

基于混合蚁群关联规则挖掘的危险源分析算法

余雅莉, 周 良

(南京航空航天大学 计算机科学与技术学院, 江苏 南京 210016)

摘 要:针对民航危险源原因分析中存在人工参与较多缺乏客观性的问题,设计了一种基于混合蚁群关联规则挖掘的危险源原因分析算法(HA-MACR),利用关联规则挖掘来探索危险源原因。该算法摒弃了传统关联规则挖掘算法重复扫描数据库导致挖掘效率较低及产生大量候选集、容易出现“组合爆炸”现象等缺点,将改进后的蚁群算法用于挖掘最大频繁项集,并由此产生质量较好的强关联规则,从而找到导致危险源的不安全事件。同时,为了避免蚁群的盲目性,混合了粒子群,借助粒子群确定蚁群的初始信息素浓度。通过上述改进,有效增强了算法的搜索能力,提高了关联规则挖掘的效率,且避免了算法陷入局部最优,从而使危险源原因分析更加快速、准确。

关键词:危险源原因分析;关联规则挖掘;蚁群算法;粒子群

中图分类号:TP183

文献标识码:A

文章编号:1673-629X(2018)11-0089-05

doi:10.3969/j.issn.1673-629X.2018.11.020

A Hazard Analysis Algorithm Based on Mixed Ant Colony Association Rules Mining

SHE Ya-li, ZHOU Liang

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, China)

Abstract: Aiming at the problem of much human participation and lack of objectivity in the hazard causes analysis of civil aviation, we design a hazard analysis algorithm based on mixed ant colony association rules mining which is used to explore the cause of hazard. This algorithm discards the disadvantages of repeated scanning database of traditional association rule mining algorithm, which leads to low mining efficiency, a large number of candidate sets and easy occurrence of “combined explosion”. It uses the improved ant colony algorithm to mine the maximal frequent item sets instead, and generates association rules with strong quality from them, thus finding unsafe incidents which lead to hazard by these rules. At the same time, in order to avoid the blindness of the ant colony, the particle swarm is mixed and the initial pheromone concentration of the ant colony is determined by the particle swarm. Through the above improvement, the search ability of the algorithm is effectively enhanced, the efficiency of the association rule mining is improved, and the algorithm is prevented from falling into a local optimum, so that the analysis of hazard cause is faster and more accurate.

Key words: hazards analysis; association rule mining; ant colony algorithm; particle swarm

0 引言

在民用航空空中交通安全管理体系中,危险源识别和风险评估是重要的组成部分。对危险源进行详细地分析并得到其产生的原因和作用机理,是相关部门对风险进行有效准确评估的前提。在传统的危险源分析体系中,使用事件树分析法、蝶形分析法、危险与可操作性分析法进行危险源的分析;现在,国内外专家和学者提出了很多不同的分析方法,但大都在传统的分

析体系上建立。Katrina Groth等提出了一种混合的分析方法,将事件树、故障树和事件序列图相结合对危险源中决定性因果路径进行分析^[1],并利用贝叶斯网络对非决定性因果关系进行分析;Leon Purton等在蝶形分析法的基础上^[2],引入技术完整性的概念,在危险链中引入技术生命周期:设计、生产和维护。这些方法可以从不同的侧面对危险源的原因进行分析,但是缺乏全面性。为此,需要一种能够较全面且客观分析危险

收稿日期:2018-01-10

修回日期:2018-05-15

网络出版时间:2018-06-29

基金项目:江苏省产学研联合创新资金项目(SBY201320423)

作者简介:余雅莉(1992-),女,硕士生,研究方向为信息系统及集成、机器学习等;周 良,博士,副教授,硕士生导师,研究方向为信息系统、知识工程。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180629.1707.072.html>

源的方法。

危险源原因分析的过程就是在危险源数据库中找到不安全事件 X , 使得 X 与危险源 Y 之间满足 $X \rightarrow Y$ 的蕴涵关系。根据关联规则的定义可知, 这恰恰就是在危险源数据库中对危险源与不安全事件进行关联规则挖掘的过程, 因此可以借助关联规则挖掘实现对危险源的原因分析。关联规则挖掘是数据挖掘中的一个重要领域, 它能够发现数据库中项目之间的隐含关系。自 1993 年被提出后, 人们提出了很多相关算法, 这些算法通过发现数据库中的频繁项目集来发现关联规则^[3]。Apriori 算法^[4]、频繁模式树算法^[5]等都是传统的经典关联规则挖掘算法, 但它们存在频繁扫描数据库、产生大量候选集等不可避免的缺陷。因此, 许多改进算法被提出。例如, Du 等^[6]提出了一种改进的 Apriori 算法, 通过重构数据库中的数据, 增强项目集之间的联系, 使得算法能够直接通过剪枝操作得到候选集, 从而减少对候选集的验证过程。不仅如此, 具有较高搜索效率的群智能算法也被应用到关联规则的挖掘中, Jitendra Agrawa 等利用基于集的粒子群优化算法挖掘数据库中的正关联规则和负关联规则^[7]; Zhou 等将粒子群优化算法与引力搜索算法相结合挖掘关联规则^[8]; 文献^[9]提出结合遗传算法和蚁群算法进行关联规则挖掘, 其中心思想是利用遗传算法优化蚁群算法的参数。该方法确实取得了较好的性能, 但是蚁群算法的初始信息素浓度仍然没有确定, 使蚂蚁在初始时刻没有方向, 消耗较多时间。

随着混合智能算法的兴起, 加上蚁群算法易于与其他算法结合的特点, 文中设计了一种基于混合蚁群关联规则挖掘的危险源分析方法。在该方法中将危险源事务数据库进行预处理, 转换为一个二进制矩阵, 利用粒子群挖掘出定量的频繁项集, 从而确定蚁群的初始信息素浓度, 进而进行最大频繁项集挖掘, 并由最大频繁项集产生强关联规则, 分析危险源原因。

1 HA-MACR 算法总体思路

在国际民航组织安全管理手册和国内民航空中交通管理安全管理体系 (SMS) 建设指导手册中, 危险源的原因分析都是对不安全事件的分析, 在人为因素分析的基础上, 对不安全行为、不安全行为的前提条件、设备设施、监督管理和组织因素等方面进行分析。关联规则有如下定义: 设 $I = \{i_1, i_2, \dots, i_n\}$ 是一个项目集, 事务数据库为 D , 事务数据库中的每个事务 T 都是项目集的一个子集 $T \subseteq I$, 关联规则具有如下形式 $X \rightarrow Y$, 其中 $X \subseteq I$, $Y \subseteq I$ 并且 $X \cap Y = \emptyset$ 。由定义可以看出, 关联规则就是发现事务数据库中项目的隐含模式, 当某些项目 X 出现时, 另一组项目 Y 也会同时出

现。这个过程可以与危险源的原因分析相结合, 因此, 文中提出利用关联规则分析危险源原因。

HA-MACR 算法将该分析过程分为 3 部分: 一是危险源状态信息预处理过程; 二是挖掘最大频繁项集; 三是产生强关联规则从而得到导致危险源的不安全事件。关联规则挖掘过程主要分为两个步骤——发现频繁项目集和产生强关联规则^[10]。其中, 发现频繁项目集是关键, 对算法的性能有着决定性的影响, 故考虑在该步骤进行优化。于是, 在第二步中先由粒子群找出固定数量的频繁项集, 从而确定蚁群算法的初始信息素浓度, 以期改善单纯用蚁群算法进行挖掘的盲目性, 再利用蚁群算法挖掘最大频繁项集。整个算法流程如图 1 所示。

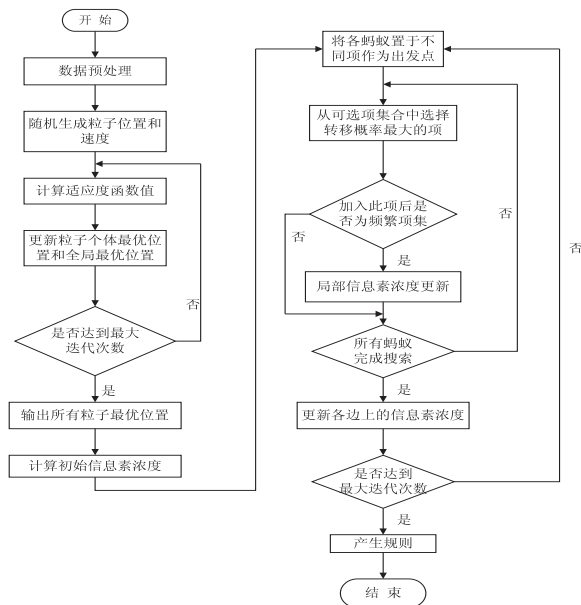


图 1 HA-MACR 算法流程

1.1 数据预处理

算法开始前, 需要对原始数据进行预处理, 将其转换为一个二进制矩阵。在危险源原因分析中, 给定事务数据库为包含危险源和不安全事件的危险源事务数据库, 将其记作 D 。在预处理阶段, 将 D 中每条记录转换成 0-1 向量, 这样便于计算支持度与置信度, 同时可以提高数据库扫描速度。下面以图 2 为例具体解释转换方法。图 2 左边是原始数据, 共有 4 条记录, 可以看出数据库中共有五个项目集, 向量长度为 5。以 T_1 为例, 它包含 I_1, I_2 和 I_5 , 因此, I_1, I_2 和 I_5 对应位置 1, I_3 和 I_4 对应位置 0。以此类推, 最终构成一个 4 行 5 列的二进制矩阵。

T_1	I_1	I_2	I_3	I_4	I_5
T_2	I_1	I_3	I_4	I_5	
T_3	I_1	I_2			
T_4	I_2	I_3			

	I_1	I_2	I_3	I_4	I_5
B_1	1	1	0	0	1
B_2	1	0	1	1	1
B_3	1	1	0	0	0
B_4	0	1	0	0	1

图 2 数据预处理

1.2 信息素初始浓度确定

蚁群算法的参数确定一直是其应用中需要解决的问题,尤其是初始信息素浓度的设置对算法性能有着重要影响。为了在一定程度上避免盲目性和提升搜索效率,HA-MACR 算法首先借助粒子群的更新迭代,得到一定量的频繁项集,从而更合理地设置本算法的初始信息素浓度。将项集看作粒子群中的粒子,对应于数据预处理后的二进制矩阵,把粒子位置设为 0-1 向量。随机生成粒子的初始位置和速度,但需要注意的是,必须保证粒子代表的项集是频繁的,若不是则需要重新生成。

综合考虑频繁项集的项目个数和支持度计数两个因素,粒子的适应度函数设计如下:

$$H(x) = W_1 \cdot |x| + W_2 \cdot \frac{\text{count}(x)}{\text{min_sup}} \quad (1)$$

其中, $|x|$ 为粒子向量维数,即对应项集的项目个数; $\text{count}(x)$ 表示该项集的支持度计数;参数 W_1 、 W_2 分别用于控制这两个因素对于适应度函数的影响程度。

由于粒子群算法中每个粒子都有记忆功能,可以记忆自己从初始时刻到当前时刻适应度最好时的位置。个体极值就是粒子彼时的适应度函数值,而全局极值就是所有粒子个体极值中的最优值。按式 1 计算粒子适应度函数并更新粒子个体极值 x_{pbest} 和全局极值 x_{gbest} 。有了个体最优位置和全局最优位置,就可以更新 t 时刻粒子位置,公式为:

$$v^j(t+1) = w \cdot v^j(t) + k_1 r_1 [x_{\text{pbest}}^j(t) - x^j(t)] + k_2 r_2 [x_{\text{gbest}}(t) - x^j(t)] \quad (2)$$

$$x^j(t+1) = x^j(t) + v^j(t+1) \quad (3)$$

其中, $v^j(t)$ 和 $x^j(t)$ 分别是第 j 个粒子迭代 t 次后的速度和位置; w 是惯性系数,随迭代次数增加而减小; k_1 、 k_2 是学习因子。

但是,这样得到的位置不是 0-1 向量,所以还需要用式 4 进行转换。

$$l(t)^j = \begin{cases} 1, r_3 < \text{sig}(l(t)^j) \\ 0, r_3 \geq \text{sig}(l(t)^j) \end{cases} \quad (4)$$

其中, r_3 为 $[0,1]$ 之间均匀分布的随机数; $\text{sig}()$ 为 sigmoid 函数。

重复上述过程,直至迭代结束或达到最大迭代次数,输出所有 $x_{\text{gbest}}(t)$ 。由此计算危险源和不安全事件组成的项集中项集 i 、 j 同时被选择的次数,即 2-项集支持度计数,由此确定信息素初始浓度,如式 5 所示。

$$\tau_{ij}(0) = \tau_{\max} - \tau_{\min} + \text{sc}_{ij} \quad (5)$$

其中, τ_{\max} 和 τ_{\min} 分别表示信息素上限和信息素下限。 万方数据

1.3 挖掘最大频繁项目集

在上一阶段,主要完成了对蚁群算法的初始信息素浓度的确定,接下来可以开始用针对危险源原因分析改进的蚁群算法挖掘关联规则中的最大频繁项目集。类似于用蚁群算法寻找最优路径^[11]的过程,用一个禁忌表(Tabu)存放蚂蚁在后续搜索中不能选择的数据项,用一个可选表(Allowed)存放它还可以选择的数据项。初始时刻,将各个蚂蚁随机地置于不同数据项作为出发点,并将该数据项添加到存放已选数据项的集合 f_k 中和禁忌表中。然后,计算出剩余所有可选数据项的概率转移函数值,如式 6 所示,并据此选择一个项。

$$P_j^k(t) = \begin{cases} \frac{\tau_j(t)^\alpha \eta_j(t)^\beta}{\sum_j \tau_j(t)^\alpha \eta_j(t)^\beta}, j \in \text{allowed}_k \\ 0, j \in \text{tabu}_k \end{cases} \quad (6)$$

其中, $\tau_j(t)$ 是 t 时刻的信息素浓度; $\eta_j(t)$ 为启发函数。

选择 $P_j^k(t)$ 最大值对应的数据项 j , 将 j 加入 f_k , 判断 f_k 是否仍为频繁项集。若是则按式 7~9 更新局部信息素浓度,同时在禁忌表中加入 j , 从 allowed 表中删除 j , 然后开始下一轮搜索;

$$\tau_{ij}(t) = (1 - \varepsilon) \cdot \tau_{ij}(t) + \varepsilon \cdot \Delta\tau_{ij}(t) \quad (7)$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^m \Delta\tau_{ij}(t)^k \quad (8)$$

$$\Delta\tau_{ij}(t)^k = \begin{cases} Q, \text{点 } i, j \text{ 在蚂蚁 } k \text{ 的 } f_k \text{ 集合中} \\ 0, \text{其他} \end{cases} \quad (9)$$

若加入 j 后 f_k 不是频繁项集,则把 j 从 f_k 中删除,并根据式 10 决定是否继续搜索。

$$\text{stop}(k) = \begin{cases} 1, p > p_0 \\ 0, \text{otherwise} \end{cases} \quad (10)$$

其中, p 是 $[0,1]$ 上的随机数; p_0 是 $[0,1]$ 间的一固定值。若 $\text{stop}(k) = 1$, 则停止搜索;若 $\text{stop}(k) = 0$, 将项目集 j 放入禁忌表并从 allowed 表中剔除后继续搜索。当所有的蚂蚁都完成了一次遍历后,记录本次遍历所找到的最大频繁项集。这样就完成了一次迭代。

但是在下一轮迭代开始之前还有一项很重要的工作要做,那就是信息素的全局更新,这正是蚁群算法具有正反馈性的关键。蚂蚁完成一次搜索后,对于本次迭代中最优频繁项集中的点所构成的边进行信息素更新^[12-13],方式如下:

$$\tau_{ij}(t+1) = (1 - \rho) \cdot \tau_{ij}(t) + \rho \cdot \Delta\tau_{ij}(t) \quad (11)$$

$$\Delta\tau_{ij}(t) = \begin{cases} \frac{Q}{L_{\text{best}}}, \text{若 } i, j \text{ 在频繁项集中} \\ 0, \text{其他} \end{cases}$$

1.4 产生规则

得到最大频繁项集后,便可以很容易地产生其对

应的强关联规则,主要分为两步:

- (1) 求出最大频繁项集的所有非空真子集;
- (2) 设 X,Y 为步骤 1 中所得任意子集,若 $X \cap Y = \emptyset$ 且 $\text{count}(X \cup Y)/\text{count}(X) \geq \text{min_conf}$,则有强关联规则 $X \rightarrow Y$ 。

2 实验结果与分析

实验采用民航空管局安全管理系统的危险源事务数据库,并按 1.1 节所述数据预处理过程进行二进制转换。从算法运行时间和产生的规则质量两个角度对算法进行对比实验,实验环境为 Intel core i7 2.6 GHz Window7 及 MatLab 2014Ra。为了保证算法初始随机性对结果的影响,取 20 次实验的平均结果进行衡量。运行时间很好衡量,而规则质量,参考文献[14]中对规则质量的计算方法,比值越大,质量越好。最终得到的危险源原因分析结果如表 1 所示。

表 1 危险源原因分析结果

危险源	危险源原因
陆空通信失效	设备故障,注意力分配不当,管制移交不及时,无线电干扰,未使用标准术语
小于最小飞行间隔	管制扇区流量过大,飞行冲突意识差,飞行计划不合理,错误指挥,进程单填写错误
小于最小飞行高度	监控不利,管制移交不及时,雷达信号失效,管制员疲劳上岗
无线电干扰	空域划设不合理,错误指挥,导航台工作不正常
飞行冲突意识差	个人准备差

由于 HA-MACR 算法中有一些参数不确定,这些参数对算法性能有一定程度的影响,故在不同取值下分别对算法执行 20 次,选取平均结果较好的参数组合,实验结果如表 2 所示。由结果可知,当 $\alpha = 2, \beta = 1$ 时,算法效果较好,故以下实验将设为该值。

表 2 参数组合

α 的取值	β 的取值	规则质量
2	2	0.692
2	1.5	0.707
2	1	0.719
1.5	2	0.671
1.5	1.5	0.677
1.5	1	0.706
1	2	0.682
1	1.5	0.709
1	1	0.712

同时根据文献[15-16],算法其余参数设置如表 3 所示。 万方数据

表 3 参数取值

参数	取值
粒子个数	50
蚂蚁个数	100
迭代次数	20,50
最小支持度	0.3,0.4,0.5,0.6,0.7
最小置信度	0.6

实验比较了不同支持度下 Apriori 算法和 HA-MACR 算法的平均规则质量,结果如图 3 所示。可以看出,HA-MACR 产生的规则质量均较好。

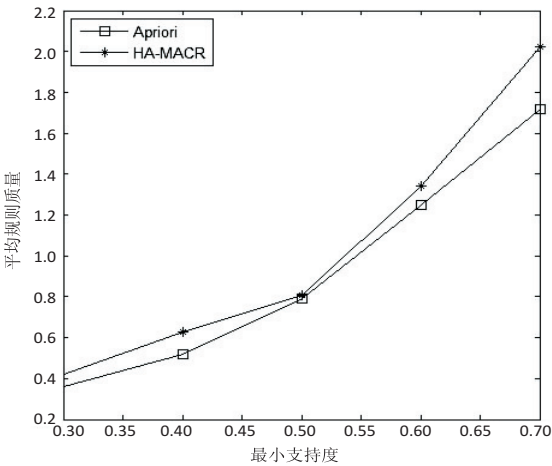


图 3 规则质量比较

另外,实验比较了不同支持度下 Apriori 算法和 HA-MACR 算法的平均运行时间,结果如图 4 所示。可以看出,HA-MACR 算法运行速度远远快于 Apriori 算法,而且 Apriori 算法随支持度降低运行时间大幅上涨,变化很大,这是由于产生的候选项集增多引起的计算量增大。而 HA-MACR 算法由于先利用粒子群合理设置了信息素浓度,而后蚁群算法挖掘最大频繁项集时也受到信息素的指导,所以不会出现这种情况,运行时间较稳定,即使支持度阈值设置得很小也不会出现运行时间剧增的情况。

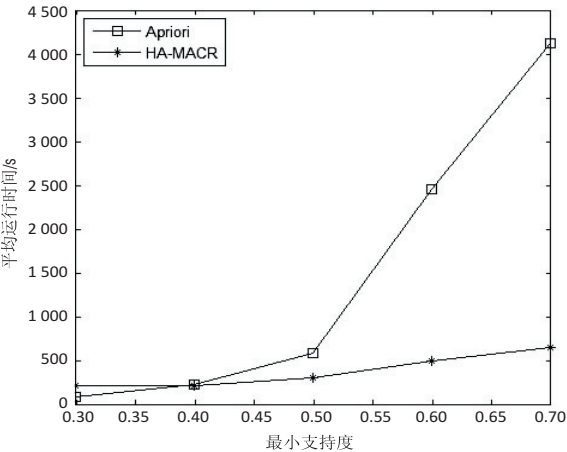


图 4 运行时间比较

3 结束语

针对危险源原因分析中存在的分析主要依赖人的参与和传统关联规则挖掘算法中候选集过多影响挖掘效率的问题,提出了一种基于混合蚁群关联规则挖掘的危险源原因分析算法 HA-MACR。该算法通过引入粒子群优化蚁群初始信息素浓度,根据频繁项集来确定,避免了蚁群算法初始时的盲目性,有效提高了危险源原因分析的效率与质量。实验结果表明,该算法充分利用了蚁群优化算法的寻优能力和正反馈性,将其应用于对危险源原因的分析过程中,通过挖掘到的最大频繁项集产生质量较好的关联规则,进而得到分析结果,不仅提高了关联规则的挖掘效率,同时也能有效提高危险源原因分析的准确性和执行效率。

参考文献:

[1] GROTH K, WANG Chengdong, MOSLEH A. Hybrid causal methodology and software platform for probabilistic risk assessment and safety monitoring of socio-technical systems [J]. Reliability Engineering & System Safety, 2010, 95(12): 1276-1285.

[2] PURTON L, CLOTHIER R, KOUROUSIS K. Assessment of technical airworthiness in military aviation; implementation and further advancement of the bow-tie model [J]. Procedia Engineering, 2014, 80: 529-544.

[3] 赵阳, 吴廖丹. 一种自底向上的最大频繁项集挖掘方法 [J]. 计算机技术与发展, 2017, 27(8): 57-60.

[4] 周发超, 王志坚, 叶枫, 等. 关联规则挖掘算法 Apriori 的研究改进 [J]. 计算机科学与探索, 2015, 9(9): 1075-1083.

[5] 郭进伟, 皮建勇. 一种基于 FP-growth 的并行 SON 算法的实现 [J]. 微型机与应用, 2014, 33(8): 60-63.

[6] DU Jiaoling, ZHANG Xiangli, ZHANG Hongmei, et al. Research and improvement of apriori algorithm [C]//Proceedings 2nd international workshop on intelligent systems and

applications. Dalian, China: IEEE, 2010.

[7] AGRAWAL J, AGRAWAL S, SINGHAI A, et al. SET-PSO-based approach for mining positive and negative association rules [J]. Knowledge and Information Systems, 2015, 45(2): 453-471.

[8] ZHOU Zhiping, ZHANG Daowen, SUN Ziwen, et al. An adaptive hybrid PSO and GSA algorithm for association rules mining [C]//International conference on cloud computing and security. [s. l.]: Springer International Publishing, 2015: 469-479.

[9] 罗茜. 遗传模拟退火算法挖掘关联规则的应用 [D]. 广州: 中山大学, 2010.

[10] 黄红星, 王秀丽, 黄习培. 挖掘最大频繁项集的改进蚁群算法 [J]. 计算机工程与应用, 2011, 47(13): 161-165.

[11] 陈昌敏, 谢维成, 范颂颂. 自适应和最大最小蚁群算法的物流车辆路径优化比较 [J]. 西华大学学报: 自然科学版, 2011, 30(3): 5-8.

[12] 许凯波, 鲁海燕, 程毕芸, 等. 求解 TSP 的改进信息素二次更新与局部优化蚁群算法 [J]. 计算机应用, 2017, 37(6): 1686-1691.

[13] LAI M, TONG X. A metaheuristic method for vehicle routing problem based on improved ant colony optimization and Tabu search [J]. Journal of Industrial & Management Optimization, 2017, 8(2): 469-484.

[14] 杨程, 范强, 王涛, 等. 基于多维特征的开源项目个性化推荐方法 [J]. 软件学报, 2017, 28(6): 1357-1372.

[15] SARIEFF N B, BUNYAMIN N. Comparative study of genetic algorithm and ant colony optimization algorithm performances for robot path planning in global static environments of different complexities [C]//IEEE international symposium on computational intelligence in robotics and automation. Daejeon, South Korea: IEEE, 2016: 132-137.

[16] 周 晓, 葛洪伟, 苏树智. 基于信息素的自适应连续域混合蚁群算法 [J]. 计算机工程与应用, 2017, 53(6): 156-161.

(上接第 88 页)

(2): 142-149.

[7] 孟祥武, 纪威宇, 张玉洁. 大数据环境下的推荐系统 [J]. 北京邮电大学学报, 2015, 38(2): 1-15.

[8] 国琳, 左万利. 基于兴趣图谱的用户兴趣分布分析及专家发现 [J]. 电子学报, 2015, 43(8): 1561-1567.

[9] 王立才, 孟祥武, 张玉洁. 上下文感知推荐系统 [J]. 软件学报, 2012, 23(1): 1-20.

[10] 刘平峰, 朱孔真, 杨柳, 等. 基于用户兴趣图谱的个性化推荐系统设计 [J]. 武汉理工大学学报: 信息与管理工程版, 2014, 36(3): 341-344.

[11] PESSEMIER T D, DHONDT J, MARTENS L. Hybrid group recommendations for a travel service [J]. Multimedia Tools

& Applications, 2017, 76(2): 2787-2811.

[12] OH J, PARK S, YU H, et al. Novel recommendation based on personal popularity tendency [C]//Proceedings of the 2011 IEEE 11th international conference on data mining. [s. l.]: IEEE, 2011: 507-516.

[13] LAI Siwei, XIANG Liang, DIAO Rui, et al. Hybrid recommendation models for binary user preference prediction problem [J]. Journal of Machine Learning Research, 2012, 18: 137-151.

[14] ADOMAVICIUS G, KWON Y O. Improving aggregate recommendation diversity using ranking-based techniques [J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(5): 896-911.