Web 用户访问路径的差异性度量方法研究

朱兴亮[1] 游中胜[2] 王 勇[3]
(重庆交通大学管理学院 重庆 400074)[1] (重庆师范大学数学与计算机学院 重庆 400047)[2]
(重庆教育学院计算机与现代教育技术系 重庆 400067)[3]

摘 要 Web 站点个性化已经成为当前研究的一个热点,人们通过各种方法,对网站内容、结构、用户行为等进行数据挖掘,建立用户兴趣模型,为网站用户提供更好的服务,加强网站的竞争力。在当前网站个性化的方法中,基于用户行为分析的方法是最具有竞争力的一类方法。对 Web 用户行为进行分析用得较多的技术是对 Web 用户访问路径进行聚类以发现有意义的模式。而良好聚类的前提是有效地度量 Web 用户访问路径的差异性。针对这个问题,提出了一种新的 Web 用户访问路径差异性度量方法,通过模拟实验也验证了方法的正确性。

关键词 Web 使用挖掘,Web 访问路径,聚类,个性化

Research on Web User Access Path's Difference Measurement

ZHU Xin-Liang[1] YOU Zhong-Sheng[2] WANG Yong[3]
(School of Management, Chong qing Jiaotong University, Chongqing 400074)[1]
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)[2]
(Department of Computer and Modern Education Technology, Chongqing Education College, Chongqing 400067)[3]

Abstract Web-site personalization has become a focus of research in recent years. People use various data mining methodologies to analyze the contents and structures of Web-sites and their user' behaviors, so as to establish user models, provide personalized services and make the Web-site more competitive. Among many Web-site personalization approaches, the user behaviors analysis based approach seems to be the most promising one. It is most used that cluster the Web user access path for user behaviors analysis to find useful patterns. Clustered well is based on the good measurement of Web user access path's difference. Thus, bring forward a new method to measure the Web user access path's difference. The experiment has proved the correctness of the method.

Keywords Web usage mining, Web access path, Clustering, Personalization

1 引言

一个大的 Web 网站,其内容极其庞杂,用户可能对站点结构(指链接结构)中互不关联或相距甚远的多个信息点(网页)感兴趣。因此,一个固定不变、缺乏推荐、没有个性化的网站结构往往会使人感到不便。我们常常会遇到这样的情况:当我们为了某个目的浏览网站时,到处找不到对于这个目的来说内容明明是相关联的信息(因为网站是按照另外一种目的来组织的)。因此,有必要提出对网站进行个性化组织的问题[1]。

当前对网站进行个性化组织用得较多的方法是对 Web 用户行为进行分析,其中又包括 Web 用户访问路径分析和网页相关性分析两种方法。本文的工作是基于 Web 用户访问路径分析,常用的方法是:首先,将类似的 Web 用户访问路径进行聚类,并认为这些聚类代表一定的用户兴趣模型;然后,当一个用户进行浏览时,将对他当前的访问路径进行跟踪,并与已知的各种兴趣模型的典型路径进行匹配,从而动态、自动地判断用户的兴趣类型,并根据典型路径对当前用户进行实时推荐或进行页面结构动态调整、页面集预存储等。

本文的主要工作是探讨如何有效地度量不同 Web 用户访问路径之间的差异性。

2 Web 用户访问路径的差异性度量方法

两条不同 Web 用户访问路径的差异性,实际上是以访问路径中所浏览的网页之间的关联性来衡量的。因此,要度量 Web 用户访问路径之间的差异性就要先计算访问路径中各网页之间的关联性。

传统的网页关联性计算是建立在简单的网页距离计算公式[2]之上的,用这种方法得到的网页关联性可以用来有效地计算那些包含了较多相同网页且在同一条路径上的用户访问路径差异性。但若用户以完全不同的访问路径来浏览相同的目的网页时,用这种方法得到的用户访问路径差异性是有很大误差的。要较准确地衡量一个网站中网页之间的关联性时,一个可行的思路是综合考虑大量 Web 用户的访问情况,从大量 Web 用户访问路径中挖掘有关网页关联性的知识。也就是说,从 Web 用户的实际活动中,找寻隐藏的网页之间的联系。基于这种"实践出真知"的思想,本文提出用一种新的方法来计算网页之间的关联性,进而在此基础上进一步较为准确地度量不同 Web 用户访问路径之间的差异性。

2.1 网页关联性的计算

首先,计算所有单一网页对应单一网页的置信度(Confidence)。

朱兴亮 硕士生,讲师,主要研究方向:数据挖掘、网络安全。

$$Confidence(A\Rightarrow B)=P(B|A)=\frac{support\_count(A\cup B)}{support\_count(A)}$$

式中，A、B 表示网站中任意两个网页，$support\_count(A\cup B)$ 表示网页 A 和网页 B 出现在一条访问路径中的次数(搜寻空间为所有的 Web 用户访问路径)。

以置信度为基础来表示网页之间的关联性可以从大量的 Web 用户访问路径中，较为准确客观地描述网页与网页之间的真实关联情况。

然后，建立网页关系矩阵。利用前面所得到的网页之间的置信度，建立一个包含所有网页关联性的矩阵，是为网页关系矩阵。步骤如下：

(1)假设网站有 n 个网页，则建立一个 $n\times n$ 的网页矩阵，将一对一网页间的置信度对应地填写到矩阵中的相应位置，即 $C_{i,j}=confidence(i\Rightarrow j)$。对计算后没有产生关联性的每一对网页，在相应位置填入 0。同时，在矩阵中的主对角线填入值 1。

$$\begin{bmatrix} C_{1,1} & C_{2,1} & \cdots & C_{n,1} \\ C_{1,2} & C_{2,2} & \cdots & C_{n,2} \\ \cdots & \cdots & \cdots & \cdots \\ C_{1,n} & C_{2,n} & \cdots & C_{n,n} \end{bmatrix}$$

(2)将矩阵中对称位置的值相加求出平均值后，再填入矩阵相应位置中。例如：矩阵元素 $C_{i,j}=0.7$ 与 $C_{j,i}=0.8$，将两者加总平均后可得值 0.75，再填入原矩阵元素位置 $R(i,j)$ 中，得到一个对称矩阵。

$$\begin{bmatrix} 1 & & & & & \\ R(1,2) & 1 & & & & \\ \cdots & \cdots & \cdots & & & \\ \cdots & R(i,j) & \cdots & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ R(1,n) & R(2,n) & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

此即是网页关系矩阵，其中的每个元素代表编号为 i 的网页与编号为 j 的网页彼此的关联性，用 $R(i,j)$ 表示：

$$R(i,j)=C_{i,j}+C_{j,i}/2, R(i,j)\in[0,1])$$

2.2 Web 用户访问路径差异性的度量

根据前面得到的网页关系矩阵，将两条 Web 用户访问路径中各自所包含的网页彼此之间的关联性加总平均后，即可求得两条 Web 用户访问路径之间的差异性值。设 $p_1=\{m_1, m_2,\cdots,m_k\}$，$P_2=\{n_1,n_2,\cdots,n_l\}$，$m_i(i=(1,2,\cdots,k))$，$n_j(j=1(1,2,\cdots,l))$为 Web 用户访问路径 p_1、p_2 中所浏览的网页集合，则 Web 用户访问路径 p_1、p_2 的差异性可以用下式来度量：

$$d(p_1,p_2)=(\sum_{i=1}^{k}\sum_{j=1}^{l}R(m_i,n_j))/(k\times l)$$

在此基础上，可形成 Web 用户访问路径差异性矩阵：

$$\begin{bmatrix} 1 & & & & & \\ d(p_1,p_2) & 1 & & & & \\ \cdots & \cdots & \cdots & & & \\ \cdots & d(p_i,p_j) & \cdots & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ d(p_1,p_2) & d(p_2,p_n) & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

3 模拟实验

网页拓扑结构如图 1 所示。

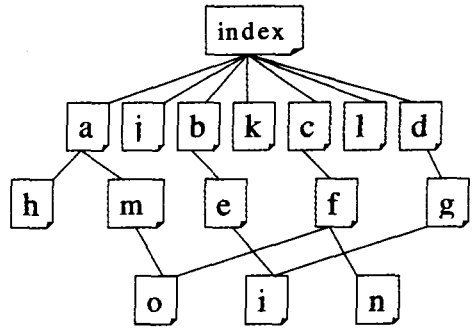

图 1 网页拓扑结构示意图

若有以下的 Web 用户访问路径：

$p_1=\{b\rightarrow e\rightarrow i\}$，$p_2=\{c\rightarrow f\}$，$p_3=\{a\rightarrow h\}$，$p_4=\{a\rightarrow m\}$，$p_5=\{d\rightarrow g\rightarrow i\}$，$p_6=\{index\rightarrow j\}$，$p_7=\{index\rightarrow k\}$，$p_8=\{index\rightarrow a\}$，$p_9=\{c\rightarrow f\rightarrow n\}$，$p_{10}=\{index\rightarrow c\}$，$p_{11}=\{index\rightarrow l\}$，$p_{12}=\{a\rightarrow m\rightarrow o\}$

当用户以完全不同的路径来浏览相同的目的网页 i 时，产生了两条 Web 用户访问路径 p_1 和 p_5。若令 p_1' 与 p_5' 分别代表路径 p_1 与 p_5 中所含网页构成的集合，则 $p_1'=\{b,e,i\}$，$p_5'=\{d,g,i\}$，那么，

$$p_1'\times p_5'=\{(b,d),(b,g),(b,i),(e,d),(e,g),(e,i),(i,d),,g),(i,i)\}。$$

这样，Web 用户访问路径 p_1 与 p_5 的差异性，就可以 $p_1'\times p_5'$ 集合中各元素所代表的网页配对之关联值来综合度量。

文[2]中使用如下的网页关联性计算方法：

$$R(i,j)=\min(1,\frac{|i\cap j|}{\max(|i|,|j|)-1})$$

式中：$|i|$表示从网站首页(入口网页)到网页的路径上，包含几条边；$|i\cap j|$表示从网站首页到网页 i 与从网站首页到网页 j 这两条路径中包含了多少条相同的边。网页之间的关联值计算如下：

$$R(b,e)=\min(1,\frac{|b\cap e|}{\max(|b|,|e|)-1}$$

$$=\min(1,\frac{1}{\max(|b|,|e|)-1}=1$$

$$R(e,g)=\min(1,\frac{|e\cap g|}{\max(|e|,|g|)-1})$$

$$=\min(1,\frac{0}{\max(|e|,|g|)-1}=0$$

同理可得：$R(b,d)=0$，$R(b,g)=0$，$R(b,i)=0.5$，$R(e,d)=0$，$R(e,i)=1$，$R(i,d)=0.5$，$R(i,g)=0$。那么，

$$d(p_1,p_5)=\frac{R(b,d)+R(b,g)+R(b,i)+R(e,d)}{3\times 3}+\frac{R(e,g)+R(e,i)+R(i,d)+R(i,g)+R(i,i)}{3\times 3}$$

$$\frac{0+0+0.5\times 0+0+1+0.5+0+1}{9}$$

$$=0.333$$

下面用本文提出的方法度量 Web 用户访问路径和的差异值：

$$C_{b,d}=confidence(b\Rightarrow d)=P(d|b)$$

$$=\frac{support\_count(b\cup d)}{support\_count(b)}=\frac{0}{1}=0$$

$$C_{d,b}=confidence(d\Rightarrow b)=P(b|d)$$

$$=\frac{support\_count(d\cup b)}{support\_count(d)}=\frac{0}{1}=0$$

$$R(b,d)=\frac{C_{b,d}+C_{d,b}}{2}=\frac{0+0}{2}=0$$

$$C_{b,i}=confidence(b\Rightarrow i)=P(i|b)=\frac{support\_count(b\bigcup i)}{support\_count(b)}=\frac{1}{1}=1$$

$$C_{i,b}=confidence(i\Rightarrow d)=P(b|i)=\frac{support\_count(i\bigcup b)}{support\_count(i)}=\frac{1}{1}=0$$

$$R(b,i)=\frac{C_{b,i}+C_{i,b}}{2}=\frac{1+1}{2}=1$$

同理可得：$R(b,g)=0$，$R(e,d)=0$，$R(e,g)=0$，$R(e,i)=1$，$R(i,d)=1$，$R(i,g)=1$。那么，

$$d(p_1,p_5)=\frac{R(b,d)+R(b,g)+R(b,i)+R(e,d)}{3\times3}+\frac{R(e,g)+R(e,i)+R(i,d)+R(i,g)+R(i,i)}{3\times3}=\frac{0+0+1+0+0+1+1+1+1}{9}=0.555$$

可见，用本文提出的方法度量 Web 访问路径的差异性比文[2]的方法更为客观（因 Web 用户访问路径 p_1、p_5 浏览了相同的目的网页，应具有较高的差异性值）。

结论 通过本文的方法，可以有效地计算出所有 Web 用户访问路径彼此的差异性值，得到 Web 用户访问路径差异性矩阵，为进一步的 Web 用户访问路径聚类操作提供可靠的数据。

将本文的方法用于实际 Web 网站中，再施以恰当的聚类操作，即可在网站运营中，根据不同用户的网页访问觉情况，对其进行个性化推荐、页面结构动态调整、页面集预存储等操作。具体的作法是：先对已有的 Web 用户访问路径进行聚类，将具有相似浏览兴趣的 Web 用户访问路径聚集成一类；然后在当前用户的浏览过程中，实时地将其访问路径与已有聚类相比较，并将其聚类到最接近的某一类别中，再根据这一类别的浏览兴趣或浏览目的对当前用户进行相应的个性化操作。

因此，作者进一步的工作就是研究好的聚类算法，在本文得到的 Web 用户访问路径差异性矩阵基础上对 Web 用户访问路径实施有效的聚类操作。

参 考 文 献

1 Buchner A G, et al. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. ACM SIGKDD Record, 1998, 27(4): 54～61

2 Xu Baowen, Weifeng, Song William , et al. Application of Data Mining in Web Pre-Fretching. IEEE Multimedia Software Engineering. In: Proceedings. International Symposium on, 2000. 372～377

3 Nasraoui O, Krishnapuram R, Joshi A. Relational Clustering Based on a New Robust Estimator with Application to Web Mining. In: Proc. Intl. Conf. North American Fuzzy Info. Proc. Society (NAFIPS 99), 1999. 596～607

4 王晔，李德毅. 自适应 Web 站点的访问数据聚类方法[A]. 中国人工智能进展[C]. 北京：清华大学出版社，2001. 402～406

5 陈恩红，徐涌，王煦法. Web 使用挖掘：从 Web 数据中发现用户使用模式. 计算机科学，2001，28(5)：85～88

（上接第 62 页）

(4)混合处理：将数据融合集处理的结果与二级预处理得到的结果进行综合处理，其处理结果可以传给二级预处理或高级处理。

(5)高级处理：将前面各相应模块传来的数据处理结果进行综合处理，通过事先制定的相应规则及逻辑推理方式得出所需要的最终数据融合结果，并输出之。

4.2 体系框架的实现

为了验证我们所提出的 WSN 中数据融合体系框架的有效性，证明其能够对 WSN 中各层次的数据融合按照不同的要求进行快速开发，我们采用一种主体框架与各个功能组件相结合的形式实现这一框架。图 3 是各个功能模块通过组件技术实现的数据融合体系框图。

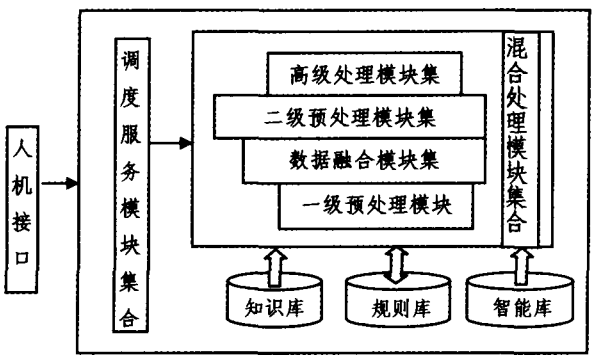

图 3 基于组件技术的处理实现框架

该数据融合架构由于采用组件的技术来实现，因此对于涉及到的模块可以动态地卸载和加载，具有很大的灵活性。在我们进行的 WSN 网络基于自组织和网内处理的数据融合仿真实验中，该框架都显示了很强的适应性。对于不同的融合要求只需对框架中的相应模块进行调整，使其满足具体的应用要求。

结束语 数据融合技术是无线传感网络中的一个关键技术，对于其中数据融合情况由于针对不同的特性和不同的分类方法存在很大差异，因此 WSN 中数据融合的处理方式也不一样，没有一个统一的处理模式。本文提出了一种基于组件和智能技术的 WSN 数据融合架构，仿真实验表明，这种系统框架可以按照 WSN 中不同的数据融合进行动态灵活的调整，显示了很强的适用性，可以作为 WSN 中数据融合的统一架构。

参 考 文 献

1 Pottie G J, Kaiser W J. Embedding the internet: wireless integrated network sensors [C] . Communications of the ACM, 2000, 43(5): 51～58

2 Kalpakis K, Dasgupta K, Namjoshi P. Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks [M]: [TR CS-02-12]. Aug. 2002

3 Zhou B, KBOSE N. Mlutitarget Tracking in clutter: Fast Algorithms for Data Association [J]. IEEE, Trans On Aerospace and Electronic System, 2003, 29(2)

4 Germain M, Voorons M, Boucher J M, et al. Fuzzy statistical classification method for multiband image fusion. Information Fusion [C]. In: Proc. of the Fifth Intl. Conf. 2002

5 Ghiasi S, Srivastava A, Yang X, Sarrafzadeh M. Optimal Energy Aware Clustering in Sensor Networks [A]. Sensors Magazine, MDPI, Issue, January. 2002. 258～269

6 Heinzelman W, Kulik J, Balakrishnan H. Adaptive Protocols for Information Dissemination in Wireless Sensor Networks [J]. In: Proc. of 5th ACM/IEEE Mobicom Conference, 1999

7 Madden S, Franklin M J, Hellerstein JM, Hong W. TAG: A Tiny AGgregation Service for ad-hoc Sensor Networks [M]. OSDI 2002

8 Cayirci E. Data Aggregation and Dilution by Modulus Addressing in Wireless Sensor Networks [A]. Communications Letters, IEEE, 2003, 7(18): 355～357

9 Cam H, Ozdemir S, Nair P, Muthuavinashiappan D. ESPDA: Energy-efficient and Secure Pattern-based Data Aggregation for wireless sensor networks [J]. Sensors, 2003. In: Proc. of IEEE, 2003, 1(22-24): 732～736