web 数据；

(2)预处理。包括数据清理、用户身份识别、网页浏览识别、事件识别等处理；

(3)模式发现。主要用到以下技术：统计分析主要用于改进系统的性能和设计方案，统计的内容有最经常访问的网页、每页平均访问时间、网站中每条路径的平均长度等，关联规则可以寻找出经常频繁访问的网页组，可用于修改网站的设计或提前缓冲页面，改进系统的性能，聚类分析可用于市场分割和个人内容定制，序列模式可用于用户的浏览趋势分析，可靠度建模不仅提供了理论化分析用户行为的框架，同时也可以用来提高网上产品销量；

(4)模式分析。目的是根据实际应用，通过用户的选择和观察，把发现的规则、模式和统计值转换为知识。

现已开发的原型系统和商用系统有：IBM 的 SpeedTracer，提供基于用户、路径和网页组的三种统计类型的报告，Simon Fraser 大学研究开发的 WebLogMiner，经过清理的 log 数据以数据立方体的形式存储在数据库中，提供包括序列模式和关联规则在内的多种数据挖掘的方法，以及由德国柏林洪保德大学商学院研究开发的 Web Utilization Miner，主要提供序列模式的发现，提供 MINT 查询语言，提供树结构的序列模式显示方式。

WUM 仍然存在一些问题，例如：服务器的日志提供的可用信息太少，动态页面的大量使用使得分析日志更为困难，Session 的分析一直是个难点，一些数据并没有被记录下来等等，因此，利用 WUM 对网页的使用进行分析，还需要一段很长的研究过程。

XML 和 WUM 的结合点目前并不是很多，最有可能的

(下转第 66 页)

表2 新算法的N皇后问题求解，数据格式：演化时间(秒)/评价函数运行次数

皇后数目	实验编号 1	2	3	4	5	6	7	8	9	10	平均值
100	0.011/95	0.024/195	0.009/55	0.010/45	0.014/115	0.013/70	0.009/45	0.006/50	0.009/65	0.008/200	0.011/93.5
1000	0.122/75	0.235/290	0.151/135	0.390/590	0.197/220	0.198/235	0.181/165	0.210/190	0.152/175	0.265/335	0.210/241
10000	9.822/195	9.322/105	9.342/125	9.767/85	10.277/220	10.619/155	10.527/155	10.596/290	12.077/210	9.300/145	10.165/168.5

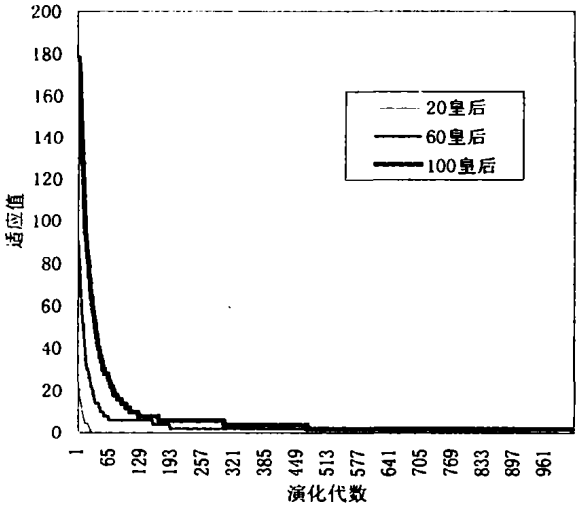

图2 一般演化算法求解N皇后问题的收敛曲线

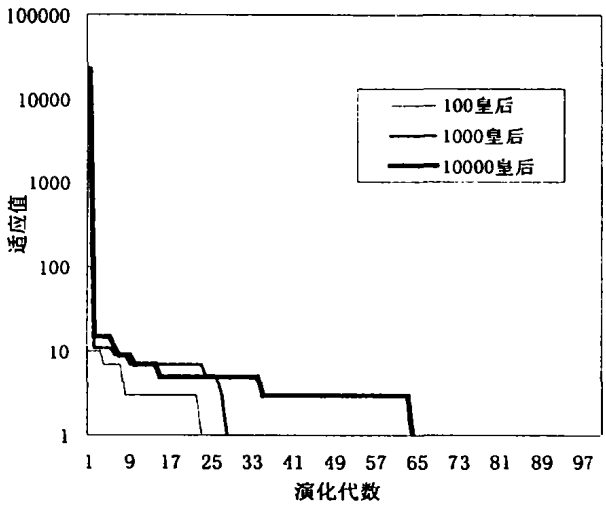

图3 新算法求解N皇后问题的收敛曲线

参考文献

1 Holland J H. Adaptation in Natural Artificial System. The University of Michigan Press, 1975

2 潘正君，康立山，陈毓屏. 演化计算. 清华大学出版社. 广西科学技术出版社，1998

3 Sosic R, Gu J. Efficient Local Search with Conflict Minimization: A Case Study of the N-Queens Problem. IEEE Transactions on Knowledge and Data Engineering, 1994, 6(5): 661～668

(上接第58页)

发展是在电子商务上，这是因为WUM现在主要应用于电子商务领域，挖掘网上用户信息，而XML在电子商务上的应用也逐渐展开。例如在个性化网站方面，WUM根据日志、Cookie等信息，利用聚类分析，发现喜好近似的用户类和同时被访问的网页类，从而根据用户特征设计用户感兴趣的网页内容和结构，而使用XML通过中间层技术则可以方便地将设计好的网页内容实时地、动态地展现给用户[13]。

结束语 当前数据挖掘和XML研究正方兴未艾，在今后的若干年将形成更大的高潮，由于Internet用户迅速增加，为了快速高效地找到网上的知识，研究在Web上的数据挖掘以及XML格式化，加强对非结构化数据如文本数据、图形图像数据、多媒体数据的挖掘，将是近期数据挖掘和XML相结合的重要课题。

参考文献

1 Extensible Markup Language. http://www.w3.org/XML/

2 Guillaume D, Murtagh F. Clustering of XML documents. Computer Physics Communications, 2000, 127: 215～227

3 Agrawal R, Imielinski T, Swani A. Database Ming: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering, Special issue on learning and Discovery in Knowledge-Based Databases, 1993, 5(6): 914～925

4 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. 机械工业出版社，2001

5 Grossman R, et al. The Management and Mining of Multiple Predictive Models Using the Predictive Modeling Markup Language. Information and Software Technology, 1999, 41: 589～895

6 XML Miner. www.metadatamining.com

7 Etainoni O. The World Wide Web: Quagmire or Gold Mine. Communications of the ACM, 1996, 39(11): 65～68

8 Madria S K, et al. Research issues in Web data mining. In: Proc. of Data Warehousing and Kno wledge Discovery, First Intl. Conf. DaWaK '99, 1999. 303～312

9 Lang K, Burnett M. XML, Metadata and Efficient Knowledge Discovery. Knowledge-Based Systems, 2000, 13: 321～331

10 Madria S K, et al. Research issues in web data mining. In: Proc. of Data Warehousing and Knowledge Discovery, First Intl. Conf. DaWaK'99, 1999. 301～312

11 Cooley R, Mobasher B, Srivastava J. Web Mining: Information and Pattern Discovery on the World Wide Web (A Survey Paper) (1997). In: Proc. of the 9th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI'97), Nov. 1997

12 Srivastava J, et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000). SIGKDD Explorations, 2000, 1(2)

13 Goldfarb C F, Prescod P. XML用户手册. 人民邮电出版社，2000