

DOI: 10.7652/xjtuxb201901023

采用语义分析的标签体系构建方法

李娜¹, 范正洁², 郝传洲¹, 王平辉¹, 陶敬¹, 林杰¹

(1. 西安交通大学智能网络与网络安全教育部重点实验室, 710049, 西安;
2. 中国移动通信有限公司研究院大数据与 IT 技术研究所, 100053, 北京)

摘要: 针对现有标签体系粒度粗和层级结构不明显的问题,提出了一种基于语义分析的自动化标签体系融合构建方法。该方法通过分析不同网站导航标签体系中标签的相似性,学习不同导航标签间的等同映射关系和上下位映射关系,进而融合不同网站的导航标签体系以得到细粒度且层级分明的标签体系。为了评估该方法在标签体系融合构建方面的精度,首次提出标签重合度和上下位关系重合度两个测试指标进行衡量。实验结果表明,与基于同义词林的标签体系融合构建方法相比,所提方法在标签重合度和上下位关系重合度上提升了 5% 以上,可以构建出精准有效且适应不同领域的标签体系,为构建精准的用户画像打下基础。

关键词: 标签体系;标签融合;用户画像

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 0253-987X(2019)01-0169-06

A Method for Building Tag Systems Based on Semantic Feature Analysis

LI Na¹, FAN Zhengjie², HAO Chuanzhou¹, WANG Pinghui¹, TAO Jing¹, LIN Jie¹

(1. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China; 2. Big Data and IT Technology Department, China Mobile Research Institute, Beijing 100053, China)

Abstract: An improved method based on semantic features analysis of tag systems is proposed to solve the problem that existing tag systems are coarse-grained and their hierarchical structures are unapparent. The method analyses similarity among multiple tag systems on different websites, learns the relationships of synonym mappings and hypernym-hyponym mappings among these systems, and further obtains a fine-grained and hierarchical tag system. In order to evaluate the performance of the algorithm, two test metrics, the tag coincidence degree and the hypernym-hyponym coincidence degree, are proposed to assess the accuracy of merging and constructing tag systems. Experimental results show that the method improves the tag coincidence degree and the hypernym-hyponym coincidence degree by more than 5%, and it efficiently builds a tag system with higher precision and applicability, meanwhile, lays a foundation for establishing user profile.

Keywords: tag system; tag merging; user profile

随着大数据技术的深入研究与应用,企业的专注点日益聚焦于怎样利用大数据来为精准营销服务,进而深入挖掘潜在的商业价值^[1-4],于是用户画像的概念也就应运而生。所谓用户画像,即用户信息的标签化,利用标签体系勾画用户的属性特征^[5]。精准、细粒度且结构化的标签体系是用户画像的基础,其

广度和粒度对用户画像的精确性有较大影响,因此标签体系的构建具有一定的研究意义和应用价值。

早期比较流行的标签体系构建方法均基于 Golder 所提出的协同标签体系^[6];Yeung 和 Tahar 等利用分众分类法构建标签体系^[7-8];Cai 等先后提出利用协同标签系统不同标签权重的结构化和非结构化标签来构建标签体系^[9-10]。协同标签体系描述众多用户以标签形式向共享内容添加元数据的过程,大多知名网站允许用户公开标签和共享内容,用户不仅可以为自己的信息分类,也可以浏览他人的分类信息,故协同标签体系的构建受到个人和公共两方面因素的影响。由于个人背景知识以及描述文字的习惯不同,不同用户对同一网页可能添加不同的标签,致使最终的标签体系存在重复和结构层次不明显的问题。随着本体的发展,利用本体构建标签体系的研究也逐渐增多,Skillen、Maleszka 和 Ferreira 等利用现有本体框架人工或自动化地构建本体以生成标签体系^[11-13],但目前基于本体所得到的标签体系准确率低,存在冗余或者缺失以及结构问题。此外,本体的构建需要大量人工参与,成本较高。Farseev 等基于主题模型^[14-15]构建标签体系,但这种标签体系粒度较粗,不利于精准画像。

针对上述问题,本文基于主流网站的导航标签,提出一种标签体系融合方法以构建细粒度结构化的标签体系。主流网站的导航标签是经过专家仔细分析,并结合广大网民的用户体验进行优化而得,故网站导航标签能够精准勾画用户的行为属性,进而为用户画像打下基础。

本文根据标签文本特征、结构特征以及标签对应的网页文本特征,提出一种基于语义特征分析的标签融合方法(TMSFA)。该方法通过识别标签间的等同和上下位这两种映射关系,对主流网站的导航标签进行融合,进而构建出精准、细粒度且结构化的标签体系。

1 标签体系构建方法

第一层级标签称作根标签,每个根标签及其所有子孙标签称作一个标签树。本文方法三阶段流程如图 1 所示,其基本思想是将标签树两两融合,将待融合标签树插入到基准标签树中,通过找到待融合标签与基准标签的等同映射关系和上下位映射关系,从而构建出融合后的标签体系。本文方法分为数据处理、标签映射以及标签融合 3 个阶段。

在数据处理阶段,首先从网站中获取原始标签

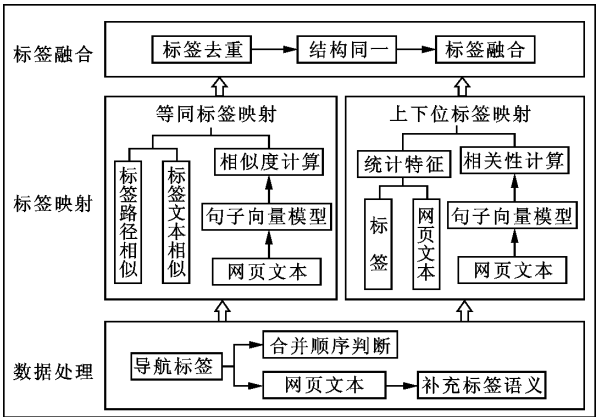


图 1 本文方法三阶段流程图

体系,并基于爬虫获取标签对应的网页文本,以丰富标签语义;其次根据标签体系内容并结合分组条件对标签树进行聚类分组,分组条件一是当两个标签树的根标签相同分为一组,分组条件二是当标签树的根标签不同但子标签内容相似则分为一组;最后根据标签树的层数、叶子标签及非叶子标签的数量判断融合顺序,标签树层数越多则标签的划分结构粒度越细,叶子标签的数量越多则标签树描述的子领域越广,非叶子标签的数量越多则标签的划分粒度越细。

在标签映射和标签融合阶段,基于标签间的等同映射关系和上下位映射关系去除标签体系中重复以及结构相同的标签,然后将待融合标签树插入到基准标签树中,进而构建出融合后的标签体系。

2 等同关系标签映射

标签间的等同关系是指两个标签语义相同,对应的网页内容相似。在识别标签的等同关系时,主要存在以下两个难点:一是不同原始标签体系的标签词义表达方式不同,例如“钓鱼用品”“垂钓用品”和“彩宝”“彩色宝石”等;二是上级标签对下级标签的语义具有约束力,例如标签“洗发水”的上级标签有“洗发护发”“男士洗护”两种,此时标签的语义截然不同。

本文基于标签体系的语义和结构提出两种方法以解决上述难点,一是通过判断标签对应网页内容是否相似;二是同时分析标签及其上级标签的语义信息,基于这两种方法可得到众多等同关系标签对,本文取这些等同关系标签对的并集作为最终识别的等同关系标签对。

四级标签体系树状图如图 2 所示,其中一级标签指所有标签树的第 1 级根标签,如标签 $a_1 \sim a_2$;二

级标签指所有标签树的第 2 层级标签,如标签 $b_1 \sim b_4$;三级标签指所有标签树的第 3 层级标签,如标签 $c_1 \sim c_8$;四级标签指所有标签树的第 4 层级标签,如标签 $d_1 \sim d_4$ 。

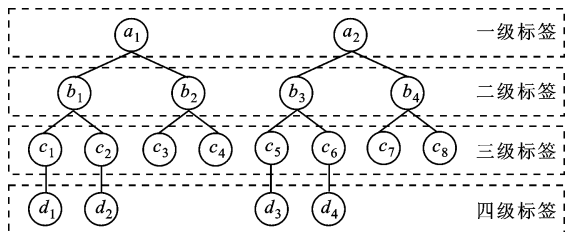


图 2 四级标签体系树状图

2.1 基于标签语义的等同关系映射方法

基于标签语义比较待融合标签与基准标签是否“相同”,若“相同”则将待融合标签和基准标签的父标签及祖父标签进行两两比较,存在一对“相同”则说明待融合标签与基准标签具有等同关系。

待融合标签 A 和基准标签 B 分别由 n 和 m 个字组成,表示为 $A_1, A_2, \dots, A_n$ 和 $B_1, B_2, \dots, B_m$,其中 $A_i (i=1, 2, \dots, n)$ 、 $B_j (j=1, 2, \dots, m)$ 分别表示组成标签 A 和 B 的字。逐个比较标签 A 和 B 的字,假设标签 A 和 B 中重复的字的个数为 h ,则标签 A 和 B 的相似度为 h/n ,若相似度大于基于统计分析设定的阈值,则说明这两个标签是“相同”的。

2.2 基于附加语义的等同关系映射方法

基于标签对应的网页文本内容的相似性来判断标签的等同关系。基于网页文本语义获得标签的向量表示,输入为每个标签对应的 k 个网页标题,通过句子向量表示模型 (Sentence2Vec)^[16]。获得每个标题的句子向量表示为 $s_1, s_2, \dots, s_n$,则标签表示为

$$S_{\text{label}} = \frac{1}{k} \sum_{i=1}^k s_i \quad (1)$$

待融合标签 A 与基准标签 B 的向量表示分别为 S_A 和 S_B ,其余弦相似度为

$$C(S_A, S_B) = \frac{S_A^T S_B}{\|S_A\| \cdot \|S_B\|} \quad (2)$$

当余弦相似度大于基于统计分析设定的阈值时,说明待融合标签 A 与基准标签 B 具有等同关系。

3 上下位关系标签映射

上下位关系是指两个标签的语义之间存在包含与被包含的关系,下位词是上位词的一个特殊实例或者一个子类。如图 3 所示,待融合标签“鲜花饼”的上级标签为“休闲食品”,标签“鲜花饼”会跟随其上级标签“休闲食品”融合到“鲜花饼 1”的位置,但

是“饼干糕点”是“鲜花饼”更为准确的上位词,故标签“鲜花饼”应融合到“鲜花饼 2”的位置。

基于标签对应的网页文本内容提出两种上下位标签映射方法:一是通过判断待融合标签对应的网页标题中包含基准标签的比例来判断上下位关系;二是通过判断待融合标签与基准标签的子标签的相关性来判断上下位关系。

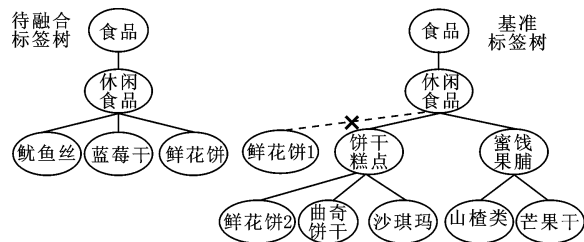


图 3 基于上下位关系的标签融合示意图

3.1 基于网页标签包含关系的上下位标签映射方法

标签对应的网页标题通常会包含其上级标签,利用这个特点判断两个标签是否存在上下位关系。待融合标签 A 的对应的网页标题个数为 p ,若网页标题中有 q 个标题包含基准标签 B ,则包含比例为

$$R(A, B) = \frac{q}{p} \quad (3)$$

当包含比例大于基于统计分析设定的阈值时,则说明待融合标签 A 与基准标签 B 具有上下位关系。

3.2 基于标签间相关性的上下位标签映射方法

如图 2 所示,“鲜花饼”与“饼干糕点”的子标签“曲奇饼干”和“沙琪玛”有一定的相关性,也就是子标签间存在着一定的相关性,据此判断两个标签是否存在上下位关系。计算待融合标签与所有包含子标签的基准标签的相关性,当最大相关性大于基于统计分析设定的阈值时,则说明待融合标签 A 与基准标签 B 具有上下位关系。基于子标签的向量表示分别计算标签 A 与标签 B 以及子标签的余弦相似度 $C(S_A, S_B), C(S_A, S_{B_1}), \dots, C(S_A, S_{B_n})$,则标签 A 和 B 的相关性为

$$G(A, B) = \frac{C(S_A, S_B) + \sum_{i=1}^n C(S_A, S_{B_i})}{n+1} \quad (4)$$

4 实验与分析

4.1 实验设置

目前用户访问的网站类型主要包括门户、金融、电商 3 类,故从这 3 类网站中分别选择两个主流网站导航作为原始标签体系进行融合,其中门户网站选择新浪网和搜狐网,金融网站选择中金在线和东方

财富网,电商网站选择天猫和苏宁易购,详情见表 1。

表 1 三类网站标签体系中各级标签分布

网站类型	网站名称	层级	一级标签	二级标签	三级标签	四级标签	标签总数
电商	苏宁	4	33	182	1 051	1 229	2 495
	天猫	4	25	201	1 143	2 175	3 544
金融	东方财富	3	8	107	351		466
	中金在线	3	26	366	99		491
门户	新浪	3	5	31	387		432
	搜狐	3	3	40	340		383

为了测试融合方法得到的标签体系的准确性,本文提出标签重合度和上下位关系重合度两个指标来验证融合方法的效果。

所谓标签重合度指标,即通过计算融合后标签体系与测试标签体系的标签重合比例来说明两个标签体系的内容差异,它是一种标签内容相似的度量。融合标签体系和测试标签体系中标签的个数分别为 N_L 、 N_T ,相同的标签个数为 N_{same} ,则标签重合度为

$$F_{\text{label}} = \frac{N_{\text{same}}}{N_T}$$

(5)

所谓上下位关系重合度指标,即通过计算融合后标签体系与测试标签体系的上下位关系相同比例来说明两个标签体系的结构差异,它是一种标签上下位关系的度量。融合标签体系和测试标签体系中树枝个数分别为 M_L 、 M_T ,测试标签体系中与融合标签体系中相同的树枝的个数为 M_{same} ,则上下位关系重合度为

$$F_{\text{branch}} = \frac{M_{\text{same}}}{M_T}$$

(6)

测试标签体系是从上述原始标签体系中随机抽取若干个标签进行分组,在 20 多位老师和同学的帮助下人工融合标签并校正,得到最终的测试标签体系。

对于电商网站,本文将“家居家纺”“母婴玩具”和“个护化妆”3 个分组的原始标签人工融合得到电商测试标签体系;对于金融网站,将“理财中心”和“股票中心”两个分组的原始标签人工融合得到金融测试标签体系;对于门户网站,将“频道”分组中的“新闻”“体育”“娱乐”“科技”“女人”“健康”“房产”“星座”和“旅游”这些小组的原始标签人工融合得到门户测试标签体系。三类网站标签体系中原始及测试标签数的统计情况见表 2,以电商为例,电商网站原始

标签体系共有 6 039 个标签,抽取其中 1 265 个标签进行融合,得到的测试标签体系共有 934 个标签。

表 2 三类网站标签体系中原始及测试标签数

网站类型	网站名称	层级	原始标签数	测试标签数
电商	苏宁	4	1 265	934
	天猫	4		
金融	中金在线	3	626	531
	东方财富	3		
门户	新浪	3	274	211
	搜狐	3		

4.2 实验结果

为了进一步说明本文方法的优越性,采用基于同义词林的融合方法(TMC)与之进行对比。TMC 方法基于同义词林和标签文本层面的相似性来判断标签映射关系,进而融合标签体系。

在同一台计算机上使用相同数据对本文方法和 TMC 方法进行对比,首先基于这两种方法分别将门户、金融、电商网站的原始标签体系进行融合,然后将融合后的标签体系分别与测试标签体系进行对比,最后计算出标签重合度和上下位关系重合度,对比结果及时间频度见表 3、表 4。

表 3 本文和 TMC 方法在评价指标上的比较

网站类型	方法	融合标签树标签数	测试标签树标签数	标签重合度/%	上下位重合度/%
电商	本文	1 041	934	89.4	88.4
	TMC	1 102	934	79.8	80.9
金融	本文	504	531	85.5	90.2
	TMC	551	531	76.7	88.4
门户	本文	232	211	90.0	95.3
	TMC	251	211	80.1	90.3

表 4 本文与 TMC 方法的时间频度对比

方法	时间频度/s		
	电商	金融	门户
本文	0.254	0.098	0.071
TMC	0.380	0.065	0.211

针对电商类网站,基于本文方法融合天猫和苏宁的标签体系得到 1 041 个标签,并与测试标签体系进行对比,标签重合度和上下位关系重合度分别为 89.4%、88.4%,而 TMC 方法的结果是 79.8%和 80.9%。电商类网站的标签体系结构复杂,经常会出现待融合标签所在的结构粒度粗而相应所属基

准标签粒度细的情况,这就需要拆分待融合标签子树并将其映射到基准标签体系中,但难点在于如何判断待融合标签与基准标签的映射关系,故造成上下位关系重合度相对标签重合度较低。针对金融类网站,基于本文方法将东方财富和中金在线的标签体系进行融合得到 504 个标签,并与测试标签树进行对比,标签重合度和上下位关系重合度分别为 85.5%, 90.2%,而 TMC 方法的结果是 76.7%和 88.4%。针对门户类网站,基于本文方法将新浪和搜狐的标签体系进行融合,得到 232 个标签,并与测试标签体系进行对比,标签重合度和上下位关系重合度分别为 90.0%、95.3%,而 TMC 方法的结果是 80.1%和 90.3%。由于词语描述差异比较大,门户类网站和金融类网站难以精准判断标签之间是否存在等同关系,故上下位关系重合度相比于标签重合度较高。

在时间复杂度方面,本文方法和 TMC 方法的时间复杂度均为 $O(n^2)$ 。从表 4 可以看出,两种方法的时间频度相差无几,但从表 3 可以看出,与 TMC 方法相比,本文方法的效果有较大的提升,并且本文方法得到的融合后标签体系相对于 TMC 方法的标签体系的标签数量较少,说明识别的融合标签对数更多。无论从等同关系识别还是上下位关系识别,本文方法都有一定的提升且具备一定的准确性,而且相对已有构建标签体系方法得到的标签体系而言更为全面、结构化。

5 结 论

本文方法将不同网站的导航标签体系进行融合得到统一且结构化的标签体系,为精准用户画像打下基础。本文方法基于标签间的等同关系和上下位关系来判断标签的融合位置,由于标签本身词语短小且包含的语义信息较少,本文不仅分析标签本身的语义信息,还结合其对应网页文本的语义信息及标签上下级结构关系信息以丰富标签语义,进而判断标签间的映射关系。本文提出标签重合度和上下位关系重合度两个指标以评估本文方法的优越性,与现有方法相比,本文方法的这两个指标至少提升 5%,证明了方法的有效性。采用本文方法可以构建出精准有效且适应不同领域的标签体系。

参考文献:

- [1] 王元卓,靳小龙,程学旗. 网络大数据:现状与展望[J]. 计算机学报, 2013, 36(6): 1125-1138.
WANG Yuanzhuo, JIN Xiaolong, CHENG Xueqi.

- Network big data: present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.
- [2] 刘海,卢慧,阮金花,等. 基于“用户画像”挖掘的精准营销细分模型研究[J]. 丝绸, 2015, 52(12): 37-42.
LIU Hai, LU Hui, RUAN Jinhua, et al. Research on precision marketing segmentation model based on mining “persona” [J]. Journal of Silk, 2015, 52(12): 37-42.
- [3] 刘光裕,任建涛. 基于客户画像的精准推荐[J]. 科技经济导刊, 2016(35): 22, 78.
- [4] 高寅欣,郭中梅. 运营商移动互联网用户大数据分析及应用[J]. 中国电信业, 2015(9): 80-83.
GAO Yinxin, GUO Zhongmei. Analysis and application of large data of mobile internet users [J]. China Telecommunication Trade, 2015(9): 80-83.
- [5] 王巍. 用户标签在基于用户画像的精准营销中的应用解析[J]. 经济管理(全文版), 2016(8): 00232.
WANG Wei. Application of user tag in precision marketing based on user portrait [J]. Economic Management Journal, 2016(8): 00232.
- [6] GOLDER S A, HUBERMAN B A. Usage patterns of collaborative tag systems [J]. Journal of Information Science, 2006, 32(2): 198-208.
- [7] 王翠英. 基于 Folksonomies 的用户偏好挖掘研究[J]. 现代图书情报技术, 2009, 25(6): 37-43.
WANG Cuiying. Research development of Folksonomies based user profiles mining [J]. New Technology of Library and Information Service, 2009, 25(6): 37-43.
- [8] TAHAR-RAFIK B, RACHID A O. Towards a new approach for generating user profile from folksonomies [C]// Proceedings of the 2014 4th International Symposium ISKO-Maghreb: Concepts and Tools for Knowledge Management. Piscataway, NJ, USA: IEEE, 2014: 1-6.
- [9] CAI Yi, LI Qing, XIE Haoren, et al. Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy [J]. Neural Network, 2014, 58(10): 98-110.
- [10] CAI Shubin, SUN Heng, GU Sishan, et al. Learning concept hierarchy from folksonomy [C]// Proceedings of the 2011 8th Web Information Systems and Applications Conference. Piscataway, NJ, USA: IEEE, 2011: 47-51.
- [11] SKILLEN K L, NUGENT C, DONNELLY M, et al. Using ontologies for managing user profiles in personalised mobile service delivery [M]// Health Monitoring and Personalized Feedback Using Multimedia Da-

- ta, Berlin, Germany: Springer, 2015: 245-264.
- [12] MALESZKA B. A method for ontology-based user profile adaptation in personalized document retrieval systems [C] // Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Piscataway, NJ, USA: IEEE, 2017: 003187-003192.
- [13] FERREIRA-SATLER M, ROMERO F P, OLIVAS J A, et al. Fuzzy ontology-based approach for automatic construction of user profiles [C] // Proceedings of the International Conference on Rough Sets and Current Trends in Computing. Berlin, Germany: Springer, 2014: 339-346.
- [14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [15] FARSEEV A, NIE L, AKBARI M, et al. Harvesting multiple sources for user profile learning: a big data study [C] // Proceedings of the ACM International Conference on Multimedia Retrieval. New York, USA: ACM, 2015: 235-242.
- [16] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C] // Proceedings of the International Conference on Machine Learning. Freiburg, Germany: IMLS, 2014: 2931-2939.

(编辑 武红江)

(上接第 105 页)

- [13] YAN J, LIU Y, WANG X, et al. An improved steady inverse method for turbomachinery aerodynamic design [J]. Inverse Problem in Science & Engineering, 2017, 25(5): 633-651.
- [14] 刘昭威, 吴虎, 梁言, 等. 多级轴流压气机反问题级间气动匹配设计方法 [J]. 推进技术, 2017, 38(9): 1987-1994.
- LIU Zhaowei, WU Hu, LIANG Yan, et al. Aerodynamic matching for multistage axial compressor based on inverse design method [J]. Journal Propulsion Technology, 2017, 38(9): 1987-1994.
- [15] 杜磊, 宁方飞. 一种基于共轭方程法求解黏性反问题的简化方法 [J]. 航空学报, 2012, 33(4): 579-606.
- DU Lei, NING Fangfei. An approximate method for viscous inverse design based on adjoint equations [J]. Acta Aeronautica and Astronautica Sinica, 2012, 33(4): 579-606.
- [16] NADARAJAH S K, JAMESON A. A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization [C] // 38th Aerospace Sciences Meeting and Exhibit. New York, USA: AIAA, 2000: 0667.

(编辑 赵炜)

(上接第 141 页)

- [7] CHILDS D W, AUTHUR S, MEHTA N J. The impact of hole-depth on the rotordynamic and leakage characteristics of hole-pattern-stator gas annular seals [C] // Proceedings of ASME Turbo Expo 2013. New York, USA: ASME, 2013: GT2013-94333.
- [8] MIGLIORINI P J, UNTAROIU A, WOOD H G. A numerical study on the influence of hole depth on the static and dynamic performance of hole-pattern seals [J]. ASME Journal of Tribology, 2015, 137(1): 259-263.
- [9] MIGLIORINI P J, UNTAROIU A, WOOD H G. A numerical study on the influence of hole aspect ratio on the performance characteristics of a hole-pattern seal [C] // Proceedings of ASME Turbo Expo 2014. New York, USA: ASME, 2014: GT2014-26581.
- [10] VANNARSDALL M, CHILDS D W. Static and rotor-dynamic characteristics for a new hole-pattern annular gas seal design incorporating larger diameter holes [J]. ASME Journal of Engineering for Gas Turbines & Power, 2014, 136(2): 022507.
- [11] CHILDS D W, SHIN Y S, SEIFERT B. A design to improve the effective damping characteristics of hole-pattern-stator annular gas seals [J]. ASME Journal of Engineering for Gas Turbines & Power, 2006, 130(1): 82-93.
- [12] NAYAK K C, DUTTA P. Numerical investigations for leakage and windage heating in straight-through labyrinth seals [J]. ASME Journal of Engineering for Gas Turbines & Power, 2016, 138(1): 545-550.
- [13] 晏鑫, 李军, 宋立明, 等. 蜂窝面迷宫密封内流动和总温升特性的研究 [J]. 西安交通大学学报, 2008, 42(7): 823-827.
- YAN Xin, LI Jun, SONG Liming, et al. Study on flow and total temperature increase of honeycomb labyrinth seals [J]. Journal of Xi'an Jiaotong University, 2008, 42(7): 823-827.

(编辑 赵炜)