

基于 URL 结构和访问时间的 Web 页面访问相似性度量^{*}

李超锋^{1,2} 卢炎生¹

(华中科技大学计算机学院 武汉 430074)¹ (中南民族大学管理学院 武汉 430074)²

摘要 Web 页面相似性是 Web 页面聚类 and Web 会话聚类的基础,其准确性直接影响聚类的质量。本文在分析基于 URL 结构的静态 Web 页面相似性度量的基础上,提出了将 URL 结构和页面访问时间结合起来度量 Web 页面对之间的相似性并给出了度量标准。实验证明,采用我们提出的度量标准得到的结果准确性更高,更接近于用户浏览兴趣。

关键词 Web 页面相似性,聚类,URL 结构,访问时间

Similarity Measurement of Web Page Access Based on URL Structure and Access Time

LI Chao-Feng^{1,2} LU Yan-Sheng¹

(Computer School, Huazhong University of Science and Technology, Wuhan 430074)¹

(Management School, South Central University for Nationality, Wuhan 430074)²

Abstract Web page similarity, whose accuracy affects the quality of clustering, is the basis of Web page clustering and Web session clustering. On the basis of analyzing the measurement of static Web page similarity based on URL structure, this paper proposes to measure the similarity between Web pages combining URL structure with page access time. The experimental results show that our measurement is more accurate.

Keywords Web page similarity, Clustering, URL structure, Access time

1 引言

研究 Web 页面聚类对于了解 Web 用户的浏览兴趣至关重要。了解 Web 用户的浏览兴趣可以提升站点性能,改进站点结构,有针对性地为用户提供个性化的 Web 页面,进一步减轻 Web 用户搜索信息的负担,从而增强企业的核心竞争力^[1]。因此,Web 页面聚类研究已经成为业界研究的焦点。Web 页面聚类的核心问题在于 Web 页面访问的相似性度量,相似性度量准确,聚类效果就好;相似性度量不准确,聚类效果就差。可以说,Web 页面访问相似性度量的选择直接决定了 Web 页面聚类结果的优劣。

然而,目前关于 Web 页面访问相似度量的研究都存在一定的缺陷,有些度量标准过于简单,如大部分研究仅仅考虑了页面之间的直接链接关系^[2,3],而忽略了页面的访问时间。还有一些度量标准又过于复杂,比如需要考虑每个 Web 页面的具体内容^[4,5],通过 HTML 或 XML 来标记其相似性,这对

于数据量巨大的用户访问来说是不可接受的。文[6]的思想与我们比较相似,但它只考虑了站点结构,而没有考虑对页面的访问情况,因而准确度不高。

本文提出了一种度量 Web 页面访问相似性的方法,该方法充分利用 Web 页面在站点结构中的位置关系,同时考虑 Web 页面被动态访问的时间。实验证明,在判断 Web 页面访问相似性时,该方法具有较高的准确性。

2 基于 URL 结构的静态 Web 页面相似性

这种相似性度量的基本思想是:不考虑 Web 页面的内容,单纯地从 Web 页面的 URL 结构观察其相似性程度。如对页面: <http://www.scuec.edu.cn/guanyuan/lab/jiaoxue.htm> 和 <http://www.scuec.edu.cn/guanyuan/lab/chengguo.htm>,它们之间的相似性很明显,即它们是关于中南民族大学管理学院实验教学具有相似主题的面。

为了计算出不同 Web 页面之间的相似性程度 SURL,可

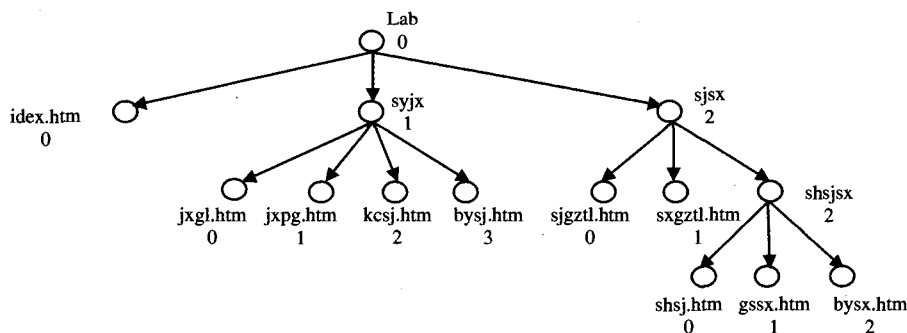


图1 中南民族大学管理学院实验中心部分站点树结构示意图

^{*}湖北省科技攻关项目(2005AA101C18)和中南民族大学自然科学基金项目(YZY06009)。李超锋 讲师,博士生,主要研究方向为 Web 使用挖掘;卢炎生 教授,博士生导师,主要研究方向为数据库系统、软件测试与数据挖掘等。

以把构成站点的 Web 页面以站点树的形式表示出来。Web 页面 URL 的层次对应于站点树的层次,每个页面的 HTML 文件对应于站点树中的一个叶子结点。图 1 所示即为中南民族大学管理学院实验中心的部分站点树结构。

这样,页面之间的相似性 S_{URL} 就可以通过比较其 URL 层次序列之间的相似性来度量。为了比较页面 URL 的每一层,可以对每层使用一个数字来表示,这样一个页面的 URL 就可以用一个数字串来表示了。比较两个页面的 URL 也就成了比较数字串。如 Web 页面“Lab/sjyx/jxgl.htm”和“Lab/sjyx/shjsjx/bysx.htm”可分别用“010”和“0222”来表示。

Web 页面的相似性 S_{URL} 计算分两步来进行:

第一步,从开头一个一个地比较两个数字串的每一个对应的数字,直到出现第一对数字不同时停止。例如,假定要比 Web 页面“/Lab/sjyx/shjsjx/bysx.htm”和“/Lab/sjyx/sxgztl.htm”。第一个 Web 页面对应的数字串是“0222”,第二个 Web 页面对应的数字串是“021”。图 2 给出了两个数字串的比较过程。从图可以看到两个数字串有 2 个相同的对应数字。

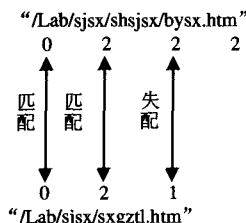


图 2 比较数字串

第二步,计算两个 Web 页面的相似性。记第一个数字串的长度是 L_1 ,第二个数字串的长度是 L_2 ,选择较长串的长度 $L_{\text{longer}} = (L_1 > L_2) ? L_1 : L_2$,然后给较长数字串的每一层一个权值:最后一层给权值 1,倒数第二层给权值 2,倒数第三层给权值 3,依此类推,直到给定权值为 L_{longer} 的第一层。两个数字串之间的相似性定义为那些匹配数字的权值之和除以总权值之和。对于给定的例子,数字串“0222”的长度是 4,“021”的长度是 3,这样 $L_{\text{longer}} = 4$,每一层的权值如图 3 所示。两个 Web 页面之间的相似性是: $S_{URL} = (4 + 3) / (4 + 3 + 2 + 1) = 0.7$ 。

数字串 1:	0	2	2	2
数字串 2:	0	2	1	
对应权值:	4	3	2	1

图 3 每一层的权值

使用上面讨论的方法,任意 URL 对之间的相似性 S_{URL} 具有如下的性质:

- 1) $0 \leq S_{URL} \leq 1$,即两个页面之间的相似性介于 0.0 和 1.0 之间;
- 2) $S_{URL} = 0$ 当且仅当两个页面完全不同,即其 URL 没有相同的对应数字;
- 3) $S_{URL} = 1$ 当且仅当两个页面精确相同,即其 URL 数字串长度相同且对应的数字相同。

3 基于用户浏览时间的 Web 页面动态访问相似性

3.1 用户浏览时间

用户浏览 Web 页面所花费的时间是度量用户对该页面

感兴趣的非常重要的信息,因此,精确地记录用户的实际浏览时间是很重要的,但同时也是很困难的。考虑图 4 所示的客户同服务器从用户请求一个页面的时间到服务器接收到用户的对下一个页面的请求时间之间的交互过程。 t_0 表示用户在地址栏输入一个 URL 地址或在某个页面中点击一个超链接从而发出一个请求的时间, t_1 表示服务器收到该请求的时间, t_2 表示服务器发出响应,将用户请求的页面发送给客户的时间, t_3 表示客户收到服务器发送过来的页面并载入到客户浏览器的时间, t_4 表示用户发出下一个请求的时间, t_5 为用户收到的用户下一个请求的时间。

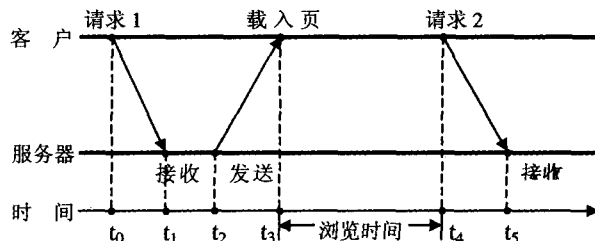


图 4 客户/服务器交互过程

从图 4 所示的客户/服务器交互过程可以看出,用户浏览一个 Web 页面的时间实际上是从 t_3 到 t_4 之间的时间,即浏览时间 $t_{\text{view}} = t_4 - t_3$ 。

有两种方法可以计算出用户访问一个 Web 页面的精确时间。一种方法是从客户端计算。该方法使用 Java applet 或 JavaScript 脚本语言开发一个远程代理^[6],当一个新的 HTML 页面被载入到客户端显示时,该远程代理将系统时间 T_{load} 发送到服务器,一旦页面退出,系统时间 T_{exit} 再一次报告给服务器。这样,服务器通过简单的计算就可以得到 Web 页面的精确浏览时间,即 $t_{\text{view}} = T_{\text{exit}} - T_{\text{load}}$ 。然而,这种方法存在一定的缺点。首先,远程代理代码需要加入到每个 Web 页面代码中,载入远程代理也需要时间,该时间计入了用户的浏览时间。其次,使用 JavaScript 实现远程代理的代价相对会小一些,但当用户使用浏览器上“back”或“forward”按钮而不是点击一个超链接时,远程代理将不被执行。最后,如果用户出于安全考虑关闭了浏览器的 Java 使能功能,远程代理将无效。

另一种方法是利用 Web 服务器的日志记录中关于页面的访问时间。然而,目前不同的 Web 服务器比如 NCSA 和 Apache 提供的是从 t_1 到 t_5 的弹性时间作为请求的时间。它们以通用或扩展通用日志文件格式存储这种信息^[7]。这样 Web 服务器日志中记录的用户“浏览时间”是 $t_{\text{log}} = t_5 - t_1$,而不是我们想要的 $t_4 - t_3$ 。因此,我们需要通过计算来获得 t_{view} 。该过程可如下进行:

首先,服务器从收到一个请求到它向客户发送页面之间的时间非常短,通常是 ms 级,因而可以忽略不计,即可以认为

$$t_1 \approx t_2 \quad (1)$$

其次,假设服务器使用 ping 命令得到的从客户返回的时间为 δ ,则可以认为用户的一个请求到达服务器的时间是服务器使用 ping 命令得到的环回测试时间的一半,即有

$$t_1 - t_0 = t_5 - t_4 = \delta / 2 \quad (2)$$

最后,服务器向客户发送页面的时间可以通过服务器所发送的分组数与一个分组从服务器发送到客户所需时间的乘积,即

$$t_3 - t_2 = PS \times \delta / 2 \quad (3)$$

其中 PS 是服务器所发送的分组数。

这样,一个 Web 页面的精确浏览时间 t_{view} 可以通过下式计算:

$$\begin{aligned} t_{view} &= t_4 - t_3 \\ &= (t_5 - t_1) - (t_3 - t_1) - (t_5 - t_4) \\ &\approx (t_5 - t_1) - (t_3 - t_2) - (t_5 - t_4) \\ &= t_{log} - PS \times \delta / 2 - \delta / 2 \\ &= t_{log} - (PS - 1) \times \delta / 2 \end{aligned} \quad (4)$$

另一个影响浏览时间信息精确性的因素是用户可能载入一个 Web 页面后因某种原因离开了。这时,我们使用式(4)计算出的很可能不是用户浏览页面所花费的真实时间。解决这个问题的一种方法是根据 Web 页面大小和人们的阅读速度计算一个阈值 t_{out} 。如果浏览时间超过该阈值 t_{out} ,那么该页面的时间信息必定是无效的并且可以将阈值 t_{out} 作为浏览时间使用。

3.2 基于浏览时间的 Web 页面动态访问相似性

第 2 节描述的 Web 页面相似性度量方法实际上是评价 Web 站点内两个静态 Web 页面之间相似程度的方法。然而,用户浏览 Web 页面的过程是一个动态的过程,比如用户打开某页面 A 可能只是为了通过 A 访问页面 B,因而实际上用户对 A 没有兴趣。再比如用户对某页面的访问时间不同,可能意味着该用户对该页面的兴趣度不同。因此在考虑两个 Web 页面之间的相似性时,我们必须考虑对这两个页面的动态浏览时间因素,据此我们提出了基于浏览时间的 Web 页面动态访问相似性。

设 A 和 B 的浏览时间分别为 t_{viewA} 和 t_{viewB} ,则 A 和 B 之间的浏览时间相似性 S_{time} 定义如下:

$$S_{time} = \frac{\min(t_{viewA}, t_{viewB})}{\max(t_{viewA}, t_{viewB})} \quad (5)$$

为了便于讨论,我们不妨设 $t_{viewA} < t_{viewB}$ 。由式(4)和(5)可得

$$\begin{aligned} S_{time} &= \frac{\min(t_{viewA}, t_{viewB})}{\max(t_{viewA}, t_{viewB})} \\ &= \frac{t_{viewA}}{t_{viewB}} \\ &= \frac{t_{logA} \times (PS_A - 1) \times \frac{\delta_A}{2}}{t_{logB} \times (PS_B - 1) \times \frac{\delta_B}{2}} \\ &= \frac{t_{logA} \times (PS_A - 1) \times \delta_A}{t_{logB} \times (PS_B - 1) \times \delta_B} \end{aligned} \quad (6)$$

对于任意两个被访问的 Web 页面,浏览时间相似性介于 0 和 1 之间,并且两个页面的浏览时间越接近,其相似性值就越大,说明二者越相似,反之亦然。如果有一个页面没有被访问,则其相似性为 0。如果两个页面被浏览的时间精确相等,则其相似性为 1。当然,由于受到诸多客观因素的影响,这种情况几乎不可能出现。不过实现时我们可以取一个阈值,当两个 Web 页面的浏览时间相似性值超过该阈值时,即可认为这两个页面的浏览时间是相等的。

4 基于浏览时间和 URL 结构的 Web 页面动态访问相似性度量

用户对某 Web 页面的访问过程是个动态的过程,因而仅仅通过比较两个 Web 页面在站点内位置结构上的相似性并不能真正揭示用户对 Web 页面的访问相似性。同样,抛开

Web 页面的静态结构,只考虑用户对两个页面的访问时间,也不能真正揭示用户对 Web 页面的访问相似性。因为所比较的两个页面不同,对其访问时间的比较毫无意义。为此,我们提出将用户浏览时间与 Web 页面的 URL 结构结合起来作为评价用户访问 Web 页面的相似性度量标准。

设两个 Web 页面 A 和 B 之间的 URL 相似性为 S_{URL} ,浏览时间相似性为 S_{time} ,则其访问相似性 S 可定义为

$$S = \begin{cases} S_{URL}, & \text{当 } S_{URL} \neq 1 \text{ 时} \\ S_{time}, & \text{当 } S_{URL} = 1 \text{ 时} \end{cases} \quad (7)$$

公式(7)表明:当所比较的两个访问页面是同一个 Web 页面时,其相似性取决于对该页面的访问时间;当两个访问是不同的页面时,其相似性取决于它们的 URL 结构,因为这时再比较两个不同的 Web 页面的访问时间已经没有意义。

5 实验及性能评价

实验数据来源于中南民族大学管理学院实验中心网站 2006 年 4 月 5 日的 Web 日志片断。经过预处理之后,我们得到了表 1 所示的信息。

表 1 经预处理的 Web 日志片断

	客户 IP	访问页面	t_{log} (秒)
1	210.42.156.123	Lab/index.htm	3
2	210.42.156.123	Lab/syix/jxgl.htm	303
3	210.42.156.71	Lab/index.htm	429
4	210.42.156.71	Lab/syix/bysj.htm	338
5	210.42.156.12	Lab/index.htm	398

表 2 是根据表 1 的内容分别采用基于 URL 结构得到的 3 对页面访问相似性及采用基于 URL 结构和访问时间相结合得到的同样 3 对页面访问相似性之间的比较。

表 2 两种不同的相似性度量标准比较结果

度量标准 \ 页面对相似性	页面对		
	<1,3>	<3,5>	<2,4>
基于 URL	1	1	0.833
基于 URL 和访问时间	0.007	0.928	0.833

从表 2 可以看出,按照基于 URL 的相似性度量标准,则页面对<1,3>和<3,5>之间的相似性均为 1,说明这三者的兴趣相同。然而,我们很明显地可以看出,1 中对 Lab/index.htm 的访问只是客户 210.42.156.123 访问页面 Lab/syix/jxgl.htm 时所用到的引用,用户的兴趣在 Lab/syix/jxgl.htm 页面上而不是 Lab/index.htm。这从用户对 Lab/index.htm 的访问时间只有 3s 这一事实可以看出。而 3 和 5 则不同,对 Lab/index.htm 的访问所花费的时间差不多都是 4min,说明客户 210.42.156.71 和 210.42.156.12 对该页面感兴趣,因而二者的兴趣是非常接近的,但同 1 的兴趣不同。基于 URL 和访问时间相结合的相似性度量准确地反映了这一差别,即<1,3>页面对之间的相似性只有 0.007,而<3,5>之间的相似性高达 0.928,没有达到 1 是因为两位不同的客户的阅读速度不同。由于<2,4>页面对两个不同的页面,因而其相似性只需要采用基于 URL 的相似性度量。

结论 Web 页面相似性是 Web 页面聚类和 Web 会话聚类的基础,其准确性直接影响到聚类的质量。本文分析了基

(下转第 286 页)

表1 实验结果

系统元件数	计算时间(毫秒)
15	1
120	7
500	30
1000	75

5 相关工作比较

Reiter^[1]给出了由冲突集来求诊断的方法,其复杂性是指数级的,难以在实际中应用。文[2,5,7,8]等在该方法的基础上做了进一步的研究,提高了诊断效率,缩减了诊断空间,但是具体的求解过程仍要通过调用定理证明器,仍然有很高的复杂性。

Fattah 和 Dechter^[9]、Stumptner 和 Wotawa^[10]给出了树型结构系统的多项式时间结束的诊断方法,但是他们没有把该方法推广到非树型结构系统上,这就限制了它的应用。文[11]给出了利用结构信息的故障诊断方法;对于一些特殊结构的系统可在多项式时间内结束,并且给出了一种直接求诊断的方法。Gerhard Friedrich^[14]等给出了多项式时间找到一个极小诊断的方法。

Console^[15]和 Heiko Milde^[16]等采用了判定树的方法应用于实际系统。利用系统模型,通过穷举所有可能的系统错误,或者穷举所有可能的观察,构造判定树。需要进行系统诊断时,仅需对判定树进行搜索即可完成。但是,对于规模较大的系统,判定树的规模将十分庞大,限制了它的应用范围。

结论和未来的工作 本文的研究尝试显式地考虑系统的值传递以实现极小诊断的高效计算。给出了一种基于有向图上值传递的系统模型,并重新定义了诊断,该定义与 Reiter 给出的基于一致性的诊断等价。在此基础上,给出了找到一个极小诊断的算法,时间复杂度是 $O(kn+n)$ 。该算法在给出诊断(可能故障的元件集合)的同时,还给出了与该诊断对应的系统当前行为的一种可能解释。特别地,如果算法返回空集合,表示系统当前行为正常。实验结果表明,在系统元件数较大的情况下,值传递诊断方法可以快速找到极小诊断。同时,本文给出了该方法与诊断测试结合实现系统修复的框架。

对于给定的诊断问题,可能存在多个值传递得到值传递结果,因此,可能得到不同的极小诊断。但是,对于大规模系统,可能的值传递的数目将会十分庞大。我们以后会对该问

题进行研究和讨论。

参考文献

- 1 Reiter. A theory of diagnosis from first principles. Artificial Intelligence, 1987, 32(1):57~96
- 2 Chittaro L. Hierarchical model-based diagnosis based on structural abstraction. Artificial Intelligence, 2004, 155:147~182
- 3 Baroni P. Diagnosis of large active systems Artificial Intelligence, 1999 (110):135~183
- 4 Console L, Picardi C, Ribaudo M. Process algebras for systems diagnosis. Artificial Intelligence, 2002, 142:19~51
- 5 李占山,姜云飞,王涛. 基于模型的诊断问题分解及其算法. 计算机学报, 2003, 26(9):1177~1182
- 6 Pencole Y. A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. Artificial Intelligence, 2005, 164: 121~170
- 7 欧阳丹彤,姜云飞. 广义因果理论的基于模型的诊断. 计算机研究与发展, 1999, 36:31~35
- 8 陈荣,姜云飞. 含约束的基于模型的诊断系统. 计算机学报, 2001, 24:127~135
- 9 Fattah Y E, Dechter R. Diagnosing tree-decomposable circuits. In: Proceedings of International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995. 572~578
- 10 Stumptner M, Wotawa F. Diagnosing tree-structured systems. Artificial Intelligence, 2001, 127(1): 1~29
- 11 栾尚敏,戴国忠. 利用结构信息的故障诊断方法. 计算机学报, 2005, 28(5):801~808
- 12 Mozetic I. A polynomial-time algorithm for model-based diagnosis. In: Proceeding of the 10th European Conference on Artificial Intelligence, Vienna, 1992. 729~733
- 13 Childress R L, Valtorta M. Polynomial-time model-based diagnosis with the critical set algorithm. In: Proceeding of the 4th International Workshop on Principles of Diagnosis, Aberystwyth, Wales, 1993. 166~177
- 14 Friedrich G, Gottlob G, Nejd W. Physical Impossibility Instead of Fault Models. AAAI, 1990. 331~336
- 15 Portinale L, Magro D, Torasso P. Multi-modal diagnosis combining case-based and model-based reasoning: a formal and experimental analysis. Artificial Intelligence, 2004, 158:109~153
- 16 Console L. Temporal decision Trees: Model-based Diagnosis of Dynamic Systems on-board. Journal of Artificial Intelligence Research, 2003, 19:469~512
- 17 Milde H. Integrating model-based diagnosis techniques into current work processes-three case studies from the INDIA project. AI Communications, 2000, 13:99~123
- 18 姜云飞,李占山. 基于模型诊断的元件替换与替换测试. 计算机学报, 2001, 24:666~672

(上接第 209 页)

于 URL 结构的静态 Web 页面相似性度量的缺陷,详细讨论了 Web 页面访问时间的计算方法,提出将 URL 结构和页面访问时间结合起来度量 Web 页面对之间的相似性并给出了相应度量标准。最后,用实验证明了采用我们提出的度量标准得到的结果准确性更高,更接近于用户浏览兴趣。Web 页面对之间的相似性度量标准的确定为研究 Web 会话之间的相似性进而研究 Web 挖掘系统打下了坚实的基础。

参考文献

- 1 Li Chaofeng, Lu Yansheng. Research on Web Usage Mining for Electronic Commerce. In: Proceedings of 2005 International Conference on Management Science and Engineering, 2005. 703~710
- 2 Bharat K, Broder A, et al. Connectivity Server. Fast Access to Linkage Information on the Web. In: Proceedings of the 7th International World Wide Web Conference, 1998. 469~477
- 3 Huang S H S, Molina-Rodriguez C H, et al. Exploring similarity among Web pages using the hyperlink structure. Information Technology: Coding and Computing, 2004, Proceedings, 2004 (1):344~348
- 4 Ester M, Kriegel H P, Schubert M. Accurate and Efficient Crawling for Relevant Websites. In: Proc. 30th Int Conf. on Very Large Databases (VLDB'04), 2004. 396~407
- 5 Kailing K, et al. Efficient Similarity Search for Hierarchical Data in Large Databases. In: Proc. 9th Int Conf on Extending Database Technology (EDBT 2004), 2004. 676~693
- 6 Wang W, Zaane O R. Clustering Web Sessions by Sequence Alignment. In: SIGMOD Conference, 2002. 394~398
- 7 李超锋. Web 使用挖掘数据源分析. 中南民族大学学报(自然科学版), 2005(4):82~85
- 8 Shahabi C, Zarski A M, Adibi V S J. Knowledge discovery from users web-page navigation. In: Proc. 7th Intl Conf on Research Issues in Data Engg, 1997. 20~29