智能手机搜索引擎的可用性评估

孔宁，张常洁
（浙江工业大学心理学系，杭州 310000）

摘要：研究目的在于评估目前市场上发展迅速却又探讨相对较少的手机搜索引擎。研究对象分别为神马、百度、搜狗和必应。首先通过实验法比较了手机搜索引擎的有效性，进而采用问卷法收集了关于被测搜索引擎的准确度、省时性与总体满意度三个维度的主观倾向性信息。实验结果表明，百度在搜索中表现最好，其它搜索引擎都在某些层面需要优化。研究有效区分了手机搜索引擎的性能，为手机引擎系统的发展与改进提供了方向。

关键词：智能手机；搜索引擎；评估
中图分类号：TB18
DOI：10.3772/j.issn.1673-2286.2015.10.012

1 引言

随着3G网络与移动互联网络的不断发展，手机搜索引擎因其便捷性、及时性以及操作简易性等特点被越来越多的智能机使用者所使用。CNIT发布的《2014年7月中国移动搜索市场研究报告》显示，截至2014年7月底，移动搜索用户规模达4.13亿，在中国手机网民中的渗透率为76.8%[1]。从主要移动搜索月活跃用户在手机网民中的渗透率来看，百度搜索（71.3%）位居首位，其次是神马（25.1%）和搜狗（23.9%）。以往搜索引擎的研究大多局限于PC端，涉及移动端的较少。中国手机用户在如何选择最佳手机搜索引擎时也缺少相关的知识与途径。基于此，本文在参考国内外相关研究的基础上提出了一种快速有效的可用性评估方法，评估手机搜索引擎的可用性。

2 相关研究

2.1 手机搜索引擎

作为一种智能手机应用（APP），手机搜索引擎被越来越多的用户所使用。尽管在技术上取得了一系列的创新与进展，但其可用性仍需进一步评估与完善[2]。国内一些研究曾对PC端搜索引擎做过评估，如刘子慧、张锋深层分析了谷歌和百度在内容有效性与直接性方面的差异[3]，但是，手机与PC平板在界面、系统运行等方面有较大不同，设计一种能够有效评估手机应用的技术方法仍是非常必要的[4]。尼尔森也在自己的专栏中指出，在移动设备上使用网页得分很低，因为这个网页不是为手机设计的[5]。笔者通过对手机搜索引擎的比较研究发现，人们最常用的几款手机搜个引擎（百度、搜狗等）的信息搜索在结构组织、具体内容、结果排列等方面也是存在差异的。

2.2 搜索结果的相关性

可用性评估可以将搜索引擎的检索表现量化分析[6]。尽管搜索系统的评估表现在诸多方面，但最重要的还是搜索结果的相关性[7]；其中，搜索结果的排列起主要作用，然而这方面却很少有研究者关注。Vaughan指出，如果不能很好地将结果排列，人们几乎不能从数以万计的结果中筛选出有用的项目[8]。因此，搜索引擎原有结果排列与用户项目排列的相关性越高，说明搜索系统越有效。以往研究中，相关性的主要测量方法是二元相关与等级相关。二元相关指在判断一个项目是否相关时，只有是/否两种选择；等级相关表示在判断

一个项目时，可按等级来评价相关程度，如分为非常相关、相关、部分相关与不相关四个等级。二元相关常用于评估具有明确性或者唯一性的项目[9]。等级相关应用更为广泛，Tang发现在评估项目时，七级的等级评定具有最高的可信度[10]。李珏伶设计并应用5分等级进行等级评定，实验结果证明5分等级评定要好于4分等级[11]。本研究不是对单个项目进行独立评定，而是通过对搜索结果进行连续的等级评定，即将搜索结果项目按照相关性从高到低排列，进而将用户给出的等级评定与原有搜索结果的排列进行进一步的数据分析。

2.3 可用性评估的理论模型

可用性指特定产品在特定使用背景下用于特定目的时所具有的有效性、效率和用户主观满意度[12]。如果一个产品的可用性不好，那么它会导致产品出现一系列不同程度的问题[13]。可用性评估模型不仅指出可用性包含的要素，更重要的是它阐释了这些因素之间的作用关系，从而针对这些要素进行评估。在实际操作中，研究者须结合具体实验的评估目标、研究对象、实验条件等因素选择具有针对性的评估方法及操作流程。

（1）EASON模型。1984年Kenneth Eason在信息技术领域首次提出此模型[14]，主要包括任务、用户和系统三个因素。任务包括频率和开放性两个子属性，用户有知识、动机和自由决定三个子属性，系统包括易学性、易用性和任务匹配。该模型把用户、系统和任务看作自变量作为一种输入，用户的反应则是因变量。这种模型是因果式的，认为可用性是几个相互作用变量的结果。

（2）Shackel模型。该模型包括4个因素，有效性、易学性、灵活性和态度[15]。这些属性根据实际情况具有不同的权重，强调对人的行为和态度等一些人为因素的测量。在改进版的模型中包括有效性、易学性和态度而不包含灵活性，因为出于测评真正有效的角度，Shackel认为很难确定和测量系统的灵活性。亲和性与用户感知、情感有着显著的相关性，因此也被认为是可用性重要的构成要素。

（3）Nielson模型。该模型主要强调系统的可接受性和可用性[16]。系统的设置应该是人们乐于接受的，也就意味着它能够真正符合人们的需求可用性包括5个主要因素，即易学性、效率、可记忆性、出错率和满意度。与Shackel模型相同，Nielson也没有给出各因素固定的权重，这些权重应根据不同的项目而各有差异。

3 研究方法

3.1 被试者的选择

本研究采用网络完全随机招募的方式进行被试者选取并进一步筛选。实验共选定24名被试者，男女各半，年龄为19~23岁，平均年龄20岁，均为大学本科在校学生。所有被试者目前所使用手机均为iphone4/iphone4s，使用时间都在半年以上，且均有操作手机搜索引擎的经验。

3.2 实验条件

测试所用手机为iphone4/iphone4s。其他设备和工具包括联想电脑（Win7系统，带有E-PRIME系统）以及手机固定架。

3.3 实验材料与任务

实验材料包括自编的手机使用满意度调查问卷和测试相关材料。实验测试任务包括两类：

第一类任务：按照电脑显示任务说明找到相关网页，对搜索到的项目按照相关程度从高到低进行等级排列。实验选用的query主要来源于Heting Chu 的研究[17]。采用这些query的原因是它们复杂程度不同，并且经过美国长岛大学图书馆员大量真实案例提炼，本文亦是对该研究[17]的拓展与延续。这些搜索项目包括三种类型，即词语搜索、句子搜索与通过布尔搜索词“和”将词语连结组成的搜索。其中，考虑到实验涉及的具体情况，笔者对其中的句子搜索做了相应修改。具体的搜索项如下：

①社会志愿者（词语搜索）

②抄袭（词语搜索）

③记忆和神经生物学（布尔词语搜索）

④作家海子的心理分析（句子搜索）

第二类任务：以问卷形式从三个维度考察手机搜索引擎可用性，分别是准确性、省时性以及总体满意度。问卷采用李克特7分量表，例如，在省时性维度上，1代表极小程度的节省，7代表被试者认为该搜索引擎能够帮助自己节省了大量的时间去完成搜索任务。

3.4 实验设计

实验的自变量为手机搜索引擎类型，包括神马、百度、搜狗和必应四种。

任务类型一的因变量是相关性等级排列，任务类型二的因变量为主观评定分数。

为排除被试间效应，实验采用被试内设计，24名被试者随机分为4组，每名被试者要完成4种搜索引擎的所有任务操作，共16个任务。拉丁方法被用来消除实验之间的顺序效应。

3.5 实验流程

实验开始首先由主试者向被试者详细介绍实验指导语，然后是手机搜索引擎熟悉的过程，大约10分钟左右，进入练习阶段。练习结束后，被试者完成基本信息的填写，正式进入实验。

被试者按照电脑显示的任务流程进行操作，要求每名被试者按照自己的判断标准独立完成每个搜索任务的排列（例如，“快乐”这个搜索词，输入后点击搜索，被试者要对这一页的前十个搜索项完成相关性从高到低排列），并且记录下自己的判别标准。然后小组成员对自己的等级评定和判定标准进行讨论。这种小组讨论的目的是为了提高排列质量，在以往的研究中这种团体共识的方法被证明是有效的[18]。被试者可以基于小组的讨论改变他们的排序，个人的排序结果会聚合到小组的平均水平，这一过程将减少因个别差异所造成的影响。

最后，所有搜索任务完成后，以问卷的形式获得被试者对每组搜索引擎的准确性、省时性和总体满意度水平。

4 实验结果及分析

4.1 搜索结果的等级排序

通过使用斯皮尔曼等级相关的方法，我们得出搜索引擎的等级排列与用户等级排列间的的相关程度。相关度越高，则搜索引擎的结果排列越优秀。具体内容参见表1。百度在四个搜索中均达到显著性水平，都在0.7以上。神马在布尔搜索中得到0.754的高显著性，其余搜索则不显著。搜狗和必应都在布尔搜索和句子搜索中达到高显著性，词语搜索并不显著。

表1 手机搜索引擎排序与用户等级排序的相关度

搜索引擎 \ 搜索项	社会志愿者	抄袭	记忆和神经生物学	作家海子的心理分析	平均值
神马	0.510	0.424	0.754*	0.213	0.368
百度	0.740*	0.723*	0.831**	0.814**	0.794
搜狗	0.518	0.515	0.708*	0.719*	0.615
必应	0.114	0.253	0.738*	0.777*	0.470

注：*表示 $p<0.05$，**表示 $p<0.01$

表2 召回率（搜索结果的前50%）

搜索引擎 \ 搜索项	社会志愿者	抄袭	记忆和神经生物学	作家海子的心理分析	平均值
神马	68.3	63.5	87.3	34.3	63.5
百度	80.2	80	83.3	85.6	82.2
搜狗	69.5	62.8	82.2	73.9	72.1
必应	57.1	52.7	88.5	80.6	69.7

4.2 召回率

召回率是指搜索到的相关项目与搜索到的所有项目的百分比[19]。召回率一直以来被认为是评价搜索引擎功能指标之一。本文计算的是前50%搜索结果的召回率，具体内容见表2。

百度在四个query搜索中均达到80%的召回率。所

有搜索引擎在布尔搜索中召回率都较好，皆在80%以上。神马与搜狗在前两个搜索词中表现是60%以上，必应是50%左右。

4.3 用户主观满意度

被试者使用四款搜索引擎的满意度评价结果如表3所示，在测试的三个可用性维度上，都达到了显著性水平。在准确性维度上，$F(3,92)=7.431$，$P=0.02$。对于省时性，$F(3,92)=4.191$，$P=0.019$。最后，用户总体满意度也达到显著性，$F(3,92)=4.85$，$P=0.01$。这表明，被试者对四款引擎的主观评价是存在明显差异的。

4.4 结果分析

由表1我们可以看出，百度表现出了最佳的排列。在四个query搜索中都表现出了显著相关，搜狗与必应在布尔搜索与句子搜索中都表现出了显著相关，但词语搜索的排列结果显示它们并没有很好地与人们的搜索习惯保持一致。相比较而言，在词语搜索方面，搜狗要好于必应。在小组讨论中，测试用户反应他们在应用必应搜索词语时，很难找到与搜索词有效且直接相关的信息。例如，在搜索“社会志愿者”时，搜索结果大都局限于地方性的志愿者招募信息，并且这些信息很多已经不具有时效性。

尽管神马搜索是目前市场上应用排名第二的手机搜索引擎，但在四个query中只有布尔搜索达到了显著相关，同其他搜索引擎相比，在句子测试中表现出较低的相关。我们对神马引擎搜索结果进行了分析，发现在句子搜索结果中，只有第十个选项与“心理分析”有关，前九个选项都只显示了与海子有关的信息。没能充分的分析句子中的主要成分或许是造成其在句子搜索中表现不理想的重要原因之一。需要说明的是，本文所用测评句子只有一例，可能会因为句子特殊性而造成误差，需要在今后研究中进一步讨论。

被试者主观满意度的实验结果中，在有效性、省时程度（效率）和主观满意度层面体验最佳的分别是百度（M=5.166）、神马（M=5.375）、百度（M=5.333）。笔者进一步对这四款手机搜索引擎做两两比较，发现在这三个维度上，神马、百度以及搜狗间的差异性皆不显著，这可能说明被试者对三款引擎满意度上没有绝对的判定结论，而只是在不同维度上有相应的满意倾向性。对于必应，则提示可能需要开发者做出更多的改进，以满足人们的使用需求。

表3 手机搜索引擎主观满意度

搜索引擎 / 评价指标	神马		百度		搜狗		必应	
	M	SD	M	SD	M	SD	M	SD
准确程度	4.667	0.816	5.166	0.408	4.320	0.967	3.230	0.894
省时程度	5.375	0.744	4.875	0.353	4.500	0.547	3.625	1.224
总体满意度	5.166	0.752	5.333	0.816	4.466	1.129	3.500	1.043

5 结语

本文采用实验法与问卷法分别评估了手机搜索引擎的有效性及主观倾向性。实验提供的针对排列结果的等级相关使研究者能在短时间内判断一个搜索引擎的有效性，相对于以往研究针对单个项目的评价具有更高的效率。另外，这种技术提供给用户更多的可比较的信息及线索，用户通过结果比较所提出的意见可以提供给研究者更多的用户评价标准或操作习惯等信息，而这些内容对于丰富开发者对产品可用性的认识及进一步优化产品是非常重要的。当然，这种技术也有其不足之处，等级排列相对于单个结果的评价施加给人们更多的认知负荷，进而产生疲劳影响到结果排列的质量。合理安排实验程序、适当地增加休息能够有效减少认知负荷所带来的影响。结合问卷法，研究者可以高效地收集手机搜索引擎可用性综合信息，做出较全面的评估。结合所有研究结果，本研究得到以下结论：

（1）从被试者等级排列结果来看，百度搜索引擎的表现最好，搜狗在四个搜索中有三个是显著相关，可以认为是令人满意的。神马在本研究中句子搜索中表现较差，但需要更多的测试与研究来验证。必应虽然在布尔搜索与句子搜索中有较好的表现，但被试反应其

词语搜索结果令人失望。

（2）从主观满意度角度来看，被试者对神马、百度和搜狗的评价没有明显的差异，必应评价较低，在准确度以及满意度等方面还有待提高。

参考文献

[1]中国IT研究中心. 2014年7月中国移动搜索市场研究报告[EB/OL].[2015-09-08]. http://www.cnit-research.com/content/201408/612.html.

[2] Nah F F-H, Siau K, Sheng H. The Value of Mobile Applications: a Utility Company Study [J]. Communications of the ACM, 2005, 48(2): 85-90.

[3] 刘子慧, 张锋, 陈硕. 基于用户体验的谷歌和百度搜索有效性比较研究 [J]. 浙江大学学报（理学版）, 2010, 37(5): 605-610.

[4] Barnard L, Yi J S, Jacko J A, et al. An Empirical Comparison of Use-in-motion Evaluation Scenarios for Mobile Computing Devices [J]. International Journal of Human-Computer Studies, 2005, 62(4): 487-520.

[5] Nielsen J. Mobile Usability. Jakob Nielsen's Alert Box [M]. 2009.

[6] Croft W B, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice [M]. Addison-Wesley Reading, 2010.

[7] Kent A, Berry M M, Luehrs F U, et al. Machine Literature Searching VIII. Operational Criteria for Designing Information Retrieval Systems [J]. American documentation, 1955, 6(2): 93-101.

[8] Vaughan L. New Measurements for Search Engine Evaluation Proposed and Tested [J]. Information Processing & Management, 2004, 40(4): 677-691.

[9] Kantor P B, Voorhees E M. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text [J]. Information Retrieval, 2000, 2(2-3): 165-176.

[10] Tang R, Shaw Jr W M, Vevea J L. Towards the Identification of the Optimal Number of Relevance Categories [J]. Journal of the Association for Information Science and Technology, 1999, 50(3): 254.

[11] 李珏伶. 搜索引擎网页相关性评估方法设计及其在 rank 模型上的应用 [D]. 北京:北京交通大学, 2011.

[12] Standardization I O f. ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability [M]. 1998.

[13] Jordan P W. An Introduction to Usability [M]. CRC Press, 1998.

[14] Madan A, Dubey S K. Usability Evaluation Methods: a Literature Review [J]. International Journal of Engineering Science and Technology, 2012, 4(2):590-599.

[15] Dubey S K, Rana A. Analytical Roadmap to Usability Definitions and Decompositions [J]. International Journal of Engineering Science and Technology, 2010, 2(9): 4723-4729.

[16] Nielsen J. Usability Engineering [M]. Elsevier, 1994.

[17] Chu H, Rosenthal M. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology [C].Proceedings of the PROCEEDINGS OF THE ANNUAL MEETING-AMERICAN SOCIETY FOR INFORMATION SCIENCE, 1996.

[18] Zhang X M. Collaborative Relevance Judgment: A Group Consensus Method for Evaluating User Search Performance [J]. Journal of the American Society for Information Science and Technology, 2002, 53(3): 220-231.

[19] Kumar R, Suri P, Chauhan R. Search Engines Evaluation [J]. DESIDOC Journal of Library & Information Technology, 2003, 25(2).

作者简介

孔宁，男，1988年生，硕士，研究方向：应用心理学、信息检索，E-mail：kongning52678@126.com。
张常洁，女，1965年生，硕士，浙江工业大学副教授，研究方向：应用心理学。

Usability Evaluation of Search Engines for Smart Phone

KONG Ning , ZHANG ChangJie
（Zhejiang University of Technology,Hangzhou 310000,China）

Abstract: The purpose of this study is to assess the smart phone search engines, which are rapidly developed but little researched. The subjects are God horse, Baidu, Sogou and Bing. First, we compared the effectiveness of the search engine by experiment. Then, the accuracy of the search engines, saving time and overall satisfaction were collected through questionnaire. Experimental results show that Baidu shows the best performance, while other search engines require some optimization in different aspects. The results effectively distinguish the performance of search engines, which also provides a direction for improvement and development.

Keywords: Smart Phone;Search Engine;Evaluation

（收稿日期：2015-09-09；编辑：雷雪）