文章编号:1005－0523(2006)02－099－02

一个基于 Web 挖掘的网站优化系统

徐晓玲

(华东交通大学 电气与电子工程学院,江西 南昌 330013)

摘要:利用 web 数据挖掘方法对网站进行优化,提出了网站优化系统的基本框架,并给出相应的 web 数据挖掘算法:频繁模式发现及序列模式挖掘两个主要算法以找到网站访问者感兴趣的页面集合以及频繁使用的链接序列,该结果以改进网站的设计,提高网站的访问率.

关 键 词:web 挖掘;网站优化

中图分类号:TP391.41　　**文献标识码**:A

1 引言

在竞争日益激烈的网络经济中,只有赢得用户才能最终赢得竞争的优势.网站是企业进行信息发布的平台,是企业对外的形象和窗口.针对网站目前的访问信息,对网站的设计进行优化,建设一个更合理、更易用、注重个性化和相关性的网站,以此提高用户满意度,提升网站知名度,带来切实的经济和社会效益.网站优化后不仅使得用户易于使用,并获得个性化服务;而且对于网站运营者而言,能为用户提供更加合理的网站结构、美观安全的页面、优秀的内容和良好的运行环境.

现阶段,智能化网站优化的主要途径为:利用数据挖掘方法进行网站优化(即 Web 挖掘).Web 挖掘是数据挖掘在 Web 上的应用,它利用数据挖掘技术从与 WWW 相关的资源和行为中抽取感兴趣的、有用的模式和隐含信息,涉及 Web 技术、数据挖掘、计算机语言学、信息学等多个领域,是一项综合技术,许多软件公司和研究机构在此方面已投入了许多研究和开发工作.具有影响力的 Web 使用挖掘系统有:德国柏林 Humboldt 大学的 Web Utilization Miner、加拿大 Simon Fraser 大学的 WebLogMiner、IBM 公司的 SpeedTracer、美国明尼苏达大学的 WebSIFT 等.由于这些系统主要数据源是 Web 日志,用途是为网站的设计者和管理者提供网站的使用情况分析,从而为网站优化提供帮助,对商业决策的帮助很少.Büchner 等人实现了一个面向客户关系管理的 Web 日志挖掘系统.此系统的功能是客户关系管理,但是只管理注册用户.其它面向电子商务的 Web 数据分析工具有 IBM 公司的 SurfAid,BlueMartini,ECOM-MINER 等.现有的此类 Web 数据分析工具的功能主要注重信息统计,而不是数据挖掘功能.本文利用 web 挖掘技术设计一个网站优化系统,并给出频繁项集发现算法用以挖掘网站中被浏览者频繁访问的网页集合以及序列模式挖掘算法用以挖掘网站中频繁访问的页面序列,这两种结果用于改进网站设计,提高网站运营效益.

2 Web 挖掘

Web 挖掘是应用数据挖掘技术自动从 Web 文档和服务中发现和抽取感兴趣的、潜在的、有用模式和隐藏的信息.Web 挖掘可在很多方面发挥作

收稿日期:2005－09－8

作者简介:徐晓玲(1962－),女,江西南昌人,副教授.

用,如对搜索引擎的结构进行挖掘、确定权威页面、Web 文档分类、Web 日志挖掘、智能查询等.

Web 挖掘可分解为以下几个子任务:

1) 资源发现:从 Web 上检索期望的 Web 文档和服务.

2) 信息抽取和预处理:从已检索的 Web 资源中自动抽取和预处理指定信息.

3) 一般化:在 Web 站点自动发现通用模式.

4) 分析:已挖掘模式的确认与解释.

根据使用的 Web 数据种类的不同,Web 挖掘可分为三类研究:Web 内容挖掘(Web Content Ming),Web 结构挖掘(Web Structure Ming)和 Web 使用挖掘(Web Usage Mining).

2.1 Web 内容挖掘

Web 内容挖掘是从文档内容或其描述中抽取知识的过程.Web 内容挖掘揭示网页的主题,但并不关心谁会真正阅读它.包括两种策略:Web 文档挖掘和搜索结果挖掘.采用第一种策略的方法是直接挖掘文档的内容;采用第二种策略的方法主要是对搜索引擎的查询结果进行进一步的处理,得到更为精确和更为有用的信息.常见的 Web 内容挖掘技术主要有对 Web 上大量文档集合的内容或搜索结果的文本摘要、分类、聚类、关联分析,以及利用 Web 文档进行趋势预测等.

2.2 Web 结构挖掘

Web 内容挖掘仅将 Web 看作是一个平面文档的集合,而忽略了其中的结构信息,然而,Web 不仅由页面组成,而且由链接页面的超链接组成,超链接环境的网络结构具有非常丰富的信息,包含了大量的潜在的人工注释.Web 结构挖掘是从 Web 的组织结构和链接关系中推导知识,揭示了哪些页面通过当前页面可以两步内到达,但并不关心多少人会实际用到这条通路.通过挖掘 Web 结构可以发现页面的结构和 Web 的结构,在此基础上对页面进行分类和聚类从而找到权威页面.这方面工作的代表有 PageRank 和 CLEVER.

2.3 Web 使用挖掘

Web 使用挖掘是通过分析和探究 Web 访问记录中的规律,从中抽取感兴趣的模式.

主要包括两个方面:一般的访问模式追踪和个性化的使用记录追踪.一般的访问模式追踪通过分析使用记录来了解用户的访问模式和倾向,以改进站点的组织结构,而个性化的使用记录追踪则倾向于分析单个用户的偏好,其目的是根据不同用户的访问模式,为每个用户提供定制的站点.

Web 使用挖掘的整个过程可以分为:数据预处理,模式发现和模式分析等任务.数据预处理阶段根据数据挖掘要求把原始数据转换成挖掘算法可用的数据,必要时把它加载到数据库中.模式发现阶段使用各种数据挖掘方法分析可用的数据,这些方法从简单的统计分析,到计算量很大的关联规则、序列模式、分类、聚类等.模式分析阶段是由于模式发现的结果往往包含许多无用的模式,需要采用各种方法(如兴趣度、模式可视化)分析和过滤出需要的模式.

3 网站优化系统基本框架

网站优化系统一共包括三个主要模块:日志预处理、结构抽取和结构优化.日志预处理属于数据预处理阶段,该模块包括将日志导入数据库表中,数据库表中的字段根据日志格式中的字段进行选择定义(见第 2 小节);数据清理从 web 日志文件中过滤无关的页面请求(如图形等)以及不成功的页面请求;识别用户采用 IP+Agent 机制实现;路径修补将由于缓冲而造成日志文件中丢失用户访问链接记录根据站点结构补充完整以便后续结构优化模块提供准确的用户访问记录.

结构抽取模块是指抽取网站链接结构,输出网站结构拓扑图,可用于辅助结构优化、日志解析中,如路径修补等.

结构优化是模式发现和模式分析阶段,该模块包括频繁访问模式分析和序列模式分析两部分.频繁访问模式分析给出频繁访问的页面集合,展示页面的重要程度,展示访客来源分类;页面的访问情况,访客百分比,停留时间,离开百分比,以图表形式展现.序列模式分析任务是找到方便用户易于使用的搜索合理的页面链接组合.

图 2 为网站结构优化系统的基本框架图.

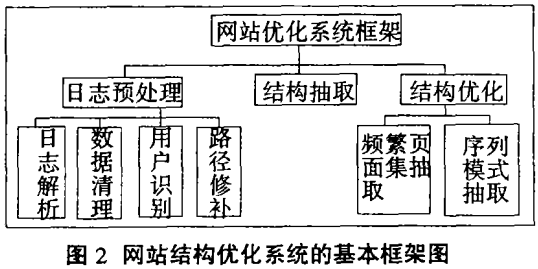

图 2 网站结构优化系统的基本框架图

(下转第 110 页)

The Application of Genetic Algorithm in Reactive Power Optimization

HUANG Ying, CHEN Ken

(Information Engineering School, Nanchang University, Nanchang 330029, China)

Abstract: The problems of reactive power optimization with genetic algorithm are discussed in detail in this paper. The crossover, mutation and inversion operations are proposed which not only reduces the search space, but also avoids the infeasible solutions produced during initialization and gene operations. The proposed genetic algorithm has been tested in an IEEE 30—bus power system. At the same time, based on the above genetic algorithm, network less of electric power systems can be effectively reduced, and then reactive power optimization can also be realized.

Key words: reactive power optimization; genetic algorithm

(上接第 100 页)

4 结论

Web 挖掘研究是几种研究领域, 如信息检索 IR、人工智能 AI 等会合的研究领域. Web 数据挖掘并不是一个单一的活动, 而是许多活动的集合. 频繁模式和序列模式挖掘算法应用于网站优化系统中表现出很好的效果以发现用户感兴趣的页面集合以及页面链接序列. Web 挖掘应用于网站优化系统使得网站访问率提高, 为网站赢得更多利润.

参考文献:

[1] 韩家炜, 孟小峰, 王 静, 李盛恩. Web 挖掘研究[J]. 计算机研究与发展, 2001, (4): 14.

[2] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques [M]. Copyright by Morgan Kaufmann Publishers, Inc, 2001

[3] On creating Adaptive Web Servers using Weblog Mining Technical report CS—TR—00—05, CSEE Department, UMBC, 2000.

[4] Low Complexity Fuzzy Relational Clustering Algorithms for Web Mining, IEEE Trans. Fuzzy Systems, 9: 4, pp 596——607, 2001.

[5] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," in SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, ACM Press, 2000, pp. 1——15.

[6] 宋擒豹, 沈均毅. web 日志的高效多能挖掘算法[J]. 计算机研究与发展, 2001, 38(3): 328～332.

[7] 柳炳祥, 盛昭翰. 一种基于 web 挖掘的网站性能评价方法[J]. 计算机工程与应用, 2003, (4)189—191.

A Website Optimize System Based on the Web Mining

XU Xiao-ling

(School of Electrical and Electronic Engineeing, East Chin Jiaotong Univousity, Nanchang 330013, China)

Abstract: This paper uses Web mining to optimize website and propose a basic framework of website optimize system. We also develop corresponding web mining algorithm which are frequent pattern mining and sequential pattern mining to find the webpage set and frequent link webpage sequences which visitors who explore the website are interesting in. These results will improve the website design.

Key words: Web mining; website optimize