

Web 应用数据挖掘可视化界面布局 的设计方法

米娜瓦尔·努拉合买提¹, 玛依拉·别克强塔依娃¹, 朱 静¹,
张太红¹, 曾 明², Osmar. R. Zaiane³

(1. 新疆农业大学 计算机与信息工程学院, 新疆 乌鲁木齐 830052;

2. 西安交通大学 软件学院, 陕西 西安 710049;

3. 加拿大阿尔伯塔大学 计算机科学系, 埃德蒙顿 T6G 2E1)

摘 要:在对 Web 应用挖掘的基本步骤作系统性研究的基础上, 设计了一个 Web 应用挖掘可视化系统。该系统能够对用户访问 Web 时服务器方留下的访问记录进行挖掘, 从中得出用户的访问模式和访问兴趣, 并对所得出的结果进行可视化的处理。为了识别用户浏览模式利用 Apriori 算法对 Web 应用挖掘过程中预处理阶段所产生的用户会话文件进行了挖掘。采用 Web 图可视化了 Web 站点的拓扑结构以及各节点访问计数和登录计数信息。Web 图的新颖之处在于两点: 首先, 为了将 Web 拓扑结构映射到 Web 图上, 利用了站点拓扑结构数据和站点应用数据; 其次, 在绘制表示用户登录计数的信息层时允许通过使用动态布局的方法, 以及为每一层的节点重新分配 360 度周长的方法来解决节点之间的冲突问题。文中较详细地阐述了该系统对 Web 应用数据挖掘可视化界面布局的具体措施。

关键词: Web 可视化; Web 图; 磁盘树; 路径图; 模式

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2011)08-0030-04

Designing Method on Visual Web Usage Mining Interface Layout

NULAHEMAITI · Mi-nawaer¹, BIEKEQIANGTAYIWA · Ma-yila¹, ZHU Jing¹,
ZHANG Tai-hong¹, ZENG Ming², Osmar. R. Zaiane³

(1. Computer and Information Engineering Institute, Xinjiang Agricultural University, Urumqi 830052, China;

2. Software Engineering School, Xi'an Jiaotong University, Xi'an 710049, China;

3. Department of Computer Science, Alberta University, Edmonton T6G 2E1, Canada)

Abstract: Underlying the systematic studies on the basic steps of Web usage mining, implemented a visual Web usage mining system, which is a system mainly used to mine the Web log access file that acquired from the Web Server, get the user visiting patterns and visiting interests and then visualize the mining results. In order to identify the navigational patterns of Web site visitors, Apriori algorithm is used on the mining of the user session file that has been generated after the data pre-processing process on the Web log file. In order to effectively visualize the topology of a Web site and the visiting numbers and login numbers of per page, the Web image is used. The novelty of the system is two: Firstly, adopt a new strategy for collapsing the Web topology into a tree structure exploiting Web usage data, in addition to the structure of the site. Secondly, deal with the problem of occlusion by using a dynamic layout that redistributes the nodes of each level along the 360 degree perimeter when visualized information layer that represents the user login numbers. It illustrates in detail concrete measures that have applied on visual Web usage mining interface layout.

Key words: Web visualization; Web image; disk tree; path image; pattern

0 引 言

伴随着 Internet 技术的发展, 网络资源同时迅速增

长, 其重要性也越来越被人们所注意到。客户浏览行为数字化, 使得通过收集大量用户行为数据来深入研究客户行为变成可能^[1,2]。如何利用这个机会, 从这些“无意义”的繁琐数据中找出大家都看得懂的、有价值的知识和信息, 并将其以用户容易理解的, 如图形、自然语言和可视化技术等方式表达出来, 成为目前面临的最紧要的问题之一^[3,4]。解决这个问题的办法之

收稿日期: 2011-01-21; 修回日期: 2011-04-27

基金项目: 新疆维吾尔自治区电子信息发展专项资金项目(XJZZXZJ20109)

作者简介: 米娜瓦尔·努拉合买提(1970-), 女(维吾尔族), 硕士, 讲师, 研究方向为 Web 数据挖掘可视化。

一就是应用 Web 数据挖掘可视化技术,即通过挖掘服务器中的日志文件,来得到用户的访问模式,并对其进行可视化处理,从而得到对改进网站的结构和服务都有用处的信息^[5~7]。

可视化的基本思想就是将隐藏在大量数据中的信息用相对直观、易于领会的图形和图像来表征,从而加快获取信息的速度。可视化技术的目标是帮助人们增强认知能力及决策能力。基于计算机的可视化技术不仅仅把计算机作为信息集成处理的工具,用计算机图形和其他技术来考虑更多的样本、变量和联系,更多的是作为与用户之间的一种交流媒介。可视化在认知激励和用户认知之间建立起一个反馈环,运用人类认知的知识,对数据进行更深入的观察和分析。通过可视化技术可避免观察出不正确模式,而采取错误的决策和行动。

Web 站点结构、Web 站点应用数据以及用户浏览行为模式的可视化工具,使人们能够更好地理解站点事件或对某一 Web 站点进行修改和设计后的结果进行分析,因此具有能够说明 Web 站点的应用状态以及 Web 站点上的活动的可视化图形是很必要的。用可视化工具表示所发现的模式,并将它们映射到其 Web 站点结构图之上,对评估和解释 Web 挖掘数据结果将会很有帮助^[8,9]。

1 Web 应用挖掘可视化图形界面布局设计

合理设计表示 Web 应用挖掘过程中所发现知识的可视化图形布局是 Web 应用挖掘可视化研究的一个重要部分^[10,11]。合理布局不仅可以清楚地观察所统计到的信息和挖掘到的知识,而且可以给用户提供更有价值的信息。目前针对 Web 信息可视化挖掘有一些布局方法。系统通过磁盘树^[10]布局法对 Web 站点的拓扑结构进行了可视化,Web 拓扑结构指的是准备挖掘的网站中各个 Web 页之间的组织关系。Web 站点的拓扑结构图的可视化方法:绘图面板中用黑色的实心圆点来表示页面节点;通过直线来连接有链接关系的两个页面节点。层次数为 2 的页面的表示方法:表示根节点(挖掘开始的页面,多数情况下为主页)被放在面板的中心,把表示与其有链接关系的页面的节点通过平均分配 360 度圆的周界放到中心节点的周围。层次数为 3 的页面的表示方法:代表第一层页面的节点被放在中间,代表第二层页面的节点(主页中的超链接节点)放在中心节点的周围,以及第三层页面的节点(各个页中的超链接节点)也是通过平

分 360 度圆的周界的方法放到了位于第二层的各自的父节点的周围。其布局效果如图 1 所示。

系统采用 Web 图可视化了通过挖掘 Web 应用数据后所得到的 Web 各节点登录计数和访问计数。Web 图是将表示 Web 应用数据的路径图映射到 Web 站点的拓扑结构上的一个复合图形。其中登录计数指的是用户直接通过该节点的 URL 进行登录的次数;访

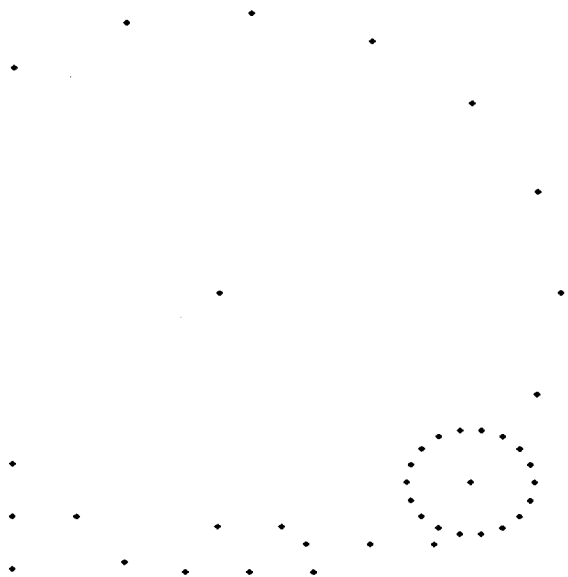


图1 层次数为3的Web拓扑结构图

问计数指的是通过某一节点中的超链接间接访问节点的次数。路径图是指用户对站点中各个 Web 页的直接登录或间接访问次数的图形化表示。由于被直接登录或通过超链接被间接访问次数较多的 Web 页在 Web 图的可视化中显示的也较突出,就像人间小径走的人多了也就成了路,因此被命名为路径图。用路径图中的这些视觉线索可获得不同的信息,诸如:哪些 Web 页被直接登录的次数多?哪些 Web 页被间接访问的多?哪些 Web 页较热?哪些 Web 页较冷?哪些 Web 页可能存在关联关系^[12]?

1.1 设计方法与措施

考虑到界面的有限性要采取以下措施达到充分利用绘图空间和尽量避免页面链接超过几层之后,其可视化的图形不断扩大所产生的图形之间的冲突问题;系统可对这种情况通过以下措施应对:

措施1:为了避免由于不同的直线厚度和节点尺寸而发生的冲突,通过删除登录计数和访问计数小于某个阈值的边和节点达到更加充分利用空间的目的。

措施2:指定某一个感兴趣的起始页,系统会将该页作为根节点,在较宽阔的空间重新绘制给定层次数范围内的路径图。

措施3:考虑到界面的有限性,显示的层数较多时,对连接父节点和子节点的线段的长度乘系数进行

调整。

措施 4: 为避免绘图界面因显示各节点的 URL 以及登录计数而变得杂乱不堪, 各个节点的 URL 以及登录计数的显示是动态的。当用户用鼠标指向某一个节点时才显示该节点的相关信息。

1.2 节点位置的计算公式

根据层次数的不同, 各个节点位置的计算公式也有所不同:

假设要在长和宽为 l 像素的正方形区域绘制 Web 图。主页或给定的一个初始页的根节点, 要放置在该正方形的中心。用以下公式计算根节点(第一层节点)的坐标:

$$\begin{aligned} x_1 &= \frac{l}{2} \\ y_1 &= \frac{l}{2} \end{aligned} \quad (1)$$

层次数为 2 时, 根节点的所有子节点要平均分配正方形内切圆的 360 度周长。计算层次数为 2 的子节点坐标的表达式为:

$$\begin{aligned} x_2 &= \frac{l}{2} \times 0.8 \times \cos\left(\frac{2\pi}{j} \times i\right) \\ y_2 &= \frac{l}{2} \times 0.8 \times \sin\left(\frac{2\pi}{j} \times i\right) \end{aligned} \quad (2)$$

式(2)中变量 l 为正方形绘图区的边长, 变量 j 表示层次数为 2 的节点个数; i 为循环变量, 其初值为 1,

终值为位于第二层的节点总数, 这些节点要平分 360° 周长; 为使绘图空间能够容纳各层所有节点, 要对正弦和余弦函数的函数值乘系数进行调整。公式中 0.8 为系数, 系数会随层次数发生变化。

当层次数为 3 时, 层次数为 1 和 2 的节点位置的计算公式同上。计算层次数为 3 的子节点坐标的表达式为:

$$\begin{aligned} x_3 &= \frac{l}{2} \times 0.15 \times \cos\left(\frac{2\pi}{k} \times h\right) + x_2 \\ y_3 &= \frac{l}{2} \times 0.15 \times \sin\left(\frac{2\pi}{k} \times h\right) + y_2 \end{aligned} \quad (3)$$

式(3)中 l 同上, k 为层次数为 2 的当前某个节点的子节点总数; h 为循环变量, 其初值为 1, 终值为层次数为 2 的当前某个节点的子节点个数; 0.15 为系数; x_2, y_2 为某个层次数为 3 的节点的父节点的中心坐标。

1.3 关联规则挖掘结果显示方法

通过 Web 图基本可以获得节点之间的关联线索, 这些线索隐含用户在访问某个节点时还可能会访问的节点信息, 即: 用户的访问兴趣。为获得更确切的信息, 采用了通过动态选择想要挖掘关联关系的节点和提供其最小支持度和置信度的方法进行挖掘。由于不能在绘图区显示太多信息, 在显示关联规则挖掘结果时, 可采用两种方法: 用图的形式或用文本信息的形式来描述关联规则的结果, 用户可以通过选项卡选择不

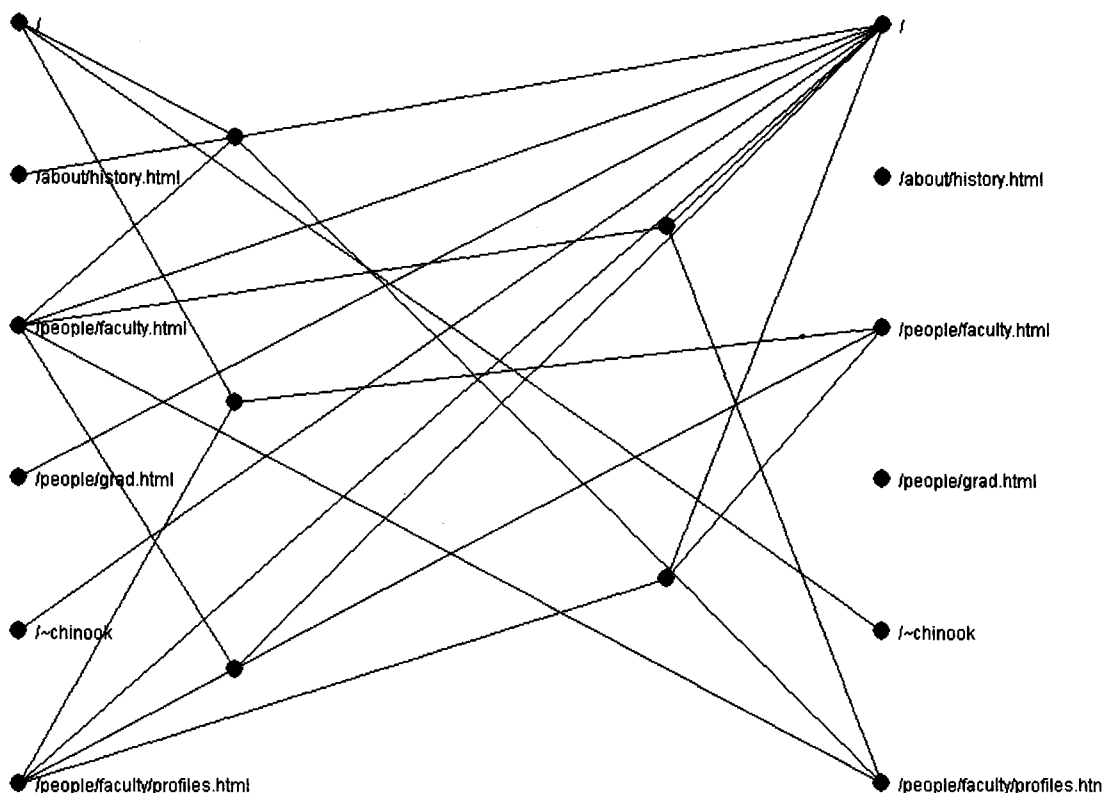


图 2 关联规则挖掘结果布局图

同的显示方式。

1.3.1 用图的形式来描述关联规则的结果

关联规则挖掘结果布局图如图2所示:其中第1列和第4列中的实心圆节点为用户选中的节点。第2列和第3列中的实心圆节点用于连接有关联关系的网页节点。

由于有关联关系的节点可能存在有四种情况:一对一、一对多、多对一以及多对多。其绘制方法为:首先将绘图面板平分为 m 行 n 列,行数 m 等于被选节点的个数,列数 n 等于4。为了用直线连接有关联关系的各节点,并且使这些直线之间不相互重叠,把被选节点以实心圆的形式绘制在第一列和第四列,这两列对应位置的实心圆表示同一个节点。第二列和第三列的实心圆节点,主要用于将具有关联关系的各组节点用直线拧结在一起,节点的个数等于挖掘出的关联规则的组数。

(1) 一对一适用环境。

如果第一列的一个节点只和第四列的一个节点有关联,就用一对一布局:直接用直线连接第一列和第四列中有关联的两个节点。

(2) 多对一适用环境。

如果第一列中有两个以上的节点跟第四列的一个节点有关联关系,就用多对一布局:用第二列节点拧结第一列中的多个节点和第四列的一个节点。

(3) 一对多适用环境。

如果第四列中有两个以上的节点跟第一列的一个节点有关联关系,就用一对多布局:用第三列节点拧结第四列中的多个节点和第一列的一个节点。

(4) 多对多适用环境。

如果第一列中有多个节点跟第四列的多个节点有关联关系,就用多对多布局:用第二列的节点拧结第一列的多个节点,用第三列的节点拧结第四列的多个节点,最后用一条直线连接第三列和第四列的节点。

1.3.2 用文本信息的形式来描述关联规则的结果

以文本的形式显示:为了能够确切地知道有关联关系的各节点的确切URL、节点之间的支持度与置信度的值,还提供了以文本的形式显示关联规则挖掘结果的方式。

2 结束语

通过Web图可以把用户的浏览数据放在其Web结构上下文中,是Web站点或其子集结构的树型表示,也就是说:Web图中蕴含Web拓扑结构图和路径图。系统采用Disktree(磁盘树)有效地对Web图进行可视化,系统的登录计数和访问计数的可视化采用的是路径图。

该系统可视化界面中的信息蕴含在信息层和模式层中:

信息层为搭设在Web图上的Web抽象数据的逻辑集合。用不同的信息层来表示不同的Web应用信息。该系统的Web图中蕴含了2个不同的信息层:

(1)登录计数层:通过该信息层可以获得每个页面节点的登录计数信息,该系统通过大小不同的实心圆来可视化该信息层。形状越大的节点,表示直接通过该节点登录的人数也越多。

(2)应用链接层:显示的是通过父节点中的超链接访问其各子节点的人数的统计计数;使用诸如直线的厚度(粗细)来可视化该信息层,如果连接某两个节点的直线较厚,表示通过直线一端的页面节点中的超链接访问直线另一端节点的人数就越多。

模式层:模式层中蕴含的是利用关联规则挖掘算法所发现的模式,可以以一种较容易理解的方式对挖掘过程中所发现的有关联关系的节点进行可视化。

利用Web应用数据挖掘结果可以改进Web站点的访问效率。面对广大用户改进Web服务器性能的一个重要手段是使Web服务器能够进行推送服务。一旦一个节点被选中,那么服务器就可以向用户推送跟其有关联的节点,以提高用户获取信息的效率。

通过该系统还可以提供被访问的较多或较少的网页的URL,用于奖励被访问次数较多的网页设计者或鼓励被访问次数较少的页面的设计者深刻分析网页访问量较少的原因,多想一些点子来提高页面的访问量。如:可通过把较冷的节点的超链接放到较热的节点中或通过提高较冷节点的信息的质量等方法,以及可以推荐适合放置商业广告的网页等。

参考文献:

- [1] Han J W, Kamber M. Data Mining: Concepts and Techniques [M]. 2nd ed. Beijing: China Machine Press, 2007: 147-155.
- [2] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C]// Proceedings of the 1993 ACM SIGMOD Conference. Washington D C: [s. n.], 1993: 207-216.
- [3] Robertson G, Mackinlay J, Card S. Cone trees: Animated 3d visualizations of hierarchical information[C]// Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems. [s. l.]: [s. n.], 1991: 189-194.
- [4] 王玉珍. Web使用模式挖掘中的几个关键问题研究[J]. 电脑开发与应用, 2003, 16(11): 18-19.
- [5] 张娥, 冯秋红, 宣慧玉. Web使用模式研究中的数据挖掘[J]. 计算机应用研究, 2001(3): 80-83.
- [6] 袁万莲, 郑诚, 翟明清. 一种改进的Apriori算法[J]. 计

(下转第38页)

本占用的空间被回收。当空间回收后,确保备份区和工作区大小相同且每个对象标识符的备份地址和工作地址相同,就完成了事务的提交。这时数据库表又恢复到一个新的一致性状态,相互交换备份区和工作区的角色后,可以开始新一轮的事务处理。

当事务失败或数据库意外重启时,系统需恢复数据库表。故障恢复等于撤销事务对数据库做出的改变,影子页面法思想是备份版本保持不变,修改总在工作版本上。因此恢复工作只需将备份区的内容拷贝回工作区起始地址即可,这时工作区与备份区内容完全相同。相比基于日志^[8,9]的故障恢复技术,影子页面法消除了日志记录和处理的开销,故障恢复速度快,适于卫星地面设备监控这种实时性要求较高的场合。

4 结束语

对已创建的表,实时数据库系统必须提供将表的存储类型灵活转换的功能,即从内存表到磁盘表或从磁盘表到内存表都能在数据库运行时转换。这是因为在实时应用里,对某些占用大量内存,但近期不太可能访问的内存表,将其转换为磁盘表能为应用程序保留更多内存。而对驻留在磁盘上,但近期需要频繁访问的磁盘表,将其转为内存表,对避免数据库的磁盘 IO 操作,保证实时的并发读写性能也至关重要。为确保数据库运行时,能动态地转换表的存储类型,文中的内存表和磁盘表采用一样的文件结构,创建表格、插入、删除、修改记录和索引的创建查询等操作方式也保持一致。

磁盘表的线性地址与文件地址的转换过程中,线程使用页面完毕后必须立即释放对页面的控制权,对数据库这种典型的高并发系统,缓冲区作为共享资源,任何线程都不能长久占有某页面,否则将导致大量的不可换出的页面驻留在缓冲区内,有时甚至会出现无法为一个磁盘页面分配缓冲区页面的现象,这将为实时数据库的稳定性带来隐患^[10]。

文中的存储引擎只能解决单机环境下的数据存储,尚不能构建一个可靠的分布式实时数据库。Er-

lang 是爱立信为开发电信交换系统而设计的语言,具有分布式、软实时、高并发、高可靠、代码热插拔等特性,适合构建大规模的“永远开启、永不停机”的分布式系统^[11,12],这使得 Erlang 成为一个理想的分布式实时系统的开发语言,下阶段将在存储引擎的基础上用 Erlang 语言实现分布式实时数据库的事务处理机制,以方便管理卫星地面设备监控中的分布式实时数据。

参考文献:

- [1] 刘云生. 现代数据库技术[M]. 北京:国防工业出版社, 2001.
- [2] 吴建强. 流程工业实时数据库研究和开发[D]. 杭州:浙江大学, 2003.
- [3] 钟宝荣, 袁文亮. 内存数据库中空闲页面管理的方法研究[J]. 计算机工程与设计, 2007(7): 37-40.
- [4] 余翔湛, 殷丽华. 动态共享内存缓冲池技术[J]. 哈尔滨工业大学学报, 2004, 36(3): 128-132.
- [5] 刘云生, 何君辉. 一种主动实时数据库的系统内存管理方法[J]. 计算机应用, 2004, 24(4): 24-26.
- [6] 卢春鹏. 一种嵌入式系统的内存分配方案[J]. 单片机与嵌入式系统应用, 2002(12): 12-16.
- [7] Shu L C, Sun H M, Kuo T W. Shadowing-Based crash recovery schemes for real-time database systems[C]//In: Proc. of the 11th Euromicro Conf. on Real-Time Systems. York, England, UK; [s. n.], 1999: 260-267.
- [8] 肖迎元. 分布式实时数据库技术[M]. 北京:科学出版社, 2009.
- [9] Panda B, Tripathy S. Data dependency based logging for defensive information warfare[C]//In: Proc. of the ACM Symp. on Applied Computing. Villa Olmo, Como; [s. n.], 2000: 361-365.
- [10] Gray J, Reuter A. 事务处理:概念与技术(影印版)[M]. 北京:人民邮电出版社, 2009.
- [11] Armstrong J. Making reliable distributed systems in the presence of software errors[D]. Sweden: Royal Institute of Technology, 2003.
- [12] Armstrong J. Programming Erlang Software for a Concurrent World[M]. USA: Pragmatic, 2007.

(上接第 33 页)

- 计算机技术与发展, 2008, 18(5): 51-53.
- [7] 赵伟, 何丕廉, 陈霞. Web 日志挖掘中的数据预处理技术研究[J]. 计算机应用, 2003(5): 62-64.
- [8] Tsay Y, Chiang J. An efficient method for mining association rules[J]. Knowledge-Based Systems, 2005, 18(3): 99-105.
- [9] 郭有强. 一种高效的关联规则维护算法研究与实现[J]. 计算机技术与发展, 2007, 17(10): 123-126.
- [10] Zhao K, Liu B, Tirpak T M, et al. A visual data mining framework for convenient identification of useful knowledge[C]//ICDM '05. [s. l.]: [s. n.], 2005: 530-537.
- [11] Oosthuizen C, Wesson J, Cilliers C. Visual web mining of organizational web sites[C]// Proceedings of the Conference on Information Visualization. [s. l.]: [s. n.], 2006: 395-401.
- [12] Chen Jiyang, Sun Lisheng, Zaia O R, et al. Visualizing and discovering web navigational patterns[C]//Seventh ACM SIGMOD International Workshop on the Web and Databases (WebDB 2004). [s. l.]: [s. n.], 2004: 13-18.