理

工

农

微信技术成瘾与用户持续使用关系的实证研究

李莎莎

摘　要：信息技术的持续使用是近几年的研究热点，以往关于信息技术持续使用的研究大多是检验传统的初期采纳和后期采纳感知对持续使用的影响，而很少有研究将用户的心理状态尤其是技术成瘾与持续使用意愿相联系。移动社交网络成瘾作为技术成瘾的特殊类型，是一种伴随着移动终端设备的普及和社交网络平台的兴起而产生的一种新的心理状态。本研究以微信平台为例，旨在探讨社交网络成瘾与用户持续使用意愿的关系，以 Gwebu 等人的双过程模型为持续使用原型，引入成瘾水平潜在变量构建了全新的持续使用模型，并采用结构方程模型对微信用户的持续使用意愿的影响因素进行分析。研究结果表明，社交网络成瘾水平通过影响微信用户的感知有用性、感知有趣性、信任度和习惯从而对用户的持续使用意愿产生影响。本项研究对拓展持续使用理论、减少和预防技术成瘾对个人及社会的危害、营造良好的移动社交网络环境以及促进互联网经济健康发展具有重要意义。

关键词：微信；技术成瘾；持续使用；双过程模型；结构方程模型

1 绪论

移动终端如智能手机、平板、笔记本电脑等在人们的日常生活中随处可见，越来越多的用户选择通过移动终端中的社交网站与家人、朋友、同事联系交流。最具代表性和使用频率最高的移动社交网络平台包括世界范围内普遍使用的脸谱和在中国盛行的微信。截至 2015 年 9 月，月内活跃使用脸谱的移动用户达到 14.4 亿，微信月内积极用户则突破 6.5 亿，同比增长 40%。

随着社交网络用户数量的快速增长，信息系统研究者对于移动社交网络使用行为的研究关注点已经从最初采纳转向持续使用和后期采纳决策。一些研究将期望确认模型（ECM）[2] 引入用户满意和技术接受理论对其进行扩展并应用于移动社交网络的研究 [1]。此外，社群意识、自我调节框架和双过程理论等也被应用于移动社交网络持续使用的研究中。

本文通过对微信用户展开调查，对用户持续使用的影响因素进行实证研究，目的是探索技术成瘾对社交网络持续使用的影响。本文使用 Turel 等人 [3] 的原始模型的外部效用作为移动社交网络成瘾的一个构念，同时将 Gwebu 等人 [4] 的双过程模型结合社交网络成瘾对该双过程模型进行拓展。技术成瘾同许多其他类型的成瘾一样，会扭曲人们做出决策的期望价值。本文的研究可以帮助研究者理解移动社交网络的成瘾水平是如何影响人们的感知和行为。它可以进一步指导辨别、管理和减少技术成瘾和有问题的技术使用等有计划地进行。

2 研究背景和研究模型

2.1 社交网络成瘾

Spira 和 Feintuch [5] 研究发现基于信息技术自身特点、个人特征和其他情境因素等，人们会花掉约

本文作者：李莎莎，江西财经大学信息管理学院研究生　研究方向：电子商务

28%的工作时间在信息技术的干扰上。Steelman[6]等人估算，每年约 28%的交通事故与开车时强迫使用手机有关。Kuss 和 Griffiths[7]指出一些人开始更倾向于与网上好友交流而非现实中的真实朋友。

Turel 等人[3]把有问题的互联网使用以及强迫性互联网使用等现象统称为技术成瘾。移动社交网络成瘾使用在许多研究都被提及。Xu 和 Tan[8]发现许多人每天花大量的时间在社交网络上，这些行为可能导致潜在的社会问题和心理问题如技术成瘾。Cheung 等人[9]认为移动社交网络的过度使用会诱发重要的行为或心理问题。例如一些人强迫性地使用脸谱并连续几个小时不断刷新朋友圈状态等。Rauch[10]等人认为社交焦虑是移动社交网络使用影响的中介。Jeong 等人[11]发现移动社交网络的使用和游戏使用是只能手机成瘾的积极前因，但移动社交网络的使用是智能手机成瘾的更强烈的前因。

2.2 社交网络持续使用

许多研究探讨了社交网络持续使用的决定性因素。整体来看，拓展的社交网络持续使用模型主要可以分为基于分析的（AB）模型、基于情感的（EB）模型和习惯视角下的（HAP）的模型[4]。

AB 模型是应用理性行为理论和技术接受模型解释移动社交网络用户持续使用意愿。感知有用性、信任信念和感知娱乐性是 AB 模型中最重要的三个采纳信念；EB 模型认为情感在决策制定过程中扮演十分重要的角色。在各种社交网络持续使用和用户保留模型中满意和忠诚是用户持续使用的直接前因；HAP 模型认为当用户持续使用移动社交网站时习惯扮演着重要角色。当一个人有规律地重复一种行为并且对其结果是表示满意时，这种行为已经成为了他（她）的习惯，实证表明习惯是后期使用信息系统意愿的直接原因。

人们的感知不是由单一过程控制的，而是由多种过程以及它们之间的相互作用所决定的[4]。Gwebu 等人[4]借鉴双过程理论，将三个后期持续使用模型进行理论结合，通过实证方法证实了双过程模型在拓展持续使用模型中的重要价值。因此，本文以 Gwebu 等人的双过程模型为基础验证用户后期持续使用微信的意愿的影响因素。相关假设见表 1。

表 1 关于移动社交网络持续使用的模型假设

H	假　　设
1	感知有用性与微信用户的持续使用意愿正向相关。
2	信任度与微信用户的持续使用意愿正向相关。
3	感知娱乐性与微信用户的持续使用意愿正向相关。
4	满意度与微信用户的持续使用意愿正向相关。
5	忠诚度与微信用户的持续使用意愿正向相关。
6	满意度与微信用户的忠诚度正向相关。
7	感知有用性与微信用户的满意度正向相关。
8	信任度与微信用户的满意度正向相关。
9	感知娱乐性与微信用户的满意度正向相关。
10	感知有用性与微信用户的忠诚度正向相关。
11	信任度与微信用户的忠诚度正向相关。
12	感知娱乐性与微信用户的忠诚度正向相关。
13	习惯与微信用户的持续使用意愿正向相关。
14	习惯在微信用户感知有用性和持续使用意愿之间有积极的调节作用。
15	习惯在微信用户信任度和持续使用意愿之间有积极的调节作用。
16	习惯在微信用户感知娱乐性和持续使用意愿之间有积极的调节作用。

2.3 社交网络成瘾与持续使用结合

Turel 等人[3]通过研究发现成瘾用户比非成瘾用户更倾向于认为信息技术是有用且有趣的。Kim[12]发现成瘾程度更大的用户更擅长于挖掘一项信息技术的内部和外部的新功能。Huh 和 Bowman[13]也认为对于同样一项信息技术成瘾程度更大的用户会做出更高的评价，即成瘾水平越高的用户对信息技术的使用更满意。已有的实证研究为支持网络成瘾与忠诚度积极相关提供了强有力的证据，在移动社交网络环境下，一旦用户使用微信成瘾，他们则更可能长时间坚持使用微信。成瘾用户往往缺乏自我约束力而导致无法有意识地控制自身行为(Thadani and Cheung, 2011)[9]。当行为没有得到有意识的控制时习惯性可能会增强[3][12]。鉴于上述分析，本文提出如下假设：

H17：社交网络成瘾水平对微信用户的感知有用性有积极影响。

H18：社交网络成瘾水平对微信用户的感知有趣性有积极影响。

H19：社交网络成瘾水平对微信用户的信任度有积极影响。

H20：社交网络成瘾水平对微信用户的满意度有积极影响。

H21：社交网络成瘾水平对微信用户的忠诚度有积极影响。

H22：社交网络成瘾水平对微信用户的习惯有积极影响。

感知有用性、信任度。感知娱乐性、满意度、忠诚度和习惯是 Gwebu 等人提出的双过程持续使用模型中的六个关键构念。本研究通过融入移动社交网络成瘾水平对该模型进行了拓展。图 1 为本文的研究模型。

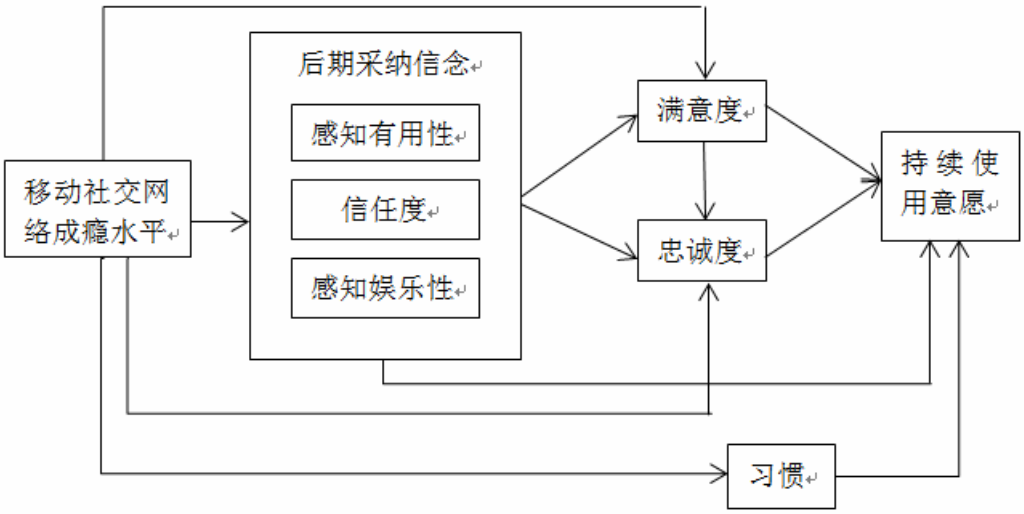


图 1 本文研究模型

3 实证分析

3.1 问卷收集和数据分析

本文采用问卷调查法利用问卷星在全国范围内发放问卷收集数据。问卷的第一部分为调查者的基本信息，包括五个题项；第二部分为对测量模型各潜变量的测量，包括：感知有趣性（PE）、感知有用性（PU）、信任度（TB）、满意度（SA）、忠诚度（LO）、习惯（HA）、成瘾水平（SAD）和持续使用意愿（CI）。为确保问卷量表的内容效度，本问卷采用的结构变量的测量均来自已有权威文献。每个测量项目均采用李克特 5 分量表。本轮调查始于 2016 年 3 月，调查截至 2016 年 5 月，期间收回问卷 579 份，剔除质量不合格问卷后得到有效问卷 495 份。

本文采用 Cronbach **α** 系数测量问卷信度。运用 SPSS21.0 软件对问卷各个测量项进行分析。结果显示各测量项的 Cronbach **α** 系数均大于 0.7，因此本问卷具有良好的信度。表 2 为问卷的信度测量表。

表 2 问卷信度测量表

变量	问项个数	Cronbach α 值	参考值
PE	3	0.926	Cronbach α ≥ 0.70
PU	5	0.890	
TB	5	0.911	
SA	4	0.922	
LO	3	0.855	
HA	5	0.956	
FOGS	8	0.915	
CI	3	0.774	

本文采用因子分析法检验问卷的构建效度。在进行因子分析之前首先测量统计量的 KMO 值和 Bartlett 球体检验。经 SPSS21.0 检测，所有潜变量的 KMO 值均大于 0.7，且均通过 Bartlett 球体检验，说明本研究的潜变量适合作因子分析，具体检测结果见表 3。在因子分析中，平均方差抽取量均大于 0.5，说明样本相对的测量误差在可以接受的范围内；而潜在变量的组合信度在 0.70 以上，表示模型内在质量理想。表 4 为因子分析结果。

表 3 KMO 值和 Bartlett 球体检验

变量	KMO	Bartlett 球体检验		
		Approx.Chi-Square	df	sig.
PE	0.758	1157.336	3	.000
PU	0.879	1306.646	10	.000
TB	0.873	1639.743	10	.000
SA	0.828	1609.595	6	.000
LO	0.726	728.133	3	.000
HA	0.910	2683.288	10	.000
SAD	0.885	2637.382	28	.000
CI	0.703	399.519	3	.000

表 4 验证性因子分析

变量	测量项目	因子负荷量	平均方差抽取量	组合信度
PE	PE1	0.875	0.808	0.926
	PE2	0.928		
	PE3	0.892		
PU	PU1	0.784	0.614	0.888
	PU2	0.767		
	PU3	0.795		

	PU4	0.755		
	PU5	0.814		
TB	TB1	0.772	0.673	0.911
	TB2	0.781		
	TB3	0.802		
	TB4	0.894		
	TB5	0.845		
SA	SA1	0.893	0.729	0.915
	SA2	0.909		
	SA3	0.845		
	SA4	0.761		
LO	LO1	0.831	0.665	0.856
	LO2	0.830		
	LO3	0.784		
HA	HA1	0.891	0.813	0.956
	HA2	0.921		
	HA3	0.903		
	HA4	0.918		
	HA5	0.874		
FOGS	FOGS1	0.653	0.582	0.917
	FOGS2	0.769		
	FOGS3	0.765		
	FOGS4	0.848		
	FOGS5	0.847		
	FOGS6	0.664		
	FOGS7	0.761		
	FOGS8	0.748		
CI	CI1	0.735	0.503	0.752
	CI2	0.667		
	CI3	0.724		

3.2 模型分析

本文采用结构方程模型 AMOS 21.0 软件对数据进行处理。表 5 为模型的拟合结果。结果显示，除卡方值比自由度未达到适配标准外，GFI、AGFI、RMSEA、CFI、TLI、IFI 均到达模型拟合要求。鉴于当样本规模较大时卡方值会相对变大，因此在判断模型拟合度时除了卡方值比自由度还必须考虑其他拟合指标，由此我们可以推断本模型拟合度较为理想。图 2 为具体的路径系数结果。

表 5 研究模型的拟合指标

拟合指标	适配标准		检验结果数据	模型适配判断
	良好	优良		
卡方值比自由度	<3.0		3.936	否

GFI	[0.7, 0.9]	>0.9	0.781	良好
AGFI	[0.7,0.9]	>0.9	0.745	良好
RMSEA	<0.08	<0.05	0.077	良好
CFI	[0.7, 0.9]	>0.9	0.890	良好
TLI	[0.7, 0.9]	>0.9	0.879	良好
IFI	[0.7, 0.9]	>0.9	0.891	良好

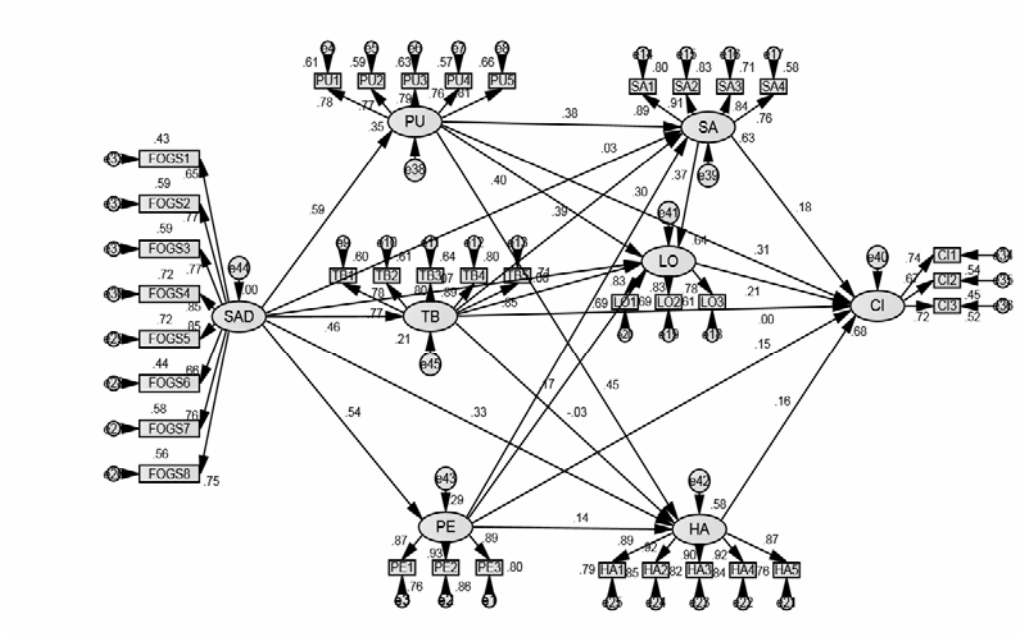


图 2 结构方程模型结果图

在显著性检验方面，数据运行的统计结果如表 6 所示，除信任度对持续使用意愿、信任度对忠诚度、习惯在信任度和持续使用意愿间的作用以及成瘾水平对满意度和忠诚度的假设之外，其余假设均得到了验证。

表 6 模型检验结果

假设/路径		路径系数	P 值	检验结果
H1	PU → CI	0.222	***	支持
H2	TB → CI	0.001	0.971	不支持
H3	PE → CI	0.093	**	支持
H4	SA → CI	0.139	*	支持
H5	LO → CI	0.145	**	支持
H6	SA → LO	0.408	***	支持
H7	PU → SA	0.354	***	支持
H8	TB → SA	0.356	***	支持
H9	PE → SA	0.230	***	支持

H10	PU → LO	0.410	***	支持
H11	TB → LO	-0.082	0.095	不支持
H12	PE → LO	0.148	***	支持
H13	HA → CI	0.083	**	支持
H14	PU→ CI	0.638	***	支持
	HA → CI	0.083	***	
H15	TB → CI	-0.046	0.923	不支持
	HA → CI	0.083	**	
H16	PE → CI	0.165	**	支持
	HA → CI	0.152	**	
H17	SAD → PU	0.459	***	支持
H18	SAD → PE	0.495	***	支持
H19	SAD → TB	0.357	***	支持
H20	SAD → SA	0.021	0.571	不支持
H21	SAD → LO	0.054	0.231	不支持
H22	SAD → HA	0.362	***	支持

注：①* − p < 0.05, ** − p < 0.01, *** − p < 0.001

②习惯变量的调节作用通过在保持其他条件不变的情况下，分别检验因变量和调节变量对果变量的单独影响来判断。

通过观察数据结果我们可以得到以下结论：

(1) 与 Gwebu 等人的双过程模型的实证结论一致，在微信用户的持续使用意愿的影响因素中，感知有用性的影响力最强，忠诚度、满意度和感知有趣性对用户的持续使用意愿有显著影响。然而，信任度对持续使用意愿的影响在本次研究中未获得证实。

(2)在感知有用性、感知有趣性、满意度、信任度、忠诚度和习惯等中间变量的关系中，除了满意度对忠诚度的影响和习惯在忠诚度和满意度间的中介作用没有得到支持外，均取得了显著结果。

(3)在本文提出的关于社交网络成瘾的 6 个假设中，社交网络成瘾水平对满意度和忠诚度的假设未获得支持。而在获得支持的四个假设中，社交网络成瘾水平对感知有用性、感知有趣性、习惯和信任度的均有较为显著的影响。

4 研究意义与总结

4.1 研究意义

4.1.1 理论意义

与技术接受理论研究相比，信息系统持续使用理论还是一个方兴未艾的研究领域，不同领域的持续使用研究尚存在理论空白有待我们进一步完善。在当前的持续使用研究大多停留在技术采纳阶段的背景下，技术成瘾水平作为持续使用的特殊潜在变量还没有得到充分证实，尽管移动社交网络的成瘾使用由于其产生的不良后果已经引起了社会的关注，但相关的理论性研究不多，研究社交网络成瘾和社交网络平台持续使用的影响的文献则还是一个比较新的领域。本研究验证了 Gwebu 等人的利用双过程模型研究用户持续使用意愿的有效性，同时融入 Turel 等人的技术成瘾变量，构建全新的移动社交网络持续使用模型，在验证前人的理论成果的同时，也验证了微信技术成瘾和用户持续使用的关系，对于拓展技术上瘾相关理论和持续使用理论均有十分重要的理论意义。

4.1.2 现实意义

技术成瘾会影响用户使用信息系统的方式，其对信息技术使用的影响不应被忽视。微信作为国内使用频率最高的移动社交网络平台，已成为人们日常生活中不可或缺的社交工具，应该获得充分关注和重视。在此背景下，本研究有着以下重要的现实意义。

第一，本研究可以帮助研究者理解社交网络的成瘾水平是如何影响人们的感知和行为的，从而帮助人们认识、控制微信技术成瘾带来的负面问题，进一步预防和减少技术成瘾和微信过度使用的不良影响。

第二，本研究对信息技术提供者增强自身法律责任有一定的警示作用，信息技术提供者在提供信息技术服务中应尽量减少技术上瘾对用户及社会的危害。

最后，本研究对营造良好的移动社交网络环境，进而对促进互联网经济健康发展具有十分重要的现实意义。

4.2 总结

社交网络成瘾和用户持续使用为相对较新的领域，本研究验证了 Gwebu 等人的利用双过程模型研究用户持续使用意愿的有效性，同时融入技术成瘾变量，构建全新的移动社交网络持续使用模型，在验证前人的理论成果的同时，也验证了微信技术成瘾和用户持续使用的关系。实证结果表明，社交网络成瘾水平通过影响微信用户的感知有用性、感知有趣性等感知以及信任度和习惯从而影响用户的持续使用意愿。鉴于信息技术已成为现代社会不可逆转的部分，识别、预防和治疗技术成瘾是必要而迫切的，因此今后在技术成瘾和持续使用领域仍需要展开进一步的研究。

参考文献：

[1]张国华,雷雳. 基于技术接受模型的青少年网络游戏成瘾机制研究[J].心理发展与教育,2015,04:437 ~ 444.

[2]Bhattacherjee A. Understanding information systems continuance: an expectation-confirmation model[J]. Mis Quarterly, 2001, 25(3):351 ~ 370.

[3]Turel, O., Serenko, A., and Giles, P. 2011. “Integrating Technology Addiction and Use: An Empirical Investigation of Online Auction Users,” MIS Quarterly (35:4), pp. 1043 ~ 1061.

[4]Gwebu, K., Wang, J., Guo, L. Continued usage intention of multifunctional friend networking services: A test of a dual-process model using Facebook. Decision Support Systems 67 (2014) 66 ~ 77.

[5]Spira J.B. and J.B. Feintuch (2005) “The Cost of Not Paying Attention: How Interruptions Impact Knowledge Worker Productivity: Executive Summary”, New York, NY: Basex Report.

[6]Steelman, Z., Soror, A., Limayen, M. & Worrell, D. (2012) Obsessive compulsive tendencies as predictors of dangerous mobile phone usage. Proceedings of the Eighteenth Americas Conference on Information Systems, Seattle, WA.

[7]Kuss, D. J., & Griffiths, M. D. (2012). Social networking on the Internet: From pastime to excess and addiction. Mental Notes, 6(3), 26 ~ 28.

[8]Xu, H., and Tan, B. C. Y. 2012. “Why do I keep checking Facebook: Effects of message characteristics on the formation of social network services addiction,” Proceedings of the 33rd International Conference on Information Systems (ICIS), Orlando (FL).

[9]Thadani, D.R., Cheung, C.M.K. (2011). Exploring the Role of Online Social Network Dependency in Habit Formation. Thirty Second International Conference on Information Systems, Shanghai, pp.1 ~ 17.

[10]Rauch, S. M., Strobel, C., Bella, M., Odachowski, Z., & Bloom, C. (2014). Face to face versus facebook: does exposure to social networking web sites augment or attenuate physiological arousal among the socially anxious? Cyberpsychology, Behavior, and Social Networking, 17(3), 187 ~ 190.

[11]Jeong S., Kim, H., Yum, J., Hwang, Y. (2016). What type of content are smartphone users addicted to?: SNS vs. games. Computers in Human Behavior 54 (2016) 10 ~ 17.

[12]Kim, S. S. 2009. “The Integrative Framework of Technology Use: An Extension and Test,” MIS Quarterly (33:3), pp. 513 ~ 537.

[13]Huh, S., and Bowman, N. D. 2008. “Perception and Addiction of Online Games as a Function of Personality Traits,” Journal of Media Psychology (13:2).

基于距离熵评估一个测试集的错误检测能力

徐常福

摘　要：在软件自动测试技术中，如何准确地评估一个测试集的错误检测能力尤为重要。然而，目前主要是采用白盒度量评估一个测试集的错误检测能力。例如，突变分析[1]是评估一个测试集错误检测能力的有效技术，但是它需要执行大量的突变集测试程序，用以获取死亡突变测试集数目信息。这在实际项目中成本太高，而且这样的白盒信息往往不可得到。例如，在某些秘密的工程中，我们无法分析源代码，甚至二进制代码也无法得到。因此，为解决实际中白盒信息常常无法得到的问题，本文提出一种基于距离熵(黑盒度量)评估一个测试集错误检测能力的方法。该距离熵度量不需要获取白盒信息就可以指导黑盒测试，从而评估一个测试集的错误检测能力。为验证该方法的有效性，我们采用 SIR[2](Software-artifact Infrastructure Repository)中的 grep、flex、sed 和 make 4 个程序进行多方面的实验测试。实验结果表明，采用距离熵度量不仅可以较好地评估一个测试集的错误检测能力，而且优于其它一般的黑盒度量。

关键词：距离熵；黑盒度量；测试集；错误检测能力

1 引言

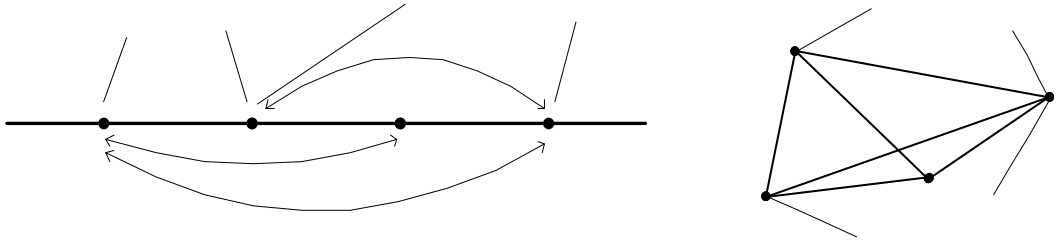
近几年，已经有许多的论文将基于多样性的技术运用于软件自动测试研究中，包括[3], [4]和[5]。这些文献表明，评估一个测试集的错误检测能力可以通过计算该测试集的多样性。这让我们设想设计一个用于计算测试集多样性的度量，通过计算测试集的多样性，从而间接地评估该测试集的错误检测能力。此外，在软件自动测试技术中，要求测试集尽可能的具有高多样性，而高多样性要求测试集最大化地覆盖整个特定的空间，如输入空间[4]、输出空间[3]。并在该同一层面上，A. Arcuri 和 L. Briand [6]通过理论分析证明：为让一维空间的测试集有最大的错误检测能力，该测试集必须能均匀地传播于整个测试空间。即测试集不仅需要均匀传播，而且各个测试用例之间距离要相同(见图 1-1 (a))。因此，基于上面的论述，实际中计算一个测试集的多样性，主要存在以下两个挑战。

挑战 1：实际中测试用例之间的相邻关系如何定义？

挑战 2：如何使用相邻测试用例间的距离评估一个测试集的多样性？

为解决上面两个挑战，本文设计出一种距离熵度量，并采用该度量计算一个测试集的多样性，从而间接评估该测试集的错误检测能力(详细见 2.3 节)。该过程主要分为两个子任务：（1）构建图模型(解决挑战 1)，即根据图论[7]，设计一个测试用例关系图，该图能够描述测试用例之间的关系；（2）定义距离熵度量用以评估测试集的错误检测能力(解决挑战 2)，即定义距离熵度量，并在该图模型的基础上，采用距离熵度量计算该测试集的多样性，从而间接评估该测试集的错误检测能力。

本文作者：徐常福，江西财经大学软件学院研究生　研究方向：可信软件技术



相邻测试用例　　不相邻测试用

W4

图 1-1 一维空间测试集的多样性图和实际测试集的多样性挑战图

在本文的实验部分，主要是采用 10 折交叉实验法进行实验。经我们的实验结果表明，该方法不仅能有效地评估一个测试集的错误检测能力，而且优于其它一般的黑盒度量。

W1　W2　W3

2 方法论

在本节中，我们分 3 个小节来论述实现基于距离熵评估一个测试集的错误检测能力：（1）对给定的一个测试集，构建一个可表示测试集的图模型；（2）在该图模型的基础上定义一种距离熵度量；（3）用该距离熵度量计算该测试集的多样性，从而间接地评估该测试集的误差检测能力。

W5

2.1 构建图模型

W6

针对给定的一个测试集T，我们定义一个测试用例关系图和最小权重集，用以描述测试用例之间的关系。

测试集的多样性：$f(\{w1,w2,w3\})$

定义 1 测试用例关系图：定义一个测试集$T = \{t_1, t_2, \cdots, t_n\}$ $(n \geq 2)$对应的测试用例关系图为一个有权无向完全图$G(T) = \langle V, E\rangle$，其中$V = T$，$E = \{(t_i, t_j) | t_i, t_j \in V, i \neq j\}$，每一条边$(t_i, t_j)$上的权重是$t_i$到$t_j$之间的距离。

(a)

在实际项目工程的测试中，一个测试集往往有成千上个测试用例，对应用例关系图$G(T)$中的边数目更是庞大。于是，为了减少时间复杂度且不影响计算测试集多样性精确，我们应该从$G(T)$中选取具有代表性边的子集。根据我们的直觉，$G(T)$中最短边连接的节点对计算测试的多样性精度影响较大。从这个角度出发，我们可以从$G(T)$中提取出一颗最小生成树$G'(T) = \langle V, E'\rangle$。又因为最小生成树中全部的边是完全图中最短的边，所以，可以用$G'(T)$近似计算该测试集的多样性。这里，我们定义$G'(T)$为$MST(T)$，并且定义$MST(T)$中全部边的权重为最小权重集。

定义 2 最小权重集：定义图$G(T)$的最小权重集为$W(G(T))$，即$MST(T)$中所有边权重的集合为$W(G(T))$。例如，一个测试关系图$G(T)$的$MST(T)$有四个节点和三条边，三条边的权重分布为 1, 2, 2，则该图的$W(G(T)) = \{1,2,2\}$。

2.2 距离熵

根据前面的论述，为保证计算一个测试集多样性的精确度，距离熵必须满足两个条件：（1）均匀扩展条件，对$\forall w_i, w_j \in W(G(T))$，有$w_i = w_j$，即每个测试用例在指定的输入空间中等间距的扩展；（2）宽度扩展条件，对$\forall w \in W(G(T))$，w要尽可能的大，即每个测试用例在指定的输入空间扩展宽度达到最大。因此，基于以上两个条件及 Shannon 熵[8]的定义，我们给出如下距离熵定义。

定义 3 距离熵：对给定的一个测试集T，其最小权重集为$\mathrm{W}(\mathrm{G}(T))$，那么定义该测试集的距离熵$\mathrm{H}(T)$为：

$$\mathrm{H}(T) = W * \left(- \sum_{w \in \mathrm{W}(\mathrm{G}(T))} \frac{w}{W} \ln \frac{w}{W} \right) \tag{1}$$

其中$W = \sum_{w \in \mathrm{W}(\mathrm{G}(T))} w$。公式(1)的距离熵主要由两部分组成：第一部分是$W$，它确保满足宽度扩展传播条件；第二部分是$-\sum_{w \in \mathrm{W}(\mathrm{G}(T))} \frac{w}{W} \ln \frac{w}{W}$，当$W$不变时，它确保满足均匀传播条件[8]。

2.3 测试集的多样性计算方法

计算一个测试集多样性的主要过程(见图 2-1)如下：

（1）根据**定义 1**，我们将一个测试集 T 建模成一个测试用例关系图$\mathrm{G}(T)$，图中的每个节点代表一个测试用例，图中边上的权值为两个测试用例之间的距离；

（2）采用最小生成树算法，将（1）中得到的图简化为一颗最小生成树；

（3）根据**定义 3**，计算（2）中最小生成树的距离熵。该距离熵代表了该测试集的多样性，从而间接地评估该测试集的错误检测能力。

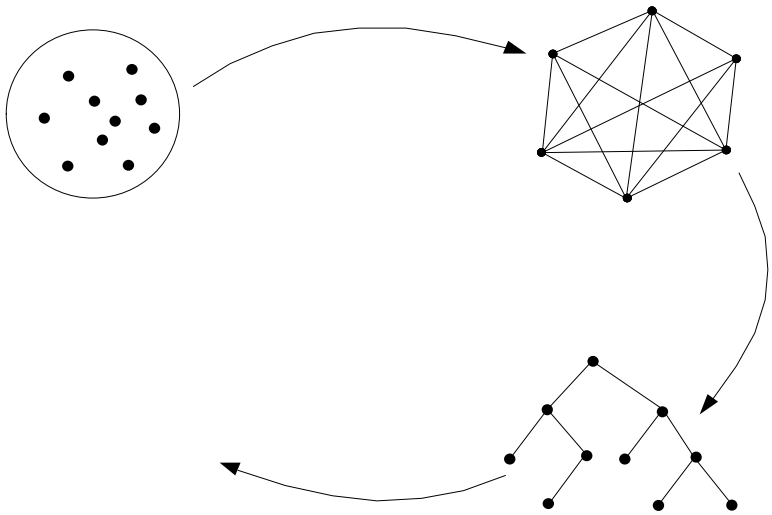


图 2-1 距离熵计算测试集的多样性过程

3 实验及有效性评估

3.1 实验数据及符号说明

根据实验的需要和保证实验的有效性，我们选择了 4 个真实的、频繁使用的程序作为本次实验的 4 个测试集。这个 4 个程序来自 SIR[2]，它们已经被广泛运用于软件测试技术研究。如表 1 所示。本次实验使用的符号如表 2 所示。

表 1 实验程序

程序类型	版本	代码行数	测试池大小
grep	2.4	10069	810
flex	2.5.1	10458	567
sed	4.0.7	14427	3622
make	3.76.1	35544	791

表 2 符号说明

序号	符号	符号说明
1	H	距离熵
2	H_i	第 i 个测试集的距离熵
3	N_i	第 i 个测试集的死亡突变用例数目
4	$\sum G$	测试用例关系图的所有权重和
5	W	最小生成树的所有权重和

3.2 实验设计

本次实验是采用10折交叉实验法进行实验,主要是通过研究以下两个问题来验证距离熵的有效性。

问题 1：当白盒信息不可获取时，距离熵能否有效地评估一个测试集的错误检测能力?

问题 2：距离熵是否优于其他一般的黑盒度量(如W 和$\sum G$)?

根据前面的论述，突变分析[1]可以较为准确地预测一个测试集检测错误能力。因此，我们采用突变分析来评估距离熵的有效性，即把突变分析过程中死亡突变的用例数目作为评估测试集错误检测能力的一种度量。本节主要实验步骤如下：

（1）对每一个指定大小的测试集T_i(共 2500 个)，计算出该测试集的距离熵H_i和死亡突变用例数目N_i（这里，我们采用 mutate.py[9]对实验程序(见表 1)进行突变，然后将执行突变程序的输出与执行原程序的输出对比，统计得到死亡突变用例的数目）;

（2）根据（1）中得到的距离熵H_i，将 2500 个测试集进行降序排序，从而得到该测试集对应降序序列;

（3）将（2）中得到的序列依次切分为 25 个子列表，每一个子列表含 100 个测试用例，并分别计算这 100 个测试用例的平均距离熵和平均死亡突变数目，从而得到一个新的序列;

（4）在新序列的基础上，求其距离熵和死亡突变用例数目的 Pearson 相关系数[10]。

3.3 评估距离熵的有效性

本次实验相关参数设置与距离类型选取:（1）设置 4 个测试集尺寸为 5、10、15、20、25、30、40、50（这里选择测试集尺寸上界为 50，是因为我们发现在该尺寸下，测试集的错误检测能力已经接近收敛），分别进行重复测验过程，且每一个尺寸的测试集包含随机选取 2500 个测试用例;（2）根据本实验测试的 4 个实验程序都是字符串，这里采用编辑距离[11]计算各个测试用例间的距离。

3.3.1 验证问题 1

本文实验采用 10 折交叉实验法，计算出每一个测试集的距离熵H_i与死亡突变用例数目N_i的 Pearson 相关系数，如表 2 所示。

表 2 H 与死亡突变用例数目的 Pearson 相关系数结果

测试集大小	flex 程序	grep 程序	make 程序	sed 程序
5	0.86	0.84	0.89	0.64
10	0.81	0.95	0.91	0.65
15	0.96	0.90	0.91	0.78
20	0.92	0.95	0.93	0.88
25	0.91	0.91	0.92	0.92
30	0.92	0.95	0.96	0.91
40	0.97	0.94	0.97	0.87
50	0.97	0.94	0.96	0.74

由表 2 可以看出，3 个实验程序的距离熵与死亡突变用例数目具有极强相关性(相关系数在 0.8~1.0 之间为极强相关，在 0.6~0.8 之间为强相关，在 0.4~0.6 之间为中等程度相关，在 0.2~0.4 之间为弱相关，在 0.0~0.2 之间为极弱相关或无相关)，只有 sed 程序的距离熵与死亡突变用例数目成强相关性。从而，可以进一步说明距离熵与突变分析一样可以有效地评估测试集的错误检测能力，即验证了距离熵的有效性。

3.3.2 验证问题 2

在该小节，我们使用W 和$\sum G$ 代替距离熵H 重复上述实验过程。计算出$\sum G$ 、W 与死亡突变用例数目的 Pearson 相关系数，如表 3 所示。为使实验结果更加显而易见，我们在一个直角坐标系中画出H 、$\sum G$ 、W 与死亡突变用例数目的 Pearson 相关系数的散点图，如图 3-1 所示。

表 3 $\sum G$ 、W 与死亡突变用例数目的 Pearson 相关系数结果

测试集大小	flex 程序		grep 程序		make 程序		sed 程序	
	$\sum G$	W	$\sum G$	W	$\sum G$	W	$\sum G$	W
5	0.71	0.73	0.80	0.80	0.80	0.74	0.54	0.40
10	0.72	0.76	0.76	0.77	0.75	0.80	0.50	0.41
15	0.62	0.80	0.78	0.80	0.80	0.84	0.41	0.42
20	0.67	0.82	0.81	0.87	0.77	0.72	0.61	0.65
25	0.53	0.87	0.81	0.87	0.76	0.76	0.62	0.60
30	0.71	0.77	0.86	0.90	0.69	0.76	0.57	0.76
40	0.79	0.85	0.75	0.85	0.71	0.80	0.57	0.68
50	0.81	0.87	0.73	0.84	0.77	0.80	0.70	0.71

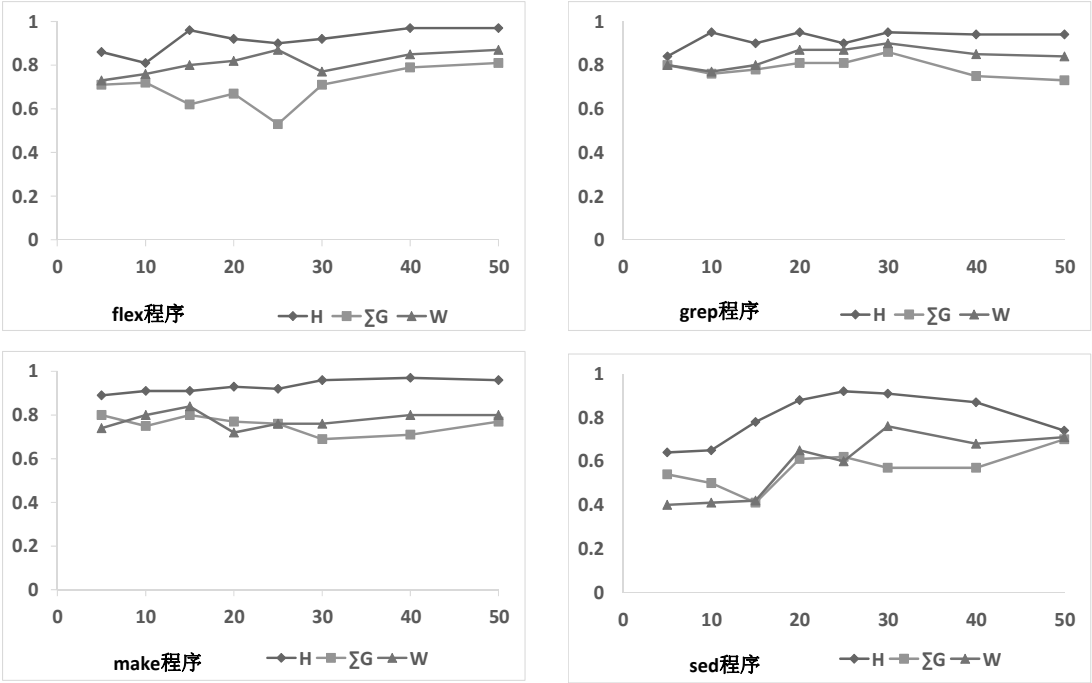


图 3-1 H、$\sum G$、W 与死亡突变用例数目的 Pearson 相关系数散点图

由图 3-1 知，随着测试集尺寸的增加，采用 H 计算得到 Pearson 相关系数始终大于等于采用 W 和 $\sum G$ 计算得到的 Pearson 相关系数。这说明，采用距离熵度量评估一个测试集的错误检测能力优于 W 和 $\sum G$ 度量，即验证了距离熵优于其它一般的黑盒度量。

4 结论

在本文中，我们通过以上的论述与实验表明：在没有白盒信息的情况下，采用距离熵方法可以有效地评估一个测试集的错误检测能力，同时，证明了距离熵优于其它一般的黑盒度量（W 和 $\sum G$）。

当前，为了满足人们在实际中各种各样的需求，导致开发出来的软件的复杂度越来越高及保证软件质量的难度也越来越大。这就要求测试人员尽可能地提高软件测试效率和检测出 bug 的能力。而我们知道，设计一个高错误检测能力的测试集是提高软件测试效率的常用方法。因此，本文提出的基于距离熵评估一个测试集的错误检测能力的方法，可以很好地帮助测试人员设计出一个错误检测能力较高的测试集，从而提高测试效率用以保证软件的质量。此外，距离熵评估一个测试集的错误检测能力不仅在实际软件开发工程中具有重大的应用前景，而且提供了软件测试学术界一个新的视角。

参考文献：

[1] Y. Jia and M. Harman. An analysis and survey of the development of mutation testing [J]. IEEE Transactions on Software Engineering, 2011, 37(5): 649–678.

[2] Software-artifact Infrastructure Repository:http://sir.unl.edu/php/previewfiles.php.

[3] N. Alshahwan and M. Harman. Augmenting test suites effectiveness by increasing output diversity [C]. In proceedings of the 34th International Conference on Software Engineering (ICSE ’12), 2012: 1345–1348.

[4] T. Y. Chen, F.-C. Kuo, R. G. Merkel, and T. Tse. Adaptive random testing: The art of test case diversity [J]. Systems and Software, 2010, 83(1): 60–66.

[5] H. Hemmati, A. Arcuri, and L. Briand. Achieving scalable model based testing through test case diversity [J]. ACM Transactions on Software Engineering and Methodology, 2012, 22(1): 1–6, 42.

[6] A. Arcuri and L. Briand. Adaptive random testing: an illusion of effectiveness? [C]. In proceedings of the 20th International Symposium on Software Testing and Analysis (ISSTA ’11), 2011: 265–275.

[7] Diestel 著.于青林, 王涛, 王光辉译. 图论(第四版) [M]. 北京: 高等教育出版社, 2013: 1–2.

[8] C. E. Shannon. A mathermatical theory of communication [J]. Bell System Tech, 1948, 27(1): 379–423, 623–656.

[9] Tools for mutation analysis. http://github.com/qingkaishi/mutation.git.

[10] J. Cohen. Statistical power analysis for the behavioral sciences [J]. Biometrics, 1988, 73(363): 19–74.

[11] Wikipedia. Hamming distance. http://en.wikipedia.org/wiki/Hamming_distance.