互联网上的数据挖掘

Data Mining on the Internet

高 飞 谢维信

（深圳大学信息工程学院 深圳518060）

Abstract Data mining technology was first applied to mine information from databases. With the rapid development of World Wide Web, the huge information stored on the Internet becomes a new important target for data mining. Different with finding and discovering sources from Internet (Resource Discovery), data mining on the Internet aims mainly at extracting implicit, previously unknown knowledge which may not be present on the surface of WWW resources. This paper overviews the development of Web mining in recent years, points out some open problems with it, and presents a prospect of its future research.

Keywords Data mining, Web mining, Resource discovery

1 引言

Web 挖掘是从 WWW 资源上抽取信息（或知识）的过程，它是将数据挖掘技术和理论应用于对 WWW 资源进行挖掘的一个新兴的研究领域。目前在该研究领域中，根据挖掘对象的不同大致可分为三个方面的挖掘研究：Web 内容挖掘、Web 结构挖掘、Web 使用挖掘（图1）。

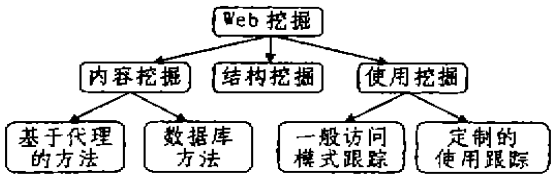

图1

数据挖掘是面向发现的数据分析技术，通过对大型的数据集进行探查，可以发现有用的知识，进而为决策支持提供有力的依据。数据挖掘中所采用的很多技术来源于过去二、三十年里所发展起来的人工智能和信息检索技术，近十年来的研究又诞生了不少新的数据挖掘技术和理论，并在生物、医疗、市场销售及金融等应用领域里获得了广泛的成功[1]。尽管数据挖掘的各种技术和理论原则上都可以直接或间接地应用于对 Web 信息的挖掘，如：基于代理的技术[2]、基于概念的信息检索技术、基于案例推理的信息检索技术[3]都已用于 Web 挖掘中。然而由于 WWW 资源的异质性、多样性、分布的广泛性，特别是其上数据的半结构化特点，导致了 Web 挖掘与对普通大型数据库所进行的挖掘有着很大的不同。Web 内容挖掘是从数以百万计的 Web 资源中发现信息或资源的过程[4]。Web 结构挖掘是从 WWW 的组织结构及引用和被应用间的链接关系中推理知识的过程。Web 使用挖掘，也称为 Web 日志挖掘是从 Web 访问日志中抽取知识的过程。其中，对于 Web 内容挖掘来说，又可以根据其挖掘策略的不同分为 Web 页内容挖掘和搜索引擎结果挖掘。为了避免混淆，我们在图中没有包含这两项的图示。尽管 Web 挖掘目前尚未有一个明确的定义，但是参照数据挖掘的定义，我们给出了 Web 挖掘的如下定义：Web 挖掘是从 WWW 资源上抽取有趣的、潜在有用的模式及隐含信息的数据挖掘过程。

2 Web 内容挖掘

近年来随着 WWW 信息的指数增加，那些只维护由关键字和超级链接所构成的数据库的搜索引擎越来越难以满足人们的需要。一个令人尴尬的事实是，搜索引擎返回了太多的结果，其中很多是无用或无关的结果，人们往往只浏览了它的前面若干个结果以后，就无奈地放弃了浏览。实际上，为了找到想要的结果，浏览

高 飞 硕士、讲师，主要研究方向：数据挖掘、模糊理论，神经网络及遗传算法等。谢维信 教授、校长，西安电子科大博导，主要研究方向：模糊理论、信号与信息处理、数据挖掘，神经网络及遗传算法等。

上几百条记录是常有的事，因此有必要开发出更为有效的技术以支持 Web 内容挖掘。根据实现的方法的不同可分为基于代理的方法和数据库方法；而根据挖掘策略的不同又可分为 Web 页概要和搜索引擎结果概要。Web 页概要直接挖掘 Web 文档的内容，搜索引擎结果概要用于增强搜索引擎的内容查询功能。

2.1 挖掘方法

2.1.1 *基于代理的方法* 代理技术是从七十年代末期发展起来的一项人工智能技术，代理可以通过一定的规则模仿人的行为，用以解决人所无法从事的大量的信息处理工作。代理是一些软件，它与传统的软件所不同的是其具有自主性，同时它具有学习功能，因此它的能力可随时间的变化进行调整。基于代理的方法包含了这样的一个人工智能系统，它可以"自主或半自主地为某个特殊的用户服务，以发现和组织基于 Web 的信息"[5]。一些智能的 Web 代理可以利用用户描述文件(User Profile)来查找相关的信息，然后组织和解释这些被查找到的信息[6]。还有一些代理则利用各种信息检索技术及开放的超文本文档的特性来组织和过滤检索到的信息[7]。另外一类代理被设计成可以学习用户的喜好，并利用这些喜好来为那些特殊的用户查找资源[8]。代理表示了一个信息检索元素构成的"分布式"网络，它们可以相互通信，并且无需人的控制。通过作用于每个 Web 服务器及相互间的通信为终端用户提供查询结果。因此，与数据库中的信息查询和过滤有所不同，代理构成的网络具有相当大的伸缩性。

使用代理的主要缺点是存在隐私泄漏的可能，由于代理具有社会化的能力，信息的交换是透明的，且代理不会通知某一用户它是否正在提交和检索该用户的信息。实际上，已经有很多关于分布式系统的研究工作，其中之一是，为了解决互联网在结构上的不足，建议在节点上存储必要的信息。(http://freenet.sourceforge.net)

2.1.2 *数据库方法* 主要集中在"对 Web 上的异质的、半结构化的数据进行整合和组织，使其成为结构化较好的、高层的资源集合。"然后再对这些组织好的资源进行访问和分析[3]。这些元数据(或泛化了的数据)于是可以组织成有结构的数据集(如关系数据库或面向对象数据库)，然后再加以分析。目前的数据库方法又可分为多层数据库和 Web 查询系统。

多层数据库是由若干层信息构成的数据库。利用多层数据库，可以提供一个与用户请求对应的指向目标文档集合的指针列表，其次允许用户交互地浏览用以指向目标文档集合的详细信息而非目标文档本身。其主要思想是：信息抽取，前提是假定用户不太愿意浏览信息的庞大细节，而更愿意扫读关于信息的一般性描述。Web 上无结构或半结构的数据被转化为较小的、结构化的和位置较近的数据库，该数据库中包含了从层次结构的前一层数据库中抽取出来的信息，其代价是牺牲了前一层数据库中的细节信息。随着数据分析、变化及泛化(generalization)技术的发展，使得把 Web 上的各种各样的原始信息变换为有一定结构的、分类的和高层的信息成为可能，其代表了多数据库层次结构中的第二层，而包含原始的、半结构化信息的 Web 则属于多层数据库中的第一层。如果需要，则可以从第二层向上建立更多的层次。

这种方法的优点是：它可以充分利用现有的数据库、数据挖掘等技术；提供高层的查询接口；信息资源的高效管理；提供关于 Web 页内容的全局视图等等。其缺点是需要额外的软件，如：数据库管理软件系统、构建层次的软件及查询系统。同时处理多媒体对象(声音、图像、视频)时也是一个难题，因为不象从文本中提取词语那么简单，图像和声音需要用其他的方式进行分类和索引。(尽管我们可以使用元信息，如图像或声音的名字，但是如果这些信息不存在时就会遇到麻烦。)有关多层数据库还有很多尚待讨论的问题，限于篇幅，本文不再详述(参阅文[10,11])。

Web 查询系统利用一个 Web 的简单关系视图，将结构和基于内容的查询准则以类似于标准的数据库查询语言(如 SQL)的方式结合起来，对 Web 上半结构化的数据进行查询，此类查询系统中采用的都是所谓的声明式查询语言[12]。从 Web 上查询半结构化的数据需要两个阶段，首先通过生成一个关联数据库来实现 Web 的关系结构，紧接着进行关键字查询及创建用于把文档特征映射为数据模型中的实例(如图或表)的外壳。目前已有多种 Web 查询系统，如：W3QL[13]、WebLog[12]、WebSQL[14]及 WebOQL[15]。这四种查询系统都是用来从 Web 上进行信息收集的，所不同的是，WebLog 和 WebOQL 主要目的是 Web 文档的重建工作，而 WebSQL 和 W3QL 则主要是并行地从搜索引擎返回的文档中发现与主题相关的文档。由于用这些查询语言所表示的查询需要翻译为数据源上的查询，而这些数据源往往具有它们自己的独立模式和查询系统，因此有必要建立简单易用的、标准化的查询语言，而现有的 Web 查询语言则缺乏统一的标准。

2.2 挖掘策略

2.2.1 *Web 页概要* 互联网上的大量信息通常隐藏于 Web 文档内部，因此一类重要的应用就是对 Web 页内容的挖掘。从由各种不同成分构成的、不规则的文档(如 Web page)中挖掘知识的研究成果中，比较杰出的是 Ahoy![16]、WebOQL[15]及 Shopbot Project[17]。Ahoy!用来发现个人主页，给定个人的信息，

Ahoy!利用互联网服务，如搜索引擎、电子邮件列表服务器等来获取与个人相关的数据，利用试探法识别文档中显示该文档作为个人主页的印刷或句法特征。WebOQL是一个用于Web页重构的查询语言，利用文档的图树表示形式，它能够从在线新闻站点或导游指南中获取信息，Shopbot所描述的购物代理通过学习来识别在线目录和电子商务站点的文档结构，抽取价格表和特定报价，该代理可以编辑从不同站点获取的信息并发现有用的交易。从Web文档内部进行有效的信息抽取的主要障碍是元数据的缺乏及没有一个标准的方法用于描述、操纵及在电子文档中交换数据。WWW协会建议的XML标准目前已被很多大公司广泛采用，这为WWW上的数据挖掘减轻了很大的负担。XML提供了灵活的数据标准，它可以对许多种类的电子文档的内容、语义及模式进行编码。其提供的通用数据格式可以将数据与文档的表示相隔离，并使文档可以利用DTD(文档类型定义)进行自解释。

2.2.2 *搜索引擎结果概要* 对搜索引擎返回的结果进行挖掘是十分必要的，这可以提供给用户更为准确的查询结果。WWW文档的异质性和缺乏结构的特点导致一些研究工作集中于挖掘已知文档的子集或与某一主题相关的文档，一个这样的子集可以是一个搜索引擎的查询结果。文[14]中利用包含最小信息的关系表提出了一种用于结果提炼的查询语言WebSQL，该系统访问搜索引擎获取的文档，并从文档内部或者从服务器提供的数据中收集诸如URL、标题、内容类型、内容长度、修改日期及链接等信息，类SQL声明式语言(SQL-Like declarative language)提出了从搜索结果中获取相关文档的能力，Zamir和Etzioni[18]提出了一种用于把搜索引擎返回的文档进行聚类的技术。该技术仅仅依赖于由搜索结果所提供的信息(如：URLs、标题、网页的第一行描述等)来归纳出聚类，并将相应的文档归入这些聚类中。这些聚类是搜索引擎返回的文档集合上的高层视图，使得在搜索引擎返回的非常大的文档列表中的过滤操作变得十分方便。

3 Web结构挖掘

由于超文本文档间的关联关系使得WWW不仅仅可以揭示文当中所包含的信息，同时也可以揭示文档间的关联关系所代表的信息。例如，指向一个文档的链接体现了该文档的被引用情况(或普及性)，而从一个文档发出的链接则体现了该文档所覆盖的主体的种类(或文档内容丰富与否)。这可以同文献的引用情况相比较，如果某篇文章经常被引用，说明它非常重要。PageRank[15]及CLEVER[20]中的方法正是利用了文档间的链接信息来查找相关的Web页。

4 Web使用挖掘

尽管WWW作为一个信息资源是繁杂、异质和庞大的，然而从局部上来说，在每一个提供信息资源的服务器上都有一个结构化较好的记录集，即Web访问日志。每当有获取资源的请求到来时，Web服务器都将记录和积累这些关于用户交互作用的数据。分析不同的Web站点的Web访问日志可以帮助人们理解用户的行为和Web的结构，从而大大提高由大量资源所构成的网站的设计工作效率。根据应用的不同，可以将Web使用挖掘分为两种主要倾向，即：一般的访问模式跟踪及定制使用跟踪。一般访问模式跟踪通过分析Web访问日志来理解访问模式及倾向，利用这些分析可以清楚地给出较好的Web结构及资源提供者的分组情况。文[21]中给出了一种对Web日志进行挖掘的工具(WebLogMiner)，并且推荐了一些如何对经过提炼和转换后的Web访问日志文件进行数据挖掘和在线分析处理的技术。把数据挖掘技术应用于Web访问日志可以获取有趣的访问模式，这些访问模式有助于网站的重构、指出Web页上有效的广告位置及研究特殊的用户行为和特殊的销售广告等[22]。定制使用跟踪可以分析个人的倾向，它的主要目的是为每个用户定制符合其个人特色的Web站点。根据个人的喜好，可以在显示的信息、网站的结构及资源的格式等方面动态地进行定制。文[23]中提出了一种自适应网站，可以通过对用户访问文件的学习自动更新网站。尽管对Web访问日志分析存在着各种各样潜在的应用，然而应该指出的是，这类应用的成功与否依赖于人们可以从大量原始数据中怎样发现和发现多少可靠的信息。目前的Web服务器上存储的访问信息是有限的，少数网站上含有为其专门定制的一些脚本，这些脚本可以存储额外的信息。当然，要进行有效的Web使用挖掘，在分析之前往往需要对这些数据进行必要的清理和变换工作[19]。

Web使用挖掘中经常采用的技术是聚类(clustering)和关联分析。聚类是按类别组织数据的方法，又被称为无监督分类；而关联分析则是挖掘所谓的关联规则。

结论和展望 总的说来Web挖掘仍处于其研究的初级阶段，互联网在技术和应用上的不断发展将会极大地推动索引技术、数据挖掘及数据库技术的发展，并直接导致Web挖掘在技术和理论上的不断发展。由基于代理技术的查询工具所构成的分布式网络将向大型和智能化发展，同样，基于数据库的查询工具也会不断发展，特别是将半结构化的查询语言与多层数据库项结合是一个研究的趋势。Web挖掘研究的进展非常

快，最近文[24]中开发出了一个比任何传统的商用搜索引擎都高效、快速的基于数据库的搜索引擎。应该指出的是，为了推动 Web 挖掘研究的发展就必须解决三个问题：1)信息安全问题；2)查询结果的质量问题；3)搜索工具的伸缩性问题。这些问题的解决依赖于代理技术、索引技术及分布式网络等技术的发展。

参 考 文 献

1 Specific Data Mining Applications. Available at: http://www.pvv.unit.no/~hgs/project/report/node80.html

2 Etzioni O. The world-wide wed: Quagmire or gold mine? Communications of the ACM, 1996, 39(11): 65~68

3 Daniels J J, Rissland E L. A case-based approach to intelligent information retrieval. In: Proc ACM SIGIR' 95 Conf., Seattle, WA, USA, 1995

4 op. cit. Available at: http://www-users.cs.umn.edu/~mobasher/webminer/survey/survey.html

5 Agent-Based Approach. Available at: http://maya.cs.depaul.edu/~mobasher/webminer/survey/node4.html # SECTION00021100000000000000

6 Harvest: A scalable, customizable discovery and access system: [Technical Report CU-CS-732-94]. University of Colorado, 1994

7 Hypursuit. Available at: http//budapest.lcs.mit.edu/publications/papers/hypertabs.htm

8 XpertRule Miner. Available at: http://www.attr.com

9 Database Approarch. Available at: http://maya.cs.depaul.edu/~mobasher/webminer/survey/node5.html # SECTION00021200000000000000

10 Han J, Zaiane O R, Fu Y. Resource and knowledge discovery in global information systems: A multiple layered database approach. [Technical Report TR94-24]. School of Computing Science, Simon Fraser University, Nov. 1994

11 Han J, Zaiane O R, Fu Y. Resource and knowledge discovery in global information systems: A scalable multiple layered database approach. In: Proc. Conf. On Advances in Digital Libraries, Washington, DC, May 1995

12 Lakshmanan L, Sadri F, Subramanian I. A declarative language for querying and restructuring the web. In: Proc. 6th Int. Workshop on Research Issues in data Engineering. New Orleans, 1996

13 Konopnicki D, Shmueli O. Weqs: A query system for the world-wide web. In: Proc. 21st Int. Conf. On Very Large Data Bases (VLDB). Zurich, Switzerland, 1995. 54~65

14 Mendelzon A, Mihaila G, Milo T. Querying the world wide web. In: Proc. PDIS' 96, Miami, Dec. 1996

15 Arocena G O, Mendelzon A O. WebOQL: Restructuring documents, databases and webs. In: Proc of ICDE Conf., Orlando, Florida, USA, Feb. 1998

16 Shakes J, Langheinrich M, Etzioni O. Ahoy! The home page finder. In: Proc. Sixth World Wide Web Conf., Santa Clara, CA, USA, April 1997

17 Doorendos R, Etzioni O, Weld D. A scalable comparison-shopping agent for the world-wide web. In: Proc. Autonomous Agents ACM, 1997

18 Zamir O, Etzioni O. Web document clustering: A feasibility demonsrration. In: Proc. ACM SIGIR' 98, 1998

19 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: 7th Int Conf. WWW. Brishbane, Australia, April 1998

20 Chakrabarti S, et al. Experiments in topic distillation. In: ACM SIGIR workshop on Hypertext Information

21 Zaiane O R, Xin M, Han J. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In: Proc. Advances in Digital Libraries ADL' 98. Santa Barbara, CA, USA, April 1998. 19~29

22 Sonny H S, et al. Electronic Commerce Technologies: Challenges and Opportunities, chapter Towards On-Line Analytical Mining on the Internet for Electronic Commerce. Prentice Hall, 1999

23 Perkowitz M, Etzioni O. Adaptive sites: Automatically learning from user access patterns. In: Proc. 6th Int. World Wide Web Conf., Santa Clara, California, April 1996

24 Techweb article on Google's 4000-nods linux setup (http://www techweb.com/wire/story/TWB20000530S001)—Valid on the 31/05/2000