基于本体与模式的网络用户兴趣挖掘

苏雪阳,左万利,王俊华
(1.吉林大学计算机科学与技术学院,吉林长春 130012;2.符号计算与知识工程教育部重点实验室(吉林大学),吉林长春 130012)

摘 要: 本文探讨了用户兴趣挖掘的新方法,首先从用户搜索日志中获取访问行为元素,并借助通用本体中的概念描述网页所体现的用户个体兴趣,然后提出了一种兴趣得分计算方法,并在此基础上从用户个体兴趣序列中识别不同的兴趣模式,判断用户的短期兴趣,并利用通用本体得出用户兴趣的集合表示,最后根据短期兴趣的增量积累推算长期兴趣.整个过程避开了以往兴趣挖掘方法中通过相似度计算和文档聚类算法进行兴趣合并的问题,为兴趣发现提供了新思路.实验结果表明,本文的方法对用户兴趣的描述更具体,取得了更优化的兴趣合并结果.

关键词: 搜索引擎; 用户兴趣; 通用本体; 兴趣模式
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2014)08-1556-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.08.015

Web User Interest Mining Based on Ontology and Patterns

SU Xue-yang, ZUO Wan-li, WANG Jun-hua
(1. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University) Ministry of Education, Changchun, Jilin* 130012, *China*)

Abstract: A novel user interest mining method is proposed. Firstly, the items of visiting behaviors are retrieved from user's search engine log, and individual user interests with every webpage are described through the concepts of common ontology. Then, a method for computing the score of interest is proposed. According to the scores, a user's interest list can be judged as different interest patterns, which can be used to find the user's short term interests. After that, a user's interest model is built with concept collection extracted from ontology. At last, based on incremental accumulation of short term interests, long term interest collection can be calculated. The whole procedure avoids the problem of using similarity computation and document clustering to merge concepts in existing interest mining methods. This paper explores a new way of thinking. And as the experiment shows, the proposed method provides a more concrete description of user interest model and obtains an optimized concepts merging result.

Key words: search engine; user interest; ntology; interest pattern

1 引言

互联网的服务模式正在逐渐向主动式和个性化等方面演进,个性化服务要求对用户的上网模式有充分的理解.搜索引擎的用户兴趣模型作为搜索个性化服务的基础,很多问题尚未很好解决,仍是研究热点.用户兴趣建模可为进一步探讨用户搜索意图、用户搜索情境等工作提供重要依据.

目前,关于用户兴趣挖掘的研究主要涉及两方面:(1)用户兴趣发现,根据用户搜索历史和访问行为,订立标准判断用户搜索兴趣;(2)用户兴趣表达,通过领域本体或开放目录提取类目,建立用户兴趣树或个性本体.Pin Zhang 等[1]提出基于用户反馈标注的概率方法处理兴趣漂移,通过实例标注数计算用户偏好的概率,以标注的概念集合描述用户兴趣并增量更新,其局限性在于缺乏兴趣概念之间语义关系.Ryen W 等[2]利用 ODP 标注查询结果、当前 session 搜索结果的点击记录及相应网页等资源,构建短期兴趣模型,很好地理解和模拟用户的信息需求,但当前 session 资源的预测范围有限.Lyes Limam 等[3]抽取查询日志的全局描述,根据查询词的语义关系组织分类目录,最后利用查询词聚类算法获得用户兴趣.该文从语义角度出发提高了分类目录中对象间联系的合理性,但仅依赖查询词聚类不够可靠.Michal Holub 等[4]根据用户访问行为,包括网页驻留时

收稿日期:2013-06-04;修回日期:2013-08-12;责任编辑:马兰英
基金项目:国家自然科学基金(No.60973040);国家自然科学青年基金(No.61300148);吉林省重点科技攻关项目(No.20130206051GX);吉林省科技发展计划青年基金(No.20130522112JH)

间、鼠标滚动次数和拷贝次数等,使用协同过滤预测用户兴趣,观察角度较全面,但对长期行为数据统计量化存在困难.Federica Cena 等[5]以本体作为用户概要的基础,从少量初始概念出发,通过其在本体中的祖先或子孙的关联路径到达其他相关领域获得兴趣,对确定用户兴趣的层级关系问题具有启发性.Bin Tan 等[6]基于语言模型和聚类,研究了长期兴趣和探索性兴趣的搜索模式.以会话数度量长期兴趣,以网页点击度量探索性兴趣,其标准值得参考.在文献[7]中,作者根据观众评分行为提出评分图和评分链,划分兴趣模式探求用户兴趣.综上,虽然在获取用户兴趣方面已探索出一些可行方法,但用户兴趣表示还缺乏恰当的理论模型.由于兴趣表示是兴趣使用的基础,因而仍有待深入系统研究.本文在现有研究基础上,根据用户访问行为,提出兴趣得分计算,定义了四种用户兴趣模式和划分标准,然后对应通用本体的片段表达短期兴趣倾向,并由短期兴趣增量推算长期兴趣,达到了良好的效果.

2　短期兴趣评分计算模型

用户兴趣分为长期兴趣和短期兴趣.本文从短期兴趣着手,建立评分计算模型.并根据短期兴趣的增量累积过程推衍长期兴趣.模型涉及两个主要问题.

(1)兴趣的基本单位　在用户搜索日志中,每个网页都是描述用户兴趣的基元,一般方法会通过提取网页的正文关键词、自带分类标签等形成描述网页内容的向量或主题,再通过计算语义距离等对网页分类或聚类,最终根据类别信息构建兴趣树或用户个性本体等兴趣模型.但此类方法的类别标签规模很大,并且用户自主性会造成相似标签冗余[8],语义计算繁琐.本文以通用本体作为网页描述和兴趣提取的标签来源,为用户兴趣识别提供了新思路.在通用本体中,语义相似度高的词汇很可能属于同义词集、上位概念相同或具有一定关联.将网页与兴趣的类别信息与通用本体中的概念相对应更利于分类处理和表示.因此,本文提出求取每个网页向量的上位概念集,作为描述兴趣的基本单位——兴趣原子.

(2)兴趣评分　为评价兴趣的强弱程度,本文将从三方面针对时效性、稳定性较高的短期兴趣进行评分:(1)在日志中反应特定兴趣的网页持续时间,包括页面停留时间之和与时间跨度;(2)相应网页在日志中的出现次数;(3)相应网页被收藏的次数.结合文献 4 及相关资料,用户基本行为可通过以上三方面描述.

2.1　基本概念

定义 1　用户兴趣原子　它是与所访问的网页相对应的上位概念集,记为 item($c_1, c_2, \cdots, c_n$),当 $i \neq j$ 时 $c_i \neq c_j$.其中,概念 $c_1, c_2, \cdots, c_n$ 来自于通用本体.

定义 2　兴趣的网页停留时间　用户在反应兴趣概念 c 的网页上停留时间总和.记为 Q,公式为:

$$Q = \sum_{k=1}^{N} q_k \tag{1}$$

其中,q_k 表示用户在反应 c 的网页 k 上的停留时间.N 为参与解析的网页总数.记 t_e 和 t_l 分别表示首尾两个反应 c 的网页时间戳,则 c 的时间跨度为 $t_l - t_e$.

定义 3　兴趣的持续时间　兴趣概念 c 的持续时间,是其时间跨度与相应网页停留时间的加权和,记为 T,公式如下:　$T(c) = t_1 - t_e + m \cdot Q$　(2)

定义 4　兴趣得分　兴趣概念 c 的得分,由兴趣持续时间、对应网页访问次数、对应网页收藏个数三者加权求和得出.记为 $S(c)$,公式如下:

$$S(c) = \lambda \cdot T(c) + \mu \cdot F(c) + \rho \cdot P(c) \tag{3}$$

其中,$F(c)$为 c 对应的网页出现的次数;$P(c)$为收藏夹中对应网页的个数.常数 λ、μ、ρ 为影响因数.

定义 5　强兴趣与弱兴趣　对于兴趣概念 c,若 $S(c) \geqslant V_1$,则为强兴趣;若 $V_0 \leqslant S(c) \leqslant V_1$,则为弱兴趣.其中,$V_0$ 和 V_1 是关于兴趣得分 S 的阀值,由实验获得.若 $S(c) < V_0$,则为非兴趣.

2.2　兴趣得分的计算

(1)获取兴趣原子

由于词语歧义性,网页向量在本体中可能有多个截然不同的上位概念,需要进行词义消歧去除无关概念.本文将采用已有的经典消歧算法,并结合上下文采用已有词法分析算法进行优化[9].

获取兴趣原子的算法 Capture_Item(Page p)综上可得.解析输入的网页 p,得出描述向量 $W(w_1, w_2, \cdots, w_n)$,将分量 w_i 提交至通用本体 O 中,结合上下文 Context 词义消歧,获得对应的上位概念,包括著名人物和地名等特定名称;判断去重后将符合条件的概念放入集合 Item$\{\cdots\}$.

算法 1　Capture_Item

```
Input: Page p;   Ontology O;
Output: Concept_Set Item;
1   Context←Extract(p);//提取网页正文
2   W←Analyze(p);//分析网页向量
3   for i←0 to W.size - 1
4       c←Disambiguate(w_i, Context, O);//有消歧算法的获得上位概念函数
5         for j←0 to Item.size - 1
6             if c = = Item_j then break;
7             end if
8         end for
9         if j = = Item.size then Put c to Item;
10          end if
11    end for
```

12 return Item;

(2)计算兴趣得分

从日志中提取某用户的访问行为序列 $B = B_1 \cdots B_n$,对应兴趣原子序列 Item = $\text{Item}_1 \cdots \text{Item}_n$. 兴趣原子 $Item_j$ 包含的兴趣概念为$(c_{j1}, c_{j2}, \cdots, c_{jk})$,经算法 1,序列 Item 扩为兴趣概念序列 $C = c_{11}\, c_{12} \cdots c_{1h} c_{21} \cdots c_{n1} \cdots c_{nk}$,如图 1. j 值相同的 c_{ji} 是由同一兴趣原子扩展所得,故它们在序列中实际位置标号相同,即 j,彼此顺序不固定.

计算兴趣得分并判断强弱的过程 Rate_Interest:输入网页 $p\{\cdots\}$,获取原子序列 item$\{\cdots\}$;分解 item_j,替换为概念列表$(c_{j1}, c_{j2}, \cdots, c_{jk})$,构成序列 C;分析计算概念得分并判断强弱. C'是辅助序列,C 中值相同的元素在 C'中对应为同一元素.

算法 2 Rate_Interest

Input: every Page p_j in the log;
Output: $T(c)$ and $S(c)$ of every concept c, List C';
1 for $j \leftarrow 0$ to p.length − 1
2 $\text{item}_j \leftarrow$ Capture_Item(p_j);
3 for $k \leftarrow 0$ to item_j.length − 1
4 $c_{jk} \leftarrow \text{item}_j$.getItem($k$);
5 Copy c_{jk} into List C;
6 end for
7 end for
8 for every concept C_x in List C
9 find all concepts $c = = C_x$;
10 $Q \leftarrow \Sigma q_k$; //Σq_k 为 C_x 对应的网页停留时间和
11 $T(C_x) \leftarrow (t_l - t_e + m \cdot Q)$;
12 $S(C_x) \leftarrow \lambda \cdot T(C_x) + \mu \cdot F(C_x) + \rho \cdot P(C_x)$;
13 Put C_x into C' without any repeater;
14 end for
15 if $S(C_x) >= V_1$ then flag = STRONG;
16 else if $V_1 > S(C_x) >= V_0$ then flag = WEAK;
17 else flag = NONE;
18 end if

3 基于模式的短期用户兴趣挖掘算法

3.1 相关定义

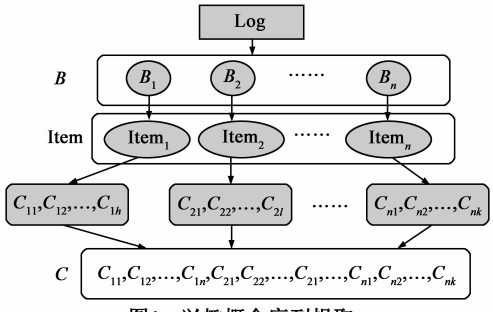

图1 兴趣概念序列提取

经算法 1、2,可算出序列 C 中元素得分,根据得分分布划分兴趣模式,决定最终用户兴趣. 结果为概念集合 $U = \{c_1, c_2, \cdots, c_n\}$. 具体流程如图 2.

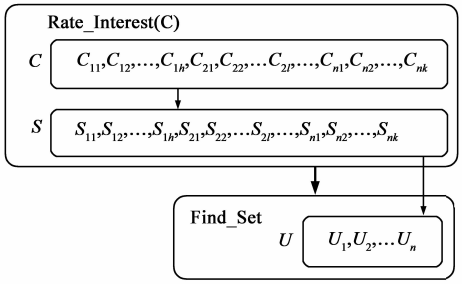

图2 兴趣评分与模式判断

本文参考文献[7],定义四种兴趣模式. 在序列 C 中:**唯一兴趣模式**(SI)有且仅有一个强兴趣;**多重兴趣模式**(MI)存在多个强兴趣;**兴趣漂移模式**(DI)无强兴趣,且弱兴趣占半数以上;**无效兴趣模式**(NI)无强兴趣,且弱兴趣占半数以下.

定义 6 兴趣跨段. c 在序列 C 中始末两位间位置段称作 c 的跨段,记为 $cover(c) = \{c.L_s, c.L_s + 1, \cdots, c.L_e\}$, $c.L_s$ 和 $c.L_e$ 为 c 的始末两位编号. 若 c_1、c_2 的跨段存在重叠段,则称在重叠段内 c_1、c_2 同时出现.

3.2 基于模式的用户兴趣挖掘

(1)唯一兴趣模式 用户在时段内只有唯一的强兴趣 c,显然 c 即代表了用户的当前兴趣.

(2)多重兴趣模式 用户在时段内同时具有多个强兴趣. 可能同时还具有一些弱兴趣,但此时弱兴趣不足以代表现阶段用户的兴趣特征.

(3)兴趣漂移模式 用户在时段内无强兴趣,关注重点随时间不断变化,此时只有最新关注的兴趣才代表当前阶段的兴趣特征.

本文借鉴文献[7]提出处理用户最新兴趣的方法. 从 C 中提出弱兴趣序列 $C* = \{c \mid c \in C \text{ and } V_0 \leqslant S(c) < V_1\}$,保留在 C 中的位置编号. 在 cover(c)范围内,概念 c 的出现是稀疏的. 函数 $y = H(L)$ 表示随着序列位置 L 的变化(本质即时间的变化)同时出现的弱兴趣的个数. 因为弱兴趣的持续时间较短,随着兴趣的生灭,$H(L)$的图像会剧烈振荡,出现峰值点,如图 3(a)所示. 峰值点后同时出现的兴趣数量减少,旧兴趣消失,新关注点尚未形成兴趣. 因此每个峰值点都代表一次兴趣漂移. 最末峰值点之后存留的概念正是用户的最新兴趣,亦即所需结果. 图 3(b)为 MI 模式的 $H(L)$图像,*MI* 下大量兴趣的持续时间较长,跨段重叠较大,概念数量变化不大,没有明显峰值点出现.

check_DI(List C)为在 DI 下求最末峰值点位置的过程,记为 L' = check_DI(C),由于函数 $H(L)$是离散

的,且客观上不存在解析式,故 L'位置的确定只能依赖图像判断,如图 3(a)的点 a.取 $U=\{c \mid c\in C*$ and $c.L_e \geqslant L'\}$,认为 $L_e < L'$的兴趣过期无效,不加入 U 中.

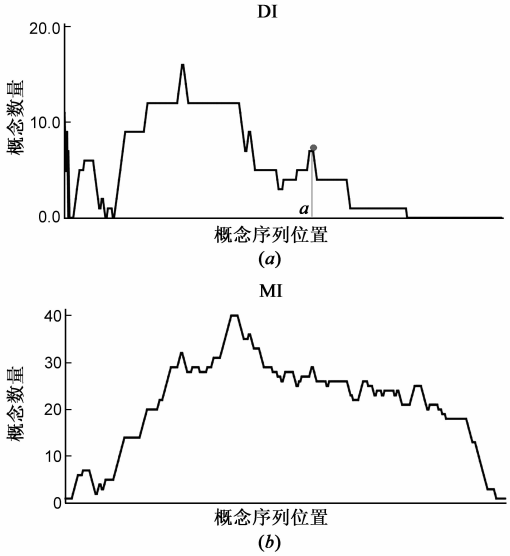

图3　兴趣个数随时间变化图像

(4)无效兴趣模式　用户在时段内无强兴趣,访问松散随机,或活动较少,不足以判断兴趣.

(5)算法 Find_Set　在求取兴趣结果集 U 的算法 Find_Set 中,定义模式变量 pat,输入为无重复元素的辅助序列 C'.结合上述分析列出四条判断规则:规则 1 若 $pat=SI$,则 $U=\{c \mid S(c)>=V_1\}$;规则 2 若 $pat=MI$,则 $U=\{c \mid S(c)>=V_1\}$;规则 3 若 $pat=$ DI,则 $U=\{c \mid c$ in $C*$ and $c.L_e>=L'\}$;规则 4 若 $pat=$ NI,则 $U=$ NONE.据此可得 Find_Set 具体过程.$C*$ 为弱兴趣序列.

算法 3　Find_Set
Input: List C', $C*$; Pattern pat;
Output: Concept_Set U;
1　switch(pat)
2　case SI: $U=\{c \mid S(c)>=V_1$ and c in $C'\}$;
3　case MI: $U=\{c \mid S(c)>=V_1$ and c in $C'\}$;
4　case DI: $L'=$ check_DI(C');
5　$U=\{c \mid c$ in $C*$ and $c.L_e>=L'\}$;
6　case NI: $U=$ none;
7　end switch
8　return U;

3.3　短期兴趣集合表示

短期兴趣集合的来源即为 U.利用通用本体检查兴趣概念集 U 中元素之间的语义关系和上下位词汇链关系,得出 $U*=u_1\cup u_2\cdots\cup u_n$,$U*$ 包含 U,且各子集间交集为空.每个子集即为时段内一个兴趣倾向.每一倾向的实质对应一个连通的本体片段.

U 中不同元素 c_x 与 c_y 间在本体中若存在以下任一关系,则称其存在相近语义关系,记作 close_to(c_x, c_y).(1)c_x 与 c_y 具有直接或间接上下位关系;(2)c_x 与 c_y 拥有共同直接上位概念;(3)c_x 与 c_y 存在共同直接下位概念;(4)c_x 与 c_y 有同义或相似关系.

逐一检查 U 中元素,将具有 close_to 关系的元素并入同一子集,将成立的 close_to 关系判定过程中涉及的共同上下位概念、间接上下位概念等衍生概念亦并入相应子集,最终得出划分 $U*=u_1\cup u_2\cdots\cup u_n$.图 4 为简单样例.

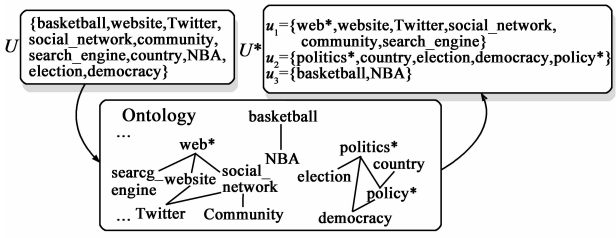

图4　根据本体片段划分兴趣集例子

该过程中,U 扩展成为了 $U*$.因 close_to 关系所涉及的元素都并入了同一子集,所以 $U*$ 各子集间交集为空.这种通过本体关联寻找概念间联系的方式,等同于以往研究中(如文献[10])对相似标签进行聚类合并的过程,而其中繁琐的兴趣标签相似度计算、聚类计算则被规避掉了.

4　长期兴趣集合的求取与表示

用户长期兴趣的时间跨度大,数据组成复杂,长期搜索日志难以直接处理.故将短期兴趣作为长期兴趣集合的子单元.用户长期兴趣集合 U_L 的得出是包括两个阶段的循环增量过程:先由当前短期兴趣集合 U 与已训练得出的初始长期兴趣集合 U_L'增量叠加得出中间结果 U_{L0},再筛选 U_{L0}中的元素确定当前时期的 U_L,该结果将成为下一时期的初始长期兴趣.

4.1　叠加规则

U 与 U_L'叠加需记录元素重复次数,并修改 U_L'划分的子集内容.集合叠加遵循以下三个规则:

规则 1　U 与 U_L'中存在相同元素,则将该元素所在的子集合并,该元素的计数加 1.

规则 2　对于 U 与 U_L'中无相同元素的集合,若 U 中某元素与 U_L'中某元素存在 close_to 关系,则将其所在子集合并,添加相应的衍生概念,在 close_to 关系中处于上位的概念元素计数值加 1;U 与 U_L'子集内元素间可能存在一对多或多对一的 close_to 关系,应将涉及的子集全部合并.

规则 3　无相同元素或 close_to 关系的子集不变,继续作为新集合的子集.

4.2 筛选过程

短期兴趣产生中间结果 U_{L0}. 设 U_{L0} 中元素 c_i 叠加次数为 α_i, α_0 为全体 α_i 的期望值, 次数 $\alpha_i < \alpha_0$ 的元素并非用户长期关注的目标, 将从 U_{L0} 中删去, 得出最终的长期兴趣集合 U_L. U_L 作为下一次叠加的初始集合, 已有的叠加次数将参与下一次计算.

对兴趣类别概括性越高的概念被衍生所得的可能性越大, 叠加次数必然越高, 较具体的实体性的低位概念叠加次数必然较低, 最终形成概括全面、低冗余的兴趣描述. 取期望值作筛选标准, 可对一时兴起的情况可以有效分辨去除; 不再关注的兴趣概念叠加次数不再增加, 长期兴趣漂移亦可由此得出.

5 实验与结果分析

5.1 实验数据集

实验采用某商业搜索引擎的开放数据作为用户搜索日志的数据来源, 包含 83566 条日志记录, 以 < 用户 *id*, 访问时间, title, URL > 的形式存放. 本文筛去导航页、网站主页、登陆页等无正文内容的网页, 以日志中 50 个不同 *id* 用户的 64273 条可用于文本处理的有效访问行为作为计算对象. 40 个用户的 48263 条记录作为训练集, 另 10 个用户的 13216 条记录作为测试集. 以维基百科知识库数据作为参考本体, 采用现有方法从维基百科中提取本体特征[11,12], 将条目和分类关系对应为本体中的概念关系, 用于实验.

用户短期兴趣以一周为评判周期. 在日志中, 网页停留时间超过 1min 为有效访问, 停留时间过短的网页则为广告或偶然访问, 不具研究价值.

5.2 参数讨论

(1)参数 m

设最小停留时间 m_0, 当停留时间之和 $\Sigma q_k < m_0$ 时是无效访问. 通过统计, 网页停留时间平均约为 $avg_m = 3$min. 若 c 有可能成为弱兴趣, 则客观上需 $F(c) > 3$: 有起始和结束两次使 $\text{cover}(c) \neq \varnothing$, 且跨段内有一次以上记录. 设最少出现次数 $F_0 = 4$, 则 $m_0 = 12$min. 在公式(2)中, 兴趣发现主要以 $t_l - t_e$ 值为依据, 但在一周范围内, $t_l - t_e$ 值远大于网页停留时间之和 $Q = \Sigma q_k$, 致使判断偏差. 因此需扩大倍数 m 进行修正. 表 1 随机列举了若干兴趣概念的 $t_l - t_e$ 值和 Q 的比较. 由表 1 可见, 用户对 Search_engine 有关网页的点击次数 $F(c)$ 最多, Q 值也较大, 但其 $t_l - t_e$ 值却与 Basketball 等在绝对数值上相差很大, 系统会误认为体育方面是主要兴趣.

若以 min 为单位, $T(c)$ 值将远大于 $F(c)$. $F(c)$ 在数值上影响太小, 但增减趋势与 Q 值相对应, 因此二者均应参与同 $t_l - t_e$ 间的平衡性比较. 为修正绝对数值偏差, m 需满足:

$$m \cdot (Q + F(c) \cdot avg\_m) \sim t_1 - t_e \quad (4)$$

$$m \cdot (Q + F(c) \cdot avg\_m) < t_1 - t_e \quad (5)$$

其中“~”号表示两端数量级相同, 以平衡参数影响力, “<”关系保证了 $t_l - t_e$ 的主导地位, $F(c) \cdot avg\_m$ 使单位统一. 式(4)两端越接近, 式(3)中各参数影响越显著. 图 5 为从训练集中随机截取的 100 个兴趣概念对应的式(4)值曲线在不同 m 值下的对比. 其中最粗曲线为 $t_l - t_e$, 几乎与水平轴重合的粗线为 $Q + F(c) \cdot avg\_m$. 其他为 $m \cdot (Q + F(c) \cdot avg\_m)$ 在 m 各典型值下的曲线, 其中 $m = 85$ 曲线与 $t_l - t_e$ 最接近, 此时式(4)两端影响程度对比最合理. 实验表明, 当 $m > 85$ 时, 不满足式(5)的区间明显增多, 不合要求. 实验抽取 12 组兴趣概念集, 每组 1000 个概念, 对各组结果求期望, 最终取 $m = 82$.

表 1 部分概念的基本数据(单位 min)

CONCEPT	Q	$t_l - t_e$	$F(c)$	CONCEPT	Q	$t_l - t_e$	$F(c)$	CONCEPT	Q	$t_l - t_e$	$F(c)$
Network	53	7827	11	Basketball	23	6530	6	Computer_science	25	5466	7
Athletic	32	1764	10	Scheme	35	6045	15	Person	16	4726	5
Search_engine	34	2451	12	Operating_system	19	3128	6	Company	23	3002	9

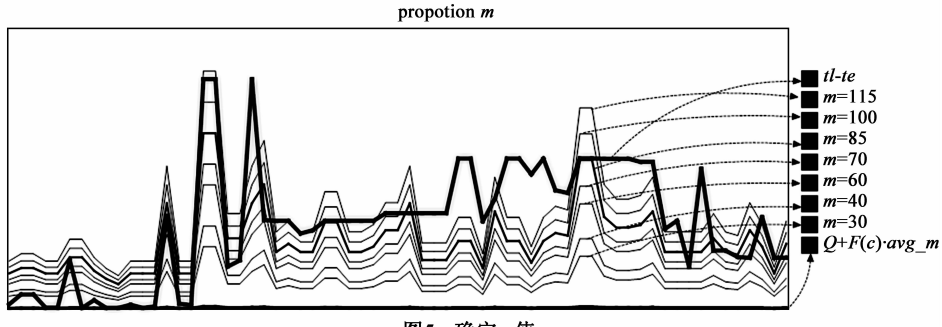

图5 确定m值

(2)参数 λ、μ 与 ρ

在公式(3)中,只要收藏网页个数 $P(c)\neq 0$,则认为 c 一定是强兴趣,ρ 值只需能分辨相应兴趣概念是否为强兴趣,与 λ、μ 无关,故令 $\rho = V_1$.对于 $\lambda\cdot T(c)+\mu\cdot F(c)$部分,由实际经验可知,点击次数对兴趣的判断影响很大,$F(c)$应具有一定影响力,但短期内的大量点击存在偶然性,并非兴趣体现,因此需兼顾以 $T(c)$为主导.取 $\lambda = 1$,数值上希望满足 $T(c)\sim\mu\cdot F(c)$.对比表 1 数据,易见可以初步估计 $\mu = 100$.

对 $T(c)$值很大的概念 μ 影响较小.$T(c)$与 $F(c)$均足够大的兴趣很可能为强兴趣,通过设定 V_1 即可保证强兴趣的查准率,μ 不影响其区分.由图 5 可见 $T(c)$值很低的概念 Q 值往往也很低,点击次数 $F(c)$从趋势上也较低,$S(c)\approx\mu\cdot F(c)$,可通过设定 V_0 来辨别.突发兴趣概念 $T(c)$较小、Q 较大($T(c)\approx Q$)且 $F(c)$也足够大,μ 的取值可能影响其取舍.因此可能归为弱兴趣的概念查准率受 μ 值影响较大.取 DI 模式下的查准率 P_{DI}作为 μ 的取值参考.如表 2:

表 2　μ 对 P_{DI}的影响

μ	P_{DI}	μ	P_{DI}	μ	P_{DI}
40	57.3%	90	61.2%	110	51.4%
60	59.1%	100	61.8%	120	60.3%
80	60.6%	105	62.1%	140	60.4%

由表 2 可见约在 $\mu = 105$ 处 P_{DI}最高.当 μ 偏小时,F 值和 T 值都较小的弱兴趣得分太低,无法提取;当 μ 偏大时,F 值较高但 T 值较低的弱兴趣会得高分而成为强兴趣,导致 P_{DI}下降.图 6 是从训练集兴趣得分统计图中随机截取的 250 个有效不重复概念得分分布,未包含网页收藏评分,参数取 $\lambda = 1$,$\mu = 105$,$m = 82$.

(3)参数 V_0 与 V_1

经统计,62% 的用户同时存在多个兴趣,其中仅个别兴趣较强烈.图 6 中概念得分呈现明显区分,数量分布状况符合客观实际与实验预期.综合多组概念得分柱状图,设定不同阀值,结果集查准率情况如下:(a)V_0 用于排除非兴趣,因此以非兴趣查准率 P_{false}和召回率 R_{false}作参考.如表 3(a).随着 V_0 值增大,更多非兴趣被筛选出,R_{false}不断升高,同时得分较低的弱兴趣概念也会被归为非兴趣,使 P_{false}降低,权衡查准率与召回率,设定 $V_0 = 1200$;(b)V_1 用于提取强兴趣,因此取值依据强兴趣查准率 P_s 和召回率 R_s 确定.如表 3(b).随着 V_1 的增大,P_s 必然增大,同时也会排除得分稍低的强兴趣,从而降低了 R_s.由表 3(b)可见,取 $V_1 = 4500$ 较为合理.

综上,实验最终训练参数为:$m = 82$,$m_0 = 12$,$F_0 = 4$,$\lambda = 1$,$\mu = 105$,$\rho = 4500$,$V_0 = 1200$,$V_1 = 4500$.

表 3

(a)根据 P_{false}确定 V_0

V_0	P_{false}	R_{false}
1100	84.7%	82.9%
1200	84.7%	85.3%
1300	81.5%	86.2%
1400	80.9%	86.2%
1500	78.7%	87.1%
1600	78.4%	87.6%
1700	78.2%	87.6%
1800	77.6%	87.6%
1900	76.8%	87.8%

(b)根据 P_s 确定 V_1

V_1	P_s	R_s
4100	56.4%	83.8%
4200	60.6%	83.3%
4300	66.0%	81.7%
4400	72.5%	81.7%
4500	77.4%	80.5%
4600	77.4%	79.9%
4700	80.2%	75.4%
4800	82.1%	71.3%
4900	82.1%	68.8%

5.3　实验结果分析

测试输入日志的测试集部分,参数取训练值.以下选取某典型的 MI 模式实验用户的兴趣结果集为例.图 7 为其一周内兴趣,参数 $m = 82$,$m_0 = 12$,$F_0 = 4$,$\lambda = 1$,$\mu = 105$,$\rho = 4500$,$V_0 = 1200$,$V_1 = 4500$.$U*$ 中带有 * 号的概念均为通过 close_to 关系衍生所得.社交网站与 java 方面的兴趣领域差距较大,当采用传统分类器方法时,二者的网页或主题相关度较低,分入了不同类别.但本例中通过 web_service、computer_science 等概念相互关联,二者进入同一划分,"社交网站相关的 java 编程"的网页会得到更高排名,更符合用户实际需求.图 8 为该用户 2 个月的长期兴趣叠加面临的历史 U_L'和短期兴趣 $U*$.篇幅所限,未列出全部结果.本次叠加 $\alpha_0 = 3.3$,即次数 $\alpha_i \geqslant 4$ 的元素将留在 U_L 中.括号中为当前叠加次数.

图 9 为叠加和筛选后得出的长期兴趣,粗体为 $U*$ 所造成的改变.其中 basketball、java 等与相同元素合并计数增加.U_L'中的 movie 与 $U*$ 中的 cinema,movie_ticket 与 ticket_price 均存在 close_to 关系,所在集合合并,movie 和 movie_ticket 计数增加.在 U_L 中:大量具体概念如 Yahoo、Oracle 被删去;突发性的 music 方面兴趣被去除;分支兴趣也得到筛选,如 movie 方面喜剧和科幻比悲剧计数更高得以保留,客观反应了用户偏好.在政治方面,用户零散关注的子兴趣都被筛去,概括性的高位概念 politics 得以保留.

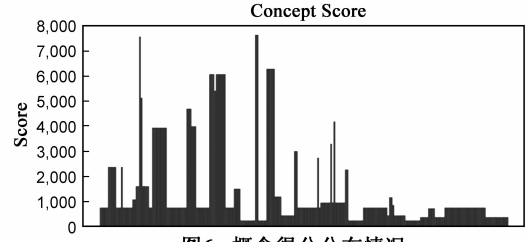

图6　概念得分分布情况

U={basketball,Olympics,sports_strategy,Apache_software_foundation,Java,linux,server,
MySQL,website,Twitter,Facebook,Iphone,mobile-phone,social-network,user,community,Ipad,
google,company,browser,game,web_page,Obama,politics,search_engine,election,country,
economics,NBA,three-point,corporation,program,government,democracy}

*U**={{basketball,three-point,basketball_terminology*,ballgame*,NBA,
basketball_leagues*,game,sports_strategy,sport*,Olympics},
{Apache_software_foundation,program,Java,program_language,linux,operating_system*,
software_platform*,server,web_service*,MySQL,database*,website,WWW*,Twitter,Facebook,
social_network,community,network*,computer_science*,web_page,google,company,search_engine,
web_browser,corporation,Apple_company*,Iphone,Ipad,smartphone*,cobile_operating_system*},
{Obama,election,politics,country,USA*,economics,democracy}}

图7 短期兴趣集合划分

U_L'={{basketball(7),ballgame(9),NBA(5),basketball_leagues(6),game(10),sports_strategy(4),sport(10),
Olympics(1),small_forward(2),basketball_position(3),three-point(1),Air_Jordan(1),rebound(1),
basketball_team(4),basketball_player(3),basketball_terminology(3),FIBA(1),shooting_guard(1),
sport_shoes(1),FIFA(1),football(2),football_leagues(1),football_positon(1)...},
{Apache_software_foundation(2),program(7),Java(6),program_language(7),linux(3),operating_system(7),
Python(2),software_platform(9),server(4),web_service(5),MySQL(2),database(5),website(6),WWW(6),
Twitter(1),Facebook(2),Yahoo(1),social_network(5),community(4),network(7),computer_science(11),
web_page(3),google(2),search_engine(6),web_browser(1),corporation(3),Apple_company(5),Iphone(1),
Ipad(2),smartphone(1),mobile_operating_system(2),Object-oriented(6),hadoop(1),open_source(2),mahout(2),
data_structure(1),program_tool(2),Oracle(1),Web_application_frameworks(7),vim(1),recommend_system(1),
machine_learning(1),Internet_search_algorithms(1),algorithms(2),search_engine_optimizaiton(1)...},
{Obama(1),election(1),politics(3),country(1),USA(1),economics(2),democracy(1),stock_market(1),
finance(1),economics_terminology(1),banks(1)},
{fashion(3),design_firm(1),clothing_brands(2),suit(1),workwear(1)},
{movie(5),Hollywood(4),movie_ticket(2),movie_star(5),director(1),sciencefiction(4),
comedy(4),tragedy(2),film_box_office(2),film_festival(1)},
...}

U*={{music,classic_music,rock_band,singer,musicalinstruments,music_ranklist},
{dunk,basketball,basketball_terminology*,shooting_guard,NBA,basketball_player*,DPOY,sport*},
{java,web_service*,Websphere,j2ee,Apache_software_foundation,program_tool,computer_science*,
machine_learning,algorithms},
{cinema,ticket_price}}

图8 已有长期兴趣与当前短期兴趣

U_{L0}={{basketball(***8***),ballgame(9),NBA(***6***),basketball_leagues(6),game(10),sports_strategy(4),sport(***11***),
Olympics(1),small_forward(2),basketball_position(3),three-point(1),Air_Jordan(1),rebound(1),
basketball_team(4),basketball_player(***4***),basketball_terminology(***4***),FIBA(1),shooting_guard(***2***),
dunk(1),***DPOY(1)***,sport_shoes(1),FIFA(1),football(2),football_leagues(1),football_positon(1)...},
{Apache_software_foundation(***3***),program(7),Java(***7***),program_language(7),linux(3),operating_system(7),
Python(2),software_platform(9),server(4),web_service(***6***),MySQL(2),database(5),website(6),WWW(6),
Twitter(1),Facebook(2),Yahoo(1),social_network(5),community(4),network(7),computer_science(***12***),
web_page(3),google(2),search_engine(6),web_browser(1),corporation(3),Apple_company(5),Iphone(1),
Ipad(2),smartphone(1),mobile_operating_system(2),Object-oriented(6),hadoop(1),open_source(2),mahout(2),
data_structure(1),program_tool(***3***),Oracle(1),Web_application_frameworks(7),vim(1),recommend_system(1),
machine_learning(***2***),Internet_search_algorithms(1),algorithms(***3***),search_engine_optimizaiton(1),
j2ee(1),***Websphere(1)***...},
{Obama(1),election(1),politics(4),country(1),USA(1),economics(2),democracy(1),stock_market(1),
finance(1),economics_terminology(1),banks(1)},
{fashion(3),design_firm(1),clothing_brands(2),suit(1),workwear(1)},
{***cinema(1)***,movie(***6***),Hollywood(4),***ticket_price(1)***,movie_ticket(***3***),movie_star(5),director(1),
sciencefiction(4),comedy(4),tragedy(2),film_box_office(2),film_festival(1)},
{music(1),classic_music(1),rock_band(1),singer(1),musicalinstruments(1),music_ranklist(1)},
...}

U_L={{basketball(8),ballgame(9),NBA(6),basketball_leagues(6),game(10),sports_strategy(4),sport(11),
basketball_team(4),basketball_player(4),basketball_terminology(4),...},
{program(7),Java(7),program_language(7),operating_system(7),software_platform(9),server(4),WWW(6),
web_service(6),database(5),website(6),social_network(5),community(4),network(7),computer_science(12),
search_engine(6),Apple_company(5),Object-oriented(6),Web_application_frameworks(7),...},
{politics(4)},
{movie(6),Hollywood(4),movie_star(5),sciencefiction(4),comedy(4)},
...}

图9 叠加与筛选

6 总结

本文通过网页主题向量在通用本体中的上位概念来表达网页访问所体现的用户个别兴趣,提出兴趣原子的概念和以兴趣得分衡量兴趣强弱.定义了四种兴趣模式求取时段内的兴趣概念集.最后借助本体将概

念集划分,表达了用户兴趣的聚类与合并过程.

本文的研究思路探索了一条新途径.首先,以通用本体中的概念代替 ODP 标签或大众分类标签表达用户兴趣,避免了标签规模问题的负面影响.其次,用户兴趣的最终表现形式不再是简单的向量,而是从通用本体映射出的具有语义层次关系的本体片段,兴趣的表示更准确,去除了冗余,避开了以兴趣标签相似度计算合并兴趣的过程.最后,相比于以往的单一数学模型,本文通过不同的兴趣模式分别得出兴趣,更加准确可靠.后续研究中,利用短期用户兴趣的长期兴趣增量学习方法,改进叠加方式以计算更准确可靠的长期兴趣将成为主要目标.

参考文献

[1] Pin Zhang, et al. A probabilistic approach for mining drifting user interests[A]. APWeb/WAIM '09 Proceedings of the Joint International Conferences on Advances in Data and Web Management[C]. Berlin, Heidelberg, 2009. 381 – 391.

[2] Ryen W. White, Paul N. Bennett, Susan T. Dumais. Predicting short-term interests using activity-based search context[A]. CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management[C]. New York, USA, 2010. 1009 – 1018.

[3] Lyes Limam, et al. Extracting user interests from search query logs: A clustering approach[A]. DEXA Workshops' 10 Proceedings of 2010 Workshop on Database and Expert Systems Applications[C]. Bilbao, Spain, 2010. 5 – 9.

[4] Michal Holub, Mária Bieliková. Estimation of user interest in visited web page[A]. WWW '10 Proceedings of the 19th international conference on World wide web[C]. New York, USA, 2010. 1111 – 1112.

[5] Federica Cena, et al. Propagating user interests in ontology-based user model[A]. AI * IA 2011: Artificial Intelligence Around Man and Beyond Lecture Notes in Computer Science[C]. Palermo, Italy, 2011. 299 – 311.

[6] Bin Tan, et al. Mining long-lasting exploratory user interests from search history[A]. CIKM '12 Proceedings of the 19th ACM international conference on Information and knowledge management[C]. New York, USA, 2012. 1477 – 1481.

[7] Huanhuan Cao, Enhong Chen, Jie Yang, Hui Xiong. Enhancing recommender systems under volatile user interest drifts[A]. CIKM '09 Proceedings of the 19th ACM international conference on Information and knowledge management[C]. New York, USA , 2009. 1257 – 1266.

[8] 张引,张斌,高克宁,等.面向自主意识的标签个性化推荐方法研究[J].电子学报,2012,40(12):2353 – 2359.
Zhang Yin, Zhang Bin, Gao Kening. Autonomy oriented personalized tag recommendation[J]. Acta Electronica Sinica, 2012, 40(12): 2353 – 2359. (in Chinese)

[9] 刘秀磊,廖建新,等.本体匹配中基于词义组合的词法分析算法[J].电子学报,2012,40(8):1624 – 1630.
Liu Xiulei, Liao Jianxin. Lexical analysis based on combining senses in ontology matching[J]. Acta Electronica Sinica, 2012, 40(8): 1624 – 1630. (in Chinese)

[10] 王李冬,魏宝刚,袁杰.基于概率主题模型的文档聚类[J].电子学报,2012,40(11):2346 – 230.
Wang Lidong, Wei Baogang, Yuan Jie. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinica, 2012, 40(11): 2346 – 2350. (in Chinese)

[11] Cui G Y, et al. Corpus exploitation from Wikipedia for ontology construction[A]. ELRA Proceedings of the Sixth International Language Resources and Evaluation (LREC08)[C]. Marrakech, Morocco, 2008. 2125 – 2132.

[12] Shirakawa M, Nakayama K, Hara K, et al. Concept vector extraction from Wikipedia category network[A]. Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (2009)[C]. New York, USA, 2009. 71 – 79.

作者简介

苏雪阳　男,1989 年 2 月出生,江苏连云港,2011 年至今于吉林大学计算机学院攻读硕士学位,从事 Web 数据挖掘、自然语言处理和搜索引擎有关研究.
E-mail: suxueyang2011@gmail.com

左万利(通信作者)　男,1957 年 12 月出生,吉林吉林人,博士,现为吉林大学计算机科学与技术学院教授、博士生导师,ACM 职业会员,从事数据库、Web 智能、网络搜索引擎、自然语言处理等有关研究.
E-mail: wanli@jlu.edu.cn

王俊华　女,1982 年 3 月出生,山东菏泽人,2010 年至今于吉林大学计算机学院攻读博士学位,从事 Web 数据挖掘和自然语言处理和搜索引擎有关研究.
E-mail: wangjunhua_1982@126.com