基于人工免疫算法的增量式用户兴趣挖掘

左万利[1,2] 韩佳育[1] 刘 露[1,3] 王 英[1,2] 彭 涛[1,2,3]
(吉林大学计算机科学与技术学院 长春 130012)[1]
(吉林大学符号计算与知识工程教育部重点实验室 长春 130012)[2]
(伊利诺伊大学厄巴纳-香槟分校计算机科学系 厄巴纳-香槟)[3]

摘 要 了解用户兴趣是为用户提供个性化服务的关键。用户兴趣有短期兴趣和长期兴趣之分,且具有不稳定性。受人工免疫系统的启发,巧妙地将免疫应答过程应用于用户兴趣挖掘。首先将概率与时间相结合,提出"概念时序动态"的概念,以更好地刻画用户在一段时间内对同一兴趣的关注程度;然后基于人工免疫原理,建立抽取兴趣标签的分类器来提取用户兴趣标签;最后针对增量式学习,建立兴趣标签的"概念时序动态",刻画出用户兴趣自首次出现以来受关注的程度,以此为依据来判断兴趣是否存在迁移及遗忘现象,并为每个兴趣标签附上权重。其主要贡献是创造性地将人工免疫原理应用于用户短期兴趣和长期兴趣的挖掘,并具有增量特性,可以很好地体现用户兴趣迁移特征,是一种自然完整的用户兴趣模型。实验结果表明,该学习模型能够很好地发现用户关注的领域,其平均精度和召回率分别达到79.5%和74.4%,是目前最贴近用户的兴趣挖掘模型。

关键词 短期兴趣,长期兴趣,兴趣遗忘,兴趣迁移,概念时序动态,增量学习,人工免疫系统
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.007

Incremental User Interest Mining Based on Artificial Immune Algorithm

ZUO Wan-li[1,2] HAN Jia-yu[1] LIU Lu[1,3] WANG Ying[1,2] PENG Tao[1,2,3]
(College of Computer Science and Technology, Jilin University, Changchun 130012, China)[1]
(Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changchun 130012, China)[2]
(Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA)[3]

Abstract Understanding user interests is the key to personalize service. User's interests can be categorized into short-term interests and long-term interests, and may evolve over time. Inspired by artificial immune system (AIS), we skillfully employed immune response process in user interests mining. By combining probability with time, we first introduced the concept temporal dynamics to describe the degree that the user pays attention to a particular interest during a specific time interval. Then, based on AIS, we built classifier to extract user interest tags. Finally, directing at incrementally learning user interest, we built concept temporal dynamics for interest tags, which characterize the attention degree since interests first appears, and on this basis judged whether interests have migrated or being forgotten and assigned weights to each interest tag. The main contribution of this paper is that we creatively applied AIS to user short-term interests and long-term interests mining with incremental feature, which can gracefully reflect the migration of user's interests. It is a natural and complete model for learning user interests. Experimental result on practical dataset indicates that the proposed learning model can effectively discover topics that user focuses on, with the average precision and recall of 79.5% and 74.4% respectively, which is the most suitable user interest mining model.

Keywords Short-term interests, Long-term interests, Interest-forgotten, Interest migration, Concept temporal dynamics, Incremental mining, Artificial immune system

随着互联网的高速发展,网页数量呈指数级增长,用户对搜索引擎检索质量的要求也日益提高。用户兴趣可以刻画出用户经常关注、感兴趣的领域。根据用户过去浏览行为建立和抽取用户兴趣模型是为用户提供个性化服务的基础,建立

到稿日期:2014-06-15 返修日期:2014-10-15 本文受国家自然科学基金项目(60903098,60973040),国家自然科学青年基金项目(61300148),吉林省重点科技攻关项目(20130206051GX)资助。

左万利(1957—),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为数据挖掘、Web挖掘、机器学习、自然语言处理、信息检索、社会化计算,E-mail:wanli@jlu.edu.cn;韩佳育(1989—),女,硕士生,CCF学生会员,主要研究方向为数据挖掘、Web挖掘、信息检索;刘 露(1989—),女,硕士生,主要研究方向为数据挖掘、Web挖掘、信息检索;王 英(1981—),女,博士,讲师,主要研究方向为Web信息检索及挖掘、本体;彭 涛(1977—),男,博士,副教授,硕士生导师,主要研究方向为数据挖掘、Web挖掘、机器学习、自然语言处理、信息检索;E-mail:tpeng@jlu.edu.cn(通信作者)。

能准确刻画用户兴趣特点的理论模型是问题的关键。

为了有效获得用户在日常浏览网页过程中所体现的兴趣，首先要确定兴趣来源。人们已经发现，用户的浏览日志中含有许多有价值的信息，且具有如下特点：(1)网页数量多；(2)更新速度快；(3)隐式地蕴含用户兴趣，是 Web 挖掘的重要对象。因此本文以用户浏览日志为数据集，提取用户兴趣。在兴趣表示方面，本文参考 ODP(Open Directory Project)类别[1)]，利用类别中的层次关系，建立 3 种用于表达用户兴趣的本体，在此基础上加入体现用户偏好的一些特征词以及用户对兴趣的关注程度，最终得到用户兴趣模型。

用户兴趣从时间上可以分为长期兴趣和短期兴趣：长期兴趣描述用户在一段较长的时间内对某些领域的关注，通常与用户的爱好、专业或从事的职业有关；短期兴趣体现用户在短时间内关注的内容，是对短期检索需求的反映，有时也与当前的热门话题有关。通过分析观察，用户兴趣具有以下几个特点：(1)长期兴趣是对短期兴趣的持续关注，由短期兴趣积累而来；(2)短期兴趣可能会随着时间的推移被遗忘；(3)在学习兴趣的过程中可能存在假兴趣，即被学习模型错误地作为兴趣。因此在提取兴趣的过程中，需要考虑到用户兴趣存在动态变化和遗忘的特点。受生物学免疫系统思想的启发，提出结合人工免疫系统的方法和过程来研究如何构建一种在保证兴趣提取准确性的同时保持实时性和高效性的增量式用户兴趣学习模型，其框架流程如图 1 所示。

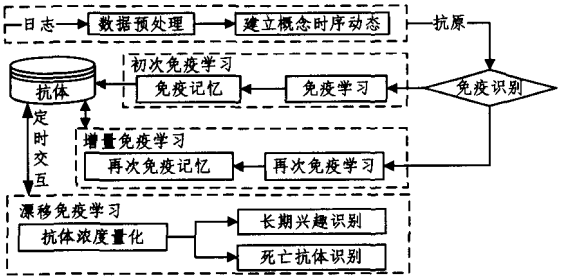


图 1　兴趣学习模型

本文第 1 节中主要介绍了有关用户兴趣挖掘以及人工免疫系统的相关工作；第 2 节描述了如何有效地提取用户兴趣标签，以及利用人工免疫的相关过程抽取用户兴趣模型；第 3 节通过实验对本文提出的方法进行验证，并对结果进行说明；最后对全文进行总结。

1　相关工作

近几年，对于用户兴趣的研究已逐渐成为 Web 挖掘领域的热点。林鸿飞[1]等人将文本段落作为识别用户兴趣的基本要素，在聚类的基础上，建立用户兴趣模型。LyesLimam 等人[2]通过计算词语之间的语义距离，根据查询词的语义关系组织成分类目录，利用查询词聚类算法获得用户兴趣，利用语义使得分类目录中对象间的联系更为合理，但在推测用户兴趣时仅用查询词聚类可靠性不够，同时缺乏实时性。Zhang Pin 等人[3]在解决用户兴趣漂移的问题上提出一种基于数值化用户反馈标注的概率方法，即通过考虑实例标注的数量计算用户喜欢某实例的概率，以概念集合数值特征描述用户兴趣，并基于 ERWA 算法增量更新用户的兴趣。此方法的局限性在于缺乏兴趣概念之间的语义关系表示。文献[4]提出了一种含隐私保护的门户个性化兴趣获取机制，实现了不同兴趣访问行为的隐式获取以及操作语义分析，并且给出了带有门户个性化兴趣描述的隐 Markov 模型扩展。文献[5]将当前查询、当前查询所处 session 的点击记录以及网页访问等资源结合起来构建短期兴趣模型。结果显示，该方法构建的兴趣模型能更好地理解和模拟用户的信息需求，但是所使用的情境信息仅来自于当前查询所处的 session，情境信息利用不够充分，同时对 session 划分的准确度有很强的依赖性。Michal Holub 等人[6]依据用户访问网页的行为预测用户兴趣，其中用户行为包括网页驻留时间、鼠标滚动次数和拷贝到剪切板上的次数等，观察角度全面，可以很好地预测短期行为，但对长期行为进行统计量化存在一定困难。Federica Cena 等人[7]以本体作为用户概要的基础，从少量的初始概念出发，通过其在本体中的父类或子类的关联路径到达其他相关领域概念从而获得兴趣，对确定本体中用户兴趣与用户行为的层级关系问题有很大的启发性。Bin Tan 等人[8]对用户长期兴趣和探索性兴趣的搜索模式进行深入研究，以日志中符合此兴趣的会话数作为长期兴趣的度量，在判断探索性兴趣时，以网页点击作为度量。

在兴趣表示方面，文献[9]利用 ODP 的主题词定义用户文件，通过计算检索时返回的网页覆盖的主题词与用户文件中主题词之间的距离来对网页进行额外的排名，从而对用户进行智能推荐。实验表明，该方法在为用户提供智能化搜索时有一定的帮助。Seulgi So 等人[10]在 ODP 已有类别的基础上，利用动词对存在于 ODP 中的名词进行扩展，将用户兴趣表示为(noun，verb)的形式。这种表示方法能够使 ODP 类别更加丰富，同时在实验过程中取得了比较好的结果，对本文有一定的启发。

在基于人工免疫系统的应用方面，Liu Tao 等人[11]提出一种基于人工免疫系统的新型聚类算法，该算法将数据作为抗原，相应的聚类中心点作为抗体，不仅可以避免局部最优，同时还可以加快收敛速度。对本文有一定影响的是基于有限人工免疫系统的免疫分类器 AIRS(Artificial Immune Recognition System)，文献[12]对 AIRS 进行介绍，并在模拟数据集上进行实验以说明算法的基本性能。文献[13]介绍的基于免疫算法的监督学习方法和文献[14]介绍的基于免疫的语义分类对本文工作很有启发，Anderw Watkins 等人[13]在基础算法上对免疫模型进行了修正，去除了免疫过程中不必要且较为复杂的环节，实验对两个版本的算法进行讨论，数据表明修改的算法不但没有牺牲算法的精度，而且还提高了数据精炼的能力。Julie Greensmith 和 Steve Cayzer[14]的动机来源于利用用户的语义结构减轻网页导航的压力，从而将基于 AIRS 的分类方法应用于复杂的多类别分类问题上，分类信息可以应用到领域本体映射上，对主动搜索有着深层次的影响。

通过上述分析发现，目前用户兴趣挖掘方法分别针对短期兴趣和长期兴趣建模，没有形成统一的框架模型，未将兴趣

1) http://www.dmoz.org/World/Chinese_Simplified/

迁移特征融入整体模型，每隔一段时间需重新对数据进行处理，不能有效地进行增量学习，效率低下。因此本文综合考虑用户兴趣特征，提出一种以人工免疫系统为基础的增量式学习方法，对短期兴趣和长期兴趣统一建模，以更有效地提取用户兴趣。

2 增量式用户兴趣挖掘

人工免疫系统的生物原型是生物免疫系统。从计算的角度来看，生物免疫系统是一个高度并行、分布、自适应和自组织的系统，具有很强的学习、识别、记忆、特征提取和遗忘等能力。受生物免疫系统的启发，产生了人工免疫系统(Artificial Immune System，AIS)。免疫系统通过学习和记忆的方式对抗原进行应答，实质上是处理抗原所包含信息的过程。免疫系统是一种高效、强大的信息处理系统，已发展成为计算智能研究的一个崭新的分支，已被应用于数据挖掘、数据聚类和数据分析等领域。免疫系统的核心是通过一些机制生成抗体及变换抗体对抗原进行分布式处理，具有很好的鲁棒性、错误耐受性和自适应性。

2.1 数据预处理

用户的浏览历史以 URL 序列的形式记录在 log 日志中，首先对网页进行简单的数据预处理，主要包括 session 划分、网页解析和特征词提取 3 个部分。合理的 session 划分可以更好地将时间和概率的关系体现出来，从而能提高兴趣分析的准确度。考虑用户兴趣需要一定时间的积累且更新频率不需要太快，本文指定一个 session 为用户在一天内访问的网页集合，用 $Day_i=\{P_1,P_2,\cdots,P_n\}$ 表示。通过网页解析过滤掉网页中的广告、图片、导航链接及其他与正文无关的噪音信息，以提高特征词提取的准确度。特征词提取是数据预处理的核心，特征词能够准确精炼地表明网页的内容，通过对网页正文进行分词，计算每个词的 TF-IDF[15] 的权值，提取网页的特征词。

2.2 利用免疫监督学习模型建立兴趣抽取分类器

通过分析用户兴趣的特点，在描述用户兴趣时需要包含两部分内容，一部分能够表明用户兴趣大的类别(如“旅游”、“体育”、“健康”)，另一部分能够进一步详细地说明用户具体兴趣点(如“马尔代夫”、“NBA”)。为了满足以上需求，本文采用 ODP 类别表示对用户兴趣进行描述。

ODP 是一种开放式目录搜索系统，是目前互联网上最大的人工编制的分类检索系统，其内容会定期进行更新。ODP 的类别是一种树形结构，它充分考虑了信息检索的环境以及用户检索的需求和特点，因此可以很好地描述领域中概念之间的关系。ODP 的中文部分包含 15 个顶级类别，延伸到叶子节点一共包含 5082 个子类别，这些类别可以比较充分地描述用户兴趣大的方向。依据 ODP 类别中概念之间的关系构建一个能够描述用户兴趣大方向的本体，由于 ODP 类别数量大，只选择一个局部来说明兴趣本体的结构，图 2 是以计算机为顶级类别构建的局部本体。在兴趣本体中，每一个节点都对应一个兴趣标签，通过箭头的连接，确定概念间上下位关系，由有向边将相关概念连接起来，构建出的本体类似于树的结构，可以体现出概念之间的关系。例如，节点“电子邮箱”表示为兴趣标签“计算机/互联网络/电子邮箱”的形式，其中“计算机”是顶级类别，“互联网络”是次级类别，“电子邮箱”为子类别，用这种方式能够比较细致地描述用户的兴趣取向。

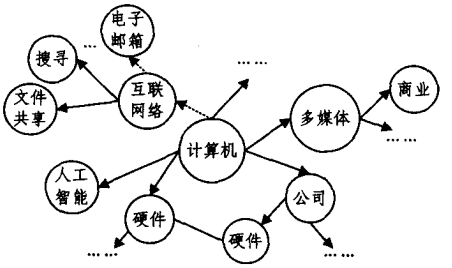


图 2 局部兴趣本体

为了获得用户兴趣标签，本文建立了一种免疫监督学习模型[16]。免疫监督学习模型是为模拟免疫系统学习记忆能力而建立的一种分类器系统，待分类网页对应于抗原，按照一定机制(如资源分配、亲和力成熟、克隆选择等)产生记忆细胞(ARB、抗体、记忆细胞等)，最终将成熟记忆细胞用作分类器，最后结合 K-NN 方法[17] 完成分类任务。用 $Interest\_tag=f(Train\_Ag,MC,\Theta)$ 表示免疫监督模型，其中，$Interest\_tag$ 表示用户兴趣标签；$Train\_Ag$ 是分类对象，对应抗原；MC 是经过免疫机制处理后得到的对 $Train\_Ag$ 的记忆细胞，反映 $Train\_Ag$ 数据特征；Θ 表示获得 MC 所采用的免疫学习过程。

基于传统的免疫过程扩展出来的免疫学习模型有多种形式，其中 AIRS 是一种鲁棒性好、效率较高、数据压缩能力强，同时保持较好分类性能的方法。利用 AIRS 免疫分类器可将用户浏览的网页映射到兴趣本体中的相应节点。AIRS 分类算法通过衡量抗原与抗体之间的亲和力，对抗原进行分类。算法 1 描述了如何构建 AIRS 分类器，算法的核心在于利用克隆机制以及资源控制机制产生对待分析数据的记忆数据，结合 K-NN 方法，对数据进行分类。

算法 1 构建分类器

```
Algorithm  Interest_tag extraction
Input:{Train_Ag},MC
Output:Interest_tag classifier
1. foreach(ag in Train_Ag)
2. do
3.   max_stim←0;
4.   foreach(mc in MC)
5.   do
6.     stim←Stimulation(ag,mc)
7.     if(stim>max_stim)
8.     max_stim←stim
9.   done
10.  if(stim>threshold )
11.    mc_match←mc
12.    update mc. source
13.    ARB=ARB∪ARB_generation(mc_match,stimulation)
14.    else
15.    mc_match←ag
16.    MC←MC∪{ag}
17.    update mc_match. source
```

```
18.     ARB=ARB∪ARB_generation(mc_match, stimulation)
19.   fi
20. done
21. allocate resources according the stimulation;
22. remove abs which do not have adequate resources,
23. sort ARB, evaluate items in ARB, determine if the item join in MC
24. cut←Aff_Thres×para //para is pre-defined
25. foreach(ab in ARB, mc in MC)
26. do
27.   if(ab. stim>mc. stim)
28.     MC←MC∪{ab}
29.   fi
30.   if(affinity(ab, mc)<cut)
31.     remove mc from MC
32.   fi
33. done
34. return Interest_tag classifier
```

定义 1 训练抗原 $Train\_Ag=(t_1,t_2,\cdots,t_n;C)$，其中 t_j 表示网页经预处理之后提取的文本特征项，C 为网页对应到兴趣本体中的节点表示。

定义 2 抗体 $Ab=(t_1,t_2,\cdots,t_m;C)$，其中 t_j 为表示网页的文本特征项，C 为抗体对应的兴趣标签。$ARB_i=\{Ab_1,Ab_2,\cdots,Ab_n\}$，$ARB=ARB_1\cup ARB_2\cup\cdots\cup ARB_n$，$ARB_i$ 表示所有类别为 i 的抗体集合，ARB 表示所有抗体的集合。

定义 3 记忆细胞 $mc=(t_1,t_2,\cdots,t_m;C;source)$，其中 t_j 为表示网页的文本特征项，C 为记忆细胞对应的兴趣标签，$source$ 表示记忆细胞所占有效资源数。$MC_i=\{mc_1,mc_2,\cdots,mc_n\}$，$MC=MC_1\cup MC_2\cup\cdots\cup MC_n$，$MC_i$ 表示所有类别为 i 的记忆细胞集合，MC 表示所有记忆细胞的集合。记忆细胞是分类的依据。

算法初始化阶段首先将训练数据在[0,1]之间正规化，计算抗原之间的欧几里得距离[1]，欧几里得距离是一种最基本的度量。对数据正规化之后，计算亲和力阈值 Aff_Thres，亲和力阈值表示的是所有训练集合中抗原的平均亲和力。最后利用抗原数据对抗体及记忆细胞进行简单的随机初始化。

初始化完成后，对抗原数据逐一处理，对每个 $Train\_Ag$ 用 $Stimulation(x,y)$ 计算抗原与抗体之间的刺激值，刺激值越大，表示抗体与抗原结合得越充分，越符合要求。受刺激最大的记忆细胞被选中，标记为 mc_{match}；若不存在符合要求的记忆细胞，用 $Train\_Ag$ 作为 mc_{match}，并对其进行克隆变异，产生 N 个抗体，加入到 ARB 中，具体过程见算法 2 和算法 3。算法事先规定一个资源总量 S，基于刺激值对抗体分配有限资源，没有资源的 $ARBs$ 从记忆细胞池中清除。当平均刺激满足事先定义的刺激阈值时，ARB 停止克隆。通过该过程的循环执行，得到分类所需的抗体集。最后结合 K-NN 方法，将待分析的兴趣抗原映射到兴趣标签上。

$$Aff\_Thres=\frac{\sum_{i=1}^{n}\sum_{j=n+1}^{n}Affinity(Train\_Ag_i,Train\_Ag_j)}{n(n-1)/2} \tag{1}$$

式中，n 表示训练抗原的数目，$Affinity(x,y)$ 表示正规化后两个训练抗原特征向量之间的欧几里得距离。

$$Stimulation(x,y)=1-Affinity(x,y) \tag{2}$$

$$N=Stimulation\times Clone\times Hypermutation \tag{3}$$

式中，$Clone$ 表示抗体的克隆率，$Hypermutaion$ 表示变异率。

算法 2 克隆抗体

```
Algorithm ARB_generation
Input: mc, stimulation
Output: ARB_i
1. MU←Ø
2. MU←MU∪makeARB(mc)
3. N←stimulation×Clone×Hypermutation
4. While(|MU|<N)
5. do
6.   flag←false
7.   mc_clone←mc
8.   mc_clone←mutate(mc_clone, flag)
9.   if(flag==true)
10.     MU←MU∪makeARB(mc_clone)
11.  fi
12. done
13. ARB_i←ARB_i∪MU
14. returnARB_i
```

算法 3 利用交叉机制进行变异

```
Algorithm: mutate
Input: mc, flag, MC_i
Output: ARB_i
1. foreach(MC_i in MC)
2. do
3.   if(i==mc. c)
4.     flag←true
5.   fi
6.   foreach(mc_1 in MC_i)
7.   do change←random() // the function of random is generating
     some random places in the feature of mc, mc_1
8.     exchange mc, mc_1 in the specified location
9.     add new items to ARB_i
10.    calculate average mutation
11.    if(average mutation>mutation_rate)
12.      stop
13.    fi
14.  done
15. done
16. if(flag==true)
17.   return ARB_i
18. fi
```

2.3 利用免疫过程增量学习用户兴趣

人工免疫系统的核心是免疫应答过程，包括免疫识别、免疫学习、免疫记忆等子过程。免疫系统信息处理模式如图 3 所示。免疫系统主体通过模式识别、学习、记忆来探测、感知、识别“外部”对象的内在信息，通过多种巧妙的信息融合方法

[1] http://zh.wikipedia.org/wiki/

对“外部”对象产生认知，区分出自己和非己，最后根据认知的结果产生免疫应答[16]。免疫应答的实质是一个识别、学习和记忆的过程。学习兴趣的过程实质也是一个识别、学习和记忆的过程，因此将人工免疫系统应用到学习兴趣的过程中是十分合理的。

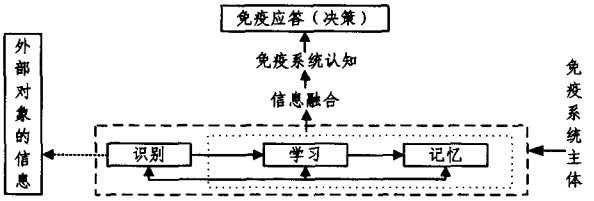


图3 免疫信息处理模式

2.3.1 免疫识别

免疫识别的本质是利用模式识别区分自己和非己。这一过程与用户兴趣挖掘中判断一个兴趣标签是否已存在于用户的兴趣集中相吻合。

用户浏览的网页经过预处理之后，通过特征提取，网页内容将由特征向量 $P_i=\{w_1,w_2,\cdots,w_n\}$ 来表示。利用 *Interest_tag* 分类器将每个网页映射到一个兴趣标签 $term_i$ 上，得到兴趣抗原 $Antigen=(w_1,w_2,\cdots,w_m;term_i)$，利用式(2)计算抗原对抗体集中抗体的刺激值 *Stimulation*，对抗原进行免疫识别，若刺激值超过阈值 k，表明免疫系统曾经受过相似抗原的刺激，抗体集中存在与此类抗原相匹配的抗体，对抗原的应答属于二次免疫应答，否则属于初次免疫应答，然后依据应答方式对抗原进行学习。

2.3.2 免疫学习

免疫学习贯穿于整个免疫应答过程，当免疫系统对一个抗原进行初次免疫应答时，免疫系统会利用克隆变异的方式产生大量的抗体，对抗原进行特征提取、分析和学习。那些与抗原亲和力高的抗体会被保留下来，当相似的抗原再次“入侵”免疫系统时，免疫系统根据记忆信息可以快速识别抗原，自动触发二次免疫应答。下面详细说明如何准确提取加工抗原中所包含的信息。

当前大多依据兴趣标签出现的频度衡量一个兴趣标签是否是用户兴趣，如文献[18]分析了用户一段时间的查询检索情况，但仅将用户在这一个行为周期内对网页的点击映射到一个频率上，不能充分描述用户的行为。因此为了更好地描述用户对一领域的关注程度与时间之间的关系，根据免疫系统本身的特点，本文引入概念时序动态。

定义 4（概念时序动态） 将概率与时间相结合用于描述用户在某行为周期内对特定领域的关注趋势，将其定义为关注程度序列：$D(c)=\langle P^{session_1},P^{session_2},\cdots,P^{session_n}\rangle$，其中，$c$ 表示兴趣标签，$P^{session}$ 表示用户在第 i 个 session 对兴趣 c 的关注程度，利用式(4)对其进行计算，$t_1,t_2,\cdots,t_n$ 为行为周期内连续的离散时间点序列。概念时序动态的特征是一旦兴趣 c 被用户关注，在随后的 m 个 session 内将持续对其关注，若在之后的某个 $session_j$ 用户没有关注兴趣 c，则 $P^{session_j}=0$。

$$P^{session_i}=\frac{tf(term,Day_i)}{\sum_{k=1}^{m}tf(term_k,Day_i)} \tag{4}$$

式中，$tf(term,Day_i)$ 表示在第 i 天浏览的网页集 Day_i 中，兴趣标签 *term* 出现的频率。

对经过免疫识别得到的抗原进行进一步特征提取，针对兴趣标签 $term_i$ 建立概念时序动态，得到与抗原亲和度较高的抗体 *Antibody*（定义 4）。若是初次免疫应答，则有 $D(term_i)=(P(term_i))$；若是二次应答，则免疫系统存在兴趣标签 $term_i$ 的概念时序动态 $D^{\sim}(term_i)$，因此更新其概念时序动态，加入新 session 的概率即可，$D(term_i)=(D^{\sim}(term_i),P(term_i))$。

通过对兴趣的特点进行分析，对兴趣设定 3 个层次，分别是用户兴趣、潜在兴趣和假兴趣。其中用户兴趣是能为用户个性化服务提供依据的真正兴趣；潜在兴趣是需要系统继续观察的一些兴趣标签，它有可能转为短期兴趣，也有可能是假兴趣；而假兴趣是用户虽对这个特征词在一个时间点有很高的关注，但不是用户真正的兴趣。3 者之间可以随着时间的推移而互相转换，如图 4 所示。

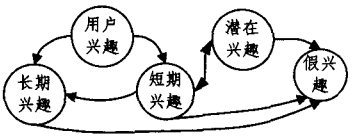


图4 兴趣转换

用户兴趣、潜在兴趣和假兴趣之间的主要区别在于一段时间内它们被关注的走势，因此可利用概念时序动态对其区分，如图 5 所示。图中给出在连续 5 个 session 中兴趣 1、2、3 的概念时序动态，易见兴趣 1 虽没有一个时间点具有很高的关注度，但是一直都有一个比较平稳的关注。通常来说，用户关心的兴趣在一个学习周期内不只出现一次，因此兴趣 1 可以表示用户经常关注的兴趣；兴趣 2 在一个周期的后半段有较高的关注，但由于推进的时间不充分，暂时不能认为是用户兴趣，可将兴趣 2 作为潜在兴趣继续观察；兴趣 3 虽然在一个时间点有很高的关注，但随后关注度降低，存在一定的偶然性，因此作为假兴趣被学习模型忽略。

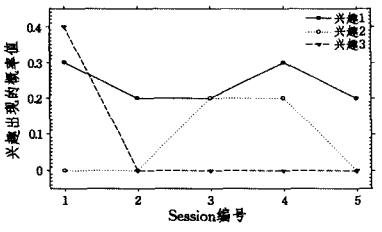


图5 概念时序动态示意图

定义 5 $Antibody=\langle C;W_1,W_2,\cdots,W_n;D(C);Source\rangle$，其中 C 为 ODP 兴趣本体中的兴趣标签，$W_1,W_2,\cdots,W_n$ 是网页中 TF-IDF 值比较高的特征词，这些特征词可以详细地描述兴趣标签，辅助说明兴趣，$D(C)$ 为兴趣标签的概念时序动态，*Source* 表示抗体所拥有的资源数。

当对一个 $term_i$ 累计观察至 k 个 session 后，为防止兴趣过多而对系统带来压力，可利用皮尔逊相关系数[1)]衡量 $term_i$ 是一个兴趣的可能性。通过分析实际兴趣发现，在一段时间内用户对不同的兴趣有着不同的关注程度，因此通过概率时序动态的走势判断一个 *term* 是一个兴趣有些困难，但通过走势能够发现一些与兴趣差距较大的非兴趣。因此利用式(5)计算 *term* 与典型非兴趣的概念时序动态之间的皮尔逊相关系数，若一致性系数 *Cons* 的值大于过滤系数 m，则说明 *term*

1) http://baike.baidu.com/view/3891263.htm

不是兴趣的可能性比较大。利用这种方式过滤掉一部分兴趣标签，从而减轻免疫系统的计算压力。

$$Cons(I_1,I_2)=\sum_{i=1}^{N}I_1I_2-\frac{\sum_{i=1}^{N}I_1\sum_{i=1}^{N}I_2}{N}/\sqrt{\sum I_1^2-\frac{(\sum I_1)^2}{N}}\sqrt{\sum I_2^2-\frac{(\sum I_2)^2}{N}} \quad (5)$$

式中，I_1 表示 *term* 的概念时序动态，I_2 表示非兴趣的概念时序动态。

2.3.3 免疫记忆

免疫学习和免疫记忆是免疫系统能够智能运转的关键。为了下次对抗原产生更好的处理，对于一个抗原，系统在产生抗体处理抗原的同时也学习了抗原的知识，并利用免疫记忆过程记住抗原的特点。免疫记忆中包含的联想式记忆是 AIS 区别于其他进化算法的重要特征之一。联想式记忆的优势在于免疫系统可对结构相似的抗原进行快速识别，为下次处理结构相同或者结构相似的抗原时带来方便，提高应答速度。

在挖掘用户兴趣过程中发现，一个兴趣点可以有多种表示方法，即表示同一类别的抗体存在多个，因此每经过一次免疫学习，首先将新得到的抗体与存在的抗体集进行合并。为了实现联想式应答，利用克隆变异算法对已有的相同类别抗体进行交叉变异，这样对同一兴趣 C 将产生多种与抗原结合较好的变异抗体，从而在与抗原结合的时候具有更大的灵活性。

免疫系统还有一个重要特性——遗忘特性。对比抗体在一段时间占有抗原的数量，占有抗原数量多的抗体比较活跃，说明用户对此类别比较感兴趣；反之说明刺激相应抗体的抗原在逐渐减少，表明用户对其关注度也在降低，清除多余的抗体。通过以上步骤得到最新抗体集合，免疫系统将以此为依据更新用户兴趣集。

2.3.4 获得用户兴趣模型

通过以上的处理，免疫系统的抗体集中包含了反映用户最近行为的兴趣标签，因此对抗体集进行特征提取、特征精炼后能够得到用户兴趣模型，同时为每个兴趣赋予一个表达重要性的权值，用以表明一个兴趣标签被关注的程度。在定义重要性权值时要考虑两方面因素：1)在一段时间内被关注了几次，为了防止局部概率过高使学习模型误将某些兴趣标签作为兴趣；2)在一段时间内概率的整体水平如何，考虑到用户行为与时间相关的特点，越接近当前时间点的兴趣概率越重要，表达的兴趣也更加接近用户的真实兴趣。结合上面两个因素，定义一个重要性权值 I 来计算兴趣标签的权重，见式(6)。

$$I(term_i)=a*\frac{occur}{N}+b*\sum_{i=1}^{|D(term_i)|}\frac{P^{session_i}(term_i)}{|D(term_i)|}*e^{-\frac{\log 2}{l}(d-d_{init})} \quad (6)$$

其中，*occur* 表示在连续观察的 N 个 session 中用户对 $term_i$ 有关注的 session 数，即 $P^{session}$ 不为零的个数。$D(term_i)$ 为 $term_i$ 的概念时序动态。由于随着时间的推近，用户可能对一兴趣标签的关注会减弱，因此引入遗忘因子 $e^{-\frac{\log 2}{l}(d-d_{init})}$[18]，$d_{init}$ 表示该兴趣标签最开始出现的时间点，d 表示当前计算权值的时间点，l 表示跨度参数，表示经过 l 个 session 后，用户对一兴趣的关注会衰减一个层次。a 和 b 为常量，并且满足 $a+b=1$。

用户兴趣包含短期兴趣和长期兴趣，短期兴趣可分为两种，一种可以转为长期兴趣，另一种只是用户在某一个时间段对其比较感兴趣。短期兴趣通过几个 session 的观察就可以获得，而长期兴趣则需要一段时间的积累逐渐学习，长期兴趣是从短期兴趣过渡而来的，因此在获取短期兴趣时要求具有较高的准确度。

学习模型难免会因一些兴趣标签局部概率过高而将一个不属于用户兴趣的兴趣标签作为用户兴趣，这样的错误兴趣如果不能在之后的学习中过滤掉，将会给系统带来很大的压力。具有遗忘特性的学习模型可以很好地解决此问题，既可以遗忘那些因“误会”加入到兴趣集中的兴趣标签，也可以遗忘随时间的推进不被用户关注的兴趣。

最后结合以上特征，利用 ODP 兴趣本体模型，共同构建出相应的用户兴趣模型，如图 6 所示。每个节点来自于兴趣本体的标签，附着在标签旁的集合是进一步描述兴趣节点的特征词，通过赋予每个兴趣标签权值来说明兴趣被关注的程度。

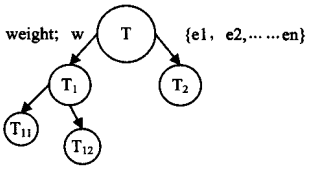
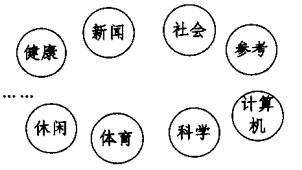


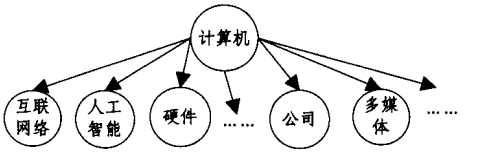
图 6 局部兴趣模型

3 实验结果与分析

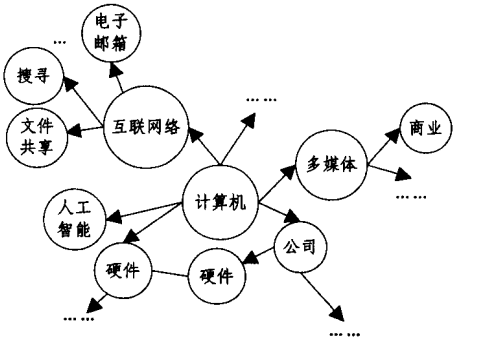
使用本文提出的方法建立一个用于抽取用户兴趣的学习模型。由于用户浏览日志存在着隐私性，因此在该研究上暂时没有可供比较的公共数据集。本文利用的数据集是实验室 10 名成员 8 周以来所收集的浏览网页的记录，共 17233 个网页，每位成员首先在兴趣本体中选取自己比较关注的兴趣标签，以更好地评价系统的准确性。



(a) 兴趣本体 A



(b) 兴趣本体 B



(c) 兴趣本体 C

图 7 兴趣本体

在依据 ODP 类别建立兴趣本体时，选取的类别级数不同，建立的本体也不相同。本文依据不同级数建立 3 个兴趣本体，依据不同本体分别构建兴趣抽取分类器，分别进行测试。图 7 表示 3 个本体的局部图，其中(a)为利用顶级类别构

建的兴趣本体 A，可以看出本体之间无交集；(b)为利用顶级类别和次级类别构建的兴趣本体 B；(c)为利用 3 层类别构建的兴趣本体 C。

依据 AIRS 算法构建 3 个分类器，对网页进行预处理之后，利用构建好的分类器将网页映射到本体中相应的兴趣标签上，得到抗原，判断此次应答是初次应答还是二次应答，计算对抗体的刺激值，若刺激值大于阈值 k，说明系统接受过类似抗原的刺激。为确定更好的 k 值，定义两个参数作为参考，见式(7)和式(8)，ag_num 表示需要二次应答的抗原数，其中系统对 *correct* 个抗原正确地采取二次应答，*wrong* 表示系统应对抗原进行初次应答实际却采取二次应答的抗原数，*num* 表示所有待处理的抗原数。从表 1 可以看出，当阈值 k 为0.7 时，两个参数的值是最理想的。

$$find\_corret = correct/ag\_num \tag{7}$$

$$find\_wrong = wrong/num \tag{8}$$

表 1　阈值 k 确定

k	find_correct	find_wrong
0.1	100%	42.5%
0.2	100%	41.6%
0.3	99.1%	34.5%
0.4	99.1%	30.5%
0.5	89.7%	28.7%
0.6	87.4%	26.8%
0.7	85.6%	20.7%
0.8	82.1%	18.9%
0.9	68.3%	13.4%
1.0	60.7%	9.8%

随后计算相应的概念时序动态，统计资源占有量，得到相应抗体。当 session 数累计到一定数量时对兴趣进行过滤，表 2 说明皮尔逊相关系数与相关程度的关系。通过实验分析，过滤阈值 m 为 0.8 时，可以过滤掉 89%的假兴趣，从而减轻了系统的负担。对过滤后的兴趣标签进行进一步分析，抽取用户兴趣模型。在计算兴趣标签所占权重时，$a+b=1$，当 $a=0.5$、$b=0.5$ 时，所得的权重值比较符合用户对兴趣关注的程度。

表 2　皮尔逊相关系数与相关程度对照

皮尔逊相关系数 c	c=0	c≤0.3	0.3<c≤0.5	0.5<c≤0.8	0.8<c	c=1.0
相关程度	零相关	微相关	切实相关	密切相关	高度相关	完全相关

本文提出该学习模型的主要目的是从用户的浏览记录中提取用户主要关心的领域。为了衡量系统提取的兴趣是否准确，定义两个标准，分别是兴趣提取的准确率(见式(9))和兴趣提取的召回率(见式(10))，其中 *Correct_Interest* 是系统正确提取出来的兴趣数，*Interest* 是系统提取出来的兴趣总数，*Exist_Interest* 是用户实际存在的兴趣数量。这两个参数能够反映出系统在提取用户兴趣方面的精确程度，通过对这 8 周的数据进行分析处理，表 3－表 5 分别列出了依据不同的本体关于每个用户兴趣的提取结果。

$$P = \frac{Correct\_Interest}{Interest} \tag{9}$$

$$R = \frac{Correct\_Interest}{Exist\_Interest} \tag{10}$$

表 3　依据兴趣本体 A 抽取用户兴趣

用户	浏览网页数	涉及的兴趣标签数	正确提取的兴趣数	提取出的兴趣数	P	R
1	1669	3	3	3	1.0	1.0
2	1694	6	5	6	0.83	0.83
3	1795	4	4	5	0.8	1.0
4	1926	7	5	6	0.83	0.71
5	1852	7	6	8	0.75	0.85
6	1764	5	4	4	1.0	0.8
7	1885	6	6	6	1.0	1.0
8	1893	8	7	8	0.875	0.875
9	1968	9	7	10	0.77	0.7
10	1687	4	4	4	1.0	1.0

表 4　依据兴趣本体 B 抽取用户兴趣

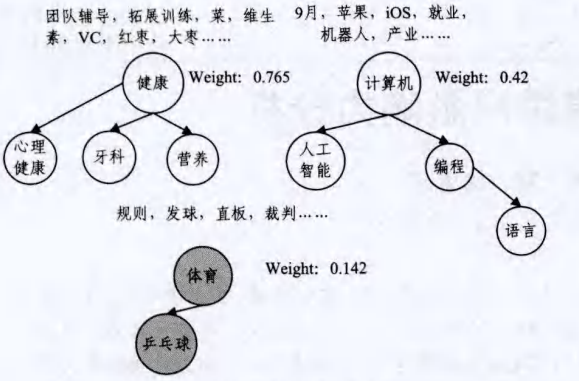
用户	浏览网页数	涉及的兴趣标签数	正确提取的兴趣数	提取出的兴趣数	P	R
1	1669	11	9	11	0.81	0.81
2	1694	18	14	20	0.7	0.77
3	1795	15	12	16	0.8	0.75
4	1926	22	16	21	0.76	0.72
5	1852	21	16	22	0.72	0.76
6	1764	16	11	15	0.73	0.68
7	1885	18	13	15	0.86	0.72
8	1893	25	17	30	0.56	0.68
9	1968	27	19	25	0.76	0.7
10	1687	16	13	18	0.72	0.81

表 5　依据兴趣本体 C 抽取用户兴趣

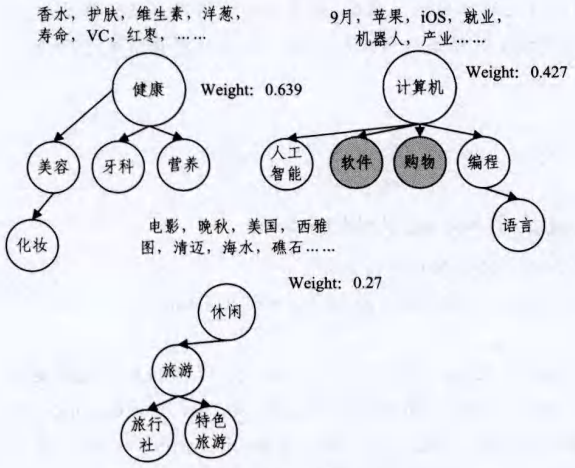
用户	浏览网页数	涉及的兴趣标签数	正确提取的兴趣数	提取出的兴趣数	P	R
1	1669	17	13	15	0.86	0.76
2	1694	22	16	23	0.69	0.72
3	1795	20	13	18	0.72	0.65
4	1926	27	19	26	0.73	0.70
5	1852	26	18	23	0.78	0.69
6	1764	20	14	19	0.73	0.7
7	1885	25	18	22	0.81	0.72
8	1893	30	22	27	0.81	0.73
9	1968	34	23	30	0.76	0.67
10	1687	23	16	21	0.76	0.69

结合本文提出的方法特点对实验数据分析可知，兴趣获取是否准确与两方面有关：1) AIRS 分类器对兴趣标签抽取是否准确；2)随着时间的推移，概念时序动态的计算是否准确。可以看出，兴趣本体构建得越详尽，包含的兴趣越多种多样，对 AIRS 分类器的要求也就越高。对比表 3－表 5 可以看出，当构建本体只考虑顶级类别时，免疫系统的兴趣提取准确率和兴趣提取召回率还是比较理想的，随着本体构建的逐层深入，当达到 3 层子类别时，由于兴趣本体中的词语存在着上下位的关系，概率会聚集在上位词、分散在子节点中，兴趣提取准确率和兴趣召回率都有小幅度的下降。

学习模型的实验数据是用户 8 周以来的浏览记录，有一定的时间跨度，因此也可以观察到用户兴趣的转移和遗忘。那些只是用户随意浏览而不是兴趣的标签在经过几个 session 后将从抗体集合中被清除，同时也从用户兴趣集中去掉。以用户 1 为例，图 8 表示在两个时间点利用本文提出的方法对用户 1 建立的用户兴趣模型，其中图(a)表示在某一时刻 T_i 所抽取的用户兴趣模型，图(b)表示在图(a)的用户模型基础上经过 7 个 session 后建立的新的用户模型，其中阴影部分为学习模型学习错误的兴趣标签。对比图(a)和图(b)可以看出，在前一时刻学错的兴趣在随后的学习过程中可以被学习模型遗忘。



(a) T_i 时刻的用户兴趣模型



(b) T_{i+7} 时刻的用户兴趣模型

图 8　用户 1 的兴趣模型

结束语　本文讨论的是如何从大量的用户日志中准确地抽取出用户兴趣，为解决此问题提出了一种基于人工免疫系统的学习模型。传统的兴趣挖掘过程对用户兴趣的抽取通常没有考虑到用户兴趣变化的特点，本文提出的学习模型具有增量和遗忘的特点，利用概念时序动态可动态地记录兴趣出现的概率，结合人工免疫过程的特点对兴趣完成增量学习，对于系统学错的兴趣或不被用户继续关注的兴趣，学习模型也可通过衡量抗体的浓度实现对“无用”信息的遗忘，从而使学习模型抽取的信息更加精炼可靠。

实验结果表明，学习模型能够建立合理的兴趣本体，构建与之相对应的分类器，将用户浏览的网页对应到相应的兴趣标签上，增量地对兴趣标签进行量化，再用特征词及权重对兴趣标签进行丰富。本文建立的用户兴趣模型与用户的意愿比较接近，用户兴趣模型既可以表现出用户兴趣的细节，同时也体现了用户对每个兴趣的关注程度；并且随着时间推移，学习模型也体现出了遗忘的特性。因此本文提出的方法是合理、有效、可行的。

将学习到的用户个性化信息应用到个性化搜索、网上购物等领域中可以更好地为用户服务，让用户对系统返回的结果更加满意。作者下一步将对算法进行进一步改进，进一步明确界定长期兴趣和短期兴趣，并结合这些兴趣为用户进行个性化推荐。

参 考 文 献

[1] 林鸿飞，杨园生. 用户兴趣模型的表示和更新机制[J]. 计算机研究与发展，2002，39(7)：843-847

[2] Limam L，Coquil D，Kosch H. LionelBrunie：Extracting User Interests from Search Query Logs：A Clustering Approach[C]// DEXA. 2010：5-9

[3] Zhang Pin，Pu Ju-hua，Liu Yong-Li，et al. A probabilistic approach for mining drifting user interests[C]// APWeb/WAIM. 2009：381-391

[4] 吴晶，张品，罗辛，等. 门户个性化兴趣获取与迁移模式发现[J]. 计算机研究与发展，2007，44(8)：1284-1292

[5] White R W，Bennett P N，Dumais S T. Predicting short-term interests using activity-based search context[C]// CIKM. 2010：1009-1018

[6] Holub M，Bieliková M. Estimation of user interest in visited web page[C]// WWW. 2010：1111-1112

[7] Cena F，Likavec S，Osborne F. Propagating User Interests in Ontology-Based User Model[C]// Proc. of XIIth International of Conference of the Italian Association for Artificial Intelligence. Berlin：Springer，2011：299-311

[8] Tan Bin，Lv Yuan-hua，Zhai Cheng-xiang. Mining long-lasting exploratory user interests from search history[C]// CIKM. 2012：1477-1481

[9] Chirita P-A，Nejdl W，Paiu R，et al. Using ODP Metadata to Personalize Search[C]// SIGIR. 2005：178-185

[10] So S，Lee J-H，Jung D，et al. Extending Open Directory Project to Represent User Interests[C]// Proc. of 27th Annual ACM Symposium on Applied Computing. New York：ACM，2012：354-359

[11] Liu Tao，Zhou Yan，Hu Zhi-feng. A New Clustering Algorithm Based on Artificial Immune System[C]// Fuzzy Systems and Knowledge Discovery. IEEE，2008：347-351

[12] Watkins A B，Boggess L C. A Resource Limited Artificial Immune Classifier[C]// Proceedings of the 2002 Congress on Evolutionary Computation Evolutionary Computation，2002(CEC '02). 2002：926-931

[13] Watkins A，Timmis J. Artificial Immune Recognition System (AIRS)：An Immune-Inspired Supervised Learning Algorithm [J]. Genetic Programming and Evolvable Machines，2004，5(3)：291-317

[14] Greensmith J，Cayzer S. An Artificial Immune System Approach to Semantic Document Classification [C]// Proc of Second International Conference on Artificial Immune Systems. LNCS 2787，Berlin：Springer，2003：136-146

[15] Manning C D，Raghavan P，Schutze H. 信息检索导论[M]. 北京：人民邮电出版社，2010：81-83

[16] 莫宏伟，左兴权. 人工免疫系统[M]. 北京：科学出版社，2009：96-100，400-410

[17] 韩家炜(加)，堪博. 数据挖掘概念与技术(第二版)[M]. 北京：机械工业出版社，2007：226-227

[18] Sugiyama K，Hatano K，Yoshikawa M. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users[C]// WWW'04. New York：ACM，2004：675-684