一种基于模式的兴趣挖掘算法

李俊芳
（集宁师范学院，内蒙古 乌兰察布市 012000）

摘　要：提出一种基于模式的兴趣挖掘算法，通过查询日志获取访问序列，使用本体中的概念描述用户兴趣，提出一种计算兴趣得分公式，并根据兴趣得分将用户兴趣序列划分为不同的兴趣模式。本文算法实验结果显示，算法能够有效地实现用户兴趣挖掘。

关键词：兴趣挖掘；模式；本体；兴趣得分

中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1001-7119(2014)11-0095-05

Interesting Mining Algorithm Based on Patterns

Li Junfang
（Jining Normal University，Ulangab，Neimenggu 012000，China）

Abstract: In this paper, I propose interesting mining algorithm - based on patterns. Accessing to the elements through query log,describing user interesting with the concepts in ontology,and calculate interesting score formula,according to the interesting score and dividing interesting sequences into different interesting patterns,the experimental results display that can effective realize interesting mining.

Keywords: interesting mining; pattern; ontology; interesting score

0　引言

用户的搜索兴趣正在向着个性化、分布式、主动式方向发展，用户兴趣模型作为查询服务的基础，依然有很多障碍未解决，用户兴趣挖掘获得，将会对用户查询意图、情境搜索提供研究基础。

关于用户兴趣挖掘的研究主要有两部分：(1)用户兴趣发现，根据用户的查询日志和访问行为，使用特殊的计算形式，判断识别用户兴趣；(2)用户兴趣挖掘的表示，使用本体描述用户兴趣。文献[1]提出一种基于用户反馈标注的概率方法解决兴趣漂移问题，此方法考虑了对实例标注的数计算用户兴趣概率，以标注的形式描述用户兴趣，并用ERWA算法更新用户兴趣，但是，此法局限于概念复合形式描述用户兴趣，缺少用户兴趣之间的语义关联表示。文献[2]提出一种基于摘要获取用户兴趣挖掘的方法，从本体中获取摘要，并通过传播激活过程推导用户兴趣，摘要能够准确地描述兴趣方向。文献[3]基于ODP标注提交搜索返回前10个查询结果、session中点击日志、查询结果点击记录等构建用户兴趣模型，模型实验结果显示，整合多种信息资源兴趣模型更好地理解用户信息需求，但是该方法只限于当前的session资源。文献[4]提出一种获取搜索日志的全局描述，并计算词语之间的语义相似度，根据查询词之间的语义关系构建分类目录，再使用查询聚类算法获取用户兴趣，此算法从语义角度方向增加分类目录中查询之间合理性，但预测用户兴趣不能只靠查询聚类实现。文献[5]根据用户访问行为，基于协同过滤算法推测用户兴趣，用户行为包括：访问停留时间、粘贴复制次数、鼠标滚动次数和是否为添加到收藏夹，但是此方法对短期兴趣有效，对长期兴趣数据分析较为困难。文献[6]提出根据视频网站中用户评价图或者

收稿日期：2014-03-28
作者简介：李俊芳(1978 -)，女，汉族，内蒙古乌兰察布市人，硕士，讲师，研究方向：计算机技术。

评分链获取兴趣模型，由兴趣模型研究用户兴趣的方法。

由上述可知，兴趣挖掘方面已经提出很多研究方法，但是，模型中未曾提及到如何表示用户兴趣，在兴趣表示方面仍然缺乏理论基础，由于兴趣表示是兴趣模型的基础所在，因此，本文给出了进一步研究。

本文在目前兴趣挖掘研究现状的基础之上，提出一种基于用户访问行为，提出一种有效计算兴趣得分公式，并定义了几种兴趣模式和相关标准，使用相关本体构建语义关系图形式的用户兴趣模型，算法的实验达到了很好的效果。

1 获取兴趣粒子

用户的每一个访问序列都对应着兴趣粒子，本文对一个有效的网页进行分析，访问网页序列集 $P=(P_1,P_2,\cdots\cdots P_n)$，抽取该网页的主题向量 $S=(S_1,S_2,\cdots\cdots S_n)$，主题向量由两部分组成：(1)网页关键词；(2)网页分类标签。网页主题向量能表示网页主要内容，但是用主题向量描述用户兴趣存在重复、冗余信息。如：腾讯网发布一则“玉兔月球车已全面苏醒”消息，其关键词包括：“玉兔号、月球车、登月梦、小兔子、苏醒”，网页分类标签是“科技、探月工程、航天”，网页主题向量是“科技、探月工程、航天、玉兔号、月球车、登月梦、小兔子、苏醒”，而对应的用户兴趣“探月工程”就可以藐视。目前，随着科学技术的不断发展，本体的出现及使用解决了这个问题，本文使用本体获取网页主题向量的上位概念集合，就是用户兴趣粒子，下面给出计算兴趣平分模型。

1.1 基本定义

定义1 兴趣粒子 IP：用户兴趣粒子与其访问序列相对应，表示本次访问兴趣方向，用概念集合表示，记为 $C=(C_1,C_2\cdots\cdots C_n)$，当 $i\neq j$ 时，$C_i\neq C_j$，其中概念 C 来自与本体，每一个概念等同于网页标签。

定义2 兴趣停留时间 t：是指用户访问的一个兴趣概念 C 在所有网页停留的时间之和，记为 t，其公式为：

$$t=\sum_{i=1}^{n}P_i \quad (C\in P_i) \tag{1}$$

其中，n 表示为，C 表示兴趣概念 C，P_i 表示为用户访问的第 i 网页。

定义3 兴趣持续时间 T：是指兴趣概念 C 在上一次访问时间到下一次访问时间之差的绝对值，与兴趣停留时间加权之和，记为 T，计算公式为：

$$T(C)=|T_n-T_p|+\mu*t \tag{2}$$

其中，T_n 表示兴趣概念 C 的下一次访问时间，T_p 表示兴趣概念 C 的上一次访问时间，μ 是一个大于1的乘数因子，t 为兴趣停留时间。

定义4 兴趣得分 $S(C)$：以分值形式体现兴趣的强弱程度，兴趣概念 C 的得分由包含概念 C 的网页访问次数、包含概念 C 的网页收藏数和包含概念 C 的兴趣持续时间三者加权之和，以 $S(C)$ 表示，计算公式如下：

$$S(C)=\alpha*V(C)+\beta*F(C)+\lambda*T(C) \tag{3}$$

其中，$V(C)$ 表示访问网页数，即是用户查询日志中获取到序列、$F(C)$ 表示收藏夹数、$T(C)$ 兴趣持续时间，常数 α、β、γ 表示三个影响兴趣得分的重要参数，体现 $V(C)$ 、$F(C)$、$P(C)$ 三者之间的重要程度。

定义5 强兴趣 QI：对于 $\exists C$，若 $S(C)\geqslant\theta_1$，则称强兴趣。

定义6 弱兴趣 WI：对于 $\exists C$，若 $\theta_0\leqslant S(C)<\theta_1$，则称弱兴趣。

定义7 无效兴趣 NI：对于 $\exists C$，若 $S(C)<\theta_0$，则称无效兴趣。

定义8 唯一兴趣 OI：对于 $\exists C$，有且仅有一个 $S(C)\geqslant\theta_1$，则称唯一兴趣。

定义9 多重兴趣 MI：对于 $\exists C$，存在多个兴趣得分 $S(C)\geqslant\theta_1$，则称多重兴趣。

定义10 兴趣漂移 DI：对于 $\exists C$，在两段相继完全相等的时间内，兴趣由强到弱或者由弱到强的过程，称为兴趣漂移。

定义11 长期兴趣 LI：对于 $\exists C$，在两段相继完全相等的时间内，兴趣得分始终有 $S(C)\geqslant\theta_0$，则称为长期兴趣。

定义12 短期兴趣 SI：对于 $\exists C$，在两段相继完全相等的时间内，兴趣得分可能存在 $S(C)\geqslant\theta_0$ 或者 $S(C)<\theta_0$，则称为短期兴趣。

其中，θ_0 和 θ_1 是关于兴趣得分的阈值，假设存在 $\theta_0<\theta_1$。

1.2 获取兴趣粒子

本文获取兴趣粒子算法 get_Interest，其步骤是：一是获取用户访问序列；二是获取网页主题向量；三是基于本体上下文的词义消歧；四是获取

主题向量的上位概念,判断兴趣粒子列表中是否存在,输出访问序列的兴趣粒子。其算法 1 如下:

get_Interest (List P,Ontology O)

Input: P—set of web page,O—Ontology;//访问的网页集合 P 和本体 O

Output: I—list of interest particle;//输出一个兴趣粒子列表 I

1: initialize List S,C;//分别存储主题向量和其上位概念的列表

2: S←Extract Subjcet Vector(P);//获取访问序列网页的主题向量

3: for i←0 to S.size−1

4: C←word sense disambiguate(S_i,Context,O);

//根据主题向量、上下文和本体进行词义消歧

5: for j←0 to I.size−1;

6: if C==I_j then break 3; //如果兴趣列表中存在兴趣概念,则跳到 3;

7: end if

8: end for

9: put C into I;// 向兴趣列表 I 中添加上位概念 C

10: end for

11: return I;

2 基于模式的兴趣挖掘算法

基于模式的兴趣挖掘算法 PIM(patterns interest mining),其主要步骤有:一是获取兴趣粒子;计算兴趣得分;二是判断兴趣模式。其算法流程图 1 如下。

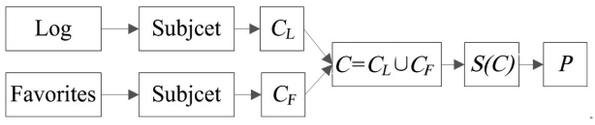

图 1 基于模式的兴趣挖掘算法

Fig.1 Interest mining algorithm based on pattern

其中,Log 和 Favorites 表示日志和收藏夹,subject 表示主题向量,CL、CF 表示日志中兴趣概念和收藏夹中兴趣概念,$S(C)$为兴趣得分,P 为兴趣模式。

2.1 计算兴趣得分

计算兴趣得分是根据算法 1 获取到兴趣粒子进行计算其分值,结合公式 3 计算每一个兴趣概念的得分,网页的访问序列是指用户查询日志中记录的访问序列,分析收藏夹中的 URL 信息,获取收藏夹包含 C 的个数。其 Compute_Score 算法详细描述过程为:

1.分析日志中的访问序列,获取不重复的兴趣概念 C_L 集合;

2.分析收藏夹中的访问序列,获取不重复的兴趣概念 C_F 集合;

3.求兴趣概念并集 $C=C_L\cup C_F$;

4.遍历兴趣概念 C,计算网页停留时间和持续时间;

5.计算兴趣概念 C 的分值 $S(C)$。其算法 Compute_Score(C)如下:

Compute_Score (Page P_1, Page P_2)

Input: P_1— page in log, P_2— page in favorites;//日志中、收藏夹中网页序列

Output: S—list score of interes particle C;//输出兴趣粒子 C 的得分列表 S

1: initialize List C_L,C_F,C,S; Ontology O

//初始化日志序列、收藏夹、全部兴趣概念 C、得分列表,本体 O

2: C_L←get_Interest (P_1,O);//获取访问序列网页的兴趣概念 C_L 及概念数

3: C_V←get_Interest (P_2,O);//获取收藏夹中网页的兴趣概念 C_V 及概念数

4: C←$C_L\cup C_F$;//合并访问序列和收藏夹中兴趣概念为 C 及全部的概念数

5: for i←0 to C.size−1;

6: if (C_i∈Page) //如果兴趣列表中存在兴趣概念,则跳到 3;

7: $t=\Sigma P_i$; $T(C_i)=|T_n-T_p|+\mu*t$; //求网页停留时间和持续时间

8: $V(C_i)$ ←obtain C_i page number in log; //获取访问序列中包含 C_i 的网页数

9: $F(C_i)$ ←obtain C_i page number in favorites; //获取收藏夹包含 C_i 的网页数

10: $S(C_i)=\alpha*V(C_i)+\beta*F(C_i)+\lambda*T(C_i)$;//计算兴趣概念 C_i 的兴趣得分

11: end if

12: put $S(C_i)$ into S;

13:end for

14:return S;

2.2 获取兴趣模式集合

首先,根据算法 2 获取得的兴趣得分;其次,

根据2.1中基本定义，本文研究唯一兴趣模式、多重兴趣模式、兴趣漂移模式、强兴趣模式、弱兴趣模式、长期兴趣模式、短期兴趣模式和无效兴趣模式；最后，将以$S(C)$兴趣得分划分不同的兴趣模式，输出兴趣模式集合P，PIM算法3实现过程如下：

PIM (List $S(C)$)

Input: $S(C)$—list of interest score obtain C ;//包含概念C的兴趣分值list

Output: P—list of interest pattern;//输出兴趣模式集合

1: for i←0 to $S(C)$.size−1

2: switch($S(C_i)$) //判断$S(C_i)$是哪一种兴趣模式

3: case *OI*: flag=1 and put C+" "+flag into P;

4: case *MI*: flag=2 and put C+" "+flag into P;

5: case *DI*: flag=3 and put C+" "+flag into P;

6: case *QI*: flag=4 and put C+" "+flag into P;

7: case *WI*: flag=5 and put C+" "+flag into P;

8: case *LI*: flag=6 and put C+" "+flag into P;

9: case *SI*: flag=7 and put C+" "+flag into P;

10: case *NI*: flag=0 and put C+" "+flag into P;

11: end switch

12: end for

13: return P;

3 实验结果及分析

3.1 实验测试数据

实验数据使用百度开放数据作为测试集，数据中包含了3万多条日志记录，其记录格式为：R={用户id,访问时间t,网页标题T,URL}，研究对象是有效访问序列和收藏夹中网页序列，访问序列有效网页停留时间是指网页打开超过1 min的阈值。

论文中研究获取不同用户id访问数据，关于100用户的有效访问序列的约4000条数据作为训练集，另选取100个用户的约5000条数据作为测试集。

3.2 参数讨论

3.2.1 μ参数

实验选取最小停留时间μ_0，当任意概念停留时间之和$t<\mu_0$时，说明该用户在本次访问的网页停留时间太短，将删除对应兴趣概念。某一个网页停留时间平均值为μ_0=3.5 min，根据这个基本情况可知，实验中设定μ_0=3 min。而公式2中，兴趣的持续时间主要是1周内，超出1周T_n-T_p值远远大于网页停留时间，因此，实验中不断扩大μ_0的值。表1随机选取了兴趣概念T_n-T_p、t值和$V(C)$值$T(C)$的比较（时间单位为min）。

表1 部分兴趣概念对比数据

Table 1 Part interest concept comparison data

兴趣概念	T_n-T_p	t	$V(C)$	$T(C)$
计算机网络	7821	25	4	
Iphone5	2211	63	15	
百度	560	18	16	
打印机驱动	1211	26	2	
新浪	1509	102	11	
集宁师范学院	1800	66	6	
java算法	2451	80	20	
搜索引擎	5214	14	14	
数据结构	7992	9	5	
腾讯微博	6300	17	13	

由表1可见，“兴趣概念”的访问次数最多，相应的t值最大；实际中，“搜索引擎”是实验获取的兴趣概念，但是，其兴趣是“计算机”。

3.2.2 α、β、λ参数

在公式2中，如果收藏网页数$F(C)\neq0$，则对应的兴趣概念是强兴趣，α参数的取值是决定了兴趣的强弱，但其与参数β、λ无关。

网页访问次数$V(C)$就是相应概念C的出现次数，如果$V(C)$值太小，则兴趣概念C不能构成兴趣，这里将兴趣概念C的最小访问次数记为V_{min}，当$V(C)<V_{min}$，则兴趣概念C是无效兴趣概念。

公式2中$T(C)$的值会远大于$V(C)$，有实验可知，访问次数对应兴趣的识别有很大影响，但是，短时间的访问次数增加不会使兴趣的体现，因此，α、β的取值不行保证$V(C)$的重要性，也要保证$T(C)$的重要程度。α的取值为1，β的取值为100，三者具有相同的数量级。

3.2.3 V_0、V_1参数

由统计分析得出，大多数用户在一段时间内存在多个兴趣，阈值V_0、V_1的取值的作用是区分兴趣强弱。概念兴趣得分分布情况可知，兴趣概念分值高的较少，绝大多数都是分值较大的。V_0的取值目的是排除非兴趣概念，V_1的取值的作用是获取强兴趣。下表中计算兴趣概念的准确率

和召回率为V_0、V_1的取值依据。表2中的V_0、V_1的取值依据,时间单位为min。

表2 V_0、V_1的取值依据/min

Table 2 V_0、V_1 value/min

序号	V_0/min	P/%	R/%	V_1/min	P/%	R/%
1	500	80.1	68.3	4500	84.9	66.7
2	600	80.6	65.1	4600	82.4	61.7
3	700	82.3	66.6	4700	81.1	62.2
4	800	83.1	60.8	4800	84.1	61.6
5	900	80.1	69.1	4900	80.5	66.0
6	1000	86.7	70.2	5000	88.7	71.3
7	1100	80.9	67.1	5100	70.9	67.5
8	1200	86.1	68.2	5200	76.1	64.2
9	1300	82.5	63.1	5300	72.6	66.1
10	1400	84.0	65.5	5400	81.0	55.5

实验中随着V_0值的增加,召回率R也随之不断提高,而若兴趣被划分到非兴趣中,反而会降低了弱兴趣的召回率,分析可得实验中将平衡准确率P和召回率R之间的大小关系,并将设置为V_0=1000最佳。V_1的取值与相似,其目的是获取分值较高的强兴趣,也依据兴趣得分的准确率和召回率获取,将设置为V_0=5000最佳。

3.2.4 *评价方法*

为了验证试验结果的准确性,本文针对兴趣挖掘数据,定义了本文体外测评方法,准确率P、召回率R的计算公式:$P=N_p/N$和$R=N_p/N_{all}$。

其中,N为本文算法获取兴趣概念数,N_p为在兴趣概念N中正确概念数,将由统计分析得出,N_{all}为测试数据总数。

3.3　实验结果分析

实验测试使用用户日志中作为测试集,使用的参数分别是前文讨论过的,研究获取不同用户id访问数据,关于选取100个用户的约5000条数据作为测试集,测试集合中计算兴趣概念 得分、兴趣强弱个数、兴趣类型和前10个兴趣概念集合。表3兴趣概念类型及数量。

表3 兴趣概念类型和数量

Table 3 Interest concept type and number

类型	QI	WI	NI	OI	MI	DI	LI	SI
结果	1240	660	88	42	250	680	1100	1400

4　结论与展望

本文提出一种基于模式的兴趣挖掘算法。首先,论文中根据本体中概念、定义可形式化表示用户兴趣,并给出多种关于兴趣的基本概念,用户兴趣概念存在于访问日志、收藏夹和访问序列中;其次,根据户访问序列获取兴趣粒子;再次,基于本体上下文特征词的词义消歧;最后,计算兴趣得分,获取兴趣主题向量的上位概念,判断兴趣粒子列表中是否存在,输出访问序列的兴趣粒子采用一种兴趣得分值的形式将划分不同的兴趣模式,输入兴趣模式集合。本文算法通过实验分析验证分析可知,获取实验数据结果很理想。

下一步的研究方向和研究内容是获取增量形式的长期兴趣研究上,这种兴趣挖掘有利于短期兴趣挖掘实现,改进兴趣得分计算公式,能够有效地计算更加准确的长期兴趣。设计出更加合理的兴趣挖掘算法,对目前的兴趣挖掘算法能有所改善。

参考文献:

[1] Mendoza M, Zamora J. Building Decision Trees to Identify the Intent of a User Query[C]//KES,2009,285-292.

[2] Mendoza M, Zamora J. Identifying the Intent of a User Query Using Support Vector Machines[C]//SPIRE, 2009, 131-142.

[3] Conde J M, Vallet D, Castells P. Inferring User Intent in Web Search by Exploiting Social Annotations[C]//SIGIR, 2010, 827-828.

[4] Yan J, Zheng Z Y, Jiang L, Castells P. A Co-learning Framework for Learning User Search Intents from Rule-Generated Training Data[C]//SIGIR, 2010, 895-896.

[5] Cheng Z C, Gao B, Liu T Y. Actively Predicting Diverse Search Intent from User Browsing Behaviors[C]//WWW, 2010, 221-230.

[6] Cao H H, Chen E H, Yang J, Xiong H. Enhancing Recommender Systems Under Volatile User Interest Drifts[C]//CIKM, 2009, 1257-1266.