计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(8):1284~1292, 2007

门户个性化兴趣获取与迁移模式发现

吴 晶 张 品 罗 辛 盛 浩 熊 璋
(北京航空航天大学计算机学院 北京 100083)
(wing.wujing@gmail.com)

Mining Interests and Navigation Patterns in Personalization on Portal

Wu Jing, Zhang Pin, Luo Xin, Sheng Hao, and Xiong Zhang
(*School of Computer Science, Beihang University, Beijing* 100083)

Abstract Personalization services pose new challenges to interest mining on portal, such as many powerful functions to customize desktops. Capturing these surfing behaviors of users implicitly and mining interest navigation patterns are the top demanding tasks. Based on the summary of personalized interest mapping on portal, a novel portal-indepeodent mechanism of interest elicitation with privacy preserving is proposed, which implements both the implicit extraction of diverse access behaviors and corresponding semantic analysis. A new schema is also presented to extend the interest representation rule efficiently in privacy preserving process. Then the legal transparent accountable interest in terms of the interest behavior entries could be implied with this interest-extended rule. Moreover, the navigation relationship among the interests can describe the user's next possible interest trends, especially benefiting the recommendation. By combining the association effect of those interests and the prediction on interest intentions, a hidden Markov model is extended with personalized interest descriptions of portal to form interest navigation patterns for different users. Then experiments are carried out in order to validate the proposed approaches with availability and feasibility. The improvement of representation accuracy and mining capability for the complex interests on portal is a feature that clearly distinguishes the proposed approaches from traditional ones.

Key words portal; interest behavior; implicit interest elicitation; privacy preserving; hidden Markov model (HMM); navigation pattern

摘 要 个性化服务技术为门户平台上的兴趣挖掘研究带来了新的挑战,如何隐式地获取门户用户兴趣行为以及发现兴趣迁移模式是其中的重要课题.在对门户个性化兴趣映射描述的基础上,提出了一种独立于门户平台的含隐私保护的门户个性化兴趣获取机制,可实现不同兴趣访问行为的隐式获取以及操作语义分析,并采用兴趣扩展规则描述方式进行了隐私保护.结合门户个性化兴趣影响以及兴趣目的预测,给出了带有门户个性化兴趣描述的隐 Markov 模型扩展,可用于发现不同用户的门户个性化兴趣迁移模式.最后通过验证实验给出了有效性和可行性的结论分析.

关键词 门户;兴趣行为;隐式兴趣获取;隐私保护;隐 Markov 模型;迁移模式

中图法分类号 TP18

随着 Web2.0 时代的到来,门户技术的迅速发展以及面向服务架构的推动,资源整合及基于门户平台的个性化服务手段已越发丰富.门户为用户提供强大的桌面定制功能的同时,也使得用户的个性化兴趣来源更加复杂多样且动态变化,主要表现在:首先,资源内容多以门户组件的形式作为频道展现在个性化桌面上,较之普通的 Web 站点页面更加生动和灵活;其次,用户的访问行为脱离了以往的被动

收稿日期:2006-09-07;修回日期:2007-05-11
基金项目:国家"八六三"高技术研究发展计划基金项目(2002AA113070)

浏览模式,增添了用户可主动编辑的操作种类,包括配置定制、浏览点击、布局外观以及编辑评论等,在行为的获取和描述方法上需要考虑更多细节;再次,门户对于用户个性化配置定制的保存相对更加分散,难以直接抽取有用的结构化信息.因此,如何跟踪、获取和学习门户个性化兴趣是门户个性化服务研究领域的重要课题和挑战.

获取用户兴趣的方式一般分为显式(explicit)和隐式(implicit)两种[1-2]:前者多依赖与用户的交互反馈,不能保证数据的完整准确;后者是动态跟踪用户访问过程进行 Web 使用挖掘(Web usage mining, WUM),提取用户兴趣特征和有用模式,可以更加直观表达用户兴趣并减少由于用户参与而带来的噪音干扰,因此在实际应用中具有重要意义.根据不同的兴趣挖掘目标,常用的隐式获取方法[1-4]有智能代理服务、动态插件加载、服务器脚本以及控制高速缓存等,多是基于普通 Web 站点的应用模式,挖掘的行为类别有限,不能完全沿用于门户领域.目前为止,关于门户个性化兴趣获取方法的探讨尚鲜见于国内外相关文献.

兴趣迁移问题[5-7]起源于 Web 站点分析用户访问兴趣,调整并改进结构设计的需求.对于门户而言,同样也需要能够动态自适应地发现用户的兴趣迁移模式,帮助提升个性化服务的质量.文献[8]给出了基于隐 Markov 模型(hidden Markov model, HMM)的兴趣迁移模式发现方法,将带有兴趣的迁移模式定义为一种关联规则,通过增量发现算法推导兴趣迁移模式.然而该方法挖掘的是群体行为特征,对于门户所关心的个体用户的不同个性化兴趣特点与使用偏好,需要进一步扩展讨论.

本文给出了门户个性化兴趣行为的映射分析,由此提出一种独立于门户平台的含隐私保护的门户个性化兴趣获取机制.其基本思想是隐式获取门户用户的不同兴趣访问行为并进行操作语义分析,以及通过 Notation3 格式描述的兴趣扩展规则,实现私有化过滤保护过程.同时,我们进行了带有门户个性化兴趣描述的 HMM 扩展,强化了不同用户个性化兴趣的权重影响以及兴趣内容间的访问目的预测,用于发现不同用户的个性化兴趣迁移模式.

1 门户个性化兴趣映射描述

1.1 基本定义

定义 1. 门户用户 u 是具有以下一些特性的网络用户:在门户注册并拥有惟一访问的账号;具有门户提供的个性化服务权限;登录门户桌面进行访问操作.则所有门户用户组成的集合定义为门户用户集 $U=\{u_1,u_2,\cdots,u_N\}$,这里忽略匿名访问的用户群.

定义 2. 门户兴趣内容集 IC 是个性化桌面上所有可进行访问操作的内容对象的并集,也可表示为门户中所有资源分类后的兴趣内容的集合:

$$IC=\{Portlet_1,\cdots,Portlet_l\}\cup\{Link_1,\cdots,Link_m\}\cup\{Tab_1,\cdots,Tab_n\}=\{InterestContent_1,InterestContent_2,\cdots,InterestContent_M\},$$

其中,$Portlet$ 是一个门户组件频道;$Link$ 是一条超链接内容;Tab 是一个标签页面;$InterestContent$ 是采用概念分层(concept hierarchy)的方法[9]分类生成的兴趣内容,则有对应的兴趣概念集:

$$\Sigma=\{\sigma_x\mid 1\leqslant x\leqslant\mathbb{Z}\},\exists\ InterestContent\mapsto\sigma_x,$$

σ_x 为兴趣内容特征概念,$\mapsto$ 表示兴趣内容到特征概念的映射.

定义 3. 门户兴趣行为集 IB 是 u 在个性化桌面上访问 IC 时所有可能的行为操作的集合.结合门户个性化服务特点,本文研究的门户个性化兴趣行为主要分 4 类:①配置定制.对兴趣内容进行的个性化配置定制,如设置 Portlet 窗口的显示模式(SetMode)、设置推荐更新的时间频度(SetRecomTime)等.②浏览点击.对兴趣内容进行的所有点击操作,如切换 Tab(Switch)、最大化(Maximize)或最小化(Minimize)或关闭(Close)Portlet 窗口、点击兴趣内容中的 Link(Click)等.③布局外观.对兴趣内容的显示外观及布局进行的修改调整,如调整 Portlet 显示分栏和顺序(Layout)、增加(Add)或删除(Delete)兴趣内容等.④编辑评论.对兴趣内容进行的编辑操作,如编辑兴趣内容(Edit)、对兴趣内容添加引用(Quote)或评论(Comment)等.则定义兴趣行为集为:

$$IB=\{\text{SetMode, SetRecomTime, Switch, Maximize, Minimize, Close, Click, Layout, Add, Delete, Edit, Quote, Comment}\}.$$

定义 4. 令 u 于同一会话 T 中的访问过程可顺序记录为一条访问事务 at,定义为多元组:

$$at=\langle at.u,\langle at.content_1,at.time_1,at.behavior_1\rangle,\cdots,\langle at.content_p,at.time_p,au.behavior_p\rangle\rangle.$$

其中,$at.u\in U$ 标识访问用户;三元组$\langle at.content,at.time,at.behaivor\rangle$标识用户的每一次访问操作,$at.content\in IC$ 标识具体兴趣对象;$at.time$($at.time_p-at.time_1\leqslant T$)标识访问时间戳;$at.behavior\in IB$ 标识具体兴趣行为.继而,将所有访问事务 at 按照会话时间顺序组成该用

户在门户上的访问事务集:$AT_u = \{at_i \mid 1 \leqslant i \leqslant |AT_u|\}$,$|AT_u|$为用户的会话总数,即访问事务集的规模.

1.2 兴趣行为分析和映射

定义3给出了门户兴趣行为的基本操作语义特点以及隐式获取的意义,可作为门户个性化兴趣程度的数值化映射和语义描述的基础.通常对于隐式获取的用户行为所反映的兴趣度的估计主要是基于页面驻留时间、点击频度等[1-2]单因素粗粒度的方法,总体来说还不够准确和全面.我们参考比对了文献[1]中给出的一般Web站点中典型兴趣行为的特点,并抽取了门户的HTTP访问请求中的重要语义特征属性("属性名(-属性值)"字符串对),再综合考虑不同行为在兴趣内容上所蕴含的意义或影享,可以得到兴趣程度的数值化映射(划分为5级),如表1所示.该映射关系为进一步的兴趣行为隐式获取对象的粒度选择和语义描述提供了有意义的参考,可以有效地精简数据预处理过程.

Table 1 Map of Personalized Interest Behaviors on Portal

表1 门户个性化兴趣行为映射

Personalized Interest Behaviors on Portal		Requests Attribute (-Value)	Similar Interest Behaviors in Web Site and Corresponding Interest Degree		Effect Factors	Interest Degree
Custom Configuration	SetMode	_mode	/	/	The state of the window	④or②
	SetRecomTime	_event	/	/	Frequency	④
Browse-click	Switch	_pageLabel	Forward/Back	④	Access transaction Set	④
	Maximize	_state-maximized	Open in a new window/Drag stroll bar	⑤	The state of the window	⑤
	Minimize	_state-minimized	Exit	②		②
	Close	_state-closed				
	Click	_URL	Click hyperlink	④	Access transaction Set	④
Layout	Layout	_windowLabel	/	/	Layout sequence & Access transaction Set	⑤or②
	Add	_windowLabel	Add the bookmark	⑤		⑤
	Delete	_windowLabel	Delete the bookmark	①		①
Edit-comment	Edit	_mode-edit	Query	⑤	Access transaction Set	⑤
	Quote	_event	/	/		④
	Comment	_event	Feedback rate	④		④

Note:①Strong negative;②Negative;③Weak positive;④Positive;⑤Strong positive

本文还引入模糊逻辑思想[10-11]对兴趣行为的影响因子的权重映射进行联合描述.

定义5. 令 $FS_{AT} = Relation(AT_u, IC \cup IB)$表示 $AT_u \times (IC \cup IB)$域上 AT_u与 $IC \cup IB$ 之间的模糊关系,描述 u 交互访问过程中的个性化兴趣行为特征及评价影响.则定义 $W_{AT}(content_k) \in [0,1]$为归一化表示 FS_{AT}所反映的个性化兴趣权重,计算方法如下:①计算 at_i 中 u 对 $at_i.content_k$($1 \leqslant k \leqslant M$)的总访问时间间隔 $d(at_i.content_k)$,以及 AT_u中 u 对 $at_i.content_k$ 的总访问时间间隔 $\sum_{j=1}^{|AT_u|} d(at_j.content_k)$;②令 w_t和 W_T分别作为单次会话以及所有会话中的短期和长期时间兴趣度阈值;③已知 r_x为每项 $at_i.behavior$ 在表1中所映射的兴趣程度,则 $at_i.content_k$ 上的兴趣度可表示为 $R_k = \frac{\sum_{x \in at_i} r_x}{|at_i|}$($r_x \in [1,5]$),得到 $W_{AT}(content_k)$的描述:

$$W_{at}^i(content_k) = \begin{cases} \left(\frac{2w_t}{W_T} - 1\right) \times R_k/5, & \frac{1}{2}W_T \leqslant w_t \leqslant W_T, R_k < 3, \\ \frac{2w_t}{W_T} - 1, & \frac{1}{2}W_T \leqslant w_t \leqslant W_T, 3 \leqslant R_k \leqslant 5, \\ 1, & w_t > W_T, 3 \leqslant R_k \leqslant 5, \\ 0, & \text{otherwise} \end{cases}$$

$$w_t = \frac{d(at_i.content_k)}{T_i}, W_T = \frac{\sum_{j=1}^{|AT_u|} d(at_j.content_k)}{\sum_{j=1}^{|AT_u|} T_j},$$

$$W_{AT}(content_k) = \frac{\sum_{at_i \in AT_u} W_{at}^i(content_k)}{|AT_u|}. \quad (1)$$

该权重也可扩展用于对多个兴趣内容上的兴趣行为特征及评价进行组合描述,需满足时间一致性,否则取单个兴趣权重中最小值者.

定义6. 令 u 的个性化桌面兴趣内容的布局变化可表示为一个桌面布局序列集:$S_u = \{at_i.sequence\}$($at_i \in AT_u$,$|S_u| = |AT_u|$),$at.sequence$ 顺序记录每一次会话个性化桌面布局调整后,兴趣内容 $at_i.content_k$ 所在的行列位置.令 $FS_L = Relation$

$(S_u, IC \cup IB)$表示 $S_u \times (IC \cup IB)$域上 S_u与 $IC \cup IB$ 之间的模糊关系,描述 u 交互访问过程中的个性化兴趣品味及关注程度.则定义 $W_L(content_k) \in [0,1]$为归一化表示 FS_L所反映的个性化兴趣权重,计算方法如下:①分别提取 S_u中每个兴趣内容 $at_i.content_k$ 向前或滞后的跃迁位移 P_k^i,即当前位置与起始位置的距离;②若对应 $Portlet$ 有宽度、尺寸等更新调整的变化,则加权调整为 $P_k^{i\prime} = \alpha \times P_k^i$($\alpha \in [0.5,1.5]$为窗口尺寸变化影响系数);③分别计算 S_u中兴趣内容 $at_i.content_k$ 的跃迁位移平均值 $\overline{P}_k$,得到 $W_L(content_k)$的描述:

$$W_L(content_k) = \frac{1}{P_k} = \frac{|S_u|}{\sum_{i=1}^{|S_u|} P_k^i} (\overline{P}_k \neq 0). \quad (2)$$

该权重也可扩展用于对多个兴趣内容的关注程度进行组合描述,一般采用均值计算方式,需满足内容自相关性,否则取单个兴趣权重中最小值者.

因此,每个 u 的结合兴趣行为和兴趣内容描述的兴趣权重可表示为 $W(content_k)$:

$$W(content_k) = \gamma_1 W_{AT}(content_k) + \gamma_2 W_L(content_k), \quad (3)$$

其中 $\gamma_1 + \gamma_2 = 1$($\gamma_1, \gamma_2 \in [0,1]$),本文中兴趣权重比例系数满足 $\gamma_1 = 0.7, \gamma_2 = 0.3$.

2　门户兴趣获取机制

2.1　兴趣行为隐式获取

由表 1 分析可得,门户兴趣行为比一般 Web 站点的访问行为更加多样化.其中,配置定制、布局外观等类别行为通常都不会被门户服务器日志记录,但可以从用户描述文件(user profile, UP)和桌面内容结构描述中获取相关信息;日志对于浏览点击、编辑评论等类别行为的记录内容也很有限,但从应用层角度看,用户在门户上的个性化访问过程仍是通过一系列 HTTP 请求/响应构成的"轨迹".

本文提出了一种独立于门户平台的门户兴趣行为隐式获取机制,可以对于门户用户的不同兴趣访问行为以及操作语义进行隐式跟踪捕获和分析统计,为访问事务集 AT_u以及桌面布局序列集 S_u提供数据准备.以下分别对文中总结的 4 类兴趣行为的获取方法予以分析说明.

1) 配置定制.可通过门户开放 API 获取不同用户的 UP,获取结果是 UP 中的"属性名-属性值"对,其本身已带有语义信息,无需进行额外语义分析.由于这些 API 特定于具体平台,因此我们增加适配主流门户平台接口的模块以实现不依赖于门户平台的通用目标.

2) 浏览点击.针对此类兴趣行为所作用的兴趣内容类别比较广泛,需要分情况讨论.对于采用 WSRP(Web services for remote portlet)机制整合内容的 Portlet,可以在 WSRP 生产者一端分析用户请求以捕获用户操作行为;对于一般 Web 应用程序访问,需要实时过滤分析服务器端收到的 HTTP 请求并进行冗余信息清洗[12-13];对于特定的门户应用程序,通过调用相应 API 可以获取响应上述门户应用程序事件的操作行为.获取结果主要用于预处理形成访问事务集 AT_u.

3) 布局外观.思路同 1),获取结果是用户个性化桌面上内容对象的层次结构,不包含用户操作语义信息,可预处理形成桌面布局序列集 S_u.

4) 编辑评论.思路同 2).

我们将表 1 的访问请求特征记录在 XML 中形成重要兴趣行为语义映射表.在实际兴趣获取过程中,先将冗余无关请求过滤清除,再结合映射表中的"属性名-属性值"比对访问请求中是否含有兴趣行为语义参数特征,若匹配则采用五元组将其定义为一次重要兴趣行为记录[10]:

〈$user, timestamp, content, behavior$ ($desktop$)〉,

其中,$user$ 为用户标识;$timestamp$ 为执行此操作行为的时间;$content$ 标识兴趣内容对象;$behavior$ 标识具体兴趣行为操作结果;$desktop$ 为可选项,则用 $desktop$ 描述当前用户个性化桌面布局结构.

2.2　隐私保护

在上述兴趣获取过程中,还需要考虑并实施有效的保护用户隐私的策略[14-15],确保收集信息的合法化,平衡个性化和隐私安全的关系.

本文借鉴文献[15]在机场网络运输事务管理中的思路,提出了一种采用兴趣扩展规则描述的隐私保护机制,可以有效地解决隐私保护中信息有效性检验以及统一数据转换两个关键性问题.其基本思想是:通过 Notation3 格式[16](N3)序列化建立兴趣扩展规则的 RDF 描述,结合重要兴趣行为语义映射表,对可能隐含重要兴趣行为语义以及桌面布局序列变化的事务数据进行私有化过滤检验,并将有效数据自动转换为重要兴趣行为记录形式.图 1 以门户兴趣行为获取过程的隐私保护为例,给出了 N3 序列化的兴趣扩展规则描述.

```
@ keywords a, is of.
@ prefix rdfs :<http ://www.w3.org/2000/01/rdf-schema #>.
@ prefix log :<http ://www.w3.org/2000/10/swap/log #>.
@ prefix string :<http ://www.w3.org/2000/10/swap/string #>.
@ prefix list :<http ://www.w3.org/2000/10/swap/list #>.
# MapList : Propertylist
# IC, IB : Classes
{?U a userID.
 ?X a timestamp.
 ?Y a content on Portal.
 ?Z a access behavior.
 @ forAll ?U, ?X, ?Y, ?Z
 log : implies ?RequestRecord.
 @ forAll ?Request Record
  {@ forSome ?AttributePair
     {?RequestRecord string : contains ?AttributePair.
      ?AttributePair string : equalIgnoringCase ?AttributeDefined.
      ?AttributeDefined list : in MapList.
     }
  }
  @forSome ?AttributePair log : implies ?Z.
  ?Y rdfs : subClassOf IC.
  ?Z rdfs : subClassOf IB.
}
⇒ {?U is a : user.
   ?X is a : timestamp.
   ?Y is a : content.
   ?Z is a : behavior}.
```

Fig. 1 Sample interest-extended rule in RDF serialized in N3.

图 1 采用 N3 序列化 RDF 的兴趣扩展规则描述实例

3 门户兴趣迁移模式发现

3.1 问题分析

兴趣间的迁移关系反映了用户在某一兴趣状态下的下一步走向,即用户对某种内容概念的兴趣目的. 文献[8]通过建立一阶 HMM 模型,反复观测初始节点到最终节点的访问状态序列的转移概率以及对节点上概念感兴趣的概率分布,在所有可能路径上求最大概率. 同时采用增量发现算法给出了兴趣关联规则反映兴趣迁移的模式,其结果是面向全体用户的统计,可用于 Web 结构挖掘中的整体站点结构调整和改进.

对于门户个性化服务而言,结合门户个性化兴趣描述挖掘不同门户用户的兴趣迁移模式,可以将每个用户最可能的访问模式提取统计出来,形成下一步个性化服务各方面所需的数据. 尤其可作为预测推荐服务时的参考,便于推荐资源推送展现的优先选择,体现个性化“因人而异”的服务特点. 因而需要考虑不同用户的个性化特点和兴趣模式区别.

本文将门户个性化兴趣迁移模式发现问题转化为寻找 u 关于某一兴趣内容的一些访问事务集 AT_u,使其访问该兴趣内容的可能性较大,并且访问其他兴趣内容的可能性较小. 这些访问序列可以看做是门户用户的兴趣迁移模式. 扩展定义后的 HMM 模型的特点主要体现在两个方面:①强化了不同门户用户的访问事务集以及对用户个性化兴趣目的预测的思想,发现的是带有个性化兴趣的不同用户的访问迁移模式,可以进行个性化预测推荐;②引入了较完备的门户个性化兴趣语义描述,提升了访问序列概率计算结果的精度,扩大了迁移模式描述的有效性范围.

3.2 带有门户个性化兴趣描述的 HMM 扩展

带有门户个性化兴趣描述的 HMM 模型扩展定义如下:

1) 个性化桌面上的 *InterestContent* 按照布局结构可以视为 HMM 的节点 $q(q \in IC)$,给定虚拟初始状态 q_1,存在兴趣内容到特征概念的映射关系 $q_i \mapsto \sigma_z^j \in \Sigma$.

2) AT_u 中,任意两节点存在访问转移概率 $P(q_i \to q_j)$,代入式(3)的兴趣权重表示如下:

$$P(q_i \to q_j) = P(q_j \mid q_i) = \frac{P(q_i q_j)}{P(q_i)} = \begin{cases} \dfrac{\gamma_1 W_{AT}(q_i, q_j) + \gamma_2 W_L(q_i, q_j)}{\gamma_1 W_{AT}(q_i) + \gamma_2 W_L(q_i)}, & i \neq j, \\ 0, & i = j. \end{cases} \tag{4}$$

3) 对于 q_j 及其对应概念 σ_z^j,存在一个观测概率分布 $P(\sigma_z^j \mid q_j)$,即 u 对 q_j 所有访问中,对概念 σ_z^j 感兴趣的概率. 由定义 4 得到 at_i 所包含的被访问节点集合为 $Q_i = \{q_1', \cdots, q_f' \mid q' \in IC\}$,则令 $Q_{i,j}$ 和 Q_{i,j,σ_z} 分别表示 at_i 中于 q_j 之后的所有被访问节点集合以及 $Q_{i,j}$ 中含有 σ_z^j 的节点集合:

$$Q_{i,j} = \begin{cases} \{q_{k+l}' \mid q_k' = q_j, l = 0, \cdots, (f-k)\}, \\ \qquad q_j \in Q_i, \\ \text{Null}, \quad q_j \notin Q_i, \end{cases} \tag{5}$$

$$Q_{i,j,\sigma_z} = \{q'' \mid q'' \in Q_{i,j}, q'' \mapsto \sigma_z^j\}, \tag{6}$$

则将 u 在 q_j 上观测概率分布 $P(\sigma_z^j \mid q_j)$ 定义为所有访问事务在 q_j 之后对概念 σ_z^j 相关的访问节点集合总数与访问事务在 q_j 之后对所有概念的访问节点集合总数之比:

$$P(\sigma_z^j \mid q_j) = \frac{\sum_{i=1}^{|AT_u|} |Q_{i,j,\sigma_z}|}{\sum_{i=1, \sigma_z' \in \Sigma}^{|AT_u|} |Q_{i,j,\sigma_z'}|}. \tag{7}$$

4) 建立带有门户个性化兴趣描述的 HMM 模型,即在 u 对于概念 σ_z^k 的所有可能访问序列中寻找一个状态序列,使其具有最大的访问概率:

$$P_{\max}(\sigma_z^k) = \arg\max \prod_{q_k \in IC} P(q_k \rightarrow q_{k+1}) P(\sigma_z^k \mid q_k), \tag{8}$$

其中,可以将 $P(q_k \rightarrow q_{k+1}) P(\sigma_z^k \mid q_k)$ 看做任意两节点间的兴趣迁移概率. 所得序列能够表征用户兴趣迁移的可能性分布,并通过关联推导得到其兴趣迁移模式.

4　实验结果分析

为了直观验证门户个性化兴趣获取与迁移模式发现方法,我们在 Weblogic Platform 8.1 平台架构部署的国内某大型工程建管门户系统上进行了相关实验. 实验数据集包括了 20 个典型门户组件和链接构成的兴趣内容集 $IC = \{q_i \mid 1 \leqslant i \leqslant 20\}$(如表 2),并定义了 3 个兴趣概念集:$\sigma_1 = \{q_1, q_2, q_3, q_4, q_5, q_6, q_{12}\}$|"工程建设",$\sigma_2 = \{q_7, q_8, q_{10}, q_{11}, q_{13}, q_{14}, q_{15}, q_{16}\}$|"信息查询"和 $\sigma_3 = \{q_9, q_{17}, q_{18}, q_{19}, q_{20}\}$|"个性化管理",以及于 2006 年 7 月至 11 月获取的 50 个用户的 76105 条有效访问事务(平均会话总数为 1522).

Table 2　Data Set About Interest Contents

表 2　实验兴趣内容集

q_i	*content*	q_i	*content*
q_1	*tab_subSystem*	q_{11}	*portlet_search . hhzh*
q_2	*portlet_subSystem . gcjs1*	q_{12}	*portlet_search . chgc*
q_3	*portlet_subSystem . gcjs2*	q_{13}	*portlet_search . jsyj*
q_4	*portlet_infoRelease*	q_{14}	*portlet_search . zhsj*
q_5	*link_infoRelease . gcjs1*	q_{15}	*portlet_search . tcqs*
q_6	*link_infoRelease . gcjs2*	q_{16}	*portlet_infoRelease . ywdt*
q_7	*portlet_subSystem . xxzx*	q_{17}	*tab_personalization*
q_8	*portlet_subSystem . szwd*	q_{18}	*portlet_calendar*
q_9	*portlet_rss*	q_{19}	*portlet_contact*
q_{10}	*tab_search*	q_{20}	*portlet_taskReminder*

4.1　兴趣获取实验结果分析

实验采用含隐私保护设计的过滤器封装实现了门户兴趣行为获取模块,UP 和门户桌面结构获取模块则采用 Servlet 的形式. 图 2 为重要兴趣行为获取记录的一个实例.

```
用户名:zhangpin
06-7-21 下午 07 时 00 分 36 秒
操作对象:portlet_calender
操作:Portlet:portlet_calendar 状态改变为:Add
Tab:tab_personalization
portlet:portlet_contact
portlet:portlet_taskReminder
portlet:portlet_calendar
Tab:tab_subSystem
portlet_subSystem . gcjs1
portlet_infoRelease
link_infoRelease . gcjs1
portlet:portlet_rss
Tab:tab_search
portlet_search . chgc
```

Fig. 2　Sample record instance of interest behaviors elicited.

图 2　重要兴趣行为获取记录实例

然后,我们在实验数据集中随机抽取一名用户的 AT_u 以及 S_u,利用式(3)分别计算在规模 $|AT_u|$ 为 100,500,1000 以及 1500 时,该用户的兴趣权重 $W(content_k)$,如图 3(a). 可以看出当 $|AT_u| = 100$ 时,折线分布较陡峭,不同兴趣内容上的兴趣分布较明显,体现了用户的突出兴趣,但由于访问频率较低,兴趣的集中稳定性还不显著,存在偶然访问因素影响;当 $|AT_u|$ 变化增长到 500 和 1000 时,兴趣分布变化较平缓且趋于稳定,基本集中在某几个兴趣内容上,能够体现用户稳定的突出兴趣;随着 $|AT_u|$ 继续增长到 1500,其兴趣分布更加平缓且变化较小,不宜再表征用户稳定的突出兴趣,需要进行兴趣模型的更新以获取更准确的兴趣信息. 从而得到结论:用户的访问事务集规模在[100,1500]时,用户偶然变化的突出兴趣会随访问频率和规模的变化逐渐

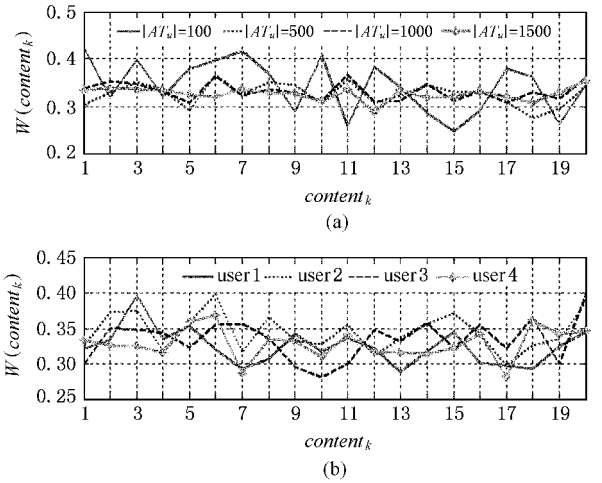

Fig. 3　Distribution of personalized interest degree. (a) Different AT_u's range and (b) Different user.

图 3　个性化兴趣分布. (a) 不同事务集规模;(b) 不同用户

逼近于稳定兴趣,可以用于门户兴趣挖掘.

基于以上结论,我们又随机抽取了 4 个用户(user1～4)的访问事务集,并将规模控制在 $|AT_u|=500$,观察不同用户的个性化兴趣分布,如图 3(b).结果显示每个用户稳定的突出兴趣确有差异,与其个性化访问动机、行为喜好等有所关联,体现了不同用户的个性化兴趣分布特点,并印证了本文个性化兴趣挖掘的意义.

此外,对门户兴趣获取机制对门户服务器的性能影响进行评测.我们分 10 次记录了加载兴趣获取机制前后,在客户端浏览器中载入门户桌面页面所需的时间,结果如图4所示.加载时间的平均增量为 120ms,对用户的正常浏览基本不会产生影响,也保证了其独立于门户平台的松耦合性.

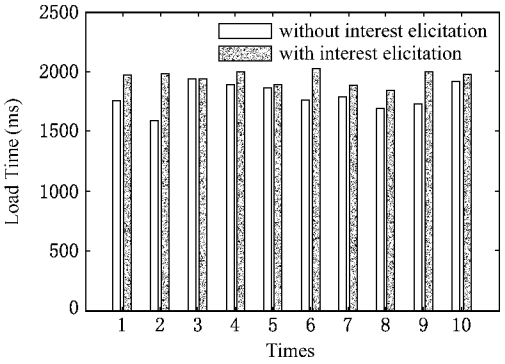

Fig. 4 Comparison in the elicitation performance.

图 4 兴趣获取机制性能影响比较

4.2 迁移模式发现实验结果分析

本实验的挖掘对象为不同门户用户,为便于观察,仍选取前面实验中的 4 个用户的访问事务集.基于门户个性化兴趣描述,分别计算其对应的兴趣内容间的转移概率 $P(q_i\rightarrow q_j)$ 以及兴趣概念集上的观测概率 $P(\sigma_z^i\mid q_i)$,并建立 HMM 模型进行迁移模式发现.

图 5 和图 6 以 user1 为例,分别给出了转移概率 $P(q_i\rightarrow q_j)$ 及其在 3 个兴趣概念集上两节点间兴趣迁移概率 $P(\sigma)=P(q_i\rightarrow q_j)P(\sigma\mid q_i)$ 的分布.图 6(a)～(c)显示了概念集 σ_1,σ_2 和 σ_3 的基本分布,不同样式的序列点刻画了计算所得的可能访问序列(Seq1～5),并按照其迁移概率升序排列(即 Seq5 具有最大的访问概率 $P_{\max}(\sigma)$),可以得到不同节点上的兴趣迁移的可能性分布.对于每个序列点,横坐标表示不同的兴趣内容节点 q_i,纵坐标表示该节点下一步迁移到节点 q_j 的访问概率 $P(\sigma)$.因而,可以根据每个序列点对应的数据对〈q_i,$P(\sigma)$〉通过简单的关联推导得到兴趣迁移的目的节点 q_j,这样一组序列的关联关系即为该用户的访问迁移模式.

P_{ij}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0.0360	0.2349	0.0432	0.0979	0.0052	0.0898	0.0443	0.0621	0.0115	0.0471	0.0011	0.1030	0.0107	0.0728	0.0638	0.0755	0.0906	0.0089	0.0742
2	0.0347	0	0.1920	0.0070	0.0598	0.0397	0.1214	0.0775	0.0252	0.0458	0.0107	0.0358	0.1342	0.0450	0.0356	0.0963	0.1076	0.1222	0.0261	0.0369
3	0.1902	0.1611	0	0.1552	0.1110	0.1944	0.2629	0.2261	0.1399	0.1995	0.1521	0.1911	0.1988	0.1312	0.2419	0.2514	0.2636	0.1830	0.1302	0.0414
4	0.0953	0.0070	0.1837	0	0.0524	0.0464	0.1275	0.0839	0.0181	0.0524	0.0037	0.0425	0.1402	0.0517	0.0284	0.1026	0.1138	0.1283	0.0329	0.0296
5	0.0891	0.0564	0.1248	0.0498	0	0.0938	0.1709	0.1295	0.0326	0.0996	0.0463	0.0902	0.1830	0.0988	0.0228	0.1472	0.1579	0.1717	0.0810	0.0216
6	0.0052	0.0413	0.2413	0.0486	0.1036	0	0.0851	0.0394	0.0676	0.0063	0.0525	0.0040	0.0984	0.0055	0.0784	0.0589	0.0707	0.0859	0.0142	0.0797
7	0.0986	0.1382	0.3567	0.1461	0.2062	0.0930	0	0.0500	0.1669	0.0861	0.1503	0.0974	0.0145	0.0869	0.1786	0.0286	0.0157	0.0009	0.1005	0.1801
8	0.0464	0.0840	0.2921	0.0916	0.1488	0.0410	0.0476	0	0.1113	0.0344	0.0956	0.0452	0.0614	0.0352	0.1226	0.0204	0.0326	0.0484	0.0557	0.1240
9	0.0505	0.0246	0.1627	0.0178	0.0337	0.0633	0.1430	0.1002	0	0.0693	0.0142	0.0595	0.1555	0.0605	0.0101	0.1105	0.1296	0.1438	0.0500	0.0114
10	0.0116	0.0480	0.2492	0.0553	0.1106	0.0064	0.0792	0.0332	0.0744	0	0.0592	0.0105	0.0926	0.0008	0.0853	0.0529	0.0648	0.0801	0.0206	0.0866
11	0.0449	0.0106	0.1794	0.0037	0.0405	0.0499	0.1307	0.0873	0.0144	0.0559	0	0.0460	0.1433	0.0551	0.0246	0.1059	0.1171	0.1315	0.0364	0.0259
12	0.0011	0.0371	0.2363	0.0444	0.0991	0.0040	0.0888	0.0432	0.0633	0.0103	0.0482	0	0.1020	0.0095	0.0740	0.0627	0.0745	0.0896	0.0101	0.0754
13	0.1149	0.1550	0.3767	0.1631	0.2240	0.1091	0.0148	0.0655	0.1841	0.1021	0.1673	0.1136	0	0.1030	0.1960	0.0438	0.0307	0.0138	0.1248	0.1975
14	0.0108	0.0471	0.2482	0.0545	0.1097	0.0056	0.0800	0.0340	0.0735	0.0008	0.0583	0.0096	0.0934	0	0.0844	0.0537	0.0656	0.0808	0.0198	0.0857
15	0.0679	0.0343	0.1511	0.0276	0.0234	0.0727	0.1516	0.1092	0.0100	0.0706	0.0240	0.0689	0.1639	0.0778	0	0.1273	0.1383	0.1523	0.0595	0.0013
16	0.0681	0.1065	0.3190	0.1143	0.1727	0.0626	0.0278	0.0208	0.1344	0.0559	0.1184	0.0669	0.0419	0.0567	0.1459	0	0.0125	0.0287	0.0777	0.1473
17	0.0817	0.1206	0.3358	0.1205	0.1875	0.0761	0.0154	0.0337	0.1488	0.0693	0.1326	0.0805	0.0298	0.0702	0.1604	0.0127	0	0.0163	0.0914	0.1619
18	0.0996	0.1392	0.3579	0.1472	0.2073	0.0940	0.0009	0.0509	0.1679	0.0870	0.1514	0.0984	0.0137	0.0879	0.1797	0.0295	0.0166	0	0.1095	0.1812
19	0.0089	0.0268	0.2239	0.0340	0.0881	0.0140	0.0979	0.0528	0.0527	0.0202	0.0378	0.0100	0.1110	0.0194	0.0633	0.0721	0.0837	0.0987	0	0.0646
20	0.0690	0.0356	0.1496	0.0288	0.0221	0.0738	0.1526	0.1103	0.0112	0.0797	0.0252	0.0701	0.1560	0.0790	0.0013	0.1284	0.1393	0.1534	0.0607	0

Fig. 5 Transfer probability $P(q_i\rightarrow q_j)$ of user1.

图 5 user1 的访问转移概率 $P(q_i\rightarrow q_j)$

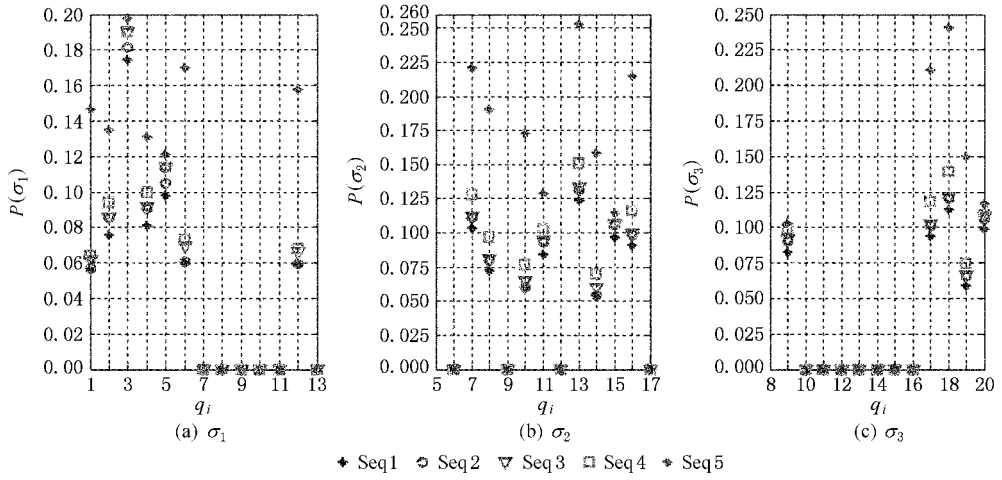

Fig. 6　Distribution of interest navigation patterns on rough interest concept sets of user1. (a) σ_1; (b) σ_2; and (c) σ_3.

图 6　user1 在不同兴趣概念集上的两节点间迁移概率分布. (a) σ_1;(b) σ_2;(c) σ_3

基于以上计算结果,表 3 给出了实验所发现的 user1～4 在不同兴趣概念集上具有最大访问概率 $P_{\max}(\sigma')$的兴趣迁移模式,为了更加直观和具体,我们结合表 2 的兴趣内容辅以说明. 观察得到,不同门户用户的兴趣迁移模式能够区分描述其个性化兴趣目的,并且可以分解得到兴趣内容两节点间的迁移概率分布,较好地刻画了访问序列的迁移状态,验证了前面分析提及的特点.

Table 3　Navigation Pattern Sequences with $P_{\max}(\sigma')$

表 3　具有最大访问概率 $P_{\max}(\sigma')$的迁移模式序列

User	Navigation Pattern Sequences
user1	*tab_subSystem*→*portlet_subSystem*.*gcjs*1→*portlet_infoRelease*→*link_infoRelease*.*gcjs*1
	tab_search→*portlet_search*.*hhzh*→*portlet_search*.*tcqs*
	portlet_rss→*tab_personalization*→*portlet_taskReminder*
user2	*tab_subSystem*→*portlet_subSystem*.*gcjs*1→*portlet_subSystem*.*gcjs*2→*portlet_infoRelease*→*link_infoRelease*.*gcjs*2
	portlet_subSystem.*szwd*→*tab_search*→*portlet_search*.*jsyj*→*portlet_search*.*zhsj*
	tab_personalization→*portlet_contact*→*portlet_taskReminder*
user3	*tab_subSystem*→*portlet_infoRelease*→*link_infoRelease*.*gcjs*2→*portlet_search*.*chgc*
	portlet_subSystem.*xxzx*→*tab_search_portlet_search*.*zhsj_portlet_infoRelease*.*ywdt*
	tab_personalization→*portlet_calendar*→*portlet_taskReminder*
user4	*tab_subSystem*→*portlet_infoRelease*→*link_infoRelease*.*gcjs*1→*link_infoRelease*.*gcjs*2
	tab_search→*portlet_search*.*hhzh*→*portlet_infoRelease*.*ywdt*
	portlet_rss→*tab_personalization*→*portlet_calendar*

从而得到结论:采用 HMM 扩展描述后的兴趣迁移模式发现方法能与兴趣行为获取机制结合,提升访问概率计算结果的准确精度和范围的同时,能够发现不同门户用户更多有效的兴趣迁移模式和兴趣目的. 为门户更加全面准确地进行个性化预测推荐提供了一定的研究根据和基础.

5　结束语

门户个性化兴趣获取与迁移模式发现在门户兴趣挖掘应用中有着重要的意义,本文总结分析了门户个性化兴趣行为的映射描述,给出了一种独立于门户平台的含隐私保护的门户个性化兴趣获取机制,以及可以进行兴趣迁移模式发现的带有门户个性化兴趣描述的 HMM 扩展. 这些方法能够表达和预测门户用户复杂多样的个性化兴趣,在完善个性化兴趣挖掘的基础上提升了描述精度和预测能力.

我们将对这些挖掘结果如何更好地应用于门户个性化推荐服务做进一步的研究,以提高实时预测的效率和精确性. 同时,完善隐私保护策略以达到跨门户平台的良好的安全性和自适应性.

参　考　文　献

[1] Zeng Chun. Concept representations and algorithms in information filtering:[Ph D dissertation][D]. Beijing:Tsinghua University, 2003 (in Chinese)
(曾春. 信息过滤的概念表示与算法研究:[博士论文][D]. 北京:清华大学, 2003)

[2] M Claypool, P Le, M Wased, *et al*. Implicit interest indicators [C]. In:Proc of the IUI '01. New York:ACM Press, 2001. 33-40

[3] M Albanese, A Picariello, C Sansone, *et al*. Web personalization based on static information and dynamic behavior [C]. In:Proc of the WIDM '04. New York:ACM Press, 2004. 80-87

[4] Xie Yi, Yu Shunzheng. A dynamic anomaly detection model for Web user behavior based on HsMM [C]. The 10th CSCWD, Nanjing, 2006

[5] J Velasquez, H Yasuda, T Aoki. Combining the Web content and usage mining to understand the visitor behavior [C]. In: Proc of the 3rd ICDM. Los Alamitos, CA: IEEE Computer Society Press, 2003

[6] L Lancieri, N Durand. Internet user behavior: Compared study of the access traces and application to the discovery of communities [J]. IEEE Trans on System, Man and Cyberneties—Part A: Systems and Humans, 2006, 36(1): 208-219

[7] M Chen, A LaPaugh, S J Pal. Categorizing information objects from user access patterns [C]. In: Proc of the CIKM '02. New York: ACM Press, 2002. 365-372

[8] Wang Shi, Gao Wen, Li Jintao, *et al*. Mining interest navigation patterns based on hidden Markov model [J]. Chinese Journal of Computers, 2001, 24(2): 152-157 (in Chinese)
(王实, 高文, 李锦涛, 等. 基于隐马尔可夫模型的兴趣迁移模式发现[J]. 计算机学报, 2001, 24(2): 152-157)

[9] H R Kim. Learning implicit user interest hierarchy for Web personalization: [Ph D dissertation] [D]. Melbourne, Florida: Florida Institute of Technology, 2005

[10] Wu Jing, Xiong Zhang. A portal-oriented personalized recommendation using meta-recommender engine [C]. The Int'l Conf on Artificial Intelligence, Beijing, 2006

[11] B Zhou, S C Hui, A C M Fong. Discovering and visualizing temporal-based Web access behavior [C]. In: Proc of the 2005 IEEE/WIC/ACM Int'l Conf on Web Intelligence. Los Alamitos, CA: IEEE Computer Society Press, 2005

[12] Zhao Liang, Zhang Shouzhi, Fan Xiaofeng. Web browsing feature mining of an anonymous user [J]. Journal of Computer Research and Development, 2002, 39(12): 1758-1763 (in Chinese)
(赵亮, 张守志, 范晓峰. 匿名用户的网络浏览特征挖掘[J]. 计算机研究与发展, 2002, 39(12): 1758-1763)

[13] Guo Yan, Bai Shuo, Yang Zhifeng, *et al*. Analyzing scale of Web logs and mining users' interests [J]. Chinese Journal of Computers, 2005, 28(9): 1483-1496 (in Chinese)
(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, 28(9): 1483-1496)

[14] J Canny. Collaborative filtering with privacy via factor analysis [C]. In: Proc of the 25th SIGIR. New York: ACM Press, 2002. 238-245

[15] D J Weitzner, H Abelson, L T Berners, *et al*. Transparent accountable data mining: New strategies for privacy protection [OL]. http://www.csail.mit.edu, 2006

[16] Notation3 [OL]. http://www.w3.org/2000/10/swap, 2005

Wu Jing, born in 1979. Ph. D. candidate. Received her B. A.'s degree in computer science and engineering from BUAA in 2001. Her current research interests include interest mining, recommendation service and portal technology.

吴 晶,1979 年生,博士研究生,主要研究方向为兴趣挖掘、推荐服务以及门户技术.

Zhang Pin, born in 1983. M. S. candidate. Received his B. A.'s degree in computer science and engineering from BUAA in 2006. His current research interests include Web usage mining.

张 品,1983 年生,硕士研究生,主要研究方向为 Web 使用挖掘.

Luo Xin, born in 1983. Ph. D. candidate. Received his B. A.'s degree in software technology from UESTC in 2005. His current research interests include personalization service and e-commerce.

罗 辛,1983 年生,博士研究生,主要研究方向为个性化服务和电子商务.

Sheng Hao, born in 1981. Ph. D. candidate. Received his B. A.'s degree in computer science and engineering from BUAA in 2003. His current research interests include media data mining and computer vision.

盛 浩,1981 年生,博士研究生,主要研究方向为媒体数据挖掘和计算机视觉技术.

Xiong Zhang, born in 1956. He is responsible professor and Ph. D. supervisor of BUAA in computer application area. His main research interests include distributed system, multimedia processing and network engineering.

熊 璋,1956 年生,北航计算机应用学科责任教授,博士生导师,主要研究方向为分布式系统、多媒体处理以及大型网络应用工程.

Research Background

Interest mining is an exciting research area that tries to enrich the personalization services on Web by using technologies from WM, AI, IR, psychology and knowledge learning. Most existing WM techniques rely on the extraction of weblog entries and observation of the structured data. There is an increasing need to both elicit personalized interests implicitly and infer the navigation trends on open portal platform in personalization. Some problems match a few of the common approach criteria, but the obtained interest descriptions are either invalid or unaccountable. Therefore, it drives us to attempt to present more effective specialized schemes in this potential area, based on our deployment and integration experiences on several large portal projects. The related papers have been published or accepted by some journals or international conferences. We gratefully acknowledge the support of the National High-Tech. R&D Program Foundation (2002AA113070).