基于搜索历史的用户兴趣模型的研究

徐 科,崔志明
(苏州大学 智能信息处理及应用研究所,江苏 苏州 215006)

摘 要:提出了一种新的基于搜索历史的用户兴趣模型,目的是解决现有搜索引擎很难考虑用户兴趣来实现用户个性化搜索以及用户兴趣很难更新的问题。提出了基于搜索历史的用户兴趣的表达方法和自动隐式学习算法。全面地描述了用户兴趣模型的建立及通过自动隐式学习算法不断更新、优化模型的处理过程,并给出了对模型的评价标准。

关键词:用户兴趣;个性化;搜索历史
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2006)05-0018-03

User Profile Model Based on User Search Histories

XU Ke, CUI Zhi-ming
(Institute of Intelligent Information Processing and Application, Soochow University, Suzhou 215006, China)

Abstract: Puts forward a new user profile model based on user search histories to solve the problem that now-used search engines can't consider the users' personal interests for their personalized search and user profile is hard to update. The expression and auto implicit learning algorithm of user profile based on user search histories are given in this paper. The construction and update of user profile are discussed thoroughly. Finally, the evaluation criteria for the system is stated.

Key words: user profile; personalization; search histories

0 引 言

Web已成为人们获取信息的一个重要途径,由于Web信息的日益增长,人们不得不花费大量的时间去搜索、浏览自己需要的信息。搜索引擎(search engine)是最普遍的辅助人们检索信息的工具,比如传统的搜索引擎AltaVista, Yahoo和新一代的搜索引擎Google等。信息检索技术满足了人们一定的需要,但由于其通用的性质,仍不能满足不同背景、不同目的和不同时期的查询请求。个性化服务技术就是针对这个问题而提出的,它为不同用户提供不同的服务,以满足不同的需求。目前Internet个性化服务主要有3种形式:个性化信息检索、个性化网站和个性化推荐[1]。

个性化信息检索是指根据用户的兴趣和特点进行检索,返回与用户需求相关的检索结果。与传统信息检索系统相比,个性化信息检索系统增加了学习/更新用户模型、优化查询和优化结果3个模块。由于个性化信息检索系统在实现个性化的时候,未考虑用户的浏览习惯、书签的添加和删除等这些用户行为,所以个性化检索系统无法证明它的搜索结果可以满足不同用户的需求。

个性化站点是指那些为不同用户提供相应内容和服务的网站,如Yahoo! 公司于1996年推出个性化入口MyYahoo!,用户可以从成百上千的栏目中选择自己感兴趣的模块。个性化站点的优点是可以提高用户的满意度,增强网站的吸引力;缺点是用户需要自己罗列其感兴趣的条目,在用户兴趣改变的时候,需用户自己改变注册信息。

个性化推荐是指根据用户的兴趣和特点,向用户推荐用户感兴趣的信息。个性化推荐系统其优点是可以减少用户寻找感兴趣信息的时间,提高用户浏览的效率;缺点是同样需要用户罗列自己感兴趣的条目。

针对以上个性化系统所存在的问题,文中提出的用户兴趣模型以自动隐示学习方式为主,无需用户输入自己的信息需求或使用机器学习技术要求用户提供大量的反馈信息来建立用户兴趣模型,在用户信息需求发生变化时,也能很好适应用户需求的变化,快速更新用户兴趣模型。

1 设计思想

很多实验证明,通过显式方式获得用户兴趣模型的方法不能及时更新,在用户兴趣改变的情况下,无法向每个用户提供其真正感兴趣的信息。因此,搜索系统应该可以在无需用户干预的情况下直接、准确地察觉到用户兴趣的改变,使得用户查得的信息更精确,更符合他的需要。文中所提出的算法正是以此为出发点,在不需要用户参与的

收稿日期:2005-08-19
基金项目:教育部高校博士学科点科研基金资助项目(20040285016);江苏省高技术研究计划项目(BG2005019)
作者简介:徐 科(1981-),男,江苏金坛人,硕士研究生,研究方向为数据挖掘、个性化服务技术;崔志明,教授,博士生导师,研究方向为智能化信息处理、计算机网络应用与数据库应用。

情况下,根据用户的搜索及浏览的历史记录得出其喜好的变化,不断调整用户的兴趣模型。

例如,在理想情况下希望得到这样的结果:根据搜索,浏览历史,发现某用户对 Java 程序设计很感兴趣,则用户输入 Java 关键字查询时,猜测可能他要查的是有关 Java 编程语言的信息,但一段时间后,用户不再对 Java 编程语言感兴趣,转而对咖啡感兴趣了,则他在输入 Java 关键字查询时,可以猜测他要查的是关于咖啡的一些信息。

2　基本架构

人们研究发现,虽然用户搜索、浏览信息的行为有一定的逻辑性[2],但他们常常在一天内做不同的工作,这就导致他们搜索和浏览不同种类的信息,因此,对用户在一天内的浏览行为需采取更为细致的分析。假设用户的兴趣是建立在历史兴趣的不断积累之上的,将用户兴趣 P 分成两个部分:长期兴趣 P^{per} 和短期兴趣 P^{today}。长期兴趣表示通过挖掘用户 N 天以来浏览网页的历史记录而形成的用户兴趣;短期兴趣则是在对当天的浏览记录分析而得到的用户兴趣。

图 1 中显示了模型的基本的架构,搜索引擎对于用户的每次查询及浏览进行记录,分别计算得出用户的长期兴趣和短期兴趣,再对用户兴趣模型进行优化、更新。

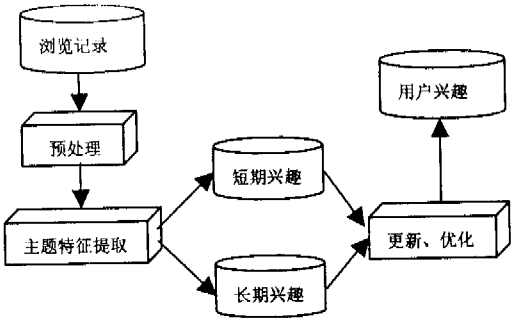

图 1　模型基本架构图

3　基于搜索历史的用户兴趣模型

为了便于说明,用图 2 表示了用户当天以及 N 天前的浏览记录。浏览记录包括用户浏览的网页,及对网页所做的动作。在更新模型的时候,需要根据用户的动作产生不同的更新。用户的动作可以是添加书签、下载文档、浏览摘要、忽略文档和删除书签等,这些动作体现了用户不同的兴趣度,具有不同的意义[3],见表 1。

表 1　用户动作对兴趣度的影响

用户动作	对用户兴趣度影响	$w_a^{hp^{(r)}}$
添加书签	很大	3
下载文档	大	2
浏览摘要	中等	1
忽略文档	无	0.5
删除书签	消极作用	0

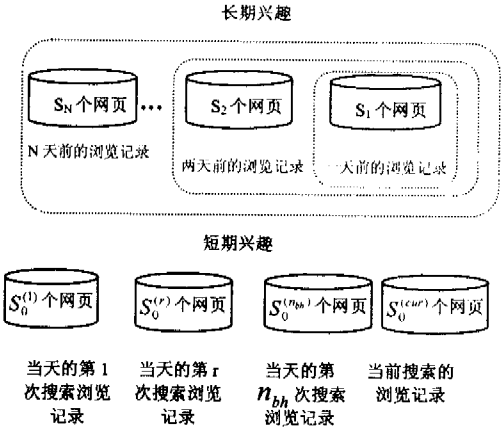

图 2　用户当天以及 N 天前的浏览记录

在这里,为了构造 P^{per},引入了窗口尺寸的概念,只有落入窗口的浏览记录,在计算用户兴趣模型时,才把它对用户兴趣的影响考虑在内,并且定义 $S_j(j = 0,1,\cdots,N)$ 为用户第 j 天浏览网页的数量,其中 $j = 0$ 表示当天的浏览记录。在该图中,假设用户在执行当前的搜索,即当天的第 cur 次搜索前已经做了 n_{bh} 次不同的搜索,因此,n_{bh} 和 cur 满足关系:cur = $n_{bh} + 1$。

3.1　短期兴趣

短期兴趣 P^{today} 分成两个部分:$P^{(\text{br})}$ 和 $P^{(\text{cur})}$。$P^{(\text{br})}$ 表示由当天的第一次到第 n_{bh} 次搜索浏览记录所获得的部分用户兴趣;$P^{(\text{cur})}$ 式根据当前最近一次搜索而获得的部分用户兴趣。P^{today} 定义如下:

$$P^{\text{today}} = xP^{(\text{br})} + yP^{(\text{cur})}, \tag{1}$$

x, y 满足 $x + y = 1$,为了强调用户当前浏览的信息对短期兴趣的影响,将 y 所赋的值大于 0.5,相应的 x 的值小于 0.5。

下面讨论如何得到 $P^{(\text{br})}$ 和 $P^{(\text{cur})}$。

首先,为了表示第 r 次搜索中浏览的网页 $hp^{(r)}$,定义一个特征向量 $w^{hp^{(r)}}$,

$$w^{hp^{(r)}} = (w_{t_1}^{hp^{(r)}}, w_{t_2}^{hp^{(r)}}, \cdots, w_{t_m}^{hp^{(r)}}),$$

其中:

m 表示网页 $hp^{(r)}$ 中不同词条的数目。

t_k 表示网页 $hp^{(r)}$ 中各个不同的词条。

$w_{t_k}^{hp^{(r)}}$ 表示词条 t_k 对网页特征的权重。

$w_{t_k}^{hp^{(r)}}$ 定义如下:

$$w_{t_k}^{hp^{(r)}} = c^{hp^{(r)}} \cdot \frac{tf(t_k, hp^{(r)})}{\sum_{s=1}^{m} tf(t_s, hp^{(r)})} \cdot w_a^{hp^{(r)}} \tag{2}$$

其中:

$tf(t_k, hp^{(r)})$ 表示词条 t_k 在网页 $hp^{(r)}$ 中出现的频率。

$c^{hp^{(r)}}$ 等于 0 或 1,当用户浏览网页 $hp^{(r)}$ 的时间经过该网页中词条的数目规格化后所得到的值大于 0.317 时[4],$c^{hp^{(r)}}$ 等于 1;否则,$c^{hp^{(r)}}$ 等于 0。

$w_a^{hp^{(r)}}$ 表示用户浏览网页 $hp^{(r)}$ 时做的动作所对应的

意义,具体对应值见表 1。

接着定义根据第 r 次搜索浏览记录获得的部分用户兴趣 $P^{(r)}=P^{(r)}=(P_{t_1}^{(r)},P_{t_2}^{(r)},\cdots,P_{t_m}^{(r)})$,$P_{t_k}^{(r)}$ 代表用户对每个特征词表示的事物感兴趣的程度,即为用户的兴趣类别,根据(2)式的定义,得到:

$$P_{t_k}^{(r)}=\frac{1}{S_0^{(r)}}\sum_{hp=1}^{S_0}P_{t_k}^{hp^{(r)}} \tag{3}$$

得到(3)式后,很容易得到 $P^{(br)}=(P_{t_1}^{(br)},P_{t_2}^{(br)},\cdots,P_{t_m}^{(br)})$,其中:$P_{t_k}^{(br)}=\sum_{r=1}^{n_{tk}}P_{t_k}^{(r)}$。

同理,可以得到 $P^{(cur)}=(P_{t_1}^{(cur)},P_{t_2}^{(cur)},\cdots,P_{t_m}^{(cur)})$

其中:$P_{t_k}^{(cur)}=\frac{1}{S_0^{(cur)}}\sum_{hp=1}^{S_0^{(cur)}}w_{t_k}^{hp^{(cur)}}\cdot w_a^{hp^{(r)}}$。

3.2 长期兴趣

用户的长期兴趣 $P^{(per)}$ 定义和短期兴趣 $P^{(today)}$ 类似,先设定窗口尺寸 $N(N=1,2,\cdots,30)$。

$$P^{(per)}=(P_{t_1}^{(per)},P_{t_2}^{(per)},\cdots,P_{t_m}^{(per)}), \tag{4}$$

其中:$P_{t_k}^{(per)}=\frac{1}{S_N}\sum_{hp=1}^{S_N}w_{t_k}^{hp}\cdot e^{-\frac{\log 2}{hl}(d-d_{tk\_init})}$,$e^{-\frac{\log 2}{hl}(d-d_{tk\_init})}$ 表示用户兴趣随时间推移而改变的程度,d_{tk_init} 为词条 t_k 初次出现的日期,d 为词条 t_k 出现的天数,hl 设置为 7,表示每过一周,用户的某种偏好就衰减一半。

3.3 用户兴趣

根据(1),(4)式定义的 $P^{(today)}$ 和 $P^{(per)}$,最终获得用户兴趣 P 如下:

$P=a\,P^{per}+b\,P^{today}=a\,P^{per}+b\,xP^{(hr)}+b\,yP^{(cur)}$,$a,b$ 满足等式 $a+b=1$;x,y 满足等式 $x+y=1$。

4 结束语

对用户模型的评估包括两个方面[5]:

(1)用户模型的收敛分析;

(2)用户模型是否反映用户的实际兴趣的分析。

这两个方面都要通过设计实验和对实验数据的分析来完成,相关文献中提供了详细的实验设计、实验数据分析的例子。

Internet 网上信息量过载已经导致人们难以利用其巨大的价值,个性化服务是一种趋势,通用的检索系统不可能满足不同背景、不同目的和不同时期的查询请求。如何从繁多的信息中将用户最需要的信息返回给用户是一个有意义的课题,这也是检索系统实现个性化服务的关键。文中提出一种基于用户搜索浏览历史的用户兴趣模型,几乎无需人机交互就准确地把握了用户的兴趣所在。如果再将本模型和 Agent 技术结合起来,就可以更好地根据用户的行为对其兴趣动态地进行跟踪,在理论上和实践上更适合于 Internet 上数量众多、信息经常改变的情形。

参考文献:

[1] 曾 春,邢春晓,周立柱.基于内容过滤的个性化搜索算法[J].软件学报,2003,14(5):999-1004.

[2] Sugiyama K. Studies on Improving Retrieval Accuracy in Web Information Retrieval[D]. Tokyo: Nara Institute of Science and Technology,2004.

[3] Bollacker K D, Lawrence S, Giles C L. Discovering relevant scientific literature on the Web[J]. IEEE Intelligent System, 2000,15(2):42-47.

[4] Sugiyama K, Hatano K, Yoshiakawa M. Adaptive web search based on user profile construction without any effort from users[A]. WWW2004[C]. New York:[s.n.],2004.152-156.

[5] 韩立新,陈贵海,谢 立.一个面向 Internet 的个性化信息检索系统模型[J].电子学报,2002,13(3):153-157.

(上接第 17 页)

而不是模拟应用软件和 OS 软件或其他信息,有其理论基础。关于硬件的模拟,目前已经实现,整个策略本身最终并未全部实现。但在解决如何存储数据,并保证长期可用性的问题上,它的目标是在未来不可预测的机器上读取原始数据文件,而不再考虑存储介质和文件编码格式的过时问题等。目前它是此类问题中具有代表性、且较为完善的解决方案之一。

4 结束语

在数据仓库环境下如何保证数据长期可用性是目前一个重要的研究课题。在数据存储和数据仓库两个方面,业界已经开始关注,并做了很多细致、深入的研究和实验,其中还有许多的实现技术需要进一步地研究和解决。基于数据库安全性的研究[5],对于数据仓库环境下数据安全性访问控制的研究将成为进一步的研究方向。关于这个问题的最终解决将会给数据仓库技术带来更大的机遇,使其更大地发挥潜在的优势,更好地为国民经济和社会服务。

参考文献:

[1] Inmon W H. 数据仓库[M].北京:机械工业出版社,2000.

[2] Kimball R. Digital Preservation[J]. Intelligent Enterprise, 2000,3(4):215-217.

[3] Rothenberg J. Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation[R]. Report for The Council on Library and Information Resources, Washington, USA: Commission on Preservation and Access, 1999.

[4] 唐光前.数字保存策略:模型法[J].情报杂志,2001,20(5):54-56.

[5] 张志勇.基于角色的两级数据库访问控制机制及其实现[J].微机发展,2004,14(1):109-111.