

文章编号:1001-9081(2009)01-0344-03

网络信息检索个性化服务的研究与设计

顾牡丹,傅秀芬,周辉奎

(广东工业大学 计算机学院,广州 510075)

(gril_gu@163.com;xfu@gdut.edu.cn)

摘要:针对目前网络信息检索个性化服务不够周全的缺点,提出并实现了一个提高服务质量的信息检索个性化服务模型。该模型主要通过用户兴趣模型的建立、多维权值排序算法 MWRA 的优化、自由方式推送用户信息模型的建立三大模块来实现。最后给出传统信息检索模式与新模式的实验结果。

关键词:个性化服务; 用户模型; 信息检索; 兴趣度

中图分类号: TP391 **文献标志码:**A

Research and design of Internet information retrieval personalized service

GU Mu-dan, FU Xiu-fen, ZHOU Hui-kui

(School of Computer, Guangdong University of Technology, Guangzhou Guangdong 510075, China)

Abstract: Since the existing information retrieval service has some deficiency, this paper proposed and implemented an improved personalized information retrieval service model. This model was achieved mainly through three modules: the establishment of Users interested model, the optimization of Multi-weight Ranking Algorithm (MWRA), the establishment of user information model freely pushing mode. At last, this paper gave out the experimental results and analyses about the traditional information retrieval model and the new one.

Key words: personalized service; user model; information retrieval; interestingness

0 引言

网络给用户提供了大量信息资源的同时,也使得用户越来越难以搜索到有用的信息。目前大型的搜索引擎如 Google、Baidu、Yahoo 等,虽具有强大的搜索功能,但难于克服以下两个缺点:其一,任何一个搜索引擎的索引也无法覆盖整个网络资源,即其检索是不完全的;其二,由于采用简单的关键词匹配方法,搜索引擎对一条检索请求可能返回数以千计的结果,需要用户在此基础上再次筛选,增加了用户的上网处理时间。例如,当用户在搜索栏中输入“网络信息检索”,普通的用户希望得到关于网络信息检索的一些概要说明等相关信息;偏技术的用户希望获得关于网络信息检索编程方面的技术信息;偏科研的用户则希望获得关于网络信息检索的研究现状与其发展趋势等相关的信息。目前的信息检索技术还不能完全满足用户个性化服务的需求,因此如何提高网络信息检索个性化服务还有待于研究。为了弥补目前网络信息检索个性化服务不够周全,本文提出了一种提高网络信息检索个性化服务的模型,并通过实验,说明新模型的优越性。

1 个性化信息服务的含义与特点

个性化是指各网站针对不同的用户需求提供有特色的服务内容,个性化服务的实质在于提供真正适应用户需要的产品。个性化信息服务首先应该能够满足用户的个体信息需求,即根据用户提出的明确要求提供信息服务;其次,个性化信息服务也应该帮助用户发现自我,表现自我、让他人了解自我的愿望;最后,个性化信息服务应该是一种培养个性、引导需求的服务,这样可以帮助个体培养个性化信息检索^[1]。

个性化信息服务^[2]主要包括三个方面的内容:服务时空的个性化、服务方式的个性化和服务内容的个性化。个性化信息服务具有如下特点:1) 用户满意是个性化服务的出发点和归宿;2) 主动服务是个性化服务的基本模式。

因此为了满足用户检索网络信息的个性化服务,就必须进一步提高网络信息搜索个性化服务。

2 个性化服务的设计与实现

通过建立新一代的网络信息个性化检索系统来实现个性化检索服务,最关键的是进行用户个性化分析,系统自动根据用户的搜索行为构建一个用户兴趣模型(user profile)^[3],根据用户搜索行为和浏览行为不断更新该兴趣模型。本文的整体设计流程如图 1 所示,针对用户提出的明确要求,利用此网络信息检索模型在海量信息库中筛选提取其最可能的信息。

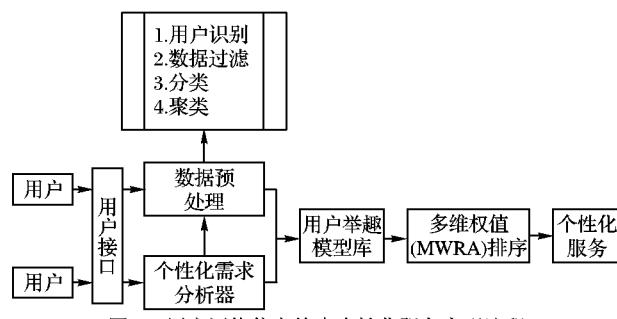


图 1 用户网络信息检索个性化服务实现流程

整个设计过程是先利用数据挖掘^[4]等技术对用户信息进行初步的筛选,再将初步筛选出的信息建立用户兴趣模型,接着利用多维权值排序算法 MWRA 对用户兴趣库中的信息

收稿日期:2008-09-09。 基金项目:广东省自然科学基金资助项目(06021484)。

作者简介:顾牡丹(1984-),女,江西高安人,硕士研究生,主要研究方向:网络安全、协同软件; 傅秀芬(1957-),女,福建上杭人,教授,CCF 会员,主要研究方向:协同软件、网络安全、数据库; 周辉奎(1983-),男,江西抚州人,硕士研究生,主要研究方向:网络安全。

以符合用户需求程度的高低为先后顺序进行排序,最后将信息以用户自由选择浏览模式反馈给用户,如用户可以选择文本、声音、图像等方式浏览所需的信息。

2.1 用户兴趣模型的建立

首先对用户兴趣进行分析与判断,实现个性化学习方法^[2]有:

1) 直接学习。提供用户可以操作的友好接口,允许用户主动地提出对用户模型的修改和维护意见。

2) 反馈学习。通过用户对检索结果的历次反馈意见进行学习。

3) 历史学习。通过用户历史查询记录的分析,经过一段时间的积累之后就可以发现用户需求的潜在规律。

4) 观察学习。对于客户端的搜索工具和系统而言,利用与客户端环境结合优势,可以从更多的方面观察并获取与用户相关的特征信息。

用户浏览器中的历史记录和收藏夹记录了用户查看过的一些网址链接,收藏夹内保存的是用户感兴趣或常用的网址链接。通过分析这些链接,可以得到用户的需求与喜好,有助于发现用户的兴趣。其次,利用文献[5]中的自动建立用户兴趣模型的方法对判断出来的信息进行建模。本文扩展地运用此方法,将用户访问一个页面的时间、网页的崭新度、入度和出度等与留言次数有关表示出来,并结合 Web 数据挖掘技术对相关信息进行筛选,尽最大可能精确地建立用户真正需要的兴趣模型。改进后的方法获得用户对网页兴趣度的最大值是一个网页访问率的 8 倍,其兴趣度公式为:

$$\begin{aligned} \text{Interest}(Page) = & \text{Frequency}(Page) \times \\ & (1 + \text{IsBookmark}(Page)) + \\ & \text{Duration}(Page) + \\ & \text{Recency}(Page) + \\ & \text{LinkVisitParent}(Page) + \\ & \text{LinkVisitChild}(Page) + \\ & \text{VisitTime}(Page) + \\ & \text{LeaveMessage}(Page) \end{aligned} \quad (1)$$

基于这些理论知识,编写出以下的用户兴趣判断算法来构建用户兴趣库模型。判断用户兴趣算法的功能是根据用户不断更新的兴趣描述文件来判断所接受到文件是否为其兴趣所在,用户兴趣库模型的构建流程图如图 2 所示,其中 $TEXT_i$ 是用户输入的实义词组集, T_i 为用户感兴趣的实义词组出现的次数, T_u 为用户无兴趣的实义词组出现的次数。 F_{ij} 为用户第 i 次输入词组中的第 j 个词, E_{ui} 为词 T_{ui} 出现在用户感兴趣词组中的次数, E_{uu} 为词 T_{ui} 出现在用户无兴趣词组中的次数。 P_g 为用户对实义词组 $TEXT_i$ 感兴趣的概率, P_s 为用户对实义词组 $TEXT_i$ 的第 j 个实义词 T_{ij} 感兴趣的概率。 D_{in} 为用户对实义词组 $TEXT_i$ 的第 n 个实义词 T_{in} 的熟词度, I_{in} 为用户对实义词组 $TEXT_i$ 的第 n 个实义词 T_{in} 的兴趣度,对兴趣度设定一个阈值 K 则有^[5]:

$$P_s = P\{TEXT_i, T_{ij}\} \quad (2)$$

$$P_g = P\{TEXT_i\} \quad (3)$$

$$D_{ij} = \left| P_s \times \log\left(\frac{P_s}{P_g}\right) - (1 - P_s) \log\left(\frac{(1 - P_s)}{(1 - P_g)}\right) \right| \quad (4)$$

实义词组的兴趣度:

$$L_i = \sum_j D(P_s, P_g); i = 1, 2, 3, \dots, k; j = 1, 2, 3, \dots, n \quad (5)$$

判断算法的过程是从库中读取用户记录,如 $TEXT_i, T_i, E_{ui}, E_{uu}$ 来计算 $P_g = E_i / (E_i + E_u)$ 。根据图 2 所示的用户兴趣判断算法得到兴趣度的大小值,将判断出来的用户感兴趣数据与其兴趣度值以捆绑的方式存入兴趣库中,以便检索。

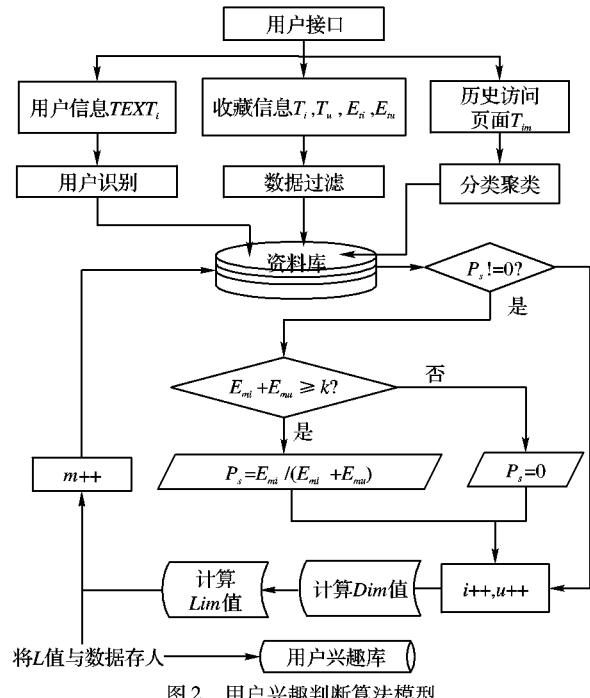


图 2 用户兴趣判断算法模型

2.2 多维权值排序算法 MWRA

当用户的兴趣模型建立完成以后,为了完善网络信息搜索的个性化服务,需要将最符合用户检索需求的网页排在最前面。目前已有一些网页排序算法来满足上述需求,如文献[6]中提出的从检索用户个体需要的角度来考虑的最新检索方法——多维权值排序算法(MWRA)。本文基于智能 Agent 的信息检索系统模型的基础上做进一步的改进,将知识库用 2.1 节获得的用户兴趣库替代。这就避免了一些不必要的检索或其他重复工作,能迅速地将兴趣模型中的所有信息按照用户兴趣度高低顺序依次推送给用户,将节省了用户寻找他们需求信息的时间。一般排在前面的信息就是用户所需求的,从而提高了检索效率。

首先对多维检索衡量权值作以下的形式化描述^[6]:

$$Mw \rightarrow \langle R, S, I, T \rangle \quad (6)$$

其中: Mw 为多维检索衡量权值(Multi-weights),其中 R 为信息本身的客观重要度权值,通过 Web 超链接的引用(reference)情况获得; S 为用户输入的检索关键词与网页的匹配度(similarity); I 为用户自身的检索偏好(inclination)与信息的归属度值; T 为信息更新的时间。

根据信息数据处理的计算公式^{[7]~(13)},分别确定 R, S, I 这三个参数的值:

$$R(d) = \sum_{s \in B(d)} \left(\frac{R(d')}{|F(d)|} \right) \quad (7)$$

$$V(d) = (t_1, w_1(d); \dots; t_n, w_n(d)) \quad (8)$$

$$w_i(d) = \frac{f_{id} \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{k=1}^n (f_{kd})^2 \times \log^2\left(\frac{N}{n_k}\right)}} \quad (9)$$

$$S(Q, d) = \frac{\sum_{i=1}^n w_i(d) \times q_i}{\sqrt{(\sum_{i=1}^n w_i^2(d))(\sum_{i=1}^n q_i^2)}} \quad (10)$$

$$Interest = In_1 \cup In_2 \cup In_3 \cup \dots \cup In_i \cup \dots \quad (11)$$

$$In_i = \{k_1, k_2, k_3, \dots, k_r, \dots\} \quad (12)$$

$$I(CurIn, d) = \alpha \sum_{i=1}^t \text{Sim}(Q, In_i) \times sLIn_i + \beta \frac{\sum_{i=1}^t w_i(b) \times bLIn_i \times \text{Sim}(Q, In_i)}{\sqrt{(\sum_{i=1}^t w_i^2(b))[\sum_{i=1}^t (bLIn_i \times \text{Sim}(Q, In_i))^2]}} \quad (13)$$

最后讨论信息更新的时间 T 值问题,这个 T 值由用户自行指定。对用户兴趣库中的数据用 R, S 和 I 值得出 Mw 值进行综合排序后,再利用 T 值对于相同的 Mw 值的页面按时间先后进行排序,最后按兴趣度 L 的大小进行排序。因此,经过三重排序后,最终反馈给用户的网络信息是最符合用户需求的并且按符合度的高低顺序先后反馈。

2.3 自由方式推送用户需求信息

经过上述双重筛选,最终检索出的信息基本上符合用户的需求,这时就需要服务器将这些信息反馈给用户。目前的反馈方法单一,只能以网页文本方式显示给用户,用户不能选择自己喜欢的浏览方式。例如,有的用户喜欢用听的方式来了解自己查到的信息,感觉在显示器上看东西眼睛比较吃力;有的用户则喜欢以生动的界面读取他们需要的信息,认为这样接受的快且有趣;不过也有一部分人喜欢浏览文字信息,他们可选择性的快速获得较多的信息。这些反馈方式各有优点与不足,关键看用户的喜好。如果能将文字、图片和声音三者快速地转换,不同的用户可以根据自己的喜好自由选择浏览方式,这种智能浏览方式能够进一步满足用户的个性化服务。

这种转换技术已经实现了,只是目前没有运用到网络信息检索中去,因此要实现自由方式推送信息给用户是完全可能的。只要在用户客户机上安装一个文字、图片、声音转换器^[8],此转换器可以运用于浏览器上,其设计模型如图 3 所示。

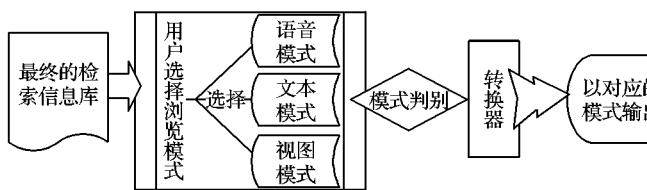


图 3 自由方式推送信息模型

3 实验分析

实验环境:SAMSUNG PC, P4, 内存 256 MB, 硬盘 80 GB, Windows XP 操作系统,C 语言,利用 Matlab 对检索信息数据进行分析,对比测试不同模型算法对检索偏好和已获得检索偏好参数之间的误差,若误差越小,说明信息检索质量越高、检索的信息越准确,即信息检索的个性化服务质量越高。

设神经元个数为 100, sigmoid 函数采用逻辑函数, α 取 0.5, 误差测试结果如图 4、5。

图 4 显示本论文提出的检索模型的检索信息误差在 0.5 趋于 0 之间,图 5 显示普通检索模型的检索信息误差在 0.5 ~

1.5, 对比两图显然得出本论文提出的新模式检索信息误差较小、准确率较高。

通过实验,得出此模型的检索方法不但速度快、准确率高,而且用户检索信息更智能化,进一步满足用户的信息检索个性化服务。

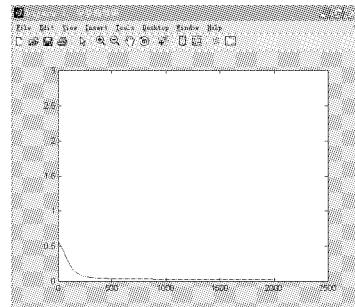


图 4 本模型算法误差曲线的输出

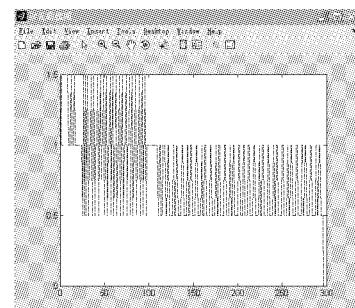


图 5 普通算法误差曲线的输出

4 结语

文中提出的网络信息检索个性化服务设计模型可以弥补目前信息检索系统的个性化服务中存在的一些不足,如检索速度加快、准确率提高、浏览方式智能化。我们下一步的主要工作有:扩展异构数据库多类型数据源的集成,实现数据源的自动注册服务;进一步优化数据查询方法和手段,使设计的数据访问与集成组件更具有实用性。

参考文献:

- [1] 王实, 高文, 李锦涛. 基于分类方法的 Web 站点实时个性化推荐 [J]. 计算机学报, 2002, 8(18): 845 - 852.
- [2] XU SAN-SAN. The research and design of user profile in personalized search engine [J]. Chinese Journal of Computers, 2007, 34 (10): 30 - 32.
- [3] 郭岩, 白硕, 杨志峰, 等. 网络日志规模分析和用户兴趣挖掘 [J]. 计算机学报, 2005, 9(311): 1483 - 1496.
- [4] HAN J. 数据挖掘概念与技术 [M]. 2 版. 范明, 译. 北京: 机械工业出版社, 2007: 30 - 184.
- [5] WANG PING . Research of personalized strategy based on small scaled search engine [J]. Computer Technology and Development, 2007, 17(11): 37 - 41.
- [6] XIAORONG XU. The agent-based information retrieval model with multi-weight ranking algorithm [J]. Journal of Electronics & Information Technology, 2008, 30(2): 483 - 485.
- [7] QI MEI-BIN, WANG DE-BAO. A search direction extensible fast search algorithm [C]// 1st International Symposium on Pervasive Computing and Applications. Urumqi IEEE Press, 2006: 32 - 35.
- [8] ISO, CCITT. ISO9594-108, CCITT X.500 - X.521, Information Processing Systems - Open Systems Interconnection, the Directory [R], 1988.