基于领域本体的用户兴趣模型构建方法研究

魏 同,李绍稳,耿凡凡,孔 晨
(安徽农业大学 信息与计算机学院,安徽 合肥 230036)

摘 要:现阶段的电子商务个性化推荐系统中,存在构建用户兴趣模型过程复杂、繁琐,蕴含的语义信息要素较少或者不完整等问题,研究基于本体的用户模型构建方法已十分迫切和必要。为此,提出了一种基于本体投影算法和概念兴趣度结合的用户兴趣模型构建方法。通过抽取数据库中商品的属性和特征值,对抽取后的属性和特征值进行处理,采用 OWL 语言表述方法手动构建茶叶领域本体,进而在此基础上采用投影算法生成用户兴趣本体;充分运用本体中的概念、属性以及实例描述用户兴趣,从语义层面解读用户个人兴趣,从而达到在个性化推荐中提高结果精度的目标。实验结果表明,该方法易于构建用户模型,且模型中的语义要素丰富,使用该模型进行推荐的精确度有所提升。

关键词:本体;投影;用户模型;茶叶
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)03-0065-05
doi:10.3969/j.issn.1673-629X.2017.03.014

Investigation on Constructed Method of User Interest Model with Domain Ontology

WEI Tong, LI Shao-wen, GENG Fan-fan, KONG Chen
(School of Information and Computer Science, Anhui Agriculture University, Hefei 230036, China)

Abstract: There are many problems in electronic commerce personalized recommendation system nowadays, such as complicated and cumbersome process in building user interest model, less and/or incomplete semantic information elements etc. So a method to construct user interest model has been presented, actually synthesis of ontology projection algorithm and concept interest degree. The tea domain ontology has been established via extracting the attribute and the characteristic value of the goods in the database and OWL language, in which user interest ontology is produced with projection algorithm. Concept, attributes and instances of the ontology to describe the user's interests, which interprets the user's personal interest from the semantic level to achieve higher accurate results in process of personalized recommendation. The experimental results show that the proposed method is prone to establish user model and semantic elements inside user interest model are rich as well as accuracy of the results has been promoted with the model established.

Key words: ontology; projection; user model; tea

0 引 言

随着网络信息资源爆炸式增长和目前信息服务机制的不健全,比如未考虑用户之间兴趣差异、缺乏语义要素支持等而导致的信息迷失、信息过载等问题日益严重,用户精确、及时和智能地筛选出与兴趣相关信息的难度越来越大[1]。在这种形势下,个性化服务应运而生,而其中的关键点是用户兴趣模型的构建[2]。

至今已知的一些用户模型构建方法主要有基于神经网络、向量空间、评价矩阵和本体等[3]。然而,基于神经网络的模型理解与使用不易、适用范围较窄;基于向量空间的模型不稳定,导致结果有一定偏差;基于评价矩阵的用户模型适应能力较差以至于难以及时地对兴趣进行更新;基于本体的用户模型相比前面几种相对较好,可以在语义层面上通过领域本体较为精确地表示用户个性化兴趣,但也有以下两点不足:

(1)大多数学者只考虑本体概念间的分类关系,很少考虑如同位关系、属性关系等非分类关系,导致模型中语义信息不完整,不能充分利用。

收稿日期:2016-05-05 **修回日期**:2016-09-01 **网络出版时间**:2017-02-17
基金项目:国家自然科学基金资助项目(31271615)
作者简介:魏 同(1991-),男,硕士研究生,研究方向为人工智能、个性化推荐;李绍稳,教授,博导,通讯作者,研究方向为人工智能、农业信息化。

(2)目前用户本体兴趣模型的更新方式大多是使用兴趣本体归并参考本体进行更新,但是本体归并过程中概念的上下位关系以及属性、实例等关系会发生错位或者遗失等问题,使得用户兴趣本体结构不完整,不能完全解读用户的需求。

针对以上问题,提出了一种改进的基于领域本体的用户兴趣模型构建方法。该方法根据《中国茶叶大辞典》以及《中国名优茶选集》结合领域专家的意见手动构建茶叶领域本体,在其中加入属性关系和同位关系等非分类关系完善其中的语义信息;通过 PC 端和移动终端获取的用户数据从领域本体投影产生用户本体;结合收集到的用户兴趣数据进行处理,结合文中提出的概念兴趣度计算公式和概念属性权重计算公式对用户兴趣模型进行初始化。实验结果表明,推荐结果对于用户兴趣的精确性有显著提高。

1 茶叶领域本体的构建

1.1 本体理论

本体被广泛引用的定义是 Gruber 提出的"本体是概念模型的明确的规范说明[4]"。一般来说,本体是用来描述某个领域中的概念以及概念之间的关系,使得这些概念和关系在一定的共享范围内具有大家共同认可的、明确的、唯一的定义[5]。

本体的 5 个基本的建模元语分别是:概念(Concept)、关系(Relation)、函数(Function)、公理(Axiom)和实例(Instance)。其结构可表示成一个五元组:ont = $\{C,R,F,A,I\}$。其中的 C、R、F、A 和 I 对应本体中概念、概念间关系、函数、公理和实例的集合[2]。

1.2 茶叶领域本体

构建领域本体就是使用手动或半自动构建方法生成在应用某一领域的本体。手动构建通常由个人完成,内容较为完善,但是其中会带有构建者的个人观点且工作较为繁琐;而半自动构建则是综合了领域专家和数据挖掘的结果,通过挖掘庞大的数据来获得相应的领域名词,工作量也较大,数据的完整和全面性会对领域本体造成一定的局限和影响。

文中采取的方法是根据《中国茶叶大辞典》以及《中国名优茶选集》结合领域专家的意见手动构建茶叶领域本体,在其中加入属性关系和同位关系等非分类关系,完善其中的语义信息要素。其中属性关系包含了数据属性(Data Property)和对象属性(Object Property),它是本体概念间重要的语义表示方法[6]。在茶叶领域本体中定义的数据属性有 hasPrice、hasLevel 等,对象属性有 Be-Made-In、Provide 等。除此之外,领域本体中包含了大量的实例,如茶叶品牌、茶叶商家等。茶叶领域本体的层次结构如图 1 所示。

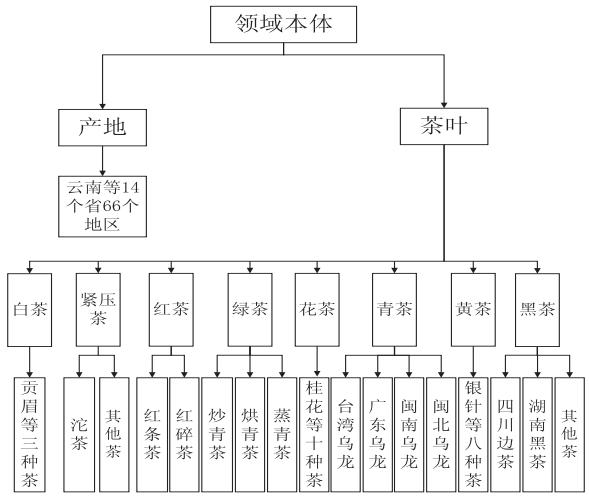

图 1 茶叶领域本体层次结构图

2 基于茶叶领域本体的用户建模

2.1 用户兴趣模型的表示

用户本体兴趣模型目前没有确切的定义,但是综合大多数学者意见,笔者认为就是通过分析用自然语言表示的用户兴趣并将其中的语义之间的关联转化为本体各概念间关系[2,7],利用本体中的概念间分类关系、非分类关系以及学习推理能力,将用户需求进行概念化、层次化、结构化的转化,达到发现用户新的兴趣,在语义层面上对信息进行表示和储存的目的[2,8]。

文中的用户兴趣模型可以明确表示为以下四元组:UserModel = (Userinfo, UserOnto, UserInt, UserTime)。其中用户的个人基本信息 UserInfo = {ID, Name, Age, Sex, Conlevel},分别描述了用户编号、姓名、年龄、性别、消费水平等信息。而用户兴趣本体 UserOnto = $\{C, R_C, R_N, F_C, A, I\}$ 中,C 表示是本体中用户兴趣概念;R_C 描述了用户兴趣本体中概念间的分类关系;R_N 描述了用户兴趣本体中的非分类关系(属性、同位等);F_C 表示函数;A 表示公理;I 表示实例。

UserInt = $\{D_C, D_P\}$ 是兴趣度的形式化描述,D_C 表示用户对特定概念的喜好程度的量化即兴趣度,且 $D_C \in [0,1]$;D_P 描述了概念中属性的权重,包括数据属性、对象属性等,而且文中定义对于同一概念 C 中全部 D_P 之和为 1。UserTime = {CreatTime, RecentTime} 表达了用户概念的时间集,CreatTime 表示概念创建的时间,RecentTime 表示概念及其所含项目(实例等)最近一次的被访问时间。

2.2 用户兴趣本体的构建

用户兴趣本体的构建需要采集用户行为数据,然后根据行为数据从领域本体通过投影算法生成用户兴趣本体。用户兴趣数据的采集大多通过对用户各种行为记录进行分析和处理。而其中主要方法有以下三

种:参考类,如点击链接;保存类,如下载、收藏等;审阅类,如页面停留时间、滚动条拖动次数、页面点击频率等。然而这三种方法都需要对网页的主题进行识别,即利用分词的方法使网页与本体中的概念相对应。本体投影就是领域本体投影于各种不同用户信息之上生成用户兴趣本体的过程,它是生成用户兴趣本体的重要方法和必要过程。文中的茶叶领域本体通过对各种不同的信息描述进行投影生成不同的投影面,进而生成用户兴趣本体[9],此过程如图 2 所示。

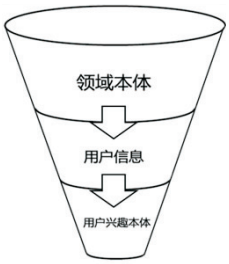

图 2　本体投影关系图

本体投影中的关键与重点就是如何创建关键词与本体中概念的关联关系,文中采用的是经典的 BPM-BM 算法[10]。文中所涉及中文本体中的概念为中文词汇,经过分词后得出中文词汇和短语,利用 BPM-BM 算法对网页分词表中的词汇和本体中概念进行匹配,本体投影算法如下:

Step1:将网页分词表中的数据赋值给字符串数组 $S[n]$,并新建栈 Z 初始化。

Step2:输出起始概念顶点(领域本体根节点),将起始顶点改为"已访问"标志,并使起始概念顶点进栈。

Step3:重复下列操作直到栈 Z 为空。

Step3.1:读取栈 Z 顶元素顶点(不出栈)。

Step3.2:若存在栈 Z 顶元素顶点未被访问的邻接点 W,则进行以下操作:

Step3.2.1:依次比较 W 与 $S[n]$ 中的各元素,若 W 与 $S[i]$ 二者匹配,则将网页作为实例加入用户本体,并将兴趣度作为数据属性加入本体初始化为 0;若二者不匹配,则继续比较 W 与 $S[i+1]$,当 $i=n$ 时 break;

Step3.2.2:将顶点 W 改为"已访问"标志;

Step3.2.3:将顶点 W 进栈。

Step3.3:否则,当前顶点退栈。

3　用户模型的学习与更新

文中概念兴趣度取决于用户本体(UserOnto)中的各种概念属性等,因此提取网页中的概念所对应的关键词是获取概念兴趣度中一项十分重要的前期工作。而网页中的内容经过分词以后所得的词语重要性各自不一,需要对提取出的词汇进一步分析,从而选择出可以代表该网页的关键词,同时也有利于降低网页的维度,为下一步的存储和分析等工作打下良好的基础。

文中采用 TF-IDF(Term Frequency - Inverse Document Frequency),"词频-逆文档频率"用来量化处理后得出关键词并得出其权重,从中选择大于预先设定阈值的词语作为特征词汇。TF-IDF 是一种用于信息搜索和信息挖掘的常用加权技术,并且被广泛应用在搜索、文献分类和其他相关领域。

TF-IDF 是一种统计方法,用来评估一个语料库或者文档集中一份文件的重要程度。假设特定词语在一个文档中出现的次数较多,同时出现在其余文档中的次数较少,则认为该词语可以用来分类,能够很好地区分不同类型的文档[11]。词语权重公式如下:

$$W(t_i,d_j)=\frac{\mathrm{tf}(t_i,d_j)\times\log_2\left(\frac{N}{n_i}+0.01\right)}{\sqrt{\sum_{j=1}^{N}\left[\mathrm{tf}(t_i,d_j)\times\log_2\left(\frac{N}{n_i}+0.01\right)\right]^2}} \tag{1}$$

其中,分母是归一化法中的归一化因子。$W(t_i,d_j)$ 表示在文档 d_j 中词语 t_i 的权重;$\mathrm{tf}(t_i,d_j)$ 表示在文档 d_j 中词语 t_i 的出现频率;N 表示文档的总数;n_i 表示在文档集中有词语 t_i 出现的文档数量。

通过公式选择出权重最大的词语后,将该词语作为该文档的代表词汇,并且与用户本体中的概念节点进行匹配,然后写入到该概念的实例之中。

3.1　概念兴趣度的学习

文中用户模型的学习与更新由采集和解析用户的行为来实现。而用户的行为数据如前文所述大致归为三种:参考、保存和审阅类。这三类行为主要都发生在用户的搜索和浏览的过程中,可以通过用户的行为数据确定其短期兴趣,然后与用户的长期兴趣相结合进而完成用户模型的学习与更新。

相关理论研究发现,人类记忆分为长期、短期和感觉记忆三种。当外部刺激作用于认得感觉器官时或产生感觉记忆,进而储存得到短期记忆,而经过一系列复杂的条件,短期记忆可以转化为长期记忆。因此文中的用户兴趣度计算公式充分考虑了用户的长期记忆和短期记忆以及用户行为数据,如下所示:

$$I=I_{\mathrm{o}}\times F_{(i)}+I_{\mathrm{N}} \tag{2}$$

其中,I 为用户兴趣度;I_{o} 为原始用户兴趣度;$F_{(i)}$ 为遗忘函数[8];I_{N} 为用户浏览新页面后产生的兴趣度变化值。

遗忘函数为:

$$F_{(i)}=e^{-\frac{\log_2\times\left|(T_{\mathrm{n}}-T_{\mathrm{v}})-S\right|}{S}(T_{\mathrm{n}}-T_{\mathrm{c}})} \tag{3}$$

其中，T_n 为当天日期；T_v 为用户本体中概念节点最近的访问时间；T_c 为用户本体中概念节点的创建时间；S 为生命周期参数，由于人的记忆在接触新知识一周后便开始衰弱，所以一般将 S 设为 7。从式中可以得出 $F_{(i)}$ 的范围是(0,1)。

在式(2)中，对影响 I_N 的因素则提出三点假设：

(1)用户对于关键词搜索的次数 S_N 和页面数量 P_N，搜索的次数越多，说明用户对这种产品主动了解程度越感兴趣，此时页面数量 P_N 与兴趣度成正比。

(2)用户消耗在页面的时间 T 及其长度 L，若长度相同，时间与兴趣成正比；若时间相同，长度与兴趣成反比。

(3)用户在某页面发生交互行为次数 C_N，很明显如果用户在网页中点击链接而进入的页面越多，和拉动滚动条次数与兴趣度成正比。

综合以上三点得出的概念兴趣度变化如式(4)：

$$I_N = W_1 \times f_1(S_N, P_N) + W_2 \times f_2(T, L) + W_3 \times f_3(C_N) \tag{4}$$

其中，f_1、f_2、f_3 是将前文的三个因素对兴趣度的影响进行量化得出的三个函数，分别表示为：

$$f_1 = 10 - (100/S_N \times P_N) \tag{5}$$

$$f_2 = 10^{-\lg\frac{L}{T}} \tag{6}$$

$$f_3 = 10^{-\frac{20}{C_N}} \tag{7}$$

其中，f_1 说明了搜索次数和页面数量与兴趣度的关系，其中分子 100 代表两个页面存在的商品数目，各电商网站每个页面含有 50 ~ 52 个左右，文中默认每个页面为 50 个商品，普遍情况下用户为寻找合适的商品一般浏览两个页面的情况占大多数；f_2 解释了停留时间和页面长度对兴趣度的影响，其中 L 的单位是字节数，T 的单位是 s；f_3 表示了交互行为与兴趣度之间的关系，其中 C_N 为交互行为的总量，分子中常数 20 则是按照统计数据，统计出大多数用户与物品的交互次数。通过以上三个公式得出 f_1、f_2、f_3 的值均在[0,1]之间。

W_1，W_2，W_3 分别是它们在兴趣度变化量中的权重，而且三者的关系满足条件：

$$W_1 + W_2 + W_3 = 1 \tag{8}$$

因此，从式(4)中可以计算出兴趣度的改变值，而且 I_N 的范围同样为[0,1]。进而可以通过式(2)计算出兴趣度 I 的值，并且 I 的范围始终为[0,1]。

3.2 概念属性权重的学习

领域本体中每个概念含有许多属性，也可以分为数据和对象两种属性，其中每个属性对应着不同的实例集合，而茶叶领域本体定义了很多不同的属性。文中的用户模型赋予了属性不同的权重，并且这些权重会随着用户的查询、浏览等交互行为而不断地学习与更新[12]。概念中属性权重的计算公式为[13]：

$$\mathrm{degree}_{(t)}(A_c) = \frac{w \times \mathrm{degree}_{(t-1)}(A_c) + \mathrm{freq}(A_c)}{w + \sum_{A_c} \mathrm{freq}(A_c)} \tag{9}$$

其中，$\mathrm{degree}_{(t)}(A_c)$ 为属性 A 在时间 t 时的权重；$\mathrm{degree}_{(t-1)}(A_c)$ 为属性 A 在时间 $t-1$ 时的权重；$\mathrm{freq}(A_c)$ 为概念 c 之中的属性 A 在用户此次的查询与浏览行为中出现的数量总和；随着时间的变化，用户的短期记忆会转变为长期记忆，w 是一个类似遗忘函数的常量，用来平衡用户的长期记忆与短期记忆[14]。

随着用户与系统的交互行为不断增加，由此产生的用户数据也不断增多。文中算法通过分析数据后可以得出概念兴趣度和概念属性权重等用户兴趣的相关知识，进而可以利用相关的算法发掘出新的用户兴趣，同时新产生的用户兴趣也会与用户产生交互行为继而产生新的用户数据，由此完成用户模型的学习与更新。通过不断地改进用户模型，完善其中的语义信息，提升模型的完整性和准确性。

4 实验结果

利用 protégé 构建领域本体，利用 Java 语言配合 Jena 对本体进行解析，进而完成系统构建。图 3 和图 4 分别为用 protégé 构建的茶叶领域本体图和茶叶领域概念节点总览图。

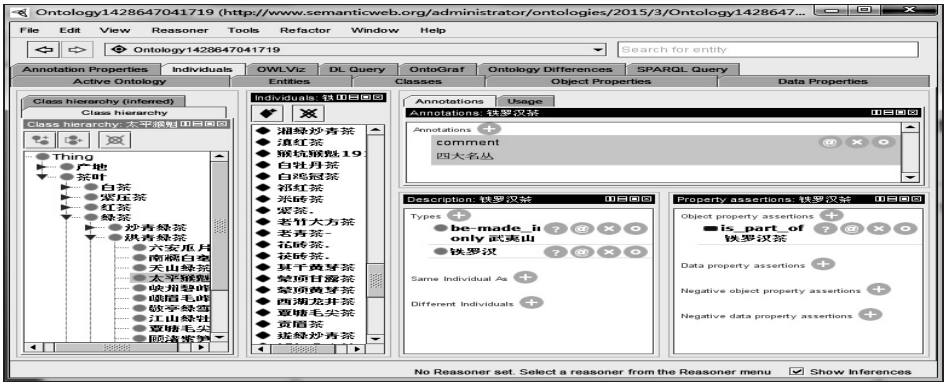

图 3 茶叶领域本体图

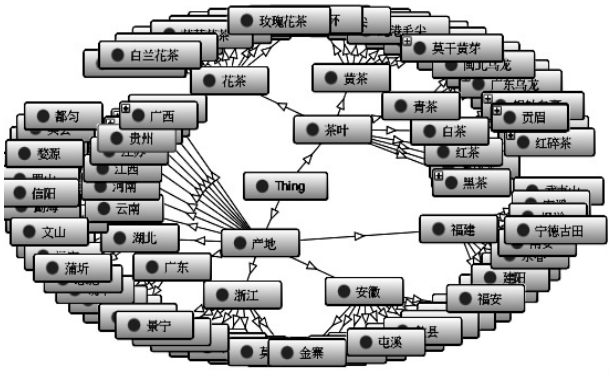

图 4　茶叶领域概念节点总览图

采用 MAE(平均绝对误差值)对实验结果进行评价,它通过比较实验的预期值与最终用户的评分,最终得出推荐结果的精确度[9]。文中采用 MAE 作为衡量用户兴趣模型的一个重要指标。MAE 的公式为:

$$\mathrm{MAE} = \sum_{i=1}^{N} | p_i - q_i | / N \qquad (10)$$

其中,N 为推荐商品的总数量;p_i 为用户对于 i 的预测评分;q_i 为真实评分。

实验采用的数据集是由 627 条数据组成,分别为用户的 ID、购买时间、物品名称以及评分。出于实验的目的,将其中的 70% 进行训练,30% 进行测试,实验结果如图 5 所示。

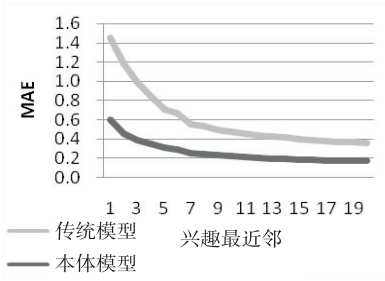

图 5　实验结果

图中的实验结果表明,文中构建的本体模型所产生的推荐结果更加精确,其 MAE 值与传统模型相比更低,但是随着兴趣最近邻的不断增加,两者之间的 MAE 值的差距逐渐变小,最后趋于稳定。

5　结束语

针对现阶段推荐系统中用户模型构建繁琐,以及其中语义要素不完整导致推荐结果精确度不高等问题,提出了结合本体投影算法和概念兴趣度的用户兴趣模型构建方法。通过构建茶叶领域本体,利用投影算法生成用户个性化本体。实验结果表明:用户兴趣模型的构建相较于传统用户模型对于个性化推荐算法有着更好的支持作用;基于本体投影算法的用户兴趣模型易于构建,能够更好地利用本体中的各种语义信息,使推荐结果更加精确。但因数据集的限制,本体中的一些语义要素还不够完整,可能会对结果造成一定的影响。下一步工作可以寻找更优的数据集,并对算法进行进一步优化,使用户模型更加完善。

参考文献:

[1] Jameson A, Paris C, Tasso C. User modeling[C]//Proceedings of the sixth international conference on UM. New York: [s. n.], 1997: 1-3.

[2] 孙雨生. 国内基于本体的用户兴趣建模研究进展(下)-模型管理[J]. 情报理论与实践, 2015, 38(1): 139-144.

[3] 陈一峰, 赵恒凯, 余小清, 等. 基于本体的用户兴趣模型构建研究[J]. 计算机工程, 2010, 36(21): 46-48.

[4] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[5] 徐济成, 李绍稳, 张友华. 农业本体及本体学习研究[J]. 计算机技术与发展, 2009, 19(8): 212-215.

[6] 张静娴. 基于网络本体语言的本体映射研究[D]. 北京: 北京工业大学, 2009.

[7] 张　瑜. 基于本体的农业科技信息用户建模系统研究[D]. 北京: 中国农业科学院, 2009.

[8] 黄彩容. 基于本体的用户兴趣模型在搜索引擎中的应用[J]. 图书馆学刊, 2009, 31(12): 100-103.

[9] 刘佳音. 基于本体的个性化信息系统的应用研究[D]. 杭州: 杭州电子科技大学, 2009.

[10] Busenberg S, Cooke K L. The effect of integral conditions in certain equations modelling epidemics and population growth [J]. Journal of Mathematical Biology, 1980, 10(1): 13-32.

[11] 杨　洁, 季　铎, 蔡东风, 等. 基于联合权重的多文档关键词抽取技术[J]. 中文信息学报, 2008, 22(6): 75-79.

[12] 陈　钰, 张功亮, 阚述贤, 等. 一种基于领域本体的用户建模方法[J]. 计算机与数字工程, 2011, 39(2): 86-89.

[13] Jiang X, Tan A H. Learning and inferencing in user ontology for personalized semantic web search [J]. Information Sciences, 2009, 179(16): 2794-2808.

[14] 蒋秀林, 谢　强, 丁秋林. 基于领域本体的用户模型的研究[J]. 计算机应用研究, 2012, 29(2): 606-608.