WEB Usage Mining:Web

使用信息挖掘的关键技术

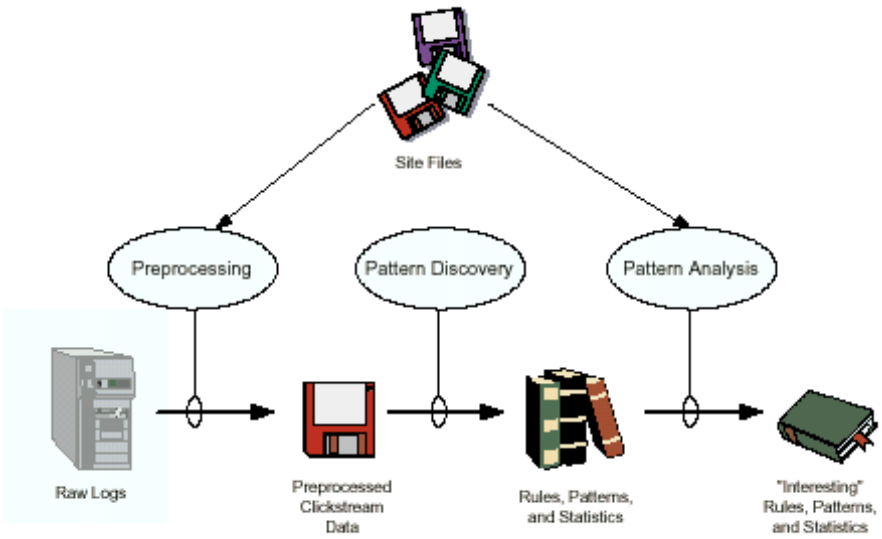
南京政治学院上海分院电教中心 赵亮

摘要:随着网络站点越来越多、站点结构越来越复杂，同时用户类型和用户需求也呈现多样性的发展趋势，从教育技术的角度看，要求信息查询服务向获取信息最优化方向发展。如何最大限度的减少用户访问的盲目性和工作量，让其“所想既所得”，是每一个系统设计人员必须解决的问题。为了体现智能性和以人为中心的原则，有必要从WEB数据中抽取出使用模式。目前国外对个性化服务和自适应网站设计等相关领域的研究正在快速发展，其中的核心技术就是WEB使用信息挖掘 （Web Usage Mining），它将数据挖掘、人工智能、图论等技术应用于WEB数据,追踪用户的访问特性，从中抽取出用户的使用模式，以便更好的为用户服务。在我们设计的非书资料管理系统中，这是关键的模块之一。我们将它分为三部分处理：预处理(Preprocessing)、模式发现(Pattern Discovery)、模式分析(Pattern Analyse)，本文概略描述了这三部分的工作，给出其潜在的巨大应用。

关键字：数据挖掘、智能、知识发现

一、介绍

在网站越来越多、站点结构越来越复杂，用户类型和用户需求越来越多样性的情况下，最大限度的提高用户访问的效率，让其“所想既所得”，提高站点的智能性的要求越来越迫切了，目前个性化服务和自适应网站设计等相关领域的研究正在快速发展，它们能够根据用户的特性、爱好自动提供相关信息，定制网页，重新构造站点，Web使用信息挖掘 （Web Usage Mining）是其中的关键技术，它被用来跟踪用户的访问特性，提取用户的使用模式。图一显示了Web使用信息挖掘的主要简化步骤：



图一：Web使用信息挖掘的主要简化步骤

二、WEB数据

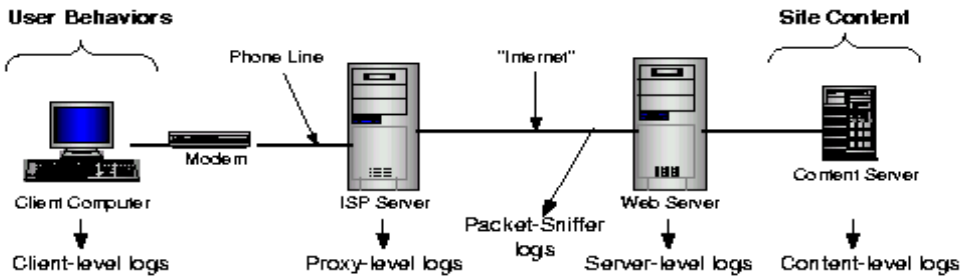
数据库中的知识发现的一个关键步骤是为执行数据挖掘任务而找到一个合适的目标数据集。在WEB挖掘中，数据

能够从几个方面采集：服务器端、客户端、代理服务器端。依据数据源位置、数据种类、实现方法的不同，每种数据的采集也不同。有许多种类的数据能够被应用到WEB挖掘中，本文将这些数据概括为以下几种：

- 内容（Content）：WEB页中的真实数据，即WEB页中被设计用来传送给用户的数据，它通常由文本和图形组成，但并不仅仅局限于此。
- 结构（Structure）：描述内容组织的数据，包括不同的Html的排列或在一个给定页面中的Xml标记。它能够被描绘成一个树结构，结构的主要信息是从一个页到另一个页的超连接。
- 使用信息（Usage）：描述WEB页的使用模式的数据，例如：IP地址、页的引用、访问时间等。
- 用户轮廓（User Profile）：关于网站用户的信息数据，包括注册信息和用户轮廓信息。

1、数据源

从不同来源采集的使用信息（Usage Data）描述了不同WEB段的访问模式，它涵盖了单用户，单站点的浏览行为到多用户，多站点的访问模式。图二显示了WEB访问中不同的来源的使用信息采集：



图二:WEB访问中不同的来源的使用信息采集

（1）、服务器端(Web Server)采集

一个WEB服务器日志是一个执行网络使用挖掘的重要来源，它记录了站点访问者的浏览记录,这些日志文件能够以常用日志格式或扩展日志格式存在。

大部分的主要工作就是对这些日志文件进行的。图三给出了扩展日志格式的例子。包侦测技术（packet sniffing）也是通过服务器日志文件获取使用信息的一种手段，它通过侦测到WEB服务器的网络流量直接从TCP/IP中抽取使用信息，

可见图二。

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

图三： 扩展日志格式

（2）、客户端采集(Client side)

客户端数据采集能够通过以下的方式来实现：

◆ 通过用一个远程代理，（如：javascript 或者 java applets）。

- 通过修改已经存在的浏览器（如：mosaic 或者mozila）以扩展它的数据采集的能力。

客户端数据采集方法需要用户的协作，或者实现javascript 和java applets的功能或者自觉的使用修改过的浏览器。

（3）、代理端(Proxy)采集

一个WEB代理在浏览用户和WEB服务器端充当一个缓冲中间件，缓冲代理能被用来减少读取时间以及减少网络流量。缓冲代理的执行效果依赖于它们对下一步用户将访问内容的预见性。代理记录了许多从多个用户到多个站点的HTTP请求，在一群共享一个代理服务器的匿名用户中，若要提取他们的共有特性时，这可以被用来当作一个数据源。

2、数据模型

上面提到的几种数据源，它们所包含的信息都能被用来构建一个数据模型，它包括以下几种数据提取：

- users(用户)：访问站点的个体。
- episodes(事件)：一次server sessions所产生的 page views的子集。
- server sessions(服务进程)：一个用户对一个站点的某次访问的click-stream 。

- click-streams(点击流)：被一个用户访问的所有page views的序列。
  - page views(页视图)：一个用户鼠标点击一次在页面上呈现的所有文件。

更详细的定义可以在W3C Web Characterization Activity (WCA) http://www.w3c.org/wca中查到。

三、WEB使用信息挖掘（Web Usage Mining）

正象图一显示的，处理WEB使用信息挖掘或WEB使用信息分析主要包括三个主要任务：预处理、模式发现、模式分析。

1. 预处理

主要指将包含在不同的数据源（usage、content、structure）中的信息转化为数据模型（users、episodes、server sessions、click-streams、page views），以便为模式发现作好准备。

（1）、Usage Preprocessing（使用信息预处理）

使用信息预处理的目标是将输入的服务器日志文件（以日志格式或扩展日志格式存在）转换为服务进程（server sessions）。当然，正向前面所说，数据源也可以是其它的应用日志文件或者包侦测（packet sniffing）。

从使用信息中获得服务进程（server session）包括4部分的内容：

- 数据清洗（data cleaning）。
- 用户/进程确定（user/session identification）
- 页视图确定（page view identification）
- 路径完整性（path completeion）

通常，一个服务进程（S）是一系列以时间为顺序的页视图（V）（单个用户在某次访问一个站点时所产生），以及一些元数据（A）。

$$S=[A:V_1,....V_n]$$

$$V=<v_i,h_j,t_f,t_1,t_e,\{d_1,....d_m\},c>$$

$$A=\{a_1,....a_k\}$$

每个页视图V由一个标志符v_i,页文件h_j，首先访问时间t_f，最后访问时间t_1，视图结束时间t_e以及一组任意值（$d_1,....d_m$），c是一个布尔操作符，用来判定该页视图是否在路径完整性处理过程中。

元数据A，依据所用的输入源的不同而不同，对于扩展日志格式文件，进程元数据将包括IP地址、代理、Cookie、嵌入的进程ID号等。

因为可用数据的不完整性，在WEB使用信息的挖掘中使用信息预处理是比较困难的任务。除非可以利用一套客户端的跟踪机制，否则仅仅用IP地址、代理、服务器端点击流（click-stream），只能用来确定users(　　用户)和server sessions（服务进程）。

（2）、Content Preprocessing（内容预处理）

内容的预处理将文本、图形、script类型文档以及其它文件如多媒体等转换为WEB使用信息挖掘过程可以使用的形式。常常用到分类（classification）和聚类（clustering）的方法，除了将页面依据主题分类或聚类外，页面还可以依据某种特定用途来分。为了对页面使用内容预处理算法，信息必须被转换成量化的形式，目前已经存在一些向量空间模型的算法可以被用来处理这步工作。

文本文件可以被拆分成字向量。关键特征或文本描述可以被用来描绘图形或多媒体。通过解析HTML和重新组织信息，静态页面的内容可以被很容易预处理，

动态页面的处理要复杂的多。内容服务器（content server）应用个性化技术和数据库技术来构建可以被预处理的页面。

（3）、Structure Preprocessing（结构预处理）

通过页面之间的超文本联结，形成了一个站点的结构。结构的预处理与站点内容的预处理类似。在结构动态的情况下，每一次*server session*都会产生一个不同站点结构。

1. 模式发现

模式发现利用了统计、数据挖掘、机器学习和模式认知等技术。

（1）、统计分析

统计技术是分析用户行为的最常用的方法，例如：可以通过求出现频率、平均、求中值等多种统计方法来分析最常访问的页面、平均访问时间、浏览路径的长度。缺点是这种分析缺乏深度。

（2）、关联规则

通过分析用户的访问页面之间的潜在联系，可以归纳出关联规则，如同买榔头的用户有可能同时买钉子一样，这里页面A->页面B（and C...），即只要访问页面A就有可能访问B（和C...），在WEB使用信息挖掘中，关联规则指：只要页面的支持度大于某个被给定的阀值，那么这些页面就都被访问。这里这些页面不允许用超联结直接连接起来。例如，用Apriori 算法可能发现在访问

电子产品页面的人和访问运动器材的人之间的一个联系。关联规则能够有助于WEB设计者重新组织站点的内容编排。

（3）、聚类

聚类是一种将具有相同共性的项集聚集在一起的技术。在WEB使用信息挖掘领域，涉及两种聚类：使用用户聚

类和页聚类。

用户聚类将具有相似访问特性的用户归在一起，在站点的个性化服务中，这种技术尤其有用。

页聚类将内容相关的页面归在一起，在搜索引擎和WEB助理的设计领域中，这种技术发挥着巨大作用。

（4）、分类

与聚类不同，分类将项集分门别类地归入预先设定好的几个类中。在WEB使用信息挖掘领域，分类主要在于发展属于特定的类的用户模型。这要求能抽取出最能反映一个给定的类的特性。通过用诱导学习机制，可以进行分类的过程，如：决策树、朴素贝叶斯算法、支持向量机等。例如，对服务器日志分类可能导出如下有趣的规则：30%的访问/product/music的用户是18-25年龄段，并且生活在西海岸。

（5）、顺序模式

顺序模式技术试图找出页面依照时间顺序出现的内在模式。

通过应用这种技术，可以预测将来的访问模式。其它类型的基于顺序模式的时间分析方法包括趋势分析或转换点检测技术。

趋势分析能够被用于随着时间的推移，发现站点使用模式的改变。

转换点检测技术可以用来确定转变何时发生。

例如：1、在过去的两个月内，对站点A的访问已经减少了

2、1-3月，对站点A的访问量增加，直到5月开始稳定，现在开始减少。

（6）、依赖模型

依赖模型是另一个有用的模式发现的任务。目的是在WEB领域的不同变量间发展出一个能够描绘他们之间依赖关系的模型。例如，可以建立一个模型来描述用户在一个网上商店浏览的不同阶段（如：从一个偶然的访问者到一个认真的潜在客户）。有几种概率学习技术能够被用来给用户浏览行为建模，如markov 模型、贝塞尔信赖网络等。

建立WEB使用信息模式的模型不光为分析用户的行为提供一个理论框架，而且在预测未来WEB资源的消耗方面也很有用，这样的信息有助于提高用户的浏览效率。

2. 模式分析

模式分析是WEB使用信息挖掘的最后步骤（见图一）。它的目的是对模式发现过程产生的规则和模式进行过滤，从中滤除不感兴趣的部分。最常见的模式分析包括知识查询机制如SQL，另一个方法与数据仓库有关，对使用信息进行联机分析处理（OLAP），同时，也需要用到可视化技术（Visualization）。

参考文献:

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB conference, pp. 487-499, Santiago, Chile, 1994.

[2] A. Buchner and M. D. Mulvenna. Discovering internet marketing intelligence through online analytical Web usage mining. SIGMOD Record, (4) 27, 1999.

[3] E. Charniak. Statistical language learning. MIT Press, 1996.

[4] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, (1) 1, 1999.

[5] R. Cooley, P-T. Tan., and J. Srivastava. WebSIFT: The Web site information filter system. In Workshop on Web Usage Analysis and User Profiling (WebKKD99), San Diego, August 1999.

[6] E-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. More. Document categorization and query generation on the World Wide Web using WebACE. Journal of Artificial Intelligence Review, January 1999.

[7] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. To appear in Proceedings of the 1999 Conference on Research and Development in Information Retrieval, August 1999.

[8] E-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraphs. In Proccedings of SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97), May 1997.

[9] E-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Hypergraph based clustering in high-dimensional data sets: a summary of results. IEEE Bulletin of the Technical Committee on Data Engineering, (21) 1, March 1998.

[10] G. Karypis, E-H. Han. Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical Report #00-016, Department of Computer Science and Engineering, University of Minnesota, March 2000.