第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于用户行为分析的网站结构优化研究综述

栗 辉 唐 萌 陈 豪

(北京科技大学自动化学院 北京 100083)

(北京科技大学计算机与通信工程学院材料领域知识工程实验室北京市重点实验室 北京 100083)

摘 要 基于用户行为分析的网站结构优化是Web挖掘领域内的主要研究方向。通过对国内外文献的归纳概括,综述了目前基于用户行为分析的网站结构优化的国内外研究现状,并通过对比分析,指出了各研究方法的优缺点,最后讨论了未来的研究方向。

关键词 用户行为分析,网站结构优化,Web挖掘

中图法分类号 TP391 文献标识码 A

Summary of Research on Website Structure Optimization Based on User Behaviour Analysis

LI Hui TANG Meng CHEN Hao

(School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China)

(Beijing Key Laboratory of Materials Science Knowledge Engineering, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract Website structure optimization based on the analysis of user behaviours has been regarded as the major research direction of Web mining. In accordance with making generalizations and summaries to literature at home and abroad, we reviewed the present situations on the research of website structure optimization based on the analysis of user behaviours. Furthermore, by the contrast analysis, we indicated the merit and demerit of research methodology respectively. Moreover, discussion on research directions in future was included as a consequence.

Keywords User behaviour analysis, Website structure optimization, Web mining

1 引言

设计人员在设计网站之初,就已经将网站结构敲定。然而,设计时的网站结构跟用户的实际期望之间,总会存在着或大或小的差异。这种差异的存在,就为网站结构的优化提供了空间。而且,随着当前互联网越来越“以人为本”,用户个性化的需求越来越高,基于用户行为分析的网站结构优化逐渐成为研究重点。

2 用户行为分析国内外研究现状

人有不同的认知风格,这些风格会影响人们组织和处理信息的方式,而且心理学的研究表明,基于用户心理特征的信息耦合方式会极大地影响用户在浏览网站时的效率[1]。正因为如此,人们在研究网站结构优化时对用户行为给予了极大的关注。

用户行为分析的过程实际上是从海量数据获得有价值信息的数据挖掘过程[2]。

2001年,Srikant和Yang提出了一种算法,根据用户当前访问路径,寻找那些不在用户期望位置的网页,他们利用人为定义的时间阈值来区分目标页面和其他页面,并依据点击率对网页作相应推荐[3]。

2006年,吴跃进认为,能够成为Web用户聚类算法评价因素的参数有且仅有3个,分别是点击次数、访问时间和访问路径;并在此基础上利用Kruskal算法衍生出了K-Bacer算法,根据访问频繁路径对用户进行聚类[4]。吴跃进将所有用户的访问序列生成无向图,通过K-Bacer算法找出其中的频繁路径。K-Bacer算法利用Kruskal算法的思想产生最小生成树,其根源是贪心算法。算法的时间复杂度依赖于排序算法,同时对所有用户生成同一个无向图,随着用户量的增加,其可维护性和可扩展性大大降低。

2007年Mihara等人将浏览时间作为用户兴趣的指标之一,提出了一种启发式方法,用来抽取用户访问模式[5]。该方法与Srikant等人所用的时间阈值的思想一致,但在阈值设定上更加贴近现实,根据不同情况设定不同的阈值,弥补了同一阈值带来的用户趋同化。

2012年,王有为等人利用序列模式分析用户访问行为,将用户的当前访问序列长短作为滑动窗口大小,从已有的频繁序列集中找出与滑动窗口大小最匹配的序列,将其后续页面作为推荐页面[6]。此种方法最大的问题在于推荐的页面严重依赖已有的频繁序列,可以说是被访问日志束缚了,根据其实验结果,所推荐的页面范围比较窄,而且由于没有对用户进行分组,导致对每一个用户都需要扫描一遍数据库,对数据库

本文受名老中医临床经验、学术思想传承研究(一),临床一线跟师人员信息采集实用软件研发课题(2013BAI13B06)资助。

栗 辉(1989—),女,硕士,助理工程师,主要研究方向为计算机应用;唐 萌(1990—),女,硕士生,主要研究方向为数据挖掘。

和服务器造成的压力较大。

2012 年,台湾的 Jia-Jiunn 等人在充分研究用户认知风格的基础上,利用多层前馈神经网络(Multilayer Feed Forward Neural Network,MLFF)算法对匿名用户认知风格进行聚类,并将结果用于网站交互优化[7]。他们认为许多学者对用户的分析都停留在对用户特点(例如目标、知识、背景等)的研究上,却没有充分考虑用户的认知风格对行为所带来的影响。在心理学者 Lipsky、Silver 等人的研究基础上,他们将用户分为控制型、交互型、理解型和自我表现型 4 类,然后通过对志愿者的测试,确定了这 4 类用户认知风格的评估标准,并将其用于 MLFF 算法。

还有部分研究者如 Dragos、余铁军等人提出了在对用户聚类时采用模糊聚类的方法,使用户不再被唯一地归为一类[8,9]。这种思想的优点在于考虑到了用户认知风格中的边界性,即用户的认知风格不是一成不变的,存在一定程度上的兴趣点漂移(User's Interests Drift)现象。

2014 年,张力平提出了一种用户行为模式的挖掘工作流程。他采用一种新的建模方式对日志库进行用户行为建模。即根据 Web 用户行为的分层特性,行为模式可分为 URL 访问、活动、会话 3 个层次[10]。再对建模后的用户行为序列进行频繁序列模式挖掘,得到频繁行为序列集后再进行行为序列的聚类分析。

通过上述总结分析,可以得出表 1。

表 1 用户行为分析方法对比

研究人员	分析方法	用户所属类个数	特点	缺点
Srikant 等	序列分析	无	根据时间阈值、点击率确定推荐页面	参数选择单一,不能充分反映用户行为
吴跃进	聚类分析 关联分析	1	根据 3 个参数确定频繁路径并以此聚类	随用户数增加维护性、扩展性降低,用户所属类单一
Mihara 等	序列分析	无	启发式时间阈值设定	参数选择单一,不能充分反映用户行为
王有为等	序列分析	无	根据滑动窗口长度确定推荐页面	计算量大,严重依赖已有序列
Jia-Jiunn 等	聚类分析 关联分析	1	充分研究用户认知风格、MLFF 算法	依赖心理学研究理论,用户认知风格可靠性有待商榷
Dragos、余铁军等	聚类分析 关联分析	多个	模糊聚类、用户分属多个类	选择参数单一,不能充分反映用户行为
张力平	序列分析 聚类分析	多个	对用户行为分层特性建模	对用户行为的建模可靠性有待商榷

此外,Jia-Jiunn 等人的工作让我们看到了用户行为分析的另一个重要方向,即从心理学的角度来分析用户。然而由于学科之间的跨度等问题,目前的研究工作更多的还是放在对现有数据的统计分析上,但侧重心理学或两者兼顾的问题解决策略势必成为将来的发展趋势。

3 网站结构优化的国内外研究现状

在网站结构优化领域,国内外的学者做出了不少努力。

1997 年,Mike Perkowitz 和 Oren Etzioni 提出了自适应网站(Adaptive Web Sites)的概念,即网站通过对用户访问模式的学习,自动地改变其组织结构和展示内容[11]。文献同时提出了 PageGather 算法,从服务器日志(也称访问日志或 Web 日志)中提取用户访问信息并构造关联图,通过对页面的聚类生成反映用户兴趣的索引页面,并以此来提高用户浏览体验。但该算法生成聚类后,需要管理员手动选取,只有选中的聚类才会成为索引页面候选集。此外,两个页面间的相似性仅限于是否有链接相连,没有考虑页面内容,也可能会破坏网站固有结构。

2001 年,香港大学的 Yen 等人将网页可达性和知名度用于网页链接,用无权有向图来描述网站,通过对网页进行分层,结合网页期望链接数和访问率,对网站结构做出调整[12]。文献简单地用网页的期望链接数来表示可达性,用访问率来表示知名度,并认为可达性和访问率应该成正比。然而从现实来讲,网页的知名度并不单纯体现在访问率上,还有访问用户数等其他因素。若考虑极端情况,某用户频繁点击某网页,则会使该网页的知名度迅速上升,从而导致网站结构出现错误调整。

2007 年,Hamed Qahri Saremi 等人利用图论定义网站模型,将二次分配问题(Quadratic Assignment Problem)扩展应用到网站链接结构分析中[13]。文献采用启发式蚁群算法求解二次分配问题,旨在对网页进行定位分配,从而为用户提供导航服务,提高网站可用性。在定义网页间距离的时候,主要考虑了访问次数、连通度和页面深度,这在一定程度上严重依赖现有网站结构,且同样没有考虑页面内容的相关性。

2008 年,黄艳欢等人提出了基于协作反馈的蚁群算法,并使用该算法对网页进行关联性分析,同时根据用户访问日志做出系统推荐[14]。文献针对传统蚁群算法中蚂蚁间互信息交换的不足,提出适合网站结构优化的改进,加强了蚂蚁之间的协作性和反馈性。文献提出的方法是一种协同过滤机制的变体,因此存在协同过滤本身固有的不足,比如在冷启动问题上的疲软;另一方面,文献未考虑页面之间的相互作用,单纯从用户的角度出发,忽略了网站结构的自身特征。

2009 年,王洪伟等人提出将 Web 挖掘与站点拓扑结构相结合的方式,利用结点连通度、结点深度、结点偏好度以及地标系数等指标筛选出网站中的重要结点,并采用高亮显示、动态地图和缓冲预取的策略为用户提供自适应服务[15]。然而,文献中的多个参数需要手动设置,使得主观因素对结果的影响较大。同时,选取访问路径长度及结点的减少率作为评价指标并不能很好地体现该方法的实际效果。

2009 年,程舒通等人将网站结构优化模型归纳为 4 个部分,分别是数据的采集、预处理、模式的发现和分析[16]。文献论述了这 4 个部分所涉及的主要算法和相关技术,并对该领域未来的发展方向做了展望。

2010 年,Shian-Hua Lin 等人将网页 HTML 文件按照内容、超链接等分割到不同的块,通过对块聚类,得到一系列具有内容、链接相关性的网页集合,再对这些集合进行分级,最终形成网站地图生成器[17]。文献在网站地图生成方面,形成了一套完善的方案,但由于没有将用户的访问行为考虑在内,无法从用户角度吸收知识和经验,导致生成的网站地图难免有失偏颇。

2012 年,M. R. Martinez-Tores 等人提出了一种基于渐进式因子分析估算的网站结构挖掘方法[18]。该方法将网站结构分解成域网(Domain Net)和页面网(Page Net)两个连通图,并将其作为社交网络[19],通过分析该网络中的多个因子,采用渐进式遗传算法,挖掘出最佳站点结构。虽然考虑了页面内容和链接两方面,但该方法相关因子数太多,且选取过程较为繁琐,在计算上复杂度较高。

2013 年,台湾的 Peng-Yeng 等人通过改进禁忌搜索(Tabu Search)算法增加自适应禁忌列表,提出了 ETS(Enhanced Tabu Search)算法,用来解决多约束条件下的网站结构优化问题,并在商业实践中予以应用[20]。文献将连通度、出度、基本链接、页面聚类、账户安全、站点深度等多个条件作为约束,对页面的访问进行分析,利用 ETS 算法搜寻最佳路径。然而 ETS 算法对网站的规模有比较严格的限制,一旦网站规模较大,页面数量增加,会导致算法的运行时间急剧上升,因此并不适合大型网站。

除此之外,Corin[21]、Lempel[22]和 Rafiei[23]等人提出了基于马尔可夫链的站点链接或网页知名度分析方法;王有为[6]等人利用改进的 PrefixSpan 算法来寻找访问序列中的频繁模式,从而生成推荐网页集合;Asllani[24]和杜华[25]等人通过分析总结多约束条件,利用遗传算法对网站结构进行优化,降低网页平均负载。

通过分析总结上述目前国内外网站结构优化领域的不同研究方法,得到了表 2。

表 2 网站结构优化的国内外研究现状对比

研究人员	数据来源	研究内容	参数	人为干扰	破坏结构	复杂度	评价指标
Perkowitz	Web 日志	页面	链接关系	是	是	低	点击率
Martinez	Web 日志	页面序列	很多	是	未知	高	很多
Hamed	Web 日志	页面	访问次数、连通度、页面深度	否	否	高	站点总连通度
Lin	HTML 文件	页面	HTML 标签、文本长度 权威值、枢纽值	否	否	高	召回率 查全率
王洪伟	Web 日志	页面 序列	页面深度、连通度 偏好度、地标系数	是	否	低	路径长度 结点减少率
黄艳欢	Web 日志	页面	序列长度、结点关联度、信息素	否	否	高	召回率、查准率
Yen	Web 日志	页面	网页可达性、知名度	否	是	低	未知
Peng-Yeng	Web 日志	页面 序列	连通度、页面出度、基本链接 页面聚类、账户安全、站点深度	否	否	高	CPU 时间
王有为	Web 日志	序列	滑动窗口	否	否	低	查准率、召回率 F 测度
杜华	Web 日志	页面	页面出度、序列长度	否	是	高	点击率、页面负载

从表 2 中可以看出,目前在网站结构优化领域,研究人员基本上都是从 Web 日志中获取数据,通过对 Web 日志进行数据挖掘,从而对页面或访问序列进行分析。

一方面,人们对页面的关注度要高于访问序列;另一方面,在对页面进行分析时,除了 Lin 等人以外,人们更多地关注其链接所带来的关联关系,很少考虑页面内容之间的相关性,这在一定程度上削弱了分析结果的可靠性。当然,也有学者将网站结构优化分为两类:1)基于用户行为评估站点结构存在的问题;2)基于站点模型的方法,而不考虑用户行为的影响[26]。

结束语 网站结构优化能够提高用户的使用体验,增加网站流量进而创造经济价值。随着互联网逐步深入国民经济的各个领域,网站类型及数量不断增多,该技术的应用场景将越来越多。本文对当前网站结构优化领域内的主要研究方法作了综合对比研究,重点综述了基于用户行为分析的研究成果及其优缺点。在未来的发展中,通过 Web 日志,结合心理学理论,对用户行为进行建模,深度挖掘用户使用心理并将其应用于网站结构优化将是重要方向。此外值得一提的是,随着网站的不断“云”化,会逐渐产生数据量更大、结构更复杂多样的网站,这对网站结构优化技术的发展既是挑战又是机遇。

参 考 文 献

[1] Carver C A, Howard. Enhancing student learning through hypermedia courseware and incorporation of student learning styles[J]. IEEE Transactions on Education, 1999, 42: 33-38

[2] 刘鹏. 网络用户行为分析的若干问题研究[D]. 北京: 北京邮电大学, 2010

[3] Srikant R, Yang Y. Mining web logs to improve website organization[C]// Proceedings of the 10th International Conference on World Wide Web 2001. Hong Kong: ACM New York, NY, 2001: 430-437

[4] 吴跃进. 综合多重评价因素的 Web 用户聚类算法[J]. 计算机工程与应用, 2006(28): 147-148

[5] Masahiro M K. A proposal of web log mining method considering page browsing time[J]. Information Processing Society of Japan SIG Notes: ICS, 2007, 67: 39-44

[6] 王有为, 张雯晶, 凌鸿. 基于序列模式的网站导航系统[J]. 系统管理学报, 2012, 21(5): 690-695

[7] Lo J J, Wang Y J. Development of an Adaptive EC Website With Online Identified Cognitive Styles of Anonymous Customers[J]. International Journal of Human-Computer Interaction, 2012, 28(9): 560-575

[8] Arotaritei D, Mitra S. Web mining: a survey in the fuzzy framework[J]. Fuzzy Sets and Systems, 2004, 148(1): 5-19

[9] 余轶军. Web 访问信息挖掘若干关键技术的研究[D]. 杭州: 浙江大学, 2009

[10] 张力平. Web 用户行为模式挖掘及其在 E-Learning 系统中的应用[J]. 软件导刊, 2014, 13(6): 117-119

[11] Perkowitz M, Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages[C]// Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference. 1998: 727-732

[12] Yen B P C, Fu K. Accessibility on Web Navigation[C]// International Conference on E-Commerce Engineering: New Challenges for Global Manufacturing in the 21st Century. 2001

(下转第 394 页)

根据程序所得排序，转换为填报后的院校名称为：

提前批一志愿：西安电子科技大学

提前批二志愿：东北大学

本科一批一志愿：山东大学

本科一批二志愿：兰州大学

本科一批三志愿：吉林大学

本科一批四志愿：华南理工大学

学生$_3$：由上两表可知，本次算例中学生$_3$的最终基因库为[31.43，37.46，27.85，19.25，27.01，17.54，20.11，27.54，18.15]。

经过程序计算，迭代$_{50000}$次后的程序结果见图$_3$。

图$_3$ 学生$_3$的实验结果

根据程序所得排序，转换为填报后的院校名称为：

提前批一志愿：燕山大学

提前批二志愿：陕西科技大学

本科一批一志愿：山东大学威海分校

本科一批二志愿：中国矿业大学

本科一批三志愿：东北大学

本科一批四志愿：北京信息科技大学

由$_3$位不同分数阶段的学生的实验结果可知，该算法切实可行，不同分数的考生冲刺、保底学校不同，且申报学校基本能够符合考生的实际需求。

结束语 由于高考录取规则在各地、每年都有所变化，因此可以根据不同地区、不同年份对最终排序值的计算方法进行调整。同时也可以加入更多院校的信息从而使得最终的计算结果更加趋近于实际。

同时也可以将提前批的院校与本科一批的院校分别形成两个基因库并分别计算，也可以在后续加入本科二批、三批的院校中进行排序。但由于作者精力、能力及针对高考院校填报的资源有限，因此本文算法的设计中并未将这些因素引入其中。

解决本问题的算法不仅可以应用于高考志愿的填报，同时可以应用于其他排序问题如任务的分配、资源的分配等等的解决，不仅对其自身的研究有重要的参考价值，而且对其应用也有重要意义。

参 考 文 献

[1] 王彬. 我国高考制度改革的价值取向研究[D]. 上海：上海师范大学，2013

[2] 李凤. 高考志愿填报与录取机制研究[D]. 成都：西南财经大学，2010

[3] 陈劲松. 高考志愿选择与未来就业的关系研究[D]. 武汉：华中师范大学，2008

[4] Holland J H. Adaptation in Natural and Artificial Systems [M]. Ann Arbor: University of Michigan Press, 1975

[5] 李敏强，等. 遗传算法的基本理论与应用[M]. 北京：科学出版社，2002：13-15，18-47，163-199

[6] 蒋冬初. 遗传算法及其在函数优化问题中的应用研究[D]. 长沙：湖南大学，2004

[7] 张志平. 基于遗传算法的汉语基本词汇自动提取研究[D]. 呼和浩特：内蒙古师范大学，2007

[8] 高浩. 适应度估算遗传算法及其应用[D]. 吉林：吉林大学，2011

[9] 高考志愿填报与录取程序解答[J]. 山西教育(高考版)，2007(9)：4-11

[10] 赵舒展. 遗传算法研究与应用[D]. 杭州：浙江工业大学，2002

[11] 张思才，张方晓. 一种遗传算法适应度函数的改进方法[J]. 计算机应用与软件，2006(2)：108-110

[12] 唐勇，唐雪飞，王玲. 基于遗传算法的排课系统[J]. 计算机应用，2002(10)：93-94，97

[13] 丁建立，慈祥，黄剑雄. 一种基于免疫遗传算法的网络新词识别方法[J]. 计算机科学，2011(1)：240-245

(上接第386页)

[13] Saremi H Q, Abedin B. Website structure improvement: Quadratic assignment problem approach and ant colony meta-heuristic technique[J]. Applied Mathematics and Computation, 2007, 195(1): 285-298

[14] 黄艳欢，何振峰. 基于协作反馈的蚁群算法的自适应网站研究[J]. 福州大学学报(自然科学版)，2008，36(6)：814-818

[15] 王洪伟，刘勰，廖雅国. 基于Web挖掘和站点拓扑的自适应网站研究[J]. 数学的实践与认识，2010，40(17)：31-39

[16] 程舒通，徐从富. 网站结构优化技术研究进展[J]. 计算机应用研究，2009，26(6)：2013-2015

[17] Lin S H, Chu K P. Automatic sitemaps generation: Exploring website structures using block extraction and hyperlink analysis [J]. Expert Syst. Appl., 2011, 38(4): 3944-3958

[18] Martinez-Torres M R, Toral S L, Palacios B. An evolutionary factor analysis computation for mining website structures[J]. Expert Systems With Applications, 2012, 39(14): 11623-11633

[19] Martinez-Torres M R, Toral S L. Web site structure mining using social network analysis[J]. Internet Research, 2011, 21(2): 104-123

[20] Yin Peng-yeng, Guo Yi-ming. Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining [J]. Applied Mathematics and Computation, 2013, 5(11): 11082-11094

[21] Anderson C R, Domingos P, Weld D S. Adaptive Web Navigation for Wireless Devices[C]// Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. 2001

[22] Lempel, Moran R S. The stochastic approach for Link-structure analysis (SALSA) and the TKC effect[J]. Computer Networks, 2000, 33: 387-401

[23] Rafiei, Mendelzon A O. What is this page known for? Computing Web page reputations[J]. Computer Networks, 2000, 33: 823-835

[24] Asllani A, Lari A. Using genetic algorithm for dynamic and multiple criteria web-site optimizations[J]. European Journal of Operational Research, 2007(176): 1767-1777

[25] 杜华，王锁柱. 网站结构优化模型及算法分析[J]. 计算机工程与设计，2008，29(21)：5591-5594

[26] 易名，杨斌. 站点结构优化方法研究综述[J]. 情报分析与研究，2008(7)：61-65