第36卷 第19期 计 算 机 工 程 2010年10月
Vol.36 No.19 Computer Engineering October 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)19—0258—03 文献标识码：A 中图分类号：TP391

基于 Lucence 的个性化搜索引擎研究

李晓丽，杜振龙
(南京工业大学电子与信息工程学院，南京 210009)

摘 要：针对通用搜索技术难以满足不同背景、不同目的和不同时期的用户查询请求的缺陷，提出一种基于 Lucence 实现用户兴趣驱动的个性化搜索引擎方法。从 Cookie 文件分析用户搜索兴趣，构造用户兴趣向量，驱动搜索引擎，产生用户关注度高的搜索结果。实验结果表明，该用户兴趣驱动的个性化搜索引擎能够搜索出用户感兴趣的搜索结果，与传统搜索引擎相比，检索准确率有一定的提高。

关键词：Cookie 文件；个性化搜索；搜索引擎；兴趣向量；用户兴趣

Investigation on Personalized Search Engine Based on Lucence

LI Xiao-li, DU Zhen-long
(College of Electronics and Information Engineering, Nanjing University of Technology, Nanjing 210009, China)

【Abstract】Conventional search engine is unsuitable for search requests from people with different cultures, different goals or periods. A method on personalized search engine is proposed in the paper, which analyzes the user search pattern from Cookie, constructs the user interests vector from search pattern, drives the engine by search pattern, hence the presented approach can index the results with high correlation to user interests. It implements the personalized search engine based on Lucence, and experiments show that user interests driven personalized search engines can significantly index the user interested results.

【Key words】 Cookie file; personalized search; search engine; interest vector; user interests

1 概述

随着计算机技术和网络技术的飞速发展，互联网已在人们日常生活中发挥着越来越重要的作用，例如，利用互联网搜索信息、购物、交流、娱乐等。面对互联网用户数量的激增和信息的爆炸性增长，如何很好地利用互联网为用户快捷地提供所需服务和信息是一个值得研究的问题。本文提出了一种利用 Cookie 发掘用户访问模式，并用于定制搜索引擎以提高检索准确性的方法。通用搜索技术已满足了人们的一定需要，然而由于其通用性，仍不能满足不同背景、不同目的和不同时期的用户查询请求[1-3]。互联网检索信息具有一定的模式，用户利用这种模式辅助搜索可使搜索结果更加精确。个性化搜索引擎[3]从用户 Cookie 文件分析用户浏览兴趣，并用来驱动搜索引擎，搜索满足用户需要的搜索结果。

Cookie 技术[4]最先由 Netscape 公司提出并应用在 Navigator 浏览器，之后，Cookie 技术被 W3C 组织接受而定义为互联规范。目前包括 IE(Internet Explorer)在内的大多数浏览器都支持 Cookie 技术。Cookie 技术记录了用户访问互联网站点的有关信息，蕴藏着用户的访问模式，本文利用 Cookie 的记录信息发现用户搜索模式。

Lucene 是 Jakarta 提供的开源全文检索引擎，由站内数据采集器、全文索引器和检索器组成，站内数据采集器和索引器和检索器提供数据。通常的搜索引擎仅根据关键词搜索，没有考虑用户的浏览模式。本文对用户搜索模式建立分词词典，由检索关键词和分词词典来检索结果，因此搜索结果具有更明显的个性化特征，搜索结果也更符合用户需求。

2 基于 Cookie 的用户浏览模式发现

互联网浏览器所赖以生存的 HTTP 是无状态、无连接的协议，不能保存一次会话的连续状态信息。随着动态交互 Web 程序的出现，HTTP 的无状态特性严重阻碍了动态页面程序的实现。Cookie 正是为了保持 HTTP 连接状态而诞生的一项新 技术。

W3 组织发布的 RFC 文档[4]定义了 Cookie 机制的 Cookie 和 Set-Cookie 报头，其中，Cookie 报头由属性-值对组成，Set-Cookie 报头由“Set-Cookie:”和属性(预定义)-值对组成。同时，RFC 文档定义了 Set-Cookie 报头的 Comment、Domain、Max-Age、Path、Secure 和 Version 等属性，其中，Version 属性在 Cookie 文件必须要有，其他属性则可选。

2.1 Cookie 工作原理

Cookie 的工作原理如图 1 所示，首先由用户向服务器发送一个请求；服务器收到请求后产生一个 Set-Cookie 报头，放在 HTTP 报文中一起发送给用户，产生一次会话；用户收到应答后，若要继续该次会话，则将 Set-Cookie 中的内容取出，形成一个 Cookie 文件。

Cookie 文件是由服务器响应浏览器 URL 请求的信息组成，以文本文件保存在客户端。浏览器未退出前，Cookie 信息保存在内存；退出浏览器后，则保存在硬盘。不同浏览器保存 Cookie 文件的位置不同。

Cookie 的主要作用是当用户再次向服务器发送请求时，浏览器对 Cookie 文件进行解释并产生相应的页面。

基金项目：江苏省高校自然科学基金资助项目(09KJB520006)；南京大学软件新技术国家重点实验室开放基金资助项目(KFKT2008B15)
作者简介：李晓丽(1971－)，女，副教授、博士研究生，主研方向：数据库，人工智能；杜振龙，副教授、博士
收稿日期：2010-04-29 **E-mail：**lixl@njut.edu.cn

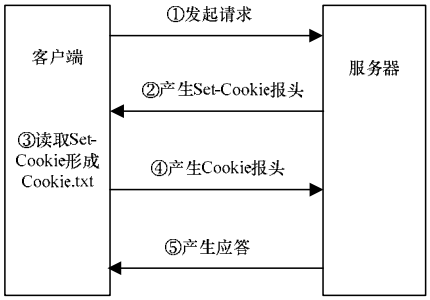

图 1　Cookie 运行过程

2.2　Cookie 信息提取

Cookie 是附属于 Response 和 Request 对象的数据集合，记录了用户所访问的 URL 和一些附加信息。Cookie 除了保存 RFC 文档定义的属性外，还记录与访问站点有关的信息。

设 Cookie 记录集合为 $\boldsymbol{C}$，$\boldsymbol{C}_i \in \boldsymbol{C}$ 是单个 Cookie 记录，$\boldsymbol{C}_i$={ d_i， u_i， v_i, …}，d_i 是域名，u_i 为 URL 访问用户，v_i 是上次访问 d_i 的时间。集合 $\boldsymbol{C}$ 蕴含了用户的搜索模式 $\boldsymbol{P}$。

用户 u_i 的浏览兴趣模式与用户访问的站点 d_i 有关，把用户一段时期内的访问站点收集起来得到 $\boldsymbol{d}=\{d_i\}$，从中所发现的用户感兴趣站点 $\boldsymbol{D}$ 具有较强的代表性。

设 t 为基准时间，$T_{\min}$、$T_{\max}$ 分别是距离 t 的最近时间、最远时间，则 $T_{\min}$、$T_{\max}$ 的计算分别如式(1)、式(2)所示：

$$T_{\min}=\{v_i \in \boldsymbol{C}_i \wedge \boldsymbol{C}_i \in \boldsymbol{C} \mid \min(t-v_i)\} \tag{1}$$

$$T_{\max}=\{v_i \in \boldsymbol{C}_i \wedge \boldsymbol{C}_i \in \boldsymbol{C} \mid \max(t-v_i)\} \tag{2}$$

用户 u_i 访问 d_i 距离基准时间 t 的长短可度量 u_i 的浏览兴趣，用概率 p 表示为

$$p(d_i \mid t, C_i)=\frac{T_{\min}}{v_i-t} \tag{3}$$

Cookie 文件集 $\boldsymbol{C}$ 记录了多个用户的访问记录，而且，即使同一登录用户也会以不同的用户名访问网站，因此，本文仅用同一登录用户下的 $\boldsymbol{C}$ 分析用户浏览兴趣模式。

设阈值 $T=\alpha\frac{T_{\min}}{T_{\max}}$，$\alpha$ 指定用户感兴趣模式的挖掘系数，则利用下式可发现用户感兴趣网站：

$$\boldsymbol{D}=\{d_i \mid p(d_i \mid t, u_i, \boldsymbol{C}_i)>T\} \tag{4}$$

3　用户兴趣向量构造

用户感兴趣网站集 $\boldsymbol{D}$ 包含了在 $T_{\max}$ 时间内用户感兴趣的网站，蕴含了用户兴趣模式 $\boldsymbol{P}$。用户兴趣模式发现是要从 $\boldsymbol{D}$ 中挖掘用户感兴趣关键词及权重，建立用户兴趣向量 $\boldsymbol{W}$。

Crawler[1]、Spider[5]等网络爬行器能够建立以 $\boldsymbol{D}$ 为基础的文档集 $\boldsymbol{R}$。从 $\boldsymbol{R}$ 中构建用户兴趣向量 $\boldsymbol{W}$ 有全自动和半自动 2 种方式：全自动方式进行全文搜索，根据 Web 页面的词语出现频率确定 $\boldsymbol{W}$；半自动方式事先给定一些词语，利用 $\boldsymbol{R}$ 重新设定每个词语权重以构成 $\boldsymbol{W}$。本文利用半自动方式构造 $\boldsymbol{W}$。

设 $\boldsymbol{W}=\{(\boldsymbol{w}_1, \boldsymbol{s}_1), (\boldsymbol{w}_2, \boldsymbol{s}_2), \cdots, (\boldsymbol{w}_n, \boldsymbol{s}_n)\}$，$\boldsymbol{w}_i (1 \le i \le n)$ 是给定一级分类词语，$\boldsymbol{s}_i (1 \le i \le n)$ 是对应的用户兴趣分量；每个 $\boldsymbol{w}_i$ 又包括多个二级分类词语，即 $\boldsymbol{w}_i=\{w_{i,1}, w_{i,2}, \cdots, w_{i,n}\}$。本文采用 Yahoo 的网站分类方法[6]设定 $\boldsymbol{w}_i$。例如，$\boldsymbol{w}_i$ 可取“常用站点”、“休闲娱乐”、“生活服务”、“实用查询”等，常用站点={中国雅虎，阿里巴巴，淘宝，……}。

计算用户兴趣向量是要根据 $\boldsymbol{D}$ 评估 $\boldsymbol{W}$ 的 $\boldsymbol{s}$ 分量。文档集 $\boldsymbol{R}$ 的分词方法和分词词典机制影响用户兴趣向量 $\boldsymbol{W}$ 的构造效率。常用的分词方法包括基于词典的分词方法和基于频度统计的分词方法[7]。分词词典机制包括整词二分的分词词典机制和 TRIE 索引树的分词词典机制。二分法分词词典以两字词为基本处理单位，处理效率高；基于 TRIE 索引树的分词词典用层次 Hash 表示分词，维护成本较高。由于本文所测试的 $\boldsymbol{R}$ 的页面数量多在 200~400 个左右，分词词典构造过程为离线方式，因此，本文采用 TRIE 索引树[7]的分词词典方式表示页面文档集 $\boldsymbol{R}$。

由于 $\boldsymbol{W}$ 的元素数较多，用数组表示 $\boldsymbol{W}$ 会引起多余的检索开销，导致检索效率低下。另外，$\boldsymbol{W}$ 的元素长度不一致，整词匹配会浪费大量的计算开销，因此，本文采用 TRIE 索引树结构表示 $\boldsymbol{W}$。

本文用算法 1 计算 $\boldsymbol{W}$，算法 1 的实质是基于树结构的匹配。

算法 1　计算用户兴趣向量 $\boldsymbol{W}$

输入　$\boldsymbol{R}$, $\boldsymbol{W}$

输出　$\boldsymbol{W}$

Begin

$\forall s_i \in W,\ 1 \le i \le n,\ s_i \leftarrow 0$;

构建 R 的 TRIE 树 $TRIE_R$;

构建 W 的 TRIE 树 $TRIE_W$;

For　∀ 叶节点 n in $TRIE_R$

For　∀ 叶节点 m in $TRIE_W$

if　n==m , S_m++

$\forall s_i \in W,\ 1 \le i \le n$,规范化 s_i

End

为了提高计算效率，在比较节点 n、m 时，若 n、m 的第一个汉字不同时，直接停止 n、m 串的比较。

算法 1 用 TRIE 树结构同时表示 $\boldsymbol{R}$ 和 $\boldsymbol{W}$，统一的结构简化了算法实现，同时也降低了检索 $\boldsymbol{W}$ 的时空开销。

4　用户模式驱动的检索

本文基于开源代码 Lucence 开发用户兴趣驱动的个性化搜索引擎。Lucence 的分析器 Analyzer 是搜索引擎开发的核心模块，包括滤波器 Filter 和分词器 Tokenizer。标准 Analyzer 根据设定的分词规则对文档进行常规分词，Filter 用来对 Analyzer 的结果进行格式转换和停止词处理等。Lucence 允许用户实现自己的分词器，为用户定制搜索引擎提供了方便。

Lucence 针对中文分词提供了 CJKAnalyzer、ChineseAnalyzer 和 IK_CAnalyzer。CJKAnalyzer 采用双字分词方法；ChineseAnalyzer 的分词结果和 StandardAnalyzer 相似，都采用一元分词；IK_CAnalyzer 采用字典方式分词，分词结果适合本文需要，因此采用 IK_CAnalyzer 分析器。

设用户给定检索词 I 和用户兴趣关注度 β_I，为了在检索过程引入用户兴趣向量 $\boldsymbol{W}$，需在 $\boldsymbol{W}$ 中根据 I 的分词结果查找用户感兴趣向量。若 $\boldsymbol{W}$ 中有匹配的用户感兴趣向量 $\boldsymbol{w}_i$，根据式(5)选取用户兴趣分量 I_W；若 $\boldsymbol{W}$ 中没有匹配 I 的感兴趣向量 $\boldsymbol{w}_i$，则 $I_W=\phi$。

$$I_W=\{w_{i,j} \wedge 1 \le i \le n, 1 \le j \le m \mid w_{i,j} \ge \beta_I\} \tag{5}$$

IndexSearcher 是 Lucence 的搜索器，根据检索词 I 进行检索，并根据 Pagerank 排序搜索结果。IndexSearcher 搜索结果没有考虑用户需求，本文对 IndexSearcher 进行了定制，把 I_W 作为 I 的同义词增加检索约束条件，搜索用户兴趣度高的结果。

本文测试数据为包含求职、兼职、软件开发、租房、培

训、考研、交友等在内的 3 512 个文本文件，测试了 Cookie、Cookie 被清理和不考虑用户兴趣的检索结果，如表 1 所示。另外，本文比较了 15 天 Cookie、30 天 Cookie 和 60 天 Cookie，所发现的用户兴趣模式对检索结果的影响如表 2 所示。

表 1　测试数据的检索结果比较　(%)

Cookie	准确率
Cookie	93.0
Cookie 被清理	81.6
不考虑用户兴趣	81.2

表 2　不同时间长度 Cookie 的检索结果比较　(%)

Cookie	15 天	30 天	60 天
Cookie	90.5	93.0	92.8
Cookie 被清理	81.2	80.7	80.3
不考虑用户兴趣	81.1	81.0	80.6

从实验结果可知，Cookie 记录了用户所访问的站点，蕴含着用户兴趣模式，发现其中的用户兴趣模式并用来约束检索结果能够改善搜索结果。同时，Cookie 文件积累的时间越长，所包含的有用信息也随之增多，从中所发现的用户兴趣模式更具有代表性，但当 Cookie 文件的时间达到一定时间时，Cookie 文件所蕴含的用户兴趣模式趋向于稳定。

5　结束语

Cookie 记录了用户访问 Web 站点的会话信息，能够保持 HTTP 的连接状态。大多数网站都在客户端生成 Cookie 文件，本文利用 Cookie 记录的 Web 站点信息发掘用户浏览兴趣，并用浏览兴趣模式建立分词词典，从而实现用户模式驱动的搜索，改善搜索结果。

值得一提的是，本文研究的是利用 Cookie 信息发现用户浏览兴趣模式，是在信息完备的 Cookie 文件得到实验结果。在有些情况下用户浏览兴趣模式挖掘会受到影响，例如，Web 站点不在客户端生成 Cookie 文件，用户删除、更改 Cookie 文件等。基于本文所提出的个性化搜索引擎方法，在未来的工作中，将利用用户群的浏览兴趣模式建立社区的定制搜索引擎。

参考文献

[1] 文振威, 秦　晓. 个性化搜索引擎的研究与设计[J]. 计算机工程与设计, 2009, 30(2): 342-344.

[2] 陈　敏, 苗夺谦, 段其国. 基于用户浏览行为聚类 Web 用户[J]. 计算机科学, 2008, 35(3): 186-188.

[3] Jansen B J, Booth D L, Spink A. Determining the Informational, Navigational, and Transactional Intent of Web Queries[J]. Information Processing & Management, 2008, 44(3): 1251-1266.

[4] Kristol D. HTTP State Management Mechanism[EB/OL]. (1997-02-16). http://www.w3.org/Protocols/rfc2109/rfc2109.

[5] 赵恒永, 沈　坚, 山　岚. 基于专业信息深度挖掘的搜索引擎 Spider 的设计与实现[J]. 计算机工程与科学, 2009, 31(6): 18-20.

[6] 张俊伟, 张　岭, 马　范. 提供个性化服务的搜索引擎页面排序算法[J]. 计算机工程, 2003, 29(19): 58-59.

[7] 李庆虎, 陈玉健, 孙家广. 一种中文分词词典新机制——双字哈希机制[J]. 中文信息学报, 2003, 17(4): 13-18.

编辑　任吉慧

（上接第 257 页）

第 2 阶段实验得到的所有用户操作各菜单项的绩效数据(该阶段实验共得到 96 个用户操作绩效数据，其中有效数据个数为 88 个)与相应的式(6)预测得到的用户操作绩效数据进行对比分析，分析结果如图 4 所示。

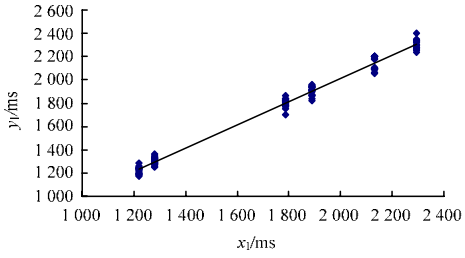

图 4　实验测试数据与模型预测数据关系图

分析得到图 4 中曲线的表达式为 $y_1=0.998x_1+6.118$，其中，y_1 代表实验测试数据；x_1 代表式(6)预测得到的用户操作绩效数据。x_1 与 y_1 关系接近于 $y=x$，且两者的相关性 R 很高($R^2=0.959$)，因此可知本文提出的菜单点击绩效模型(式(6))是有效的。

5　结束语

本文基于 Fitts'定律和操纵定律，提出了一个预测菜单点击绩效的模型，并设计实验验证了该模型的有效性。实验结果表明，该模型是有效的，为研究人员快速对交互系统菜单设计进行评估提供了依据。

参考文献

[1] Sears A, Shneiderman B. Split Menus: Effectively Using Selection Frequency to Organize Menus[J]. ACM Trans. on Computer-Human Interaction, 1994, 1(1): 27-51.

[2] Fitzmaurice G, Khan A, Pieke R, et al. Tracking Menus[C]//Proc. of UIST'03. Vancouver, Canada: ACM Press, 2003: 71-79.

[3] Bederson B. Fisheye Menus[C]//Proc. of UIST'00. San Diego, USA: ACM Press, 2000: 217-225.

[4] John B, Kieras D. The GOMS Family of User Interface Analysis Techniques: Comparison and Contrast[J]. ACM Trans. on Computer-Human Interaction, 1996, 3(4): 320-351.

[5] Byrne M. ACT-R/PM and Menu Selection: Applying a Cognitive Architecture to HCI[J]. International Journal of Human-Computer Studies, 2001, 55(1): 41-84.

[6] Walker N, Smelcer J B, Nilsen E. A Comparison of Selection Times from Walking and Pull-down Menus[C]//Proc. of CHI'90. Seattle, USA: ACM Press, 1990: 221-226.

[7] Accot J, Zhai Shumin. Beyond Fitts' Law: Models for Trajectory-based HCI Tasks[C]//Proc. of CHI'97. Atlanta, USA: ACM Press, 1997: 295-302.

编辑　任吉慧