

2000, 27(17)

自适应Web站点

万维网

数据挖掘

计算机科学 2000 Vol. 27 No. 7

自适应 Web 站点: 挑战与机遇

Creating Self-Adaptive Web Sites: The Challenges and Opportunities

张 德 董逸生

(东南大学计算机科学与工程系 南京 210096)

TP393.4

Abstract The creation of a complex Web site are a thorny problem in user interface design. First, different visitors have distinct goals. Second, even a single visitor may have different needs at different times. Much of the information at the site may also be dynamic or time-dependent. Third, as the site grows and evolves, its original design may no longer be appropriate. Finally, a site may be designed for a particular purpose but used in unexpected ways. Web servers record data about user interactions and accumulate this data over time. We believe that data mining techniques can be used to examine user access logs, in order to improve the site automatically. There are a great deal of challenges and opportunities when creating self-adaptive Web sites: sites that automatically improve their organization and presentation based on user access patterns. Success would have a broad and highly visible impact on the World Wide Web and the data mining community.

Keywords Self-adaptive Web site, Data mining

1 引言

万维网(World Wide Web)已经成为信息传播、交流与共享的主要媒体。在全球 Web 站点数目迅速增长的同时,各个 Web 站点的信息量及其复杂度也在迅速上升,包含成千上万个网页与超链接是很平常的。由于以下的因素,数据密集型 Web 站点的设计与管理也变得越来越困难

(1)各个用户具有不同的身分、性别、年龄、背景、语言、文化、习惯、爱好……,不同的用户访问某个 Web 站点时带有各自不同的目的,关注的内容也就各不相同;

(2)在不同的时期,同一个用户对 Web 信息的需求也可能不同;

(3)Web 站点随着时间的推移不断发展,内容逐渐增加,导致其初始的设计不再适合;

(4)Web 站点实际提供的信息服务往往超出了其设计的范围,甚至完全改变了定位。

目前的 Web 站点一般是被动的,用户需要自己去寻找所关心的信息在哪里,操作繁琐,非常不方便。一种改进的做法是将 Web 站点的内容组织为层次结构,访问该 Web 站点的用户可以进行定制,但这种方法只提供极其有限的类别,难以实现真正的个性化服务,而

且缺乏灵活性,为解决这些问题,计算机学术界开始研究自适应 Web 站点:通过学习用户的访问模式自动地改进 Web 站点上信息的组织与显示。

与一般的商用软件包不同,Web 站点由于其动态性、时效性、异质性,通常无法交给专门的实验室进行可用性测试。幸运的是,用户在访问某个 Web 站点时,其行为都被记录进相应服务器的日志,这些日志为创建自适应 Web 站点提供了原始数据。在某些场合下 Web 站点还可以获得访问用户的注册信息。因为日志文件与注册信息往往庞大而不直观,所以我们需要对它们进行自动的分析以发现其中蕴含的知识,即进行数据挖掘。

虽然目前所有 Web 站点都不是自适应的,但是 Web 站点已经保存了大量支持自适应性的数据。随着 Web Web Web 普及程度与复杂程度的提高,自适应 Web 站点的研究显得越来越重要。我们相信,创建自适应 Web 站点这个任务将给计算机科学诸多领域的发展带来深远的影响。

本文介绍了自适应 Web 站点研究的主要问题,叙述了自适应 Web 站点研究的主要目标,回顾了近几年这一方面研究的主要进展,探讨了自适应 Web 站点质量评价的问题,最后展望了其发展前景。

张 德 博士生,主要研究兴趣为数据库、Web 和人工智能。董逸生 教授,博士生导师,研究领域包括数据库、信息系统和软件工程等。

2 自适应 Web 站点研究的主要问题

2.1 如何设计

自适应 Web 站点应该为不同的用户提供不同的视图,所以其内容要能够进行灵活的剪裁与拼装。在设计时就需要考虑 Web 站点的哪些部分是固定的,哪些部分是可变的。例如可以为用户提供一个根据其访问模式动态生成的浏览指南(tour guide)页面^[1];或者在 HTML 页面中加入一些特殊标记,指示修改的位置以及修改的方法;或者允许系统根据对 Web 站点的语义描述进行一些推理;甚至把整个 Web 站点作为一个数据库^[5]。理想的自适应 Web 站点将使信息的显示格式、逻辑结构、物理组织都相互独立。

2.2 如何学习

Web 站点日志与关系型数据有所不同,需要为之建立合适的模型,发展新的数据挖掘算法。另外,自适应 Web 站点(每周 7 天每天 24 小时)一直在运行,用户的访问数据也在不断积累,所以对用户访问数据的学习应该是持续进行的。一些数据挖掘算法需要从静态的改成动态的、从批量式的改成增量式的,才能符合自适应 Web 站点的需要。

2.3 如何改进

根据用户的访问模式,自适应 Web 站点可以做出改进包括:(1)使用户所关心的页面更加容易访问;(2)使用户所关心的超链接更加醒目;(3)链接相关页面;(4)聚合相似页面;(5)增加缓冲预取机制,改善服务器响应时间;(6)合理设置广告;等。例如,为东南大学计算机系建立一个自适应 Web 站点,那么一个数据库研究方向的博士生经过一段时间的访问之后,会看到该 Web 站点为自己作了调整,首页上有研究生院的通知、数据库研究组学术讨论的时间表、所参与研究项目的进展等;学术会议消息的页面上与数据库有关的超链接被放在最前面或者加亮显示……。

一些简单的改进可以由系统根据学习的结果自动完成。在复杂的情况下,系统应该向站点管理员(Webmaster)提出修改的建议,解释修改的原因,由站点管理员负责作出最终决定。这时就存在着自适应 Web 站点怎样与站点管理员有效地交流协作的问题。

3 自适应 Web 站点研究的主要目标

自适应 Web 站点研究的基本目标有两个:服务个性化与性能最优化。

3.1 服务个性化

服务个性化是指 Web 站点为适应某一个特定的用户的需要而实时地调整数据的组织与显示。一种实现服务个性化的方法是允许用户手工定制 Web 站点

的显示选项,系统将记住每个用户的定制,并在该用户再次进入时进行相应的调整,例如微软的 MSN 站点(<http://www.msn.com>)允许用户只看自己所选择的新闻栏目。

与此相反,路径预测系统试图猜测用户在寻找什么信息并帮助他(她)更快地找到。一个路径预测系统至少需要回答以下几个问题。

(1)预测的内容是什么 预测的内容可能是用户的下一步操作,例如系统发现在浏览某页面时其中的某一个超链接被用户点击的可能性最大,就可以把这个超链接加亮或者提前。预测的内容也可能是用户的最终目标,例如系统发现用户在寻找某个特定的页面,就可以把这个页面直接弹出来。

(2)预测的依据是什么 预测的依据可能是单个特定用户的历史行为,也可能是对多个用户进行抽象,以便更快地收集数据。

(3)预测的影响是什么 根据预测,可以做的调整包括将某个超链接加亮、变粗、提前或者在其周围放置图片、合成一个新的页面等等。

3.2 性能最优化

性能最优化是指 Web 站点提高对所有用户的整体服务性能。为实现性能最优化目标,Web 站点需要学习所有用户的访问模式,并据此做出调整,使 Web 站点访问起来更容易,因此,即使是以前一无所知的新用户,也能够从这种改进中获益。

如果我们把所有可能的 Web 站点设计看作一个站点设计空间,那么改进一个 Web 站点就相当于在该空间中寻找更好的点。假设我们能够评价 Web 站点的“好坏”程度(请参见本文第 4 节),这就成为人工智能中经典的搜索问题。例如我们可以采用“爬山法”来搜索上述设计空间,即不断调整 Web 站点,且保证每一步调整都能够改进 Web 站点的整体质量。在这方面,如何组织庞大的搜索空间以及如何选取合适的搜索策略等等都是有待研究的问题。

4 自适应 Web 站点研究的主要进展

World Wide Web 诞生以后,人工智能、信息检索、数据库等不同领域的研究人员就从不同的角度开始涉及自适应 Web 站点的某些问题,例如机器学习^[2]、知识表示、规划识别^[7]、用户建模^[6]等等。1997 年华盛顿大学的 Mike Perkowitz 和 Oren Etzioni 在第十一届 IJCAI(国际人工智能联合会议)上正式将自适应 Web 站点作为一个挑战向计算机学术界提出,并号召相关研究人员联合起来,建立完整的系统^[16]。

在对 Web 站点日志的分析方面,已经从简单统计发展到数据挖掘,其中应用最普遍的是由 Ming-Syan

Chen 等人提出的浏览路径挖掘。浏览路径是指用户依时间顺序访问的站点页面构成的序列。浏览路径挖掘先将 Web 站点日志按照用户的 IP 地址和访问时间将 Web 站点日志划分成会话(session)集合,然后在这些会话集合中寻找出现频率大于给定阈值的浏览路径,称为热门浏览路径。浏览路径挖掘与一般的关联规则挖掘不同,热门浏览路径中各项(页面)之间是有序的,而关联规则挖掘所发现的高频项集则是无序的。自适应 Web 站点可以根据所发现的热门浏览路径作出改进,如调整链接结构、增加缓冲预取机制、设置广告页面等等。

WebWatcher (<http://www.cs.cmu.edu/~web-watcher/>)软件能够根据用户当前所在页面以及用户指定的浏览兴趣来预测用户可能会点击哪一个超链接,并将用户可能点击的超链接突出显示并置于页面顶端^[2]。用户在访问 Web 站点之前,WebWatcher 会先粗略问一下访问者需要哪个方面的信息(即浏览兴趣所在)。用户在访问 Web 站点之后,WebWatcher 会问一下访问者是否找到了所需要的信息。对那些成功地找到了所需要信息的用户,WebWatcher 将学习其访问模式。例如 WebWatcher 发现许多用户通过“people”超链接找到了所关心的个人主页,以后如果有用户也要寻找个人主页时,WebWatcher 就可以把“people”这个超链接突出显示,方便用户浏览。

AVANTI 是(<http://zeus.gmd.de/projects/avanti.html>)基于“目标识别”理论的自适应 Web 站点研究项目^[6]。“目标识别”指根据用户的动作序列确定用户的最终目标是什么^[7,8]。Lesh 和 Etzioni 提出了处理该问题的一个统一框架^[9],他们把用户动作抽象成一种运算并表示为“先决条件/事后状态”的形式,在此基础上进行推理。与 WebWatcher 相似,AVANTI 也需要用户在进入 Web 站点时提供自己的浏览兴趣等信息,不同的是 AVANTI 只学习单个用户的访问模式,它不仅能预测该用户的下一步操作,而且试图预测该用户的最终目标页面。一旦 AVANTI 对用户的最终目标页面做出预测,就会把目标页面直接送给用户。

华盛顿大学的研究人员致力于自适应 Web 站点的系统优化,他们所设计的系统从 Web 站点日志中提取频繁出现的访问模式,据此对 Web 站点的结构进行调整,从而提高 Web 站点对大多数用户的服务性能^[1]。可能的调整包括重新安排超链接顺序、突出显示某些超链接、合成新页面等等,他们所做的一个实验在某个 Web 站点的首页中自动加入一些超链接,这些超链接指向访问频率很高的页面,结果发现每天 10,000 至 15,000 次的页面访问中约有 10%是通过上述自动产生的超链接而进入的,约有 25%的访问者点击

了这些自动产生的超链接,而且这些页面由于容易访问而更加流行了。他们还研究了通过合成新页面实现 Web 站点系统优化的方法。对站点的 Web 页面集合以及用户的访问模式进行聚类分析,能够发现某些页面相互之间有很强的关联(针对同一主题或者面向同类用户),然后 Web 站点可以将这些相关链接合成到一个新的导航页面中,方便用户的浏览。

对 Web 站点内容与结构的了解程度直接限制了 Web 站点的自适应能力。例如某个页面中包含一系列超链接的列表,在列表最前面添加一个新的超链接是否可行就取决于这一系列超链接的内涵,指向学生的超链接不能加入到指向教师的超链接列表中,如果这个列表是按字母顺序排列的,新的超链接也只能放在相应的位置……。但是,从 HTML 页面中我们所能得到的关于 Web 站点的知识是极其有限的。显然,在 Web 站点中引入元数据会大大改善这种状况。

提供元数据的一种方法是以类似于数据库的方式来管理 Web 站点,精确定义 Web 站点中数据的语法和语义。STRUDEL 站点管理系统是这方面最早的研究^[3],STRUDEL 系统使 Web 站点中信息的显示格式、逻辑结构、物理组织都相互独立开来,通过高层的说明性语言查询和更新 Web 站点,而不直接涉及页面与超链接的细节。例如,东南大学计算机系的 Web 站点可以用 STRUDEL 查询语句从“学生”与“教师”两个数据库中提取数据,用 STRUDEL 视图定义语言创建一个页面,在每个研究生的后面都加上一个指向其导师的超链接。在 STRUDEL 系统的基础上创建自适应 Web 站点很方便,我们可以对 Web 站点的状态进行逻辑推理,使用说明性语言进行调整也更加容易,STRUDEL 系统的弱点在于它要求把 Web 站点的内容全部放进数据库,或者为 Web 站点的内容建立映射到 STRUDEL 系统的包装器(wrapper),这种改造对现存的 Web 站点来说代价太高,包装不太规则的内容也很费事。

提供元数据的另外一种相对简单的方法是给 Web 站点中的 HTML 页面加上元数据标记,元数据与 HTML 页面共存。元数据标记能够指出 Web 站点各部分内容之间的语义联系,说明在何处可以进行调整以及如何调整。元数据标记的覆盖范围与复杂程度视需要而定,这种方式的一个实例是 Apple 公司提出的 MCF(Meta-Content Format),MCF 试图建立 Web 元数据标记的标准(<http://mcf.research.apple.com>)。用一个 MCF 使能的浏览器访问一个以 MCF 标记过的站点时,可以从多个角度察看 Web 站点的结构。这种方式的另一个例子是 SHOE 语言(<http://www.cs.umd.edu/projects/plus/SHOE/>)。SHOE 语

言为 HTML 页面加上关于本体论 (ontology) 的知识^[1], 例如某人的个人主页中被加上其性别、年龄、职业等等情况的标记。SHOE 语言的设计是为了更准确地搜索 Web 信息, 但它同时也为创建自适应 Web 站点提供了有力的支持。使用元数据标记的方式比 STRUDEL 系统简单, 但也有许多不足。首先, 元数据标记与 Web 站点的实际内容是分离的, 它们之间的一致性难以维护。其次, 元数据标记仅仅附加在现有的页面中, 无法自动地添加到新生成的页面中, 对 Web 站点的调整也仍然要修改原始的 HTML 页面。

为了进一步减轻为 Web 站点提供元数据的工作量, Mike Perkowitz 和 Oren Etzioni 提出专门面向自适应 Web 站点的 A-HTML 语言^[10]。A-HTML 扩展了 HTML, 以说明性的语言在更高的层次上对 Web 站点进行抽象描述。例如一个按字母顺序排列的列表可以用 `<list order="alphabetical">` 来标记。在实际显示时 Web 站点把 A-HTML 翻译成 HTML, 所以 Web 站点能够解释 A-HTML 就行, 对浏览器没有要求。最近, 随着 Web 标准化组织 W3C (<http://www.w3c.org/>) 对可扩展标记语言 XML 以及资源描述框架 RDF 规范的制定, 为 Web 站点提供元数据会变得越来越容易, 自适应 Web 站点研究也必将跨上新的台阶。

自适应 Web 站点研究才刚刚起步, 还有大量问题值得研究, 例如防止站点被用户擅自修改; 防止修改导致系统的崩溃; 使自动生成的页面更加美观……。

5 自适应 Web 站点的质量评价

评价自适应 Web 站点的质量是一件令人头痛的事情, 但这对深化自适应 Web 站点的研究具有重要意义。最基本的评价指标是看用户平均作多少“努力”才能在该 Web 站点中找到自己需要的信息。用户的“努力”可以简单地定义为点击超链接的次数以及在页面中寻觅这些超链接的困难程度的函数。例如, 从一个 Web 站点的首页至少要经过五个超链接才能到达其最受欢迎的页面, 那么在首页上添加直接指向最受欢迎的页面的超链接显然将提高该 Web 站点的质量。

通过分析 Web 站点日志, 我们可以得到用户所作“努力”的近似值, 但是还不足以完全刻画用户浏览行为。例如, 日志并不能区分来自同一个 IP 地址的不同用户 (当多个用户共享一台计算机或者通过 proxy 代理时就会出现这样的情况), 也没有记录用户到底是点击当前页面中哪一个超链接而进入下一个页面的。现在已经出现一些软件, 能够让 Web 站点记录下用户在站点上浏览时的全部行为, 如 WebThread (<http://www.webthread.com>)。将 WebThread 之类的软件提

供的数据与 Web 站点自身的结构综合起来分析, 可望比较准确地度量用户的“努力”。

除此之外, 自适应 Web 站点还可以就某个主题作专门的受控测试。这种测试允许我们仔细观察用户与 Web 站点的交互过程, 得到比日志更详尽的信息。测试主题一般是完成某种任务, 如寻找特定文档、下载软件或注册信息等等。测试数据一般包括任务的完成率、任务的完成时间、经过多少步操作等等。Web 站点还可以鼓励用户以在线表单或电子邮件的方式提出意见, 作为日志分析与受控测试的补充。

结束语 自适应 Web 站点将不仅仅是 HTML 页面和超链接的集合, 而成为不断进化的智能主体。广义的自适应 Web 站点不仅能够根据用户的访问模式自动地改进 Web 站点上信息的组织与显示, 还包括自动地为市场营销提供建议, 支持电子商务的发展。IDC (国际数据公司) 的一份报告预测这方面的软件到 2002 年之前将拥有超过一亿美元的市场价值, 我们认为, 自适应 Web 站点研究不仅具有广阔的商业前景, 而且将极大地推动计算机科学的发展。

参考文献

- 1 Anantharaman T, et al. Singular extensions: adding selectivity to brute-force searching. *Artificial Intelligence* 1990, 43(1): 99~109
- 2 Armstrong R, et al. Webwatcher: A learning apprentice for the world wide web. In *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments*, Stanford University, AAAI Press, 1995. 6~12
- 3 Chen M-S, et al. Data Mining for Path Traversal Patterns in a Web Environment. In: *Proc. of the 16th Intl. Conf. of Distributed Computing System*. 1996. 385~392
- 4 Etzioni O. Moving up the information food chain: solibots as information carnivores. In: *Proc. 14th Nat. Conf. on AI*. 1996
- 5 Fernandez M, et al. System demonstration-strudel: A web-site management system. In *ACM SIGMOD Conf. on Management of Data*. 1997
- 6 Fink J, et al. User-oriented adaptivity and adaptability in the avanti project. In *Designing for the Web: Empirical Studies*. Microsoft Usability Group, Redmond (WA), 1996
- 7 Kautz H. A Formal Theory Of Plan Recognition. [PhD thesis] University of Rochester, 1987
- 8 Lesh N, Etzioni O. A sound and fast goal recognizer. In: *Proc. 14th Int. Joint Conf. on AI*. 1995. 1704~1710
- 9 Luke S, et al. Ontology-based web agents. In: *Proc. of the First Interl. Conf. on Autonomous Agents*. 1997
- 10 Perkowitz M, Etzioni O. Adaptive web sites: an AI challenge. In: *Proc. of the Fifteenth Intl. Joint Conf. on Artificial Intelligence*. 1997
- 11 Perkowitz M, Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages. In: *Proc. of the Fifteenth National Conf. on Artificial Intelligence*. 1998
- 12 Pollack M. Plans as complex mental attitudes. In: Cohen P, et al. eds. *Intentions in Communication*. MIT Press, Cambridge, MA, 1990. 77~101
- 13 Thorpe C. *Vision and Navigation: the Carnegie Mellon Navlab*. Kluwer Academic Publishing, Boston, MA, 1990