# 基于用户点击流的标签兴趣模型的研究与实现

高宏宾 刘劲飞
(五邑大学计算机学院 广东 江门 529020)

**摘 要** 分析用户的网络交互行为与用户兴趣之间的关系,针对目前的兴趣标签建模方法的不足,提出将用户的点击对象进行标签量化,通过用户的点击行为建立用户兴趣模型的方法,并在社交网络环境中对模型进行用户兴趣分析与测试。测试结果表明,该方法能有效地构建用户兴趣模型,证明了该方法的可行性。

**关键词** 标签 点击流 用户兴趣模型 权重

**中图分类号** TP39 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2013.06.046

# ON USER CLICK STREAM-BASED TAG INTEREST MODEL AND ITS IMPLEMENTATION

Gao Hongbin Liu Jinfei
(*School of Computer Science and Technology*, *Wuyi University*, *Jiangmen 529020*, *Guangdong*, *China*)

**Abstract** In this paper we analyse the relation between user's network interaction behaviours and interests. According to the inadequacy of interest tag modelling at present, we propose a new method in which users' clicking objectives are quantified by tags and the users interest model is established through the clicking behaviours of users. We also carry out users interest analyse and test on this model in social networking environment. Test results show that the proposed method can effectively construct user interest model, this proves its feasibility.

**Keywords** Tag Click stream User interest model Weight

## 0 引 言

标签(Tag)是 Web 2.0 的重要组成部分,用户在发布内容的同时可以添加一些标签,这些标签在概括内容的同时也表征了用户的兴趣。一般来说用户的兴趣爱好会在其信息交互的行为中表现出来[1],用户根据自己的兴趣阅读、收藏、评论内容等,这些行为反映了用户的兴趣所在。目前,基于标签的兴趣模型被广泛研究,如:胡昌平的基于标签词频的用户兴趣分析[2]、毛进的基于密度聚类的兴趣建模方法[3]、SHEPITSEN 的基于层次聚类的标签兴趣模型[4]等,这些基于标签的用户兴趣建模的研究方法主要包括:基于标签词频、加权树、聚类、维基法等,通过分析,发现这些方法仍有以下不足:

1) 不能在短期内对新用户很好地进行兴趣挖掘,如:基于标签词频统计的兴趣分析;

2) 用户的短期兴趣变化无常,需要实时挖掘用户信息,如:基于遗忘机制[4]的用户兴趣模型;

3) 用户兴趣可以通过多种方式表现出来,用户的浏览、搜索、评论、收藏等行为都有用户的兴趣体现;

4) 不能提供一种交互方式,让用户去慢慢建立自己的兴趣特征库[5]。

从国内外对于兴趣模型的研究来看,普遍采用对用户的历史数据进行挖掘分析,最终建立用户的兴趣模型。很明显这些方法对不产生历史数据的用户存在局限性。用户的兴趣还与用户的行为有很大关系,基于历史数据的兴趣挖掘无法挖掘用户行为所包含的兴趣特征,对此,本文从用户的访问行为进行分析,从中挖掘出用户的兴趣。

点击流数据挖掘,是挖掘用户行为广泛使用的方法。用户根据自己的兴趣浏览、搜索、评论、收藏内容等等,这些行为反映了用户的兴趣的偏好,如果将这些行为进行标签量化,那么就可以将用户的一系列页面交互行为转换为用户对自己感兴趣标签的搜集过程。

本文创新地提出了基于标签量化的点击流兴趣建模方法。根据页面的内容建立标签和标签初始权值,再针对用户的行为(如:浏览、评论、收藏等)分别定义标签权值的加权数,当有用户访问到内容并产生相应的行为时,系统就会收集到用户特定行为所产生的兴趣标签集合,最后对这个标签集合进行有效的算法转换,建立用户的持久兴趣模型。

本文最后构建一个基于这个兴趣模型的社交网络,通过分析发现该模型能有效地建立用户兴趣特征。

## 1 相关概念

基于用户点击流的标签兴趣模型,包含点击流、流标签兴

收稿日期:2012 - 07 - 08。广东省自然科学基金项目(S2011010003681)。**高宏宾**,副教授,主研领域:数据挖掘与分布式计算,算法设计与分析。**刘劲飞**,硕士生。

趣建模等相关概念,本节对这些概念进行阐述,并进行相关定义。

### 1.1 点击流

点击流[5]即访问者在网页上的持续访问轨迹。随着 Web 技术的发展,客户对网站的每一次点击都会被记录到网络服务器的日志中,由此产生了点击流数据,点击流数据简单说,就是 Web 服务器上一系列有序的日志记录。

常见的基于 Apache 服务的日志格式:

LogFormat "%h %l %u %t\"%r\" %>s %b" common

h: 为远程主机

l: 为远程登录名字(来自 identd)

u: 是远程用户(来自 auth)

t: 表示的时间(或称为标准英文格式)

r: 为请求的第一行

可以看到,一次点击产生的日志描述了访问者的相关信息,通过这些信息可以分析客户的访问行为。

基于点击流的数据分析[6],可以挖掘出用户感兴趣的内容,传统的点击流数据挖掘一般需要挖掘用户访问日志,建立用户的页面访问轨迹,以此给用户提供相关服务。

单纯的建立用户访问行为模型,并不能完全反映出用户的兴趣喜好,进而也不能提供很好的推荐服务。

### 1.2 Folksonomy 介绍

Folksonomy 是 Folks 与 Taxonomy 组合词,Folks 表示一群人,Taxonomy 指分类方法,所以也称 Folksonomy 为社会化标签分类方法 Social Classification。标签是 Folksonomy 的典型运用。

Folksonomy 可以用一个元组 $F$ 表示:$F=(U,T,R,A)$。其中 $U,T,R$ 分别表示用户、标签以及资源集合。元组中 $A$ 表示这三者的关系 $A\subseteq U\times T\times R$。$(u,r,t)\in A$ 称为标签分配,也就是用户 $u$ 给资源 $r$ 添加标签 $t$。

Personomy 是 $F$ 在单个用户 $u$ 上的约束,可以表示 $P_u=(T_u,R_u,A_u)$,$A_u=\{(t,r)\mid(u,r,t)\in A\}$ 表示用户 $u$ 的标签分配,$T_u=\{t\mid(t,r)\in A_u\}$,$R_u=\{r\mid(t,r)\in A_u\}$ 分别表示用户 $u$ 的标签和资源的集合。

基于点击流的标签兴趣模型主要对用户的交互行为(点击流)进行兴趣挖掘,标签是这个挖掘过程的基本元素。

### 1.3 流标签

结合用户点击流和 Folksonomy(标签)技术,本文提出流标签概念,流标签即用户点击流的标签量化。用户的每次点击都代表了一定的含义,这就需要系统去区分每个点击的内容。对于流标签我们定义为:

$$ST=[T_1\quad T_2\quad\cdots\quad T_i\quad\cdots\quad T_k]$$

$$T_i=[t_i\quad w_i\quad d_i\quad]^{\mathrm{T}}\quad 1\leqslant i\leqslant k$$

$t_i$: 是标签名(tag),就是网页内容的关键字标签;

$w_i$: 是标签权重(weight),通常一篇内容对应多个标签,同一个标签可以对应多个内容,标签在内容中的权值也不同;

$d_i$: 用户浏览内容的时间(date),用户对用户的短期兴趣进行分析。对于同一篇内容,$d_i$ 是相同的,公式可以简化为:

$$\begin{bmatrix}t_1&t_2&\cdots&t_i&\cdots&t_k\\w_1&w_2&\cdots&w_i&\cdots&w_k\\d_1&d_2&\cdots&d_i&\cdots&d_k\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&d\end{bmatrix}\begin{bmatrix}t_1&t_2&\cdots&t_i&\cdots&t_k\\w_1&w_2&\cdots&w_i&\cdots&w_k\\1&1&\cdots&1&\cdots&1\end{bmatrix}$$

简化处理之后,对于页面上的一个内容对象 $R$ 只定义标签名称 $t_i$ 并计算标签的权重 $w_i$(通常使用词频来定义权重),时间 $d$ 在数据持久化时处理。区分了每次点击所代表的兴趣信息,就可以对用户一系列的点击行为进行分析,挖掘出用户的兴趣。

### 1.4 会话缓冲区

会话缓冲区是一个数据存储区,用于缓存用户的点击流数据,从数学分析的角度,会话缓冲区表示为一个集合,集合当中的元素为点击流。

用户登录进入网站,系统建立该用户的会话(Session)缓冲区(定义为 $S$)。社交网络面向的是海量的用户并发的访问,如果对用户的每次点击事件,都直接将流标签更新到用户兴趣模型库中,那么服务器将不堪重负。因此,本文在模型当中引入会话缓冲区,通过会话缓冲区,将用户的兴趣标签缓存起来,可以有效地解决实际应用的高负荷问题。

同时,会话缓冲区可以过滤掉用户的一些无用访问数据,比如用户误点了某项自己并不感兴趣的内容,这种情况不能更新到用户兴趣模型;当用户频繁访问了同一个标签的时候,可以在会话缓冲区中增加该标签的权重。

## 2 点击流信息表示

前面提到了标签、内容、用户这三个对象,通过内容建立标签,通过标签建立用户的兴趣。所以要收集用户的兴趣标签就需要对页面的内容添加标签,并计算标签的初始权重。

### 2.1 标签的权重

Web 内容标签的权重可以使用 TF * IDF(Term Frequency, Inverse Document Frequency)[7]来计算,其中 $TF$ 表示标签 $t_i$ 的使用频率,$IDF$ 表示逆向词频,由总体文档树比包含标签 $t_i$ 的文档树再取对数得到。

$$TF*IDF=tf_{i,j}\times idf_i=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{|\{j:t_i\in d_j\}|}\quad(1)$$

其中 $n_{i,j}$ 表示词 $t_i$ 在文件 $d_i$ 中出现的次数,分母表示文件 $d_i$ 中所有词的出现次数之和,$|D|$ 表示语料库中文件总数,$|\{j:t_i\in d_j\}|$ 表示包含词 $t_i$ 的文件数目(即 $n_{i,j}\neq 0$ 的文件数目),因为可能出现词 $t_i$ 不出现在标签库中,就会导致被除数为零,因此一般将 $|\{j:t_i\in d_j\}|+1$。

对本文提出的点击流,单纯的 TF * TDF 计算标签的权重还不够。用户阅读文档 $d_i$ 和评论文档 $d_i$ 所代表的兴趣量是不一样的,很明显用户评论文档说明用户对文档更加感兴趣,应该适当的增加权重,由此,可以定义 Web 内容的标签权重基数为 $e(e=TF*IDF)$,用户评 J 论、收藏文档时,相应的标签权重设为 $e$ 的 $n$ 倍,这样就可以区分用户的每次点击所代表的兴趣量。

### 2.2 点击流向量空间模型

用户的页面访问会产生一系列的点击事件,并产生相应的点击流标签集合,按照 Folksonomy[8]理论,可以建立用户点击行为的标签信息向量模型。

**定义 1**　点击流向量 $V(r)=\{t_1,w(t_1,r,n);t_2,w(t_2,r,n);\cdots t_k,w(t_k,r,n);\cdots;t_n,w(t_n,r,n)\}$。$V(r)$ 表示点击对象 $r$ 的向量,$t_1,t_2,\cdots,t_k,\cdots,t_n$ 是对象 $r$ 的标签集合,$w(t,r,n)$ 是标签在 $r$ 对象中的权重,$n$ 是不同操作定义的标签权值的倍数。

权重 $w(t,r,n)$ 的计算,是按照上文所述的 TF * IDF 式(1)来计算的,将式(1)整理如下:

$$w(t,r,n) = n \times w(t,r) = n \times \frac{TF(t,r) \times \log(m/n_k + 0.01)}{\sqrt{\sum_{t \in T} [TF(t,r) \times \log(m/n_k + 0.01)]^2}} \quad (2)$$

式中:$w(t,r)$即标签初始化权值,$TF(t,r)$是标签 $t$ 标注 Web 对象 $r$ 的次数,$m$ 为 Web 对象的总数,log 底数为 10,$n_k$ 是被标签 $t$ 标注过的 Web 对象数量,式中分母为归一化因子。

点击对象 $r$ 与上文的内容资源 $R$ 区别是,同一个内容资源对应多个点击对象 $r$,如:浏览、评论、收藏等等。

用户的每次点击都会产生一个标签向量集合 $V(r)$,该集合中的标签权值按照用户的点击行为和基础标签权值计算(本文采用 TF * IDF 方法计算基础权值)。

对于同一篇内容,标签权重按照用户的行为取值,如果用户是单纯的浏览内容,则可以定义标签的权重为基本值 $e$($e$ 默认初始权重),如果用户是评论内容,说明该内容与用户的兴趣关系更加密切,可以适当增加标签的权值(如 $2e$),这样,通过综合用户一系列的点击事件,可以建立关联用户行为的标签向量集合。

## 3 兴趣模型

基于点击流的标签兴趣模型首先将用户的点击行为进行标签量化,再根据用户在页面的行为建立用户的兴趣标签集合,最后综合用户的历史兴趣更新用户的兴趣标签,建立用户的持久兴趣标签集合。

### 3.1 建立用户兴趣模型

**定义 2** 用户兴趣 $UF(s,u) = \{f(s,u) \mid s \in ST, u \in UF\}$。$s$ 是会话缓冲区中的流标签,$ST$ 是缓存区全部流标签集合,$u$ 为用户的历史兴趣标签,$UF$ 是用户历史兴趣集合,$f$ 是标签加权算法。

**定义 3** 长期兴趣 $LF$。对于 $\forall$ 标签 T,若 $w > \alpha$,且 $d > \beta$,则称标签 $T$ 为用户的一个长期兴趣 $LF$。其中 $w$ 为标签 $T$ 的权值,$d$ 为标签 $T$ 的添加时间,$\alpha$ 和 $\beta$ 为判断长期兴趣的阈值。

**定义 4** 无关兴趣 $NF$。用户的兴趣标签数量最大定义为 $\lambda$,当用户的兴趣标签数超过 $\lambda$,每次更新兴趣,剔除权值最小的超出的标签 $NF$。

结合以上定义,本文基于用户点击流的标签兴趣模型如图 1 所示。

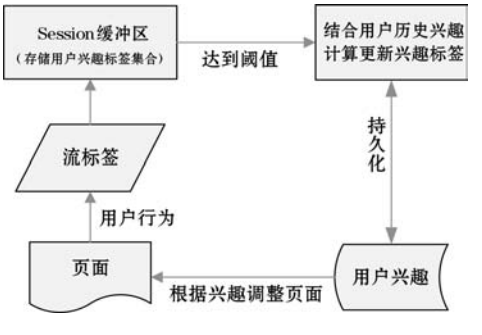


图 1 用户兴趣模型

模型运作的整个流程:

(1) 用户发布内容 $R_u$ 的同时按照自己的主观想法和兴趣添加上对应的标签 $T_u$。

(2) 系统对用户提交的内容 $R_u$,按照式(2)分别计算用户相关交互行为(浏览、收藏、评论等)的标签权值。

(3) 其他用户 $U_i$ 访问 Web 系统,系统建立会话 Session 存储区。

(4) 用户 $U_i$ 浏览、收藏、评论…Web 系统上的内容,系统收集操作对应的兴趣量,存储到 Session 存储区。

(5) 当 Session 存储区中用户兴趣集合达到阈值(比如标签数超过 10,或者标签的权值总数超过 15 等),将 Session 中的兴趣标签执行第(6) 步的更新方法,否则接着第(4)步的收集过程。

(6) 对 Session 中的每个兴趣标签与用户的历史兴趣标签进行加权,计算长期兴趣 $LF$,剔除无关兴趣 $NF$,最终生成新的用户兴趣集合 $UF$,更新完成后,清空 Session 中的数据,回到第(4)步重新收集用户的兴趣标签。

在整个过程当中,(1)和(2)为产生内容,(3) - (6)为使用内容,用户定义内容的标签,系统再计算各个标签在用户各种行为下的权值。确定了内容的流标签之后,就可以通过流标签对内容访问者进行兴趣收集,随着用户行为的积累,逐步建立起用户的兴趣标签集。

### 3.2 用户对于资源的兴趣度

在社交网络上,用户与资源的兴趣度通过标签关联,如果标注资源的标签同样存在用户的兴趣特征库中,那么可以通过计算,得出用户对资源的兴趣度。定义 $FR$ 为用户的兴趣度,$T_u^{name}$ 与 $T_r^{name}$ 分别表示用户兴趣标签与资源的标签关键字集合,对于 $\forall$ 标签 $T_i^{name}$,若 $T_i^{name} \in T_r^{name}$,且 $T_i^{name} \in T_u^{name}$,则标签 $T_i^{name}$ 被用户和资源所共有,则 $FR$ 可以表示为 $T_i^{name}$ 对应的权值的函数。

$$FR = \sum \frac{T_i^{weight}}{N(T_u^{weight})} \times T_{i,r}^{weight} \times \frac{N(T_r^{Max})}{N(T_r)}$$

其中,$T_i^{weight}$ 为标签 $T_i^{name}$ 对应的权值,$N(T_u^{weight})$ 为用户 $u$ 的标签权值总数,$T_{i,r}^{weight}$ 为资源 $r$ 的标签权值,$N(T_r^{Max})$ 为网络资源可定义标签的最大数目(用户发布内容的时候可以添加的标签数不能无限制,比如新浪轻博客、点点都限制为 5 个标签),$N(T_r)$ 表示资源 $r$ 所包含的标签个数。

$N(T_r^{Max}) / N(T_r)$,资源 $r$ 包含的标签数为 $N(T_r)$,$N(T_r)$ 越小则单个标签的权值越高,该标签表达资源内容的含义越有概括性。$T_{i,r}^{weight}$ 为资源标签的权值,其计算公式按照式(1)来计算。

## 4 实验评估

为了测试兴趣模型的实际效果,搭建了基于该模型的社交网络平台(http://sns.scofier.com),通过跟踪用户的使用数据,组织实验,得出以下结果。

### 4.1 兴趣标签使用分布统计

下面的实验数据,共统计文章 165 篇,包含科技、笑话、情感、健康等四大类,得到标签总数 247 个,通过这些数据进行用户点击流兴趣分析。

用户标签的使用频率按照指数分布(幂率,power laws),也就是常说的长尾分布,频繁使用的标签只占整个标签集合的很少部分,多数标签只被用户使用一两次。

通过对收集到的 165 篇文章的标签进行统计,得出标签使用频次分布如图 2 所示。

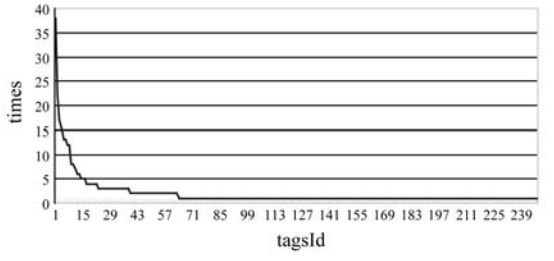


图 2　标签使用频次分布

图中横坐标是标签的 ID,纵坐标是标签的使用次数,从图中可以看出,用户标签确实服从长尾分布,26% 的标签使用频次占据了标签使用频次总数的 63% 。

## 4.2　兴趣模型的性能对比

为了检测基于用户点击流的标签兴趣模型的性能,本文采取 MAE(平均绝对误差)作为度量的标准。平均绝对误差也就是系统预计算内容与用户的兴趣度,与用户对内容的实际兴趣度的偏差。同时,为了对比性能,实验中还使用基于标签词频统计的建模方法进行对比测试。在测试集中,$n$ 个系统预测内容表示为向量 $\{s_1, s_2, \cdots, s_n\}$,实际用户评定的兴趣度向量为 $\{u_1, u_2, \cdots, u_n\}$,则 *MAE* 计算方法为:

$$MAE = \frac{\sum_i |s_i - u_i|}{n}$$

其中,$|s_i - u_i|$ 表示系统预测值 $s_i$ 与用户实际兴趣度 $u_i$ 的绝对误差,*MAE* 的值越小表示系统预测的效果越好。

实验中我们根据收集到的用户兴趣标签,将系统中的文章内容分别采用基于词频统计建模方法和基于点击流标签兴趣建模方法来进行用户与资源的兴趣度计算,根据用户与资源的兴趣程度将文章推荐给用户。再调查 10 个用户对系统推荐的文章进行实际兴趣标注。用户只采用 1 与 0 来标注自己对推荐的文章的兴趣度,1 表示喜欢,0 表示不喜欢。因为每个用户的兴趣不同,所以可以推荐的文章数 $n$ 不一样。对用户的调查统计数据,如表 1 所示。

**表 1　系统推荐文章的 MAE 值**

| 用户 | 注册时间 | 标签数 | 文章数 | 评论数 | 推荐文章 | 喜欢文章 | *MAE* |
|---|---|---|---|---|---|---|---|
| User1 | 20120215 | 50 | 48 | 0 | 52 | 40 | 0.23 |
| User2 | 20120217 | 50 | 67 | 2 | 52 | 39 | 0.25 |
| User3 | 20120229 | 50 | 12 | 2 | 14 | 12 | 0.14 |
| User4 | 20120229 | 50 | 14 | 1 | 93 | 66 | 0.29 |
| User5 | 20120326 | 7 | 7 | 0 | 12 | 8 | 0.33 |
| User6 | 20120326 | 5 | 2 | 2 | 60 | 50 | 0.17 |
| User7 | 20120326 | 0 | 0 | 0 | 0 | 0 | 0 |
| User8 | 20120327 | 0 | 0 | 0 | 0 | 0 | 0 |
| User9 | 20120415 | 15 | 12 | 4 | 33 | 25 | 0.24 |
| User10 | 20120430 | 0 | 0 | 0 | 0 | 0 | 0 |

基于词频统计的方法,主要对用户提交的内容进行词频统计分析,统计用户提交内容使用率最高的标签,将其定义为用户的兴趣标签,并以此作为内容推荐的依据。实验中对用户频繁使用的标签进行统计,根据用户最频繁使用的标签进行内容推荐,最后测试用户对于推荐内容的兴趣程度。从图 3 中可以看到基于点击流标签兴趣模型的用户的 *MAE* 值比较理想,有效性比基于词频统计要好很多。

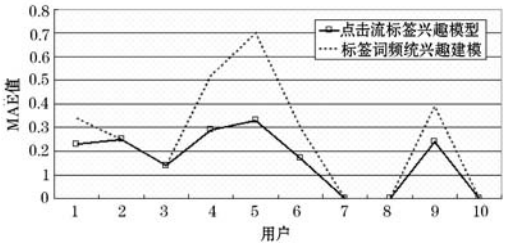


图 3　系统 *MAE* 值测试

## 4.3　用户兴趣相似性测试

用户兴趣的相似性通过用户的兴趣标签来计算的,用户的兴趣标签是一个矩阵,对于 $\forall$ 标签 $T_i$,若 $T_i \in U_i$,且 $T_i \in U_j$ 则表明 $T_i$ 是用户 $U_i$ 和用户 $U_j$ 的共同兴趣。同时需要考虑标签 $T_i$ 的权值与用户兴趣整体权值之比,比值越高说明该标签对用户兴趣的表现能力越强,我们得出如下用户兴趣相似性 *FS* 计算公式:

$$FS = n \times \sum \frac{T_i^{power}}{N(T_i^{power})} \times \frac{T_j^{power}}{N(T_j^{power})}$$

其中,$n$ 表示用户 $U_i$ 与用户 $U_j$ 包含相同标签的个数,$T_i^{power}$ 表示用户 $U_i$ 的标签 $T_i$ 的权重,$N(T_i^{power})$ 表示 $U_i$ 所有标签的权重总和,同样,$T_j^{power}$ 表示用户 $U_j$ 的标签 $T_j$ 的权重,$N(T_j^{power})$ 表示 $U_j$ 所有标签的权重总和。

通过对社交网络平台用户的兴趣数据进行相似测试,得出用户兴趣相似结果如图 4 所示。

图 4　用户兴趣相似图谱

图 4 中,中间的圈为当前登录用户,连接中间圈的 4 个圈表示与当前登录用户兴趣相似的用户,圆角矩形内的关键字表示与登录用户的共同兴趣标签。可以看到用户依山静水分别与丫丫草、阿 Niu、花苑、睇泪兴趣度相似最高,与睇泪有 5 个共同的兴趣。

# 5　结　语

基于兴趣的社交网络是目前关于社交网络研究的一个热门方向,传统的兴趣建模方法倾向于对用户已有的数据进行挖掘分析,对此,本文创新地将点击流与标签技术相结合,提出基于用户点击流的标签兴趣模型,该模型通过不断地挖掘用户的点击行为(事件)产生的兴趣标签集,建立用户的稳定的兴趣模型。最后,通过实验,论证了基于用户点击流的标签兴趣建模的可行性。

## 参　考　文　献

[1] Julia Stoyanovich, Sihem Amer Yahia. A Study of the Benefit of Leveraging Tagging Behavior to Model [J/OL]. http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.152.2719&rep=rep1&type=pdf.

[2] 林鑫,胡昌平. 基于标签词频统计的科研用户兴趣分析[C]//信息

化与信息资源管理学术研讨会,2009:84 -91.
[3] 易明,操玉杰,沈劲枝,等. 社会化标签系统中基于密度聚类的 Web 用户兴趣建模方法[J]. 情报学报,2011(1):37 -43.
[4] Shepitsen A, Gemmell J, Mobasher B, et al, Personalized recommendation in social tagging systems using hierarchical clustering[C]//Proceedings of the 2008 ACM Conference on Recommender Systems. New York: Acm, 2008:259 -266.
[5] Sule GüNDüZ, M Sc. Recommendation models for web users: user interest model and click-stream tree[D/OL]. 2003. http://web.itu.edu.tr/~sgunduz/papers/thesis.pdf.
[6] Li X, Guo L, Zhao Y H. Tag-based Social Interest Discovery. www2008/Refereed Track: Social Networks & Web 2.0-Discovery and Evolution of Communities[M]. 2008:675 -684.
[7] http://zh.wikipedia.org/wiki/TF-IDF.
[8] Au Yeung, Ching Man, Gibbins, et al. A study of user profile generation from folksonomies[BB/OL]. 2010-04-03. http://eprints.ecs.soton.ac.uk/15222/1/swkm2008_paper.pdf.
[9] Krishnan Ramanathan, Julien Giraudi, Ajay Gupta. Greating hierarchical user profiles using wikipedia [EB/OL]. 2010 -06 -22. http://www.hpl.hp.com/techreports/2008/HPL -2008 -127.pdf.

(上接第 126 页)

采用的是自主 DDD 开发框架 Takia 来支持域模型建模,与普通的领域驱动开发不同的是,Takia 不仅支持 DDD 的四个基本开发元素和相关设计模式[13],还通过扩展 Spring 的 IoC 容器,实现了域模型的透明缓存支持,并基于 AOP 机制实现了异步的声明式消息机制,Takia 同时支持 JDK Future 和 Disruptor 两种底层消息机制,实现 DDD 和传统开发模式的无缝结合。铁水联运平台的业务信息系统使用如图 6 所示的统一 DDD 开发模型。

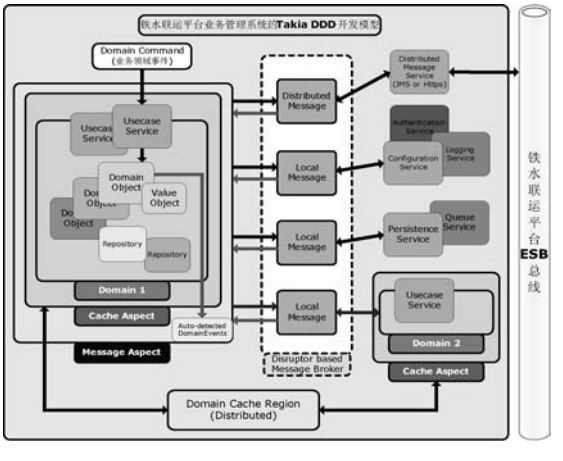


图 6 铁水联运平台业务管理系统的 DDD 开发模型设计

从图 6 中可以看出,业务管理系统的模块在本文中被划分成多个领域进行开发,领域一般由领域对象(Domain Object)、值对象(Value Object)、服务(Usecase Service)和仓储(Repository)等组件构成,领域的外围是 Takia 提供的透明缓存和消息系统,根据基于 Annotation 的领域模型配置,使用 AOP 方式注入到相关的领域对象中,缓存用于提升领域对象的性能,异步的消息模型不仅可以减小领域之间的直接耦合,还能够大幅提高系统的并发性能。Takia 的 DDD 插件支持多种缓存实现,该项目选用 EhCache 以支持集群结构。上图中领域与外界进行通信均通过消息代理(Message Broker)完成,消息代理支持本地消息和分布式消息两种方式,本地消息用于同一虚拟机内模块间的数据交换,如访问其他领域、使用系统的公共服务等,本地消息的底层实现选用基于 Disruptor 的一对多消息机制,Disruptor 作为 LMAX 的底层组件,采用 RingBuffer 结构对 CPU 的并发锁进行了有效优化,具有很好的并发性能[14],以保障铁水联运平台在大批量数据处理过程中充分地利用 CPU 性能,提高业务系统和数据交换平台的吞吐量;分布式消息用于访问外部系统。在本项目中,分布式消息基本是通过 JMS 和 Http 协议转发到铁水联运平台的 ESB 系统中进行处理的。由于消息代理对各领域进行了充分解耦,可以很方便地使用 TDD(测试驱动开发)方式完成业务系统开发,独立的领域也很容易实施单元和集成测试。对于需求的变更,可以方便地修改、删除现有领域或增加新领域,并通过消息代理进行系统集成[15]。

## 3 结 语

本文根据铁水联运平台的特点,提出了一个基于 SOA 和 DDD 的铁水联运信息平台构架方案。该方案采用分层群集的结构建设铁水联运平台,在数据交换问题上使用以 ESB 总线结构为核心的 SOA 构架替代了传统的 EDI 数据交换方式,不仅在数据结构上更加灵活高效、实现无缝的系统整合,也能跨系统实现流程整合。在业务管理系统的开发中采用 DDD 方式建模,基于领域开发构架更加切合业务流程,通过消息机制构建松散的业务模块,领域缓存能够有效地提升业务系统的整体性能,并且对测试和需求更新做出快速反应,具有较大的应用价值。

## 参 考 文 献

[1] 陈涛,黄强,倪少权. 铁水联运信息整合的框架研究方案[M]. 铁道部铁水联运关键技术研究项目,2011(9 -12):23 -106.
[2] 张骏温,许向东. 铁路国际集装箱多式联运 EDI 总体结构的研究[J]. 铁路计算机应用,2002,9(16).
[3] 刘国强. 基于 SOA 技术的企业级软件构架方法研究[J]. 山西建筑,2008,13(2).
[4] 卢致杰,覃正. SOA 构架与电子商务应用集成[J]. 计算机应用研究,2004,21(10).
[5] 唐国磊,宋向群,王文渊,等. 基于 Java 开源框架的港口信息服务系统设计[J]. 水运工程,2011,5(5).
[6] 叶霖. 我国运河物流信息系统框架体系研究[J]. 水运科学,2009,9(3).
[7] Oracle Webcenter Production Documentation[EB/OL]. http://www.oracle.com/technetwork/middleware/webcenter/suite/overview/index.html.
[8] 刘涛,侯秀萍. 基于 ESB 的 SOA 架构的企业应用研究[J]. 计算机技术与发展,2010,10(3).
[9] 周晓燕. 企业服务总线(ESB)在 SOA 中的应用研究[D]. 大连海事大学,2009:33 -53.
[10] 苏小会,刘云. 基于 ESB 的交通信息服务系统的研究[J]. 电脑知识与技术,2011,5(22).
[11] Eric Evans. Domain Driven Design[M]. Addison-Wesley, 2006: 13 -162.
[12] Abel Avram, Floyd Marinescu. Domain-Driven Design Quickly[M]. InfoQ.com, 2007:34 -68.
[13] Jimmy Nilsson. Applying Domain-Driven Design and Patterns[M]. Addison-Wesley, 2006.
[14] Martin Fowler. LMAX Architecture[EB/OL]. http://martinfowler.com/articles/lmax.html.
[15] 黄强,钱文辉. Takia-DDD 构架及详细设计说明书[M]. 上海瑞瀚科技有限公司,2012:24 -67.