Web 挖掘研究综述*)

Web Mining Research Survey

宋爱波　董逸生　吴文明　孙志挥

(东南大学计算机科学与工程系　南京 210096)

Abstract　Two important and active areas of current research are data mining and the World Wide Web. A natural combination of the two areas, referred to as Web mining, has been the fertile research area. In this paper, we survey the current research in this area, including three Web mining categories: Web content mining, Web structure mining and Web usage mining. With respect to each category, it's representation method, processing method and application area are pointed out. We also discuss the relation between Web mining and information retrieve and machine learning.

1　引言

今天 Web 已成为信息发布、交互及获取的主要工具，Web 上的信息量正以惊人的速度增加着，人们迫切需要能自动地从 Web 上发现、抽取和过滤信息的工具。同时，近年来，由于电子商务的快速发展，许多公司借助 Internet 进行在线交易，企业管理者需要分析大量的在线交易数据，从而发现用户的兴趣爱好及购买趋势，为商业决策风险投资等提供依据。具体来讲，当我们与 Web 交互时，常面临如下问题：

1. 查询相关信息。这是查询触发的过程，我们希望从 Web 上找到关于 VC++编程指南的书，关于申办奥运会的信息，甚至关于爱滋病的报道等等。可以用搜索引擎如 Yahoo Sohu 等进行关键字查找，然而，今天的搜索引擎都有两个严重问题：低查准率会返回很多不相关的结果；低查全率有很多相关的文档找不到，

2. 从 Web 数据发现潜在的未知信息。这是数据触发的过程，仅仅用关键字的查找是不能实现的，需要机器学习和数据挖掘技术，现在的搜索引擎不具备这些功能。

3. 了解用户的兴趣爱好。Web sever 能根据用户的浏览信息，自动地发现用户的兴趣爱好，即用户的 Profile。

4. 信息个性化。不同人访问 Web 的目的、兴趣、爱好是有差别的，使用户能依据自己的兴趣爱好定制网页，甚至 Web server 能根据已发现的用户 Profile 自动为用户定制网页。

最后三个问题与电子商务、Web 站点设计、自适应 Web 站点紧密联系。现在的搜索引擎仅仅能解决第一个问题，其它问题是无能为力的。当今世界上研究的热门领域—Web 挖掘能直接或间接地解决上述问题。Web 挖掘是数据库、数据挖掘、人工智能、信息检索、自然语言理解等技术的综合应用，由于 Web 是异质分布且不断增长的信息系统，对其挖掘并不是上述技术的简单综合，它需要有新的数据模型、体系结构和算法等。

2　Web 挖掘与信息检索、机器学习

Web 挖掘是从 Web 中寻找有用的潜在的以前未知的知识。在文[1]中，把 Web 挖掘分成四步：

1. 资源发现：在线或离线检索 Web 的过程，例如用爬虫(crawler)或蜘蛛(spider)在线收集 Web 页面。

2. 信息选择与预处理：对检索到的 Web 资源的任何变换都属于此过程，如英文单词的词干提取，高频低频词的过滤，汉语词的切分，索引库的建立甚至把 Web 数据变换成关系，

3. 综合过程：自动发现 Web 站点的共有模式，

4. 分析过程：对挖掘到的模式进行验证和可视化处理。

Web 挖掘与信息检索、机器学习是紧密联系的，但又有所区别。信息检索是根据用户的需求描述，从文档集中自动地检索与用户需求相关的文档，同时使不相关的尽量少，它是目标驱动，查询触发的过程，主要任务是对于给定的文档怎样建索引，怎样检索。现代信息检索研究的领域包括：建模、文档预处理、文档分类聚类、用户需求描述(查询语言)、用户界面和数据可视

*)本文得到国家自然基金资助项目资助(79970092)。

化等。Web 挖掘使用信息检索技术对 Web 页面进行预处理、分类聚类、建索引。从这一点讲，Web 挖掘是信息检索的一部分。但 Web 挖掘要处理的页面是海量、异质、分布、动态、变化的，要求 Web 挖掘采取更有效的存取策略、更新策略，同时，Web 挖掘是一个数据触发的过程，它发现的知识是潜在的用户以前未知的。

机器学习被广泛应用于数据挖掘中，而 Web 挖掘是对 Web 在线数据的知识发现，所以机器学习是一种有效的方法，研究表明[3]与传统 IR 相比，用机器学习对文档分类，效果更好，但有些 Web 上的机器学习并不属于 Web 挖掘，如搜索引擎[3]使用机器学习技术来判断下一步最佳路径。

3 Web 挖掘分类

Web 数据有三种类型：通常所说的 Web 数据如 HTML 标记的 Web 文档，Web 结构数据如 Web 文档内的超链，用户访问数据如服务器 log 日志信息。相应地，Web 挖掘也分成三类[4]：Web 内容挖掘、Web 结构挖掘和 Web 访问挖掘。在不引起混淆的情况下，第一种类型数据仍简称为 Web 数据。

3.1 Web 内容挖掘

Web 内容挖掘是从 Web 数据中发现信息。随着 Internet 进一步扩展，Web 数据越来越庞大，种类繁多。有早先的 Gopher Ftp Usenet 数据，有数字图书馆政府部门数据，以及各公司自己组建的数据仓库。这些数据既有文本数据，也有图像、声频、音频等多媒体数据；既有来自于数据库的结构化数据，也有用 HTML 标记的半结构化数据及无结构的自由文本。对于多媒体数据的挖掘称为多媒体数据挖掘[5]；对于无结构自由文本的挖掘称之为文本的知识发现[6]。在文[7]中把 Web 内容挖掘分成两大类：IR(information retrieve)方法和数据库方法。

3.1.1 IR 方法　主要应用 IR 技术，评估改进搜索信息的质量，可以处理无结构数据和 HTML 标记的半结构化数据。

1. 处理无结构数据。一般采用词集(bags of words)方法，用一组组词条来表示无结构的文本。首先用 IR 技术对文本预处理，然后采取相应的模型进行表示。若某词在文本中出现为真，否则为假，就是布尔模型；若考虑词在文本中出现的频率即为向量模型；若用贝叶斯公式计算词的出现频率，甚至考虑各个词不独立地出现，这就是概率模型。另外，还可以用最大字序列长度[6]、划分段落[5,12]、概念分类[6]等方法来表示文本。

对于词集表示，采用的处理方法有：TFIDF[5,11]、Hidden Markov Model、统计方法[12]、判决树(decision trees)[13]和最大熵(Maximum entropy)[14]等。主要应用有：文本分类[5,11,14]、层次聚类[12]和预测词的出现关系[13]等。当然，也可以综合运用上述的表示和处理方法。如文[9,10]用词集和段落表示文本，文[9]用 TFIDF、判决树、Naive Bayes、Bayes nets 方法进行文本分类，而文[10]用聚类算法、k-最近邻算法(K-Nearest Neighbor)和判决树进行事件探测。文[8]用长度不超过 n 的词条表示文本，采用无监督层次聚类、判决树、统计分析方法，对文本进行分类和层次聚类。

2. 处理半结构化数据。半结构化数据指 Web 中由 HTML 标记的 Web 文档，同无结构数据相比，由于半结构化数据增加了 HTML 标记信息及 Web 文档间的超链结构，使得表示半结构化数据的方法更丰富。词集、超链[15]、词集、URL、元信息[16]，概念、命名实体[17]，句子、段落、命名实体[14]等的处理方法上主要利用数据挖掘技术如关联规则、分类算法[17]、演义逻辑、规则学习[15]等。应用有：Web 文档的分类[14]、聚集[16]，在 Web 文档内发现模式[17]等。

3.1.2 数据库方法　数据库方法是指推导出 Web 站点的结构或者把 Web 站点变成一个数据库以便进行更好的信息管理和查询。在文[19]中把数据库管理分成三个方面：

1. 模型化与查询 Web。研究 Web 上的高级查询语言，而不仅仅是现有的基于关键字查询。

2. 信息抽取与集成。把每个 Web 站点及其包装程序(wrapper program)看成一个 Web 数据源，研究多数据源的集成，可通过 Web 数据仓库(data warehouse)或虚拟 Web 数据库实现。

3. Web 站点的创建与重构。研究如何建立维护 Web 站点的问题，可以通过 Web 上的查询语言来实现。

数据库方法的表示法不同于 IR 方法，一般用 OEM(object exchange model)[20]表示半结构化数据，OEM 使用带标记的图来表示，对象为结点，标记为边。对象由唯一的对象标记符和值组成，值可以是原子的，如整数、字符串等，也可以是引用别的对象的复杂对象。

应用主要集中在模式发现或建立数据向导(dataguides)，也有的研究者用来建立多层数据库，低层为原始的半结构化数据，较高层为元数据或从低层抽取的模式，在高层被表示成关系或对象[21~23]等。另外，还有一些 Web 上的查询系统。早期的查询系统是把基于搜索引擎的内容查询与数据库的结构化查询结合起来，如 W3QL、WebSQL 等。近来的查询语言强调支持半结构化数据，能够存取 Web 对象，用复杂结构表达查询结果，如 Lorel UnQL、WebOQL、StruQL、

Florid 等系统。

3.2 Web 结构挖掘

Web 结构挖掘研究的是 Web 文档的链接结构，揭示蕴含在这些文档结构中的有用模式，处理的数据是 Web 结构数据。文档间的超链反映了文档间的某种联系，如包含、从属、引用等。文[24]使用一阶学习的方法对 Web 页面超链进行分类，以判断页面间的 members of project，department of persons 等关系。文[25，26]分别使用 HITS 和 Pagerank 算法计算页面间的引用重要性，基本思想是对于一个 Web 页面，如果有较多的超链指向它，那么该页面是重要的，此重要性可作为 Web 页面评分（rank）的标准。这方面的算法有 HITS[25]、Pagerank[26]及改进的 HITS[27]（把内容信息加入到链接结构中）。成型的应用系统有 Clever system[27]、Google[26]等。Web 页面内部也有或多或少的结构，文[28]研究了 Web 页面的内部结构，提出了一些启发式规则，用于寻找与给定的页面集合相关的其它页面；文[29]使用 HTML 结构树对 Web 页面进行分析，得到其内部结构，用于学习公司的名称和地址等信息的页面内的出现模式。另外，在 Web 数据仓库中可以用 Web 结构挖掘检测 Web 站点的完整性[30]。

3.3 Web 访问挖掘

Web 访问挖掘是用挖掘 Web 服务器 log 日志获取的知识预测用户浏览行为的技术。由于 Web 自身的特点 异质、分布、动态、无统一结构，使得在其上进行内容挖掘较困难，它需要在人工智能自然语言理解等方面有突破性进展。然而，Web 服务器的 log 日志[31]却有完美的结构，每当用户访问 Web 站点时，所访问的页面、时间、用户 ID 等信息，在 log 日志中都有相应的记录，因而对其进行挖掘，是切实可行的也是很有意义的。Web 的 log 数据包括：server log，proxy server log 及 client 端的 cookie log 等。一般先把 log 数据映射成关系或对其进行预处理，然后才能使用挖掘算法。进行预处理包括清除与挖掘不相关的信息，用户、会话、事务的识别等[32]，对 log 数据可靠性影响最大的是局部缓存和代理服务器（proxy server）。为了提高性能，降低负载，很多浏览器都缓存用户访问的页面，当用户返回浏览时，浏览器已从其局部缓存取得，服务器却没有用户返回动作的记录。代理服务器提供间接缓存，它比局部缓存带来的问题更严重，从代表服务器来的所有请求，即使用户不同，它们在服务器的 log 中也有相同的 ID。目前解决的主要方法是 cookies 和远程 Agent 技术[33，34]。

对 log 数据挖掘采用的算法有：路径分析、关联规则和有序模式的发现、聚类分类等，为了提高精度，访问挖掘也用到站点结构和页面内容等信息[34]。

Web 访问挖掘可以自动发现用户存取 Web 的兴趣爱好（即用户 profile）及浏览的频繁路径。Web 用户希望 Web 服务器能了解他们的爱好，提供他们感兴趣的东西，要求 Web 具有个性化服务的功能；另一方面，信息提供者希望依据用户的 profile 和浏览模式，改进站点的组织性能。Web 访问挖掘获得的知识，可以帮助我们进行自适应站点设计、信息组织、个性化服务、商业决策等。

我们对上述 Web 挖掘的讨论总结如下表：

	Web 内容挖掘		Web 结构挖掘	Web 访问挖掘
处理数据类型	IR 方法 无结构数据、半结构化数据	数据库方法 半结构化数据	Web 结构数据	用户访问 Web 数据
主要数据	自由文本、HTML 标记的超文本	HTML 标记的超文本	Web 文档内及文档间的超链	Serverlog、proxy serverlog、client log
表示方法	词集、段落、概念、IR 的二种经典模型	OEM 关系	图	关系表、图
处理方法	IDIDF、统计、机器学习、自然语言理解	数据库技术	机器学习、专有算法如 HITS Pagerank	统计、机器学习、关联规则
主要应用	分类、聚类、模式发现	模式发现数据向导、多层数据库、站点创建与维护	页面权重、分类聚类、模式发现	用户 profile 自适应 Web 站点、商业决策

结论 近年来，电子政务、电子图书馆、远程教育等已成为 Web 的主要应用，这使得 Web 挖掘成为国际上的热门研究领域。本文给出了 Web 挖掘研究的三种分类，针对每一种分类介绍了其表示形式、处理方法、应用领域及最近的研究情况，讨论了 Web 挖掘与信息检索、机器学习的联系。但是，限于篇幅，本文没有进行深入细致的探讨，有兴趣读者可以查阅所列文献。应该重视的是，近来，越来越多的研究者，不仅仅是学术界的，更多的来自于企业界，他们把目光集中在 Web 内容挖掘的数据库方法上，试图把整个 Web 集成于一个数据库系统中，以提供完善的查询、优化和维护的机制，这是一个极具挑战性的研究课题。Web 的广泛应用使之成为必要，XML 的出现使之成为可能。

参考文献

1 Etzioni O. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11): 65～68

2 Mladenic D. Text-learning and related intelligent agents.

IEEE Intelligent Systems, 1999, 14(4): 44～54
3 Rennie J, McCallum A. Using reinforcement learning to spider the web efficiently In: Proc. of the 16th Int. Conf. on Machine Learning ICML-99, 1999
4 Madria S K, et al. Research issue in web data mining. In: Proc. of Data Warehousing and Knowledge Discovery, First Intel. Conf. ,DaWak'99, 1999 303～312
5 Zainane O R, et al. Multimediaminer: a system prototype for multimedia data mining. In: Proc. ACM SIGMOD Int Conf. On Management of Data, 1998. 581～583
6 Feldman R, Dagan I. Knowledge discovery in textual databases. In: Proc. of the first Int. Conf. On Knowledge Discovery and Data mining, Montrel, Canada, 1995. 112～117
7 Kosala R, Blockeel H. Web mining research: a survey
8 Kargupta H, Hamzaoglu I. Distributed data mining using an agent based architecture. In: Proc. of Knowledge Discovery and Data Mining: AAAI Press, 1997. 211～214
9 Dumais S, Platt D J. Heckerman Inductive learning algorithms and representations for text categorization. In: Proc. of the 1998 ACM 7th Int. Conf. On information and knowledge management, Washington United States, 1998. 148～155
10 Yang Y, et al. Pierce Learning approaches for detecting and tracking news events. IEEE Intelligent System, 1999, 14(4): 32～43
11 Billsus D, Pazzani M. A hybrid user model for news story classification. In: Proc. of the 7th Int. Conf. On User Modeling (UM'99), Banff, Canada, 1999
12 Honfmann T. The cluster abstraction model. Unsupervised learning of topic hierarchies from text data. In. Proc. of 16th Int. Joint Conf. On Artificial Intelligences IJCAI-99, 1999. 682～687
13 Nahm U Y, Mooney R J. A mutually beneficial integration of data mining and information extraction. In: Proc. of the 17th National Conf. On AI, 2000
14 Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: Proc. of the IJCAI-99 Workshop on Machine Learning for information filtering. 61～67
15 Furnkranz J. Exploiting structural information for text classification on the www. In Advances in Int. Data Analysis, Third Int. Symposium, IDA-99, 1999. 487～498
16 Crimmins F, Smeaton A. Information discovery on the internet. IEEE Intelligent Systems, 1999, 14(4): 55～62
17 Singth L, Chen B, Height R. A robust system architecture for mining semistructured data. In: Proc. of the Second Int. Conf. On Knowledge Discovery and Data Mining, 1998. 329～333
18 Soderland S. Learning information extraction rules for semistructured and free text. Machine Learning, 1996, 34: 233～272
19 Florescu D, Levy A Y. Database techniques for the world wide web. A survey. SIGMOD Record, 1998, 27(3): 59～74
20 Abiteboul S, et al. The Lorel query language for semistructured data. Int. J. on Digital Libraries, 1997, 1(1): 68～88
21 Zaiane O R, Han J Resource and knowledge discovery in global information systems: A preliminary design and experiment. In: Proc. of the First Int. Conf. On Knowledge Discovery and Data Mining. Montreal, Qubec, 1995 331～336
22 Khosla I, Kuhn B, Soparkar N. Database search using information mining. In: Proc. of 1995 ACM-SIGMOD Int Conf. On Management of Data, 1995
23 Merialdo P, Atzeni P, Mecca G. Semistructured and structured data in the web: Going back and forth. In. Proc. of Workshop on the management of Semistructured Data (in conjunction with ACM SIGMOD) 1997
24 Craven M, Slattery S, Nigam K. First-order learning for web mining. In: Proc. of 10th Europen Conf On Machine Learning. Chemnitz, 1998
25 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: 7th Int. World Wide Web Conf., Brisbane, Australia, 1998
26 Kleinberg J M. Authoritative source in a hyperlinked environment. In: Proc. of ACM-SIAM Symposium on Discrete Algorithms. 1998. 668～677
27 Chakrabarti S, et al. Mining the link structure of the world wide web. IEEE Computer, 1999, 32(8): 60～67
28 Spertus E. Mining structural information on the web. In. Proc. of the sixth Int. World Wide Web Conf. 1997. http. // decweb. ethz. ch/www6/technical/paper205/paper206. htm
29 DiPasquo D. Using HTML formatting to aid in natural language processing on the World Wide Web. School of Computer Science, Canegie-Mellon University, 1998
30 Luotonen. The common log file format. http: // www w3. org/pub/www/. 1995
31 Cooley R, Mobasher B, Srivastara J. Data preparation for mining World Wide Web browsing patterns
32 Elo-Dean S, Viveros M. Data mining the IBM official 1996 Olympics Web site: [Technical report] IBM T. J. Watson Research Center
33 Shahabi C, et al. Knowledge discovery from user Webpage navigation. In Workshop on Research Issues in Data Engineering Birmingham, England, 1997
34 Spiliopoulou M. Data mining for the web. In Principale of Data Mining and Knowledge Discovery, Second European Symposium, PKDD'99, 1999. 588～589

(上接第 39 页)

2 Aleksander I, Thomas W V, Bowden P A. WISARD—a radical step forward in image recognition, Sensor Review, July, 1984. 120～124
3 Yang Guoqing, Chen Songcan, Lu Jun. Multilayer Parallel Distributed Pattern Recognition System Model Using Sparse RAM Nets. IEE PROCEEDINGS-E, 1992, 139(2): 144～146
4 Kolcz A, Nigel M. Allinson: N-tuple Regression Network. Neural Networks, 1996, 9(5): 855～869
5 Tattersall G D, Foster S, Johnston R D. Single-layer lookup perceptrons. IEE PROCEEDINGS-F, 1991, 138(1): 46～54
6 Kanerva P. Sparse Distributed Memory. MIT Press, Cambridge, Massachusetts, 1988
7 Rohwer R, Morciniec M. The Theoretical and Experimental Status of the n-tuple Classifier. Neural Networks, 1998, 11(1): 1～14