Web挖掘的现状和展望

郑 弦

(四川大学 计算机学院,四川 成都 610065)

摘 要:概述了如何在网络中获取有用信息,Web挖掘的基础知识及其相关比较。阐述了Web挖掘的发展过程、现状,以及未来趋势。介绍了Web内容挖掘、Web日志挖掘,还有被视为Web挖掘未来的云计算挖掘。

关键词:Web挖掘;Web内容挖掘;Web结构挖掘;Web日志挖掘;云计算挖掘

doi:10.3969/j.issn.1006-8554.2013.03.036

0 引言

互联网的广泛应用从根本上改变了人们的生活方式,包括沟通、获取信息、开展业务、购物。当年万维网和电子邮件使用频率的暴涨,促使计算机科学家和物理学家们急切想研究这一新现象。最初让他们惊讶的是互联网的庞大和多样性,但很快他们发现了一个普遍的规律:网络中包含大量的小元素,大元素较少。少数网站包含上万个网页,但更多的网站仅包含几个页面。大部分的网民集中到少许特定的网站,而大量的网站却鲜有人问津。

万维网的发展让大量信息可以被用户免费访问,不同的数据类型必须加以管理和组织,以方便不同用户有效的访问。因此,数据挖掘技术在Web上的应用现在成为越来越多人的研究重点。有些数据挖掘的方法已经用于挖掘Web中的隐藏信息,然而,Web挖掘不只是用数据挖掘的技术在Web中存储数据,还必须修改算法来更好满足网络的特殊需求[1]。新方法应该更适合Web中数据的属性。此外,不单是数据挖掘,人工智能、信息检索、自然语言处理技术都可以使用起来。因此,Web挖掘已经发展成一个独立的研究领域。

1 Web挖掘的历史演变

Web挖掘技术是经过漫长研究和产品发展的结果。这种演化从人们把商业数据存储在电脑和互联网上开始,之后产生了更多更新的数据获取和实时处理技术,方便了我们驾驭这些信息。不同地域、不同网络和调查产生的数据共同定义了数据集合。数据存储又涉及了软件、检索,存储介质等问题。

在商业信息的演化过程中,每一次发展都跟前一次息息相关。从用户的角度来看,表1中列出的5步是具有革命性的,因为它们让新的商业问题得到了快速而准确的解决。例如,数据库的巨大存储能力对Web挖掘来说就至关重要。

数据挖掘技术基本上被用在了Web挖掘中,Web挖掘是数据挖掘的扩展版本。数据挖掘离线操作,而Web挖掘在线操作。在数据挖掘中,数据存储在数据仓库中,而Web挖掘则存储在服务器或Web日志中。

Web挖掘主要技术的形成经过了数十年的发展,在研究领域,诸如人工智能、机器学习都有涉及。如今,技术本身的成熟,加之关系型数据引擎的高性能和数据集成的进步,使得这些技术在当前数据仓库环境中变得可行。

表1 Web挖掘的演化过程

演化过程	商业问题	可行技术	产品提供商	特征
数据收集(1960年代)	过去五年公司的总收入?	计算机、磁带、磁盘	IBM、CDC	可回溯、静态数据传输
数据访问(1980年代)	去年3月份公司在上海的销售额是多少?	关系型数据库、SQL、ODBC	Oracle、Sybase、Informix、IBM、Microsoft	可回溯、记录层面的动态数据传输
数据仓库&决策支持(1990年代)	去年3月份公司在上海的销售额是多少?探讨下北京的情况。	在线分析处理、多维数据库、数据仓库	Pilot、Comshare、Arbor、Congnos、Microstrategy	可回溯、多层面的动态数据传输
数据挖掘(2000年代)	公司在北京下个月销售额估计是多少?为什么?	预测算法、多处理器计算机、大型数据库	Pilot、Lockheed、IBM、SGI、众多创业型公司	预测性、积极的信息传递
Web挖掘(如今)	公司在北京N年后(N年前)的销售额是多少?	WWW、互联网、巨大规模数据库	RockWare、IBM、Web Trends、SPS、NetGenesis	功能强大、高效快速、用经济实惠的工具挖掘大型数据仓库和关系型数据库

现有Web挖掘方法存在的缺点:

1)用户感觉响应时间过长。

2)Web的爆炸式增长加重了对网络要求。

3)资源和Web服务器。

4)有一个明显提高网络质量的方法:增加带宽。但也会增加经济成本。

5)Web缓存方案有两个显著缺点:如果代理服务器没有正确更新,用户可能收到过期的数据;当用户数量增多后,原始服务器通常会成为瓶颈。

6)几个削弱Web缓存效果的因素。最显著的是缓存系统资源有限(即:内存空间,磁盘存储,I/O带宽,处理器能力和网络资源)。即便缓存空间是无限的,也无法避免一些问题,特别是更新巨大的Web对象集合时,管理起来十分困难。

7)加强系统的主要缺点:预取策略可能不是用户的最终

请求，而预取方案增加了网络流量和Web服务器的负载。

2 Web挖掘及其分类

Web挖掘是一种基于数据挖掘在网络中发现隐藏信息的技术。Web上所有页面都是节点，且有超链接相互连接。Web挖掘是有效提取信息、图像、文字、声音、视频、文件和多媒体的方式。现在我们搜索任何话题都能轻易的从网上获取相关信息，在以前想得到相关准确信息是很难的[1][2]。Web挖掘被认为是数据挖掘的一个特定应用，但值得单独提出来研究。

Web挖掘的流程是：数据抽取、信息选择和预处理、模式发现、模式分析。基于这4个流程，Web挖掘可以看作是使用数据挖掘技术自动从网络文件和服务器上检索、提取和分析信息来做知识发现。

Web挖掘根据用途不同被分为3类：Web内容挖掘；Web结构挖掘；Web日志挖掘。

2.1 Web内容挖掘

Web挖掘主要从网络上提取信息，如果其过程是访问网络上的信息，则属于Web内容挖掘。打开网页来获取网络上的信息，属于Web内容挖掘。打开搜索页面和在搜索页面上浏览信息一样，都是Web内容挖掘的最新定义。

2.2 Web结构挖掘

我们用图来定义Web结构挖掘，一个网页代表一个节点，一个链接代表图的一个边。它反映了网络页面间的关系。Web结构挖掘的动机是理清网络间的关系。它反映了从一个网页到另一个网页的链接。

2.3 Web日志挖掘

它用于发现用户在网络中不同位置产生信息的规律。它自动搜集存储在服务器上的用户使用日志，代理日志，客户端缓存，用户资料，网页个性内容，网站结构[3]。

Web日志挖掘目的是利用数据挖掘技术来探寻面对Web上不同应用时用户的使用模式。当用户上网时，它是预测用户行为的技术[4]。

Web日志挖掘分为3步：

1)预处理：根据客户端，服务器，代理服务器，它首先检测网络资源中那些未加工的数据并处理它们。本步骤自动转换这些原始数据。

2)模式发现：在这步中，根据不同数据使用诸如机器学习、数据挖掘等技术来发现知识。

表2 Web挖掘和数据挖掘的比较

比较	Web挖掘	数据挖掘
规模	规模不大，网络数据库中有千万数量级的任务。	规模大，在数据库中有上亿数量级的任务。
权限	数据是公开的，不隐藏网络数据库数据，但要经网络日志监视器允许。	数据具有私有性，只有授权用户可获得数据库数据。
结构	从结构化，非结构化和半结构化网页中提取信息，即从广大的数据库中获取信息。	从特定的结构中获取信息，相比Web挖掘它不能从广大的数据库中得到信息。

3)模式分析：模式分析在模式发现之后。它检查模式是否正确，还有如何实现在web中提取信息。

3 云计算与Web挖掘

云计算显然是当前最引人注目的技术之一，因为它有经济性、高效性和灵活性等优势。尽管人们在云计算方面的兴趣和参与性在增多，仍存在一个重要的问题：云计算会因为发展的阻碍，最终妥协成一种新的IT采购模型吗?[7]“云”这个词是互联网的一种象征，是互联网潜在基础设施的一种抽象，同时还代表用户与外部供应商的关系。

基本上，云计算挖掘是一种新型数据处理技术。SaS(软件即服务)已经实现，它减少了Web挖掘的开销，并尝试为用户提供安全性，这已经是云计算挖掘技术。现在我们准备改变Web挖掘的框架来满足云计算的需求[5]。考虑到“挖掘”云，Hadoop和MapReduce社区开发了一个强大的框架来预测分析复杂的分布式信息源。

4 结语

本文提供了Web挖掘领域当前现状和未来趋势的研究。指出了数据挖掘和Web挖掘间的区别。网络数据量正在飞速增长，Web挖掘是有前途的研究领域，许多成功的应用已经出现。本文还提出了Web挖掘的步骤和未来趋势。现在我们正在研究将Web结构挖掘和日志挖掘两者结合起来。也关注着云计算中的数据挖掘。实际上，通过云计算，Web挖掘的代价大大减少，所以我们也相信云计算挖掘是Web挖掘今后的趋势。

参考文献：

[1] Virgilio Almeida, IEEE International Conference in Parallel and Distributed Information Systems, December 1996.

[2] Pei, J. Han, "Mining Access Patterns efficiently from Web Logs", Knowledge discovery and Data Mining, 2000.

[3] Etzioni, O, "The, World Wide Web: quagmire or gold mine", Communication of the ACM , No. 11, 65 - 68, 1996.

[4] Wu, K. L. Yu, "A Web usage minging and analysis tool", IBM Systems Journal, 2010.

[5] Ajay Ohri, "Data mining through Cloud Computing", http://knol.google.com/k/data - mining - through - cloud - computing, 2010

[6] Michael Jennings, "What are the major comparisons or differences between Web mining and data mining?", Information Management Online, June 25, 2002.

[7] Deyi Li, "Mining association rules with linguistic cloud models", lecture Notes in Computer Science, 1998.

作者简介：

郑弦(1988 -)，男，重庆人，硕士研究生在读，研究方向：云计算与数据挖掘。