计算机科学 2000Vol. 27№. 4

Web 数据挖掘

Web Mining

王 实 高 文 李锦涛
(中国科学院计算技术研究所 北京 100080)

Abstract Web Mining is an important branch in Data Mining. It attracts more research interest for rapidly developing Internet. Web Mining includes: (1)Web Content Mining; (2)Web Usage Mining; (3) Web structure Mining. In this paper we define Web Mining and present an overview of the various research issues, techniques and development efforts.

Keywords Data mining, Web mining, World-Wide Web(WWW)

1 引言

当前 WWW 正在深度和广度方面飞速地发展着，Internet 也正在前所未有地改变我们的生活。WWW 上的一些主要工作，例如 Web 站点设计、Web 服务设计、Web 站点的导航设计、电子商务等工作正变得越来越复杂和越来越繁重。

从站点经营方来说，他们需要好的自动辅助设计工具，可以根据用户的访问兴趣、访问频度、访问时间动态地调整页面结构，改进服务，开展有针对性的电子商务以更好地满足访问者的需求。从访问者来说，他们希望看到的是个性化的页面，希望得到更好的满足各自需求的服务。这种需求从某种意义上说，访问者本身也未必清楚。因此，欲解决这两方面需求的一个有利工具就是 Web 数据挖掘，即利用数据挖掘的思想和方法，将其利用到 Web 上，进行 Web 挖掘，挖掘出有用的信息。数据挖掘的对象将不仅是传统的关系数据库，而且也包括 WWW 上的各种有价值的信息。

ORACLE 公司最近推出的 Oracle 8i 数据库处理的数据已不仅是传统的关系数据，而且也包括各种 Web 页面数据。从这里我们也可以看出，数据挖掘的对象正在从传统的关系数据，进一步扩展到 Web 信息。

通过 Web 数据挖掘，我们可以从数以亿计存储大量多种多样信息的 Web 页面中提取出我们需要的有用的知识。

通过 Web 数据挖掘，对总的用户访问行为、频度、内容等的分析，可以得到关于群体用户访问行为和方式的普遍知识，用以改进我们的 Web 服务方设计，而更重要的是，通过对这些用户特征的理解和分析，可以有助于开展有针对性的电子商务活动。

通过 Web 数据挖掘，对每个用户访问行为、频度、内容等的分析，能提取出每个用户的特征，给每个用户个性化的界面，提供个性化的电子商务服务。

Web 数据挖掘的定义是：针对包括 Web 页面内容、页面之间的结构、用户访问信息、电子商务信息等在内的各种 Web 数据，应用数据挖掘方法以发现有用的知识来帮助人们从 WWW 中提取知识，改进站点设计，更好地开展电子商务。

2 Web 数据挖掘的分类

Web 数据挖掘总的来说分为内容挖掘、访问信息挖掘和结构挖掘 3 类。

2.1 Web 内容挖掘

Web 内容挖掘是对 Web 页面内容进行挖掘。它包括：

1)传统的从 WWW 上提取信息的搜索引擎，包括：Lycos, Vista, WebCrawler, ALIWEB, MetaCrawler。

2)新近的从 WWW 上更智能地提取信息的搜索工具，包括 Intelligent Web Agent, Information Filtering/Categorization, Personalized Web Agents。

3)数据库方法：把半结构化的 Web 信息重构得更结构化一些，然后就可以使用标准化的数据库查询机制和数据挖掘方法进行分析。

4)对 HTML 页面内容进行挖掘，对页面中的文本进行文本挖掘[1]，对页面中的多媒体信息进行多媒体信息挖掘[2]。包括对页面内容摘要、分类、聚类以及关联规则发现。

2.2 Web 访问信息挖掘(Web Usage Mining)

Web 访问信息挖掘是对用户访问 Web 时在服务

器方留下的访问记录进行挖掘[3~5]，即对用户访问Web站点的存取方式进行挖掘。挖掘的对象是在服务器上的包括Server Log Data等日志。挖掘的手段是：①路径分析；②关联规则和序列模式的发现；③聚类和分类。

Web访问信息挖掘可以从Web服务器那里自动发现用户存取Web页面的模式，得出群体用户或单个用户的访问模式和兴趣。

2.3 Web结构挖掘

Web结构挖掘是对Web页面之间的结构进行挖掘。整个Web空间里，有用的知识不仅包含在Web页面的内容之中，而且也包含在页面的结构之中。例如，如果我们发现一个论文页面经常被引用，那么，这个页面一定是非常重要的。发现的这种知识可以被用来改进搜索引擎。现有的方法有：PageRank[6]，CLEVER[7]。

3 Web数据挖掘对象

1 服务器日志数据 个人浏览Web服务器时，服务器方将会产生3种类型的日志文件[8]：Server logs，Error logs，Cookie logs，这些日志用于记录用户访问的基本情况。

1)Server log：

表1 Server log文件格式

Field	Description
Date	Date，time，and timezone of request
Client IP	Remote host IP and/or DNS entry
User name	Remote log name of the user
Bytes	Bytes transferred (sent and received)
Server	Server name，IP address and port
Request	URI query and stem
Status	http status code returned to the client
Service name	Requested service name
Time taken	Time taken for transaction to complete
Protocol version	Version of used transfer protocol
User agent	Service provider
Cookie	Cookie ID
Referrer	Previous page
…	…

2)Error log：存取请求失败的数据，例如：丢失连接、授权失败，或超时。

3)Cookie，由Web server产生的记号并由客户端持有，用于识别用户和用户的会话。Cookie是一种标记用于自动标记和跟踪站点的访问者。在电子商务的环境中，存储在Cookie log中的信息可以为交易信息。

通过对这三种日志的分析和挖掘，就可以开展访问信息挖掘。

2 在线市场数据 这是传统的关系数据库结构数据，用于电子商务站点，存储电子商务信息。

3 Web页面 满足HTML标准的Web页面，现有的Web数据挖掘方法大都是针对Web页面开展的。如2.1节所述的各种方法。

由于HTML页面包含文本和多媒体信息(包括图片，语音，图像)，所以涉及到数据挖掘领域中的文本挖掘和多媒体挖掘，现有的HTML页面内容缺乏标准的描述方式，难以挖掘。为了解决这个问题，1998年WWW社团提出了XML语言标准(eXtensible Markup Language)[3]。该标准通过把一些描述页面内容的标记(tag)添加到HTML页面中，用于对HTML页面内容进行自描述，例如对一个内容为科技论文的页面添加相关标记描述其作者、关键字等。XML的标记并不限制死，可由页面的创立者自己安排给出和定义，但要遵循一定的规范。

4 Web页面超链接关系 对Web页面之间的超链接关系是一种重要的资源，页面的设计者把他认为是重要的页面地址添加到自己的页面上，显然，如果一个页面被很多页面引用，那么它一定是重要的。这就是一种需要被挖掘的知识。

5 其它信息 这些信息主要包括用户注册信息等一系列信息。为了更好地实现挖掘任务，适当地附加信息(如描述用户的基本情况和特征的信息)是必要的。

4 Web数据挖掘方法

由于Web页面空间的不确定性，当要用传统的数据挖掘方法对Web进行挖掘时，在数据预处理阶段，需要把Web页面空间需要挖掘的特征组织到关系数据库二维表中，然后就可以用通用的数据挖掘方法进行处理。

4.1 Web内容挖掘方法

Web页面信息主要包括文本信息和多媒体信息，所以分为对Web页面文本信息的挖掘和对Web页面多媒体信息的挖掘。

4.1.1 对Web页面内文本信息挖掘 挖掘的目标是对页面进行摘要和分类。在对页面做摘要时，对每

一个页面应用传统的文本摘要方法可以得到相应的摘要信息。在对页面进行分类时，分类器输入的是一个Web页面集(训练集)，再根据页面文本信息内容进行监督学习，然后就可以把学成的分类器用于分类每一个新输入的页面。

对分类挖掘而言，在前处理阶段，要把这个Web页面集合文本信息转化成一个二维的数据表，其中列集为特征集，每一列是一个特征；行集为所有的页面集合，每一行为一个Web页面的特征集合。在文本学习中常用的方法是TFIDF向量表示法，它是一种文档的词集(bag-of-words)表示法，所有的词从文档中抽取出来，而不考虑词间的次序和文本的结构。这种构造二维表的方法是：

1)每一列为一个词，列集(特征集)为辞典中的所有有区分价值的词，所以整个列集可能有几十万列之多。

2)每一行存储一个页面内的词的信息，这时，该页面中的所有词对应到列集(特征集)上。列集中的每一个列(词)，如果在该页面中不出现，则其值为0；如果出现k次，那么其值就为k；页面中的词如果不出现在列集上，就说明该词不具有区分价值，可以被放弃。这种方法可以表征出页面中词的频度。

3)对中文页面来说，还需先分词然后再进行以上两步处理。

这样构造的二维表表示的是Web页面集合的词的统计信息，最终就可以采用朴素贝叶斯方法或k-最近邻方法进行分类挖掘。在挖掘之前，一般要先进行特征子集的选取，以降低维数。

4.1.2 对Web页面内多媒体信息挖掘 总的挖掘过程是先要应用多媒体信息特征提取工具，形成特征2维表，然后就可以采用传统的数据挖掘方法进行挖掘。

在特征提取阶段，利用多媒体信息提取工具进行特征提取。文[2]中信息提取工具能够抽取出image和video的文件名、URL、父URL、类型、键值表、颜色向量等。对这些特征可以进行如下挖掘操作：

1)关联规则发现：例如：如果图像是"大"的而且与关键字"天空"有关，那么它是蓝色的概率为68%。

2)分类：根据提供的某种类标，针对特征集，利用决策树可以进行分类。

4.2 Web访问信息挖掘方法

4.2.1 预处理任务 Web访问信息挖掘在预处理阶段主要的工作是从Server Log Data中识别事务，尔后就可以利用对事务数据库进行挖掘的方法，如关联规则和序列模式的发现等挖掘技术进行数据挖掘。预处理任务是这种挖掘方法的一个关键，它包括：

1)数据清洗：①消除不相关的项，如带有后缀gif，jpeg，zip，ps等。②判断在存取日志中，有无重要的存取未被记录下来。③判断用户是一个还是多个。

2)事务识别：在对Web日志数据进行数据挖掘之前，需要把对Web页的访问序列组织成逻辑单元以表征事务或用户会话。一个用户会话是在该用户访问一个站点时，访问的全部页的参照序列，在一个用户会话中，依赖于识别事务的标准，事务可以是一页，也可以是全部页。一种访问事务识别的方法为：

$$t = (ip_i, uid_i, \{(l_1^i.url, l_1^i.time), \cdots, (l_m^i.url, l_m^i.time)\})$$

其中，对于$1 \leqslant k \leqslant m, l_k^i \in L, l_k^i.ip = ip_i, l_k^i.uid = uid_i, l_k^i.time - l_{k-1}^i.time \leqslant C$。这里C是一个固定的时间窗。

对Log进行处理，找到每一个事务，然后就可以对这些事务进行关联规则和序列模式的发现。寻找访问事务的简单算法是：

①对日志进行预处理。

②根据每一个访问者IP，划分日志。即在Log中找到每一个访问者的访问记录集。

③对每一访问者的访问记录集，根据C进行分割，找到每一个访问者的每一次访问记录集，这时，每一个访问者的每一次访问记录集就构成了一个访问事务。

④最终所有的事务按时间戳有序构成我们进行关联规则发现的基础。

4.2.2 Web访问事务中的挖掘技术 一旦用户会话和事务已被识别，那么就可以用如下的挖掘技术进行挖掘。

1)路径分析：它可以被用于判定在一个Web站点中最频繁访问的路径。还有一些其他的有关路径的信息通过路径分析可以得出：

①70%的客户端在存取/company/product2时，是从/company开始，经过/company/new，/company/products，/company/product1。

②80%的客户存取这个站点是从/company/products开始的。

③65%的客户在浏览4个或更少的页面后就离开了。

利用这些信息就可以改进站点的设计结构。

2)关联规则和序列模式的发现：使用关联规则发现方法从Web访问事务中，我们可以找到如下的相关性：

①40%的用户访问Web页面/company/product1时，也访问了/company/product2。

②30%的客户在访问/company/special时，在/company/product1进行了在线定购。

利用这些相关性，我们可以更好地组织站点内的Web空间，实行有效的市场战略。

在时间戳有序的事务集中，序列模式的发现就是指找到那些如"一些项跟随另一个项"这样的内部事务模式。例如：

①在访问/company/products的顾客中，有30%的人曾在过去的一星期里用关键字w在yahoo上做过查询。

②在/company/product1上进行过在线定购的顾客中，有60%的人在过去15天内也在/company/product4处，下过订单。

发现序列模式，能够便于预测用户的访问模式，有助于开展针对这种模式的有针对性的广告服务。依赖于发现的关联规则和序列模式，能够在服务器方动态地创立特定的有针对性的页面，以满足访问者的特定需求。

3)分类和聚类：发现分类规则可以给出识别一个特殊群体的公共属性的描述。这种描述可以用于分类新的项。例如：

①政府机关的顾客一般感兴趣的页面是/company/product1。

②在/company/product2进行过在线定购的顾客中有50%是20—25岁生活在西城区的年轻人。

聚类分析可以从Web访问信息数据中聚集出具有相似特性的那些客户。在Web事务日志中，聚类顾客信息或数据项能够便于开发和执行未来的市场战略。这种市场战略包括：自动给一个特定的顾客聚类发送销售邮件，为一个顾客聚类动态地改变一个特殊的站点等。在PageGather方法[10~12]中，采用聚类技术以发现在一起被访问的Web页面，并把它们组织到一个组里，以帮助用户更好地访问。

4.3 Web结构挖掘方法

在设计搜索引擎时，一种新的方法PageRank对Web页面的链接结构进行挖掘以得出有用的知识。Web页面的链接类似学术上的引用，因此一个重要的页面可能会有很多页面的链接指向它，也就是说，如果有很多链接指向一个页面，那么它一定是很重要的。

在搜索引擎中存贮了数以亿计的页面，很容易得到它们的链接结构。需要做到的是寻找一种好的利用链接结构来评价页面重要性的方法。在PageRank方法中PageRank被定义为：设u为一个Web页。F_u为所有u指向的页面的集合，B_u为所有指向u的页面的集合。设$N_u=|F_u|$为从u发出的链接的个数，$c(<1)$为一个归一化的因子(因此所有页面的总的PageRank为一个常数)，那么u页面的PageRank被定义为(简化的版本)：

$$R(u)=c\sum_{v\in B_u}R(v)/N_v$$

即一个页面的PageRank被分配到所有它所指向的页面中；每一个页面求和所有指向它的链接所带来的PageRank以得到它的新的PageRank。该公式是一个递归公式，在计算时可以从任何一个页面开始，反复计算直到其收敛。

对于搜索引擎的键值搜索结果来说，PageRank是一个好的评价结果的方法，查询的结果可以按照PageRank从大到小依次排列输出。CLEVER方法本质上和PageRank方法是一致的。

4.4 其他一些挖掘方法

文[13]提出一种能够从Web中发现迁移模式的模型：g-序列，并给出相应的挖掘工具WUM[14]。

文[15]给出WEBKB系统，是一个可训练的信息抽取系统，输入为：1)定义好的感兴趣的类和关系；2)已标注的具有这些类和关系的训练例。给定这些输入，系统学习完成后，就可以用于自动创建一个计算机可理解的WWW知识库。

文[16]提出一种新的在电子商务环境下对Web访问信息和在线零售信息进行挖掘的方法，用于发现行销智能。

5 研究意义和方向

Web数据挖掘在当前是前沿的研究领域，是把Internet，WWW和数据挖掘结合起来的一种新兴的技术，国外对此也处于刚起步阶段，几个非常有用的研究方向是：

1)聚类分析在Internet DVB接入过程中，动态地构造和组织DVB页面广播内容的研究。

2)路径分析在建立Internet虚拟社区的研究。

3)关联规则和序列模式的发现在构造自组织站点的研究。

4)分类在电子商务市场智能提取中的研究。

Web数据挖掘的应用领域非常广阔，不但涉及页面信息提取、站点分析、设计，而且在即将广阔蓬勃发展的基于Internet的电子商务方面也会有良好的应用前景。例如我们现在正在把聚类方法应用到DVB页面中的广播中，已取得了一定的研究成果。

参 考 文 献

1 Mladenic D. Machine Learning on non-homogeneous, distributed text data, doctoral dissertation. University of Ljubljana, 1998

2 Osmar R Z. Resource and knowledge discovery from the

(下转第41页)

```
    END
    Text_Summary=""
    FOR I=1 To n DO IF t[I] THEN Text_Summary=
        Text_Summary+s[I]
END.
```

结束语 本文提出的文本挖掘模型是建立在文本特征提取基础上的，面对大量的电子文本，良好的响应能力和不依赖具体领域的特性是首要关注的方面。其基本结构如图1。

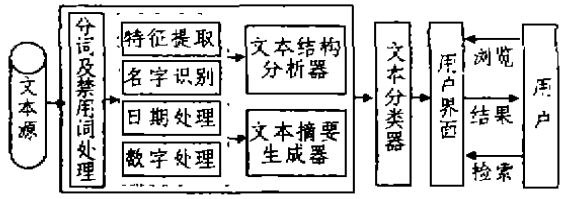

图1 中文文本挖掘模型结构示意图

在文本挖掘模型中，本文提出了基于概念的文本分类方法，初步解决了概念的影射和概念消歧问题。在文本特征提取方面，提出了一般特征项的筛选方法和中文姓名的识别算法以及基于模糊语义的数字特征（日期、时间、数字和货币等）表示方法，在文本摘要方面，提出了基于统计的摘要生成算法。面对海量文本，采用统计方法有着较强的适应性和良好的反映能力，不依赖于具体领域知识，但是，随着需求的深入，引入基于自然语言理解的语法、语义、语用分析势在必行，以便挖掘更深入的知识。但如何协调适应性和精确性的关系，文本的来源多样化与领域知识库的关系是关键问题，也是下一步重点。

参考文献

1 姚天顺，等. 自然语言理解. 清华大学出版社，1995
2 麻志毅，林鸿飞，姚天顺 基于情境的文本中时间信息分析. 东北大学学报，1999，13(6)
3 吴立德，等. 大规模中文文本处理. 复旦大学出版社，1997
4 刘开瑛，等. 中文文本中抽取特征信息的区域与技术，中文信息学报，1998，12(2)
5 Udo Klemens, Schnattinger. Deep Knowledge Discovery from Natural Language Texts. In Proc of the 3rd Conf on Knowledge Discovery and Data Mining. 1997，175～178
6 Salton G, et al. Automatic Structuring and Retrieval of Large Text Files. Communications of the ACM，1993，37(2)：97～108

（上接第31页）

internet and multimedia repositories, doctoral dissertation. University of Simon Fraser，1999
3 Cooley R, et al. Web Mining: Information and Pattern Discovery on the World Wide Web. In: Proc. of Intl. Conf. on Tools with Artificial Intelligenc, Newport Beach, IEEE，1997
4 Cooley R, et al. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. In: Proc. of Knowledge and Data Engineering Workshop. Newport Beach, CA, IEEE. 1997
5 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems，1999，1(1)
6 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: 7th Int. Conf. WWW. Brisbane, Australia, April 1998
7 Chakrabarti S, et al. Experiments in topic distillation. In: ACM SIGIR Workshop on Hypertext Information Retrieval on the Web. Melbourne, Australia，1998
8 Luotonen A. The common log file format. Available at: http: // www.w3.org/pub/WWW/，1995
9 World Wide Web Consortium. Available at: http: //www.w3.org/XML/，1998
10 Perkowitz M, Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages. In: Proc. of AAAI-98. 1998
11 Perkowitz M, Etzioni O. Adaptive web sites: an AI challenge. In: Proc. 15th Int. Joint Conf AI. 1997a
12 Perkowitz M, Etzioni O. Adaptive web sites: Automatically learning from user access patterns. In: Proc. of the Sixth Int. WWW Conference. 5 1997b
13 Spiliopoulou M. The laborious way from data mining to web mining. Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web", Mar. 1999
14 Spiliopoulou M, Faulstich L C. WUM: A Web Utilization Miner. 1998
15 Craven M, et al. Learning to Extract Symbolic Knowledge from the World Wide Web. Same to [10]
16 Buchner A. et al. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. SIGMOD Record，1998，27(4)
17 Fink J, et al. User-oriented Adaptivity and Adaptability in the AVANTI Project. In Designing for the Web: Empirical Studies. 1996
18 Wexelblat A, Maes P. Footprints: History-rich web browsing. In: Proc. Conf. Computer-Assisted Information Retrieval(RIAO). 1997. 75～84
19 Chen M S, et al. Data mining for path traversal patterns in a web environment. In ICDCS, 1996. 385～392