个性化信息服务中基于 Tag 的用户兴趣模型①

王卫平，杨金侠

(中国科技大学 管理学院，合肥 230026)

摘 要：随着 web 信息爆炸增长，个性化信息服务成为人们研究的热点，用户兴趣建模是个性化服务的关键，针对当前用户建模的缺点和 tag 的广泛应用，对基于 tag 的用户兴趣建模进行研究，首先通过实验证明 tag 中蕴含用户稳定的兴趣及 tag 分布的其他特征，然后提出加权树形结构由粗到细的粒度表示用户模型，为提高服务时效性，对用户频繁一起使用的 tag 建立加权频繁项集表，该模型避免了提取关键词的复杂过程，而且从用户的角度表达用户的兴趣，实验表明该模型能提高个性化服务的质量。

关键词：tag；用户建模；个性化服务；加权树；频繁项集

Model of User Profile Based on Tag in Personalized Information Service

WANG Wei-Ping, YANG Jin-Xia

(Department of Management Science and Engineering, University of Science and Technology of China, Hefei 230026, China)

Abstract: With the explosive growth of web message, personalized information service becomes a focus for researchers, user interest model is a key technology in personalized service, this paper research tag-based User Modeling contrary to the shortcomings of the current user modeling and the extensive use of tag. Firstly, through some experiments prove that the tags which user using contains stable interest and other characteristics about tags distribution; then presents a weighted tree from coarse to fine granularity describes the user model. To improve the timeliness of service establish weighted frequent itemset table for often used together tags for each user. The model avoids the complex process of extracting key words and expresses user's interest from the user's point of view, experiments show that the model can improve the quality of personalized service.

Keywords: tag; user profile; personalized information; weighed tree; frequent itemset

1 引言

当前 web 信息的爆炸增长，导致人们迷失在信息过量的海洋里。传统的基于关键词匹配的搜索引擎虽然在很大程度上提高了查询速度，但没有分析用户偏好，使得不同背景、不同目的的人输入相同的关键字得到的查询结果相同，从而降低查询的准确性。个性化服务通过收集和分析用户信息来学习用户兴趣和行为，从而实现主动推荐的目的[1]。用户兴趣建模是实现个性化服务的关键，模型的准确性、时效性直接决定个性化服务质量的优劣。

用户兴趣建模的方法很多，这些方法在一定程度上反映了用户兴趣，提高了个性化信息服务的质量但都存在一定的缺陷，如基于关键词列表法[2]仅用几个简单的关键词来表示用户的兴趣，没反映出用户对每个关键词感兴趣的程度；基于神经网络[3]和遗传算法的用户建模[4]，虽然在动态获取用户兴趣方面有一定的优势，但基于神经网络的方法依赖于模型学习过程所采用的神经网络类别和算法，适应范围比较窄、不易理解且需要用户提供一定数量的反馈，在一定程度上干扰了用户的生活，此外，很多参数设置的难度也比较大。目前在推荐系统和信息检索领域最流行的基于向量空间模型(Vector Space Model VSM)，把用户模型 n 表示成一个维特征向量 $\{(t_1,w_1),(t_2,w_2),...,(t_n,w_n)\}$ 其中 w_i 是特征词 t_i 的权重，该法隐式获取用户信息，反映了用户对不同概念的喜爱程度，但关键词的提取比较复杂，容易引入

① 收稿时间:2010-06-08;收到修改稿时间:2010-07-17

许多无用的词，而且，不同用户浏览相同文档提取的关键词是相同的，没有考虑不同用户对同一文档可能有不同侧面的理解，使得表达结果不准确。

针对以上用户建模方法的缺陷和 tag 的广泛应用，提出了基于 tag 的用户兴趣建模，首先通过实验证明用户使用的 tag 中确实隐藏着用户的稳定兴趣，然后提出基于加权树形结构和加权频繁项集表表示用户模型的方法，实验表明该模型能提高个性化服务的质量。

2 基于tag的用户模型表示

2.1 关于 tag

Tag 总的来说是一种分类系统(结构或类型)[5]，是一种开放、灵活的分类方式，用户可以对自己喜欢的图片、文字、视频等贴上自己喜欢的 tag(众多用户的标注行为称为社会化标注)， tag 不同于一般的关键词标记，它可以用文章中根本没有的词来标记文章，Tag 的意义不仅在于分类，更在于它可以体现出用户各人的思想、生活和情感[5]，因此，用户用 tag 标记自己感兴趣的东西更加贴近自己的理解，和自己兴趣的表达方式。

近些年来，随着 WEB2.0 、WEB3.0 的应用，tag 被广泛应用到不同的社会网络中，在国外 tag 技术已经被广泛应用到如 Flickr, Del.icio.ous，此外一些图书馆如 Lewis&Clark 也把此技术应用到图书管理中，并且取得了成功；国内如 365key.com 博客托管服务商 BlogBus 等都推出了标签服务，虽然广泛接触的比较晚，但是标签服务已经受到广泛欢迎。

亚马逊首席科学家 Andreas S.Weigend 认为互联网经历了三个发展时期：前两个时期分别是通过类似 FTP 服务和目标搜索的超链接获得服务信息，第三个时期就是 tag 技术，他认为 tag 技术将带来一些集成大量高效功能的产品，帮助用户快速有效的获取信息。随着网络的发展和信息的增加，tag 技术将会越来越广泛的应用于各个互联网上。

2.2 用户浏览文档与 tag 的关系

用户浏览网页时，对于自己喜欢的网页为方便下次查找，在保存文档对应的 URL 时会对其贴上自己喜欢的标签即 tag，这个标签是用户自由选择文本的关键字，定义了用户思维中的概念与对应网络资源之间的关系[6]。人们倾向于使用描述性标签标注他们感兴趣的内容，并且对他们感兴趣的网页使用的 tag 与该网页的内容相关，而 tag 更简洁、更贴近用户对文章的理解和对内容的判断[7]，因此用户频繁使用的标签可以表征和捕捉用户感兴趣的主题。

2.3 关于 tag 的几个实验

基于 tag 建立用户兴趣模型的前提是 tag 中蕴含用户的稳定兴趣，为证明此假设以及更多的了解 tag 的特性，本文选取数据集进行分析。

2.3.1 数据集 D_1 的选取

本文实验数据均来自 del.icio.ous 网站，为保证取到足够数据和研究同一 URL 中某一特定用户和其他用户使用的 tag 的关系，按如下要求选取实验数据：

(1) 被研究用户 U_i 对其 bookmark 中的每个 url_j 使用的 tag 数 $|U_t^i| \geq 5$ 。

(2) url_j 至少被 100 个人标注过。

(3) 为了研究 tag 数目的变化规律及鉴于实际每个 URL 的 tag 个数考虑，按每个 tag 被标记次数多少，选取每个 URL 中被所有用户标注的 top-30 个 tag(少于 30 的则全部选择)。

(4) 按照时间先后顺序从每个用户的 bookmark 中选取 150 个 URL 的 tag。

(5) 为了保证选取到满足上述条件的足够数据，每个被选取的用户 U_i 管理的 bookmarks 数量 $|U_b^i| \geq 100$ 。

2.3.2 数据预处理

因为 tag 标注的随意性，搜集到的数据包含大量噪声和错误的信息，为了建立高质量的用户模型，首先除去非英语和错误的词语，然后把诸如 computer/computers，program/programming 之类的表示相同意思但词形不同的词合并。

2.3.3 实验分析

按照上述条件随机选取 100 位用户数据预处理后，对每个用户的数据分别做了如下实验，虽然每位用户的数据有所差距，但是都遵循同样的规律，为描述方便和篇幅限制，仅列出 100 位用户的平均实验结果。

实验 1：tag 数 $|T|$ 随 URL 数 $|URL|$ 增长的关系：设用户 U_i，$i \in [1,100]$ 的每个 URL 中取 top-30 个 tag 得到的总数为 $|T|$，其中包含用户 U_i 的 tag 总数为 $|ti|$，对数据预处理后发现，$\text{Max}\{|T_i|\}=912$，$\text{Min}\{|T_i|\}=783$，$\text{Max}\{|t_i|\}=421$，$\text{Min}\{|t_i|\}=280$，实验可知，很多 tag 是被重复使用的，平均每个 URL 的 tag 数 $|T|$ 随 URL

数 |URL| 增长的关系如图 1 所示：

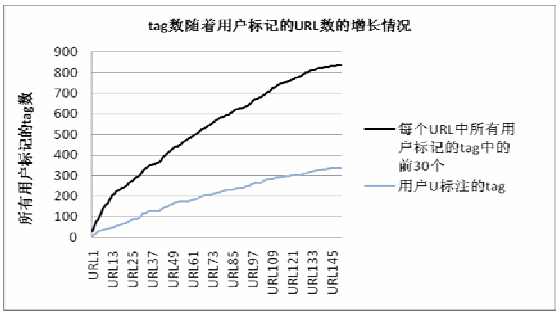

图 1　|T| 和 |URL| 的关系

由图 1 可以看出 |T| 随着 |URL| 的增长而增长，并且增长的速度越来越慢，统计数据集中每个 U_i 的每个 tag 被 U_i 所贴的次数发现，几乎半数的 tag 被用户至少使用 2 次，有的甚至被使用几十次，这说明 tag 中蕴含用户的稳定兴趣，用户频繁浏览感兴趣的信息，频繁使用表示自己兴趣的 tag。文献[7]通过大量 tag 分析发现，所有用户对每一个 URL 贴标签词汇，足够丰富的描述此 URL 对应网页内容的主要自然概念，而 tag 的个数远远小于从网页中抽取关键词的个数，并且不同用户对同一 URL 表达的主题有不同的判断，这种属性不是从相应网页中提取关键词所拥有的，因此用户的 tag 可以作为联系用户和网页的桥梁，且标签更加贴近用户自己对文章内容的理解和其对文章需要的一部分，更能体现用户需求的个性化。

实验 2：其他用户 U_n 与某一特定用户 U_i 使用的 tag 关系。给出两个定义：

U_n：一个整体，设 U_i 标注过的 URL 集合为 $URL_i=\{url_j\,|\,(1\le j\le m)\}$，则 U_n 为除 U_i 外标注过 URL_i 的所有用户的整体。

协同预测率：设 $|U_n|$ 的 tag 总数为 $|T|$，数据集中 U_i 的 URL 总数为 m，则协同预测率 $C_{(\alpha/m-\alpha)}$ 为：

$$C_{(\alpha/m-\alpha)}=\frac{|T_{U_i}^{|url|=m-\alpha}|}{|T_{U_n}^{|url|=\alpha}|}\times 100\%$$

其中，$|T_{U_n}^{|url|=\alpha}|$ 是前 α 篇文章中 U_n 使用的 tag 总数；$|T_{U_i}^{|url|=m-\alpha}|$ 是后 $(m-\alpha)$ 篇 URL 中被 $|T_{U_n}^{|url|=\alpha}|$ 包含的 U_i 的 tag 总数，100 位用户的实验结果平均值如表 2 所示：(m=150)

表 1　U_n 对 U_i 使用 tag 的协同预测率

α	80	100	120
$C_{(\alpha/m-\alpha)}$	65.16%	72.61%	80.17%

由表 1 知 U_n 的 tag 对预测 U_i 的兴趣有影响，并且取的 URL 数越多，包含 U_i 的兴趣词条越多，建立 U_i 的兴趣模型时不仅要考虑其自己使用的 tag 也要考虑标注同一 URL 的其他用户使用过的 tag，以更准确和全面的建立用户个性化模型，达到提高个性化服务的目的。

实验 3：tag 使用分布。文献[7]通过所有人使用 tag 分布发现受欢迎的 tag 被大多数用户使用。我们按照标注次数分别取数据集中每个 URL 的 top-10 、top-30 受欢迎的 tag，为描述方便我们给出几个定义：

覆盖率：$COV_i=\dfrac{|ti|_n}{|ti|}\times 100\%$

利用率：$UTI=\dfrac{|ti|_n}{|T_i|_n}\times 100\%$

其中，$|ti|_n$ 为每个 URL 的 top-N 中包含用户 U_i 使用的 tag 数，$|ti|$ 是 U_i 使用的 tag 总数，$|T_i|_n$ 是每个 URL 的 top-N 中被所有用户标记的总数。结果如表 2

表 2　tag 分布情况

提取的 tag	Top-10	Top-30
平均覆盖率	89%	92.6%
平均利用率	76.1%	39.4%

由表 2 得知，总体来说 U_i 在所有 URL 中使用的 tag 绝大部分被所有用户使用的 top-N 个 tag 覆盖，说明对于每个 URL，大多数用户都喜欢用比较受欢迎的 top-N 个 tag 来标注，通过分析发现每个 URL 被贴的 tag 的次数差别很大，同一个 URL 中的 tag top-N 个被标注几百次甚至几千次，而不受欢迎的仅被标注一次，总体来看覆盖率越大，利用率越低，因此考虑使用同一 URL 的其他用户的 tag 建立用户模型时，为了减少数据量，降低噪音，仅选取每个 URL 中的 top-N 个 tag 来建立用户模型即可。

3　用户兴趣建模

3.1 tag 的权重

通过上述实验得知研究用户 U_i 兴趣时，用来建立模型的 tag 包含两部分：U_i 的 tag 和 URL_i 中 top-N 个 tag,使用向量空间模型（VSM）来描述 URL，在 VSM

中使用 tag-URL 矩阵 $A=(a_{ij})\in R^{t\times u}$，每个向量 $a_j\ (1\le j\le u)$ 代表一个 url_j，则基于 TF 计算 tag$_i$ 在 url_j 中的权重为：

$$a_{ij}=\frac{Fre_{ij}}{\sqrt{\sum_{k=1}^{t}f_{kj}^2}}$$

在每个 URL 中若 tag$_i$ 被用户 U_i 使用，为了突出 U_i 使用 tag 的重要性令 $\alpha_{ij}^u=\beta\bullet\alpha_{ij}$，$\beta$ 通常取大于 1 的数，如 $(\beta=1.1)$，计算出每个 tag 在每篇 URL 中占的权重，可以得出每个 tag 的总权重 W_i：

$$W_i=\frac{\sum_{j=1}^{u}a_{ij}}{|T_i|}$$

其中 $|T_i|$ 为出现 tag$_i$ 的所有 URL 数。

3.2 用户模型的表示

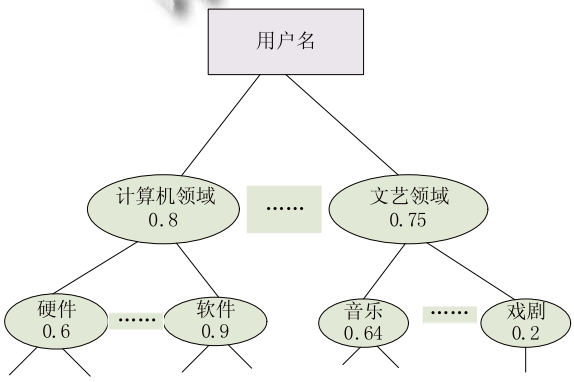

图 2 用户模型的加权树形结构

用户u的频繁项集表	
Frel	{w$_i$ | i ∈Fre 1}
Fre2	{w$_j$ | j ∈Fre 2}
…	…

图 3 用户的 tag 频繁项集表

对每一位用户本文使用加权树形结构和频繁项集表共同表示用户模型。加权树形结构中(如图 2 所示)，最上层(第 0 层)是一个虚的用户模型的标识节点，第一层是特定的领域知识，这一层在建立用户模型初期每位用户都有所有领域的知识，初始化时领域知识权值均为 0，通过下层各概念的学习调整相应权值，往下每一层都是上一层具体概念的进一步划分，每个节点都有相应的权重，本文结合特定领域知识使用概念聚类算法以分类树的形式来构建用户模型。

在用户加权树形模型中，领域概念的权重是各分支中的权重加权平均值来决定，第二层开始细分的各节点权重由 3.1 中公式计算得出。

使用 Apriori 算法找出各用户的频繁 2 项集开始的频繁项集，并标上相应的权重，建立频繁项集表(如表 3)，表左边是各频繁项集，右边是各词条对应的权重。

3.3 个性化信息服务的实现

本文建立的用户模型更细粒度的划分了用户的兴趣，对用户进行信息推荐时分两种情况：

(1) 已经被贴上 tag 的文档，我们首先查询频繁项集，与频繁项集中重合的词条越多，越被优先推荐给用户，然后扫描加权树，重合词条越多，词条权重越大，越优先推荐。

(2) 从未被标注的文档，则首先判断文档属于哪个领域，根据各领域权重判断推荐次序，在领域内根据各概念权重决定推荐的先后顺序。

4 过滤实验

我们在 del.icio.ous 网站上随机选取了 10 位标注的 URL 数大于 100 的用户，每位用户按照时间顺序选取 100 个 URL，组成数据集 D_2，对每个用户做了如下实验：

(1) 因为数据集 D_1 的选取添加了很多限制条件，为了验证上述实验结论的普遍性，我们对 D_2 数据集中的用户也做了 2.3 中的实验，实验结果表明，2.3 中得到的结论推广到所有用户都是成立的。

(2) 从数据集 D_1 中随机抽取 10 位用户按时间顺序取其标记的前 100 个 URL 组成数据集 D_{11}，在数据集 D_{11} 和 D_2 中每个用户的 URL 中前 60 个作为训练集，后 40 个作为预测集，经预处理后，分别选取每个 URL 的 top10 和 top30(若不足，则全部选择)来进行过滤仿真实验，设每个 URL 与用户模型中 tag 重合数为支持数 Sup，数据集 D_{11} 当 $Sup\ge 5$，D_2 中当 $Sup\ge 2$ 时，则此 URL 推荐给用户，因实验环境

限制，仅用查全率来检验模型质量，得到查全率平均值如表 3：

$$\text{查全率}(R)=\frac{\text{过滤结果中用户标注过的URL数}}{\text{信息源中用户标注过的全部URL数}}$$

表 3 模型平均查全率

查全率（R） / 选取的 tag 数	D_{11} 平均查全率	D_2 平均查全率
Top10	0.884	0.750
Top30	0.895	0.762

从实验中 D_2 看出的查全率并不是很高，分析发现 D_2 中有的 URL 被标注的次数较少，甚至仅被要研究的用户标注过一次。导致那些仅被用户标注过一次的 URL 不能被过滤出来，但是可以通过调整 Sup 克服此缺点，通过 D_{11} 和 D_2 实验结果对比看出对于每个 URL 被标注的人数越多，贴的 tag 数越多，查全率越高，推荐的准确性越高，个性化服务的质量也就越高。

5 结论

本文首先通过实验证明用户长期使用 tag 中含有稳定的兴趣，即建模的可行性，并分析了协同预测率和 tag 的分布规律，最后提出基于加权树形结构和加权频繁项集表共同表示用户兴趣，实验证明模型具有很高的查全率又避免了向量空间模型中提取关键词的复杂过程，并且从用户角度反映不同用户对同一文档感兴趣的不同侧面，使得个性化服务质量更高。

以后的工作中主要深入研究如何使用聚类算法更精确和更好的建立加权概念树。

参考文献

1 曾春,邢春晓,周立柱.个性化服务技术综述.软件学报,2002,13(10):1952－1961.

2 Joachims T, Freitag D, Mitchell T. WebWacher: A Tour Guide for the World Wide Web. Proc. of 15th International Joint Conference on Artificial Intelligence (IJCAI-97). Nogoya, Japan, August, 1997:770－775.

3 Shepherd M, Watters C, Marath AT. Adaptive User Modeling for Filtering Electronic News. Proc. of the 35thAnnual Hawaii International Conference on System Sciences. 2002. 1180－1188.

4 Lupez-Pujalte C, Guerrero-Bote VP, Moya-Aneg FD. Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback. Journal of the American Society for Information Science and Technology, 2003,54(2):152－160.

5 http://baike.baidu.com/view/728.htmfr=ala0_1

6 Marieke Guy.Folksonomies Tidying up Tags. http://www.dlib.org/dlib/january06/guy/01guy.html,2009-4-20.

7 Li X, Guo L, Zhao YH.Tag-based Social Interest Discovery. www2008/Refereed Track:Social Networks&Web 2.0-Discovery and Evolution of Communities. 2008. 675－684.

(上接第 90 页)

4 结论

通过进行 Matlab 仿真实验，PSO-BP 模型的收敛速度、泛化能力都优于 BP 模型。PSO-BP 模型克服了 BP 模型收敛速度慢、易陷入局部最优、初始值难以确定等内在缺陷，达到了很多的训练效果。由于在高校学生个人信用评价方面的研究还很不成熟，在数据取得、指标建立等方面还存在一定的缺陷，本文建立的评价体系只是一个初步的尝试。但是在建立起统一的数据平台，成立权威的执行机构之后，本文的模型以及相关的体系建立方法与流程，对未来建立高校学生个人信用评价体系有一定的借鉴和参考作用。

参考文献

1 Yi D, Ge XR. An improved PSO based ANN with simulated annealing technique. Neuron computing, 2005,63(11):527－533.

2 Kennedy J, Eberhart RC. Swarm Intelligence. San Francisco: Morgan Kaufmann Publishers. 2001.

3 彭宇,彭善元,刘兆庆.微粒群算法参数效能的统计分析.电子学报,2004,32(2):209－213.

4 Jiang CW, Bompard E. A self-adaptive chaotic particle swarm algorithm for short term hydroelectirc system scheduling in deregulated environment. Energy Conversion and Management, 2005,46:2689－2696.

5 艾永冠,朱卫东,闫冬.基于 PSO-BP 神经网络的股市预测模型.计算机应用,2008,28(12):105－107.

6 潘昊,侯清兰.基于粒子群优化算法的 BP 网络学习研究.计算机工程与应用,2006(16):41－43.

7 冯俊青,郁志宏. PSO-BP 网络模型在数据分类中的应用.自动化技术与应用,2007,26(11):13－15.

8 艾永冠. 粒子群优化神经网络在股市预测中的建模与应用[硕士学位论文].合肥:合肥工业大学,2009.19－21.