·计算机及网络技术 ·

基于 Web的可视化数据挖掘工具综述*

秦甲拓**, 张 浚
(电子科技大学 成都 610054)

摘要: 在互联网存储的信息中, 对于含有有效信息的数据挖掘工作具有极高的价值, 而数据可视化工具又为挖掘工作以及对于信息的分析提供了更直观方便的方法。文章介绍了数据挖掘、数据可视化、Web挖掘的基本概念、基本方法及流行技术; 比较了常见的可视化数据挖掘工具, 并且对数据挖掘技术的发展做出了适当的展望。

关键词: 数据挖掘; 数据可视化; 可视化数据挖掘; Web数据挖掘; 知识发现
中图分类号: TP311·13 **文献标识码**: A **文章编号**: 1672-4550(2006)07-0065-04

On Visual Data Mining Tools Based on Web

QIN Jia-tuo, ZHANG Jun
(University of Electronic Science and Technology of China Chengdu 610054)

Abstract: The work of mining the data that contains valid information on the internet is highly regarded for its value. And the data visualization tools provide visual and convenient methods to get data and process information. In this paper, the fundamental concepts, methods and popular technology of data mining, data visualization and Web mining are summarized; usual tools of visual data mining are compared; and the prospect of data mining technology is also presented here.

Key words: data mining; data visualization; visual data mining; Web mining; knowledge discovery

1 引 言

在 20世纪后期, 计算产生大量的数据, 其规模巨大。在商业活动中, 每一个记录所包含的数据, 其价值则取决于对其理解的程度。利用数据挖掘工具, 可使这些数据具有应有的竞争价值。

互联网络的规模在不断增长, 早在 2004年底, Google就宣称索引了的有效网页数量突破了 80亿。将近百亿的网页使互联网本身成为了一个庞大的数据库。Web挖掘就是挖掘互联网当中的数据, 使这些有效存储的数据发挥其应有的价值。

当人类被数据包围时, 在生活当中随处可见的数据可视化则有利于数据的理解。无论是金融市场当中的变化趋势, 还是气象预报的地形图, 数据可视化工具可以将那些复杂的数据直观地表示出来, 从而使得数据更容易被解释, 并且提升用户的洞察力。直观的二维或者三维的数据可视化, 可以更方便地将数据集当中真正有价值的信息挖掘出来, 帮助用户发现新的模式和趋势, 并将发掘的结果与决策人员沟通和交流。有效地将 Web数据挖掘与数据可视化结合, 可以更好地利用互联网络当中海量的数据, 并将其转化为商业的赢利和投资回报率。

2 Web与可视化数据挖掘

2.1 数据挖掘

数据挖掘就是对观测到的数据集 (经常是很庞

* [收稿日期] 2006-07-25
** [作者简介] 秦甲拓 (1985-),男, 本科生, 就读于计算机科学与工程学院。

大的)进行分析。

数据挖掘通过各种不同的算法来实现不同的任务目标。其算法由模型、偏好和搜索三部分组成。其中算法的目的就是找到适合于数据的模型。数据挖掘模型在本质上分为预测性模型和描述性模型两类：预测性模型利用从不同数据中发现已知的结果，从而对数据进行预测；描述性模型则提供了一种探索被分析数据的性质的方法，从而对数据中的模式或关系进行辨识。另外，必须使用一些标准来进行模型选择，而且所有的方法都要使用一些技术对数据进行搜索。

数据挖掘的八个基本任务包括：分类、回归、时间序列分析、预测、聚类、汇总、关联规则及序列发现等。

数据库中的知识发现（Knowledge Discovery in Databases, KDD)是一个包含了很多步骤的过程，数据挖掘仅仅是其中一个基本步骤。数据库中的知识发现（KDD）是从数据中发现有用的信息和模式的过程。数据挖掘则是使用算法来抽取信息和模式，是 KDD过程的一个步骤。

数据挖掘过程当中，衡量其有效性和有用性是比较困难的，在商业应用上，投资回报率（Return On Investment, ROI）用来衡量由于使用数据挖掘而增加的收入与使用数据挖掘的成本之比，是一个比较优秀的度量指标。

2.2 可视化数据挖掘

数据可视化是将巨量数据转化为有意义的图像的过程。数据可视化工具被用来创建业务数据集的二维或三维的图形。可视化数据挖掘工具则帮助用户创建可视化的数据挖掘模型，并且利用这些模型发现业务数据集中存在的模式，从而辅助决策支持以及预测新的商机。

数据可视化工具包括多维可视化和专门的层次及地形可视化两类。其中，多维数据可视化工具能够让用户直观地在空间坐标系上比较一个数据维和其他数据维之间的关系。常用的可视化图表类型有：柱形图和条形图、分布图和直方图、箱式图、折线图、散点图和饼图。层次、地形和其他特殊的数据可视化工具，是为了探索和提高对业务数据集自身结构的理解，包括树型可视化和地图可视化。

2.3 Web数据挖掘

Web数据挖掘是挖掘跟万维网有关的数据，既可以是网页包含的数据，也可以是 Web操作产生的数据。Web挖掘活动的分类如图 1所示：

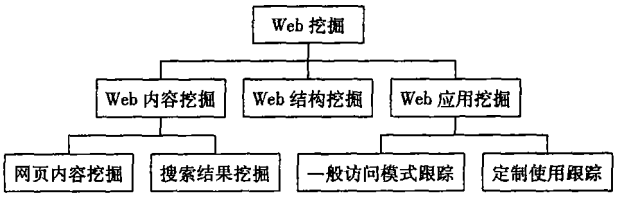

图 1 Web挖掘活动的分类

2.3.1 Web结构挖掘

Web结构挖掘是从网页的实际组织结构中提取网络的拓扑信息，即网页之间的链接信息。通过对于链接的分析以及对于引用的计算，可以将网页分类，并对网页建立相似性度量。其中比较流行的技术有：

（1）PageRank

PageRank算法用于提高搜索引擎的搜索效果和效率，度量网页的重要性以及为传统搜索引擎使用关键字搜索的结果进行优先级排序。网页的 PageRank指通过指向它的网页计算，这实际上是基于网页后向链接的一种度量。后向链接不是该网页链出去的链接，而是指向该网页的链接。PageRank值的计算不光考虑后向链接的数目，而是对来自重要网页的链接给于更高的权值。给定网页 p，用 Bp表示指向 p的网页集合，Fp表示由 p指向其他网页的链接集合。网页 p的 PageRank值定义为：

$$PR(p) = c\sum_{d\in B_p}\frac{PR(q)}{N_q}$$

这里 $N_q = |F_q|$。常数 c介于 0和 1之间，用于归一化。

（2）Clever

IBM公司开发的 Clever系统，其目标是发现权威网页和中心网页。权威网页支队请求的信息是“最好的源”的网页，含有指向权威网页链接的是中心网页。Clever系统用加权技术识别权威网页和中心网页。搜索的目标可以看作是寻找最好的权威网页和中心网页。该系统使用的是由 Kleinberg提出的 HITS（Hyperlink - Induced Topic Search）算法。

2.3.2 Web应用挖掘

Web应用挖掘用来提取关于客户如何运用浏览器浏览和使用这些链接的信息。其研究对象就是 Web使用数据或者 Web日志。Web应用挖掘实际包

含预处理、模式发现和模式分析三种类型的工作。

用于 Web挖掘的有效的最低级的数据就是点击流。每次客户点击网站时生成的数据 (也就是点击流数据)可以被收集、存储、净化和管理，以用于进一步的分析。通过分析客户的点击流数据，公司可以很快了解到更多的客户行为 ——包括他们喜欢或不喜欢的行为，从而提高其 Web投资的有效性。

点击流的分析始于网络日志。当服务器获得大量的点击之后，还有许多过滤和整理的工作要做，包括：过滤、反蜘蛛化、客户验证、会话、路径补全。因此，应用挖掘可以帮助确定使用模式并且对使用提出改进的建议，从而提高网站的可用性。

2.3.3 Web内容挖掘

Web内容挖掘通过研究网页本身的内容以及 Web搜索的结果来提取文字、图片或其他组成网页内容成分的信息。可将其看作对基本搜索引擎所完成工作的扩展。Web挖掘的一种分类体系把 Web内容挖掘分为基于代理的挖掘和基于数据库的挖掘。基于代理的挖掘有软件系统 (代理)负责内容挖掘；基于数据库的内容挖掘则把互联网上的数据看作是属于数据库的数据。目前比较流行的 Web内容挖掘技术有：

(1) 爬虫

爬虫 (又称机器人、蜘蛛)是指遍历网页超文本结构的程序。由于互联网规模巨大，产生了专用爬虫。专用爬虫只访问与特定主题相关的网页。与传统的爬虫相比，使用许多专用爬虫能够覆盖更多的网页，并且随着 Web规模的增长有更好的扩展性。专用爬虫结构包括超文本分类器、提取器和爬虫三个主要组成部分。专用爬虫性能的目标是高的准确率，或者称为收获率。

(2) Harvest系统

Harvest系统使用缓存、索引和爬虫技术，实际上是一组工具，用于从众多来源收集信息。Harvest系统的设计集中在搜集器和代理的使用，其索引和代理是面向特定主题的，用于搜索网页文本非常有效。

(3) 虚拟 Web视图

对于互联网上大量无结构数据的处理，可以在网页数据上建立规模宏大并且分布式的多层数据库 (Multiple Layered DataBase, MLDB)。MLDB为互联网提供了一种抽象的精简视图，成为虚拟 Web视图 (Virtual Web View, VWV)。

(4) 个性化挖掘

使用个性化挖掘，网页访问或者网页的内容可以被更改从而更好地适应用户的需求。这种设计为每个用户创建独特的网页或者根据用户的要求决定搜索哪些网页。

3 常见 Web与可视化数据挖掘工具

通常情况下，数据挖掘工具一次一般只能解决一个问题或者任务，例如分类、估计、预测、关联分析、聚类和细分。通过解决不同类型的任务，可以将数据挖掘工具分成两大类：有监督和无监督的学习。

3.1 有监督的学习工具

有监督的学习工具包括：(1) 决策树和规则集模型；(2) 用于分类的神经网络模型；(3) 线性回归模型；(4) Logistic回归。

3.2 无监督的学习工具

无监督学习把一组记录的集合作为输入，然后试图从中发现一些模式。各个工具之间的差别在于发现的模式和搜索的过程，主要包括以下三种：(1) 关联规则；(2) 聚类；(3) SOM (Kohonen自组织映射)。

4 常见数据挖掘工具对比

不同的数据挖掘工具能够解决不同的数据挖掘任务，如表 1所示。

同样，不同的数据挖掘工具具有不同的优点和缺点，如表 2所示。

表 1 数据挖掘工具功能对比

数据挖掘工具	分类	估值	预测	相关分组	聚类和细分	解释
决策树	√		√			√
神经网络	√		√			
线性回归		√	√			√
Logistic回归	√		√			√
关联规则			√	√		
聚类			√		√	√
SOM			√		√	

表 2中效力：对于有监督学习来说，判断效力的标准就是工具的准确率；对于无监督学习来说，判断有效性的标准就是数据挖掘工具发现的模型的有用程度。

表 2 数据挖掘工具优缺点对比

数据挖掘工具	效力	可解释性	易于实施	产生模型时间	是否可信任	能否可视化	适合概念证明
决策树	Good	Excellent	Good	Fast	Yes	Yes	Yes
全体决策树	Excellent	Not Good	Not Good	Slow	Yes	No	No
神经网络	Excellent	Not Good	Not Good	Slow	Yes	No	No
线性回归	Good	Excellent	Excellent	Fast	No	No	Yes
Logistic回归	Good	Excellent	Excellent	Fast	Yes	No	Yes
关联规则	Good	Excellent	Good	Slow	Yes	Yes	Yes
聚类	Good	Excellent	Good	Fast	Yes	Yes	Yes
SOM	Good	Good	Not Good	Slow	No	No	No

可解释性：指的是一个领域专家或者一个不具有数据挖掘相关知识的人员理解数据挖掘模型的难易程度。

易于实施：指的是在生产和测试环境中部署模型的难易程度，直接和模型的复杂程度相关。

产生模型时间：数据挖掘工具通过搜索模式来形成最后的模型，不同工具的搜索速度不同。

是否可信任 (可信程度)：对于有监督学习任务来说通过利用可信程度，可以对预测结果进行排序，从而可以使用其中最准确的一部分结果；对于无监督学习任务来说，比如聚类，相关的可信程度表示是否能够计算出记录隶属于一个聚类的程度或者到某聚类的距离。

能否可视化：可以了解模型对未知例子打分的过程，对于模型部署之后的监督尤为重要。

适合概念证明：用于证明数据挖掘是否能带来利润，从而展示在特定问题上数据挖掘的价值。

4 发展前景与结束语

在数据库系统的发展过程中，数据挖掘只是很多工具的综合体，利用这些工具可以发现很多隐藏在数据库当中真正有价值的信息。虽然在 KDD过程中有很多数据挖掘工具，但是还没有一个能够很好地包含所有工具的模型或者方法。随着人工智能技术的不断进步，数据挖掘工具的不断完善，知识发现技术的不断发展，数据库当中人工参与的比重也将不断降低。

目前已经提出了一个基于 SQL的数据挖掘查询语言 (Data Mining Query Language, DMQL)。DMQL允许存取注入概念层次之类的信息，并不是数据的简单汇总。其复杂程度要求必须明确要挖掘的知识的类型，而且要挖掘的信息应该服从阈值或必要的重要度。

DMQL语句的核心是规则说明部分，因为这一部分要描述数据挖掘请求。通常有四种数据挖掘请求：泛化关系、特征规则、判别规则以及分类规则。

知识与数据发现管理系统 (Knowledge and Data Discovery Management System, KDDMS)被用来描述下一代数据挖掘系统，其中包括了数据挖掘工具以及管理数据的技术。还为特定的数据挖掘查询提供了存取入口。为了有效地存取，需要对数据挖掘查询进行优化。

数据挖掘的跨行业标准过程 (Cross Industry Standard Process for Data Mining, CRISP - DM) 可应用于许多不同的领域，强调 KDD过程中的所有步骤。其生命周期包括如下几个步骤：商业需求理解、数据理解、数据准备、建模和评价使用。可将这些步骤总结为“5A”：评估 (Access)、访问 (Access)、分析 (Analyze)、行动 (Act)和自动化 (Automate)

参 考 文 献

[1] Margaret H. Dunham. Data Mining Introductory and Advanced Topics[M]. Upper Saddle River, NJ: Prentice Hall, 2003.

[2] Gordon S Linoff, Michael J.A. Berry. Mining the Web: Transforming Customer Data into Customer Value. Hoboken[M]. NJ: John Wiley & Sons, Inc, 2001.

[3] Jaideep Scrivastava, Robert Cooley, Mukund Deshpande, et al. Web usage mining: Discovery and applications of usage patterns from web data[J]. Philadelphia: SIGKDD Explorations, January 2000(1): 12 - 23.

[4] Osmar RachidZaiane. Resource and knowledge discovery from the internet and multimedia repositories[M]. Burnaby, B.C. : Technical report, PhD Dissertation, Simon Fraser University. March 1999.

阅读此文的还阅读了：

1. 可视化数据挖掘技术
2. 数据挖掘过程可视化与交互式一般内容的研究
3. 基于applet的数据挖掘信息可视化
4. 岩土工程可视化及砂土液化评价可视化的探讨
5. 海难事故的数据挖掘
6. JSPMaker完全征服(上)
7. Trawling算法在Web结构挖掘中的应用
8. Web数据挖掘研究
9. 虚拟天文台:科学、工具及应用
10. 基于异几率属性的可视化关联规则挖掘
11. 图像的聚类和可视化方法研究
12. 关联规则的可视化人机交互系统应用研究
13. 聚类结果可视化研究
14. 可视化CSS工具
15. 数据挖掘在提高web用户网络访问速度上的研究
16. 基于平行坐标的关联规则挖掘技术可视化研究与实现

17. Web数据挖掘综述

18. 数据挖掘在企业中的应用

19. 浅论学习C语言编程的必要性

20. 关联规则下数据挖掘可视化技术的探讨与实现

21. 应用MedlineR挖掘基因间关系的尝试

22. 分布式智能公安GIS设计

23. TranEd第二版软件有助于复杂信号设计的研究

24. 基于ECharts的数据可视化分析组件设计实现

25. 数据挖掘结果的可视化问题

26. 浅谈工厂可视化管理

27. 信息可视化在技术监测中的应用

28. 基于Web的数据挖掘方法综述

29. 福建省地图制图学与地理信息工程学科发展研究

30. 基于数据挖掘的三峡水库调度自动化系统水位数据质量研究

31. 基于Web的数据挖掘研究综述

32. 7^th International Imaging Genetics Conference

33. Web数据挖掘综述

34. 基于CORBA的Web数据挖掘工具的设计及应用

35. 基于Web日志的数据挖掘技术在Web机器人识别中的研究

36. 基于Web的可视化数据挖掘工具综述

37. 谈谈数据仓库及数据挖掘工具

38. 基于WebGIS的计生药具数据挖掘系统

39. 数据挖掘在企业中的应用

40. 电子商务数据挖掘可视化系统模型研究及应用

41. 基于WEB的数据挖掘研究综述

42. 数据挖掘软件产品综述

43. 基于Web的数据挖掘技术研究及其应用

44. 数据挖掘在足球运动中的应用

45. 福建省地图制图学与地理信息工程学科发展研究

46. DAMP, an acidotropic pH indicator, can be used as a tool to visualize non-esterified cholesterol in cells

47. 基于CORBA的Web数据挖掘工具的设计及应用

48. 数据挖掘工具的评判

49. 图像的聚类和可视化方法研究

50. 基于web的数据挖掘系统的研究与设计