

基于 UES 模型的信息检索系统评价方法探讨

庄梦蝶

(南京大学信息管理学院, 江苏南京 210093)

摘要: 本文通过对信息检索系统评价实验方法以及主要评价参数进行收集整理, 并以国际范围内热门项目为例, 分别从面向系统评价方法以及面向交互评价方法两个角度指出信息检索系统评价研究所面临的挑战以及主要局限。通过对 UES 模型参数与从系统日志中抽取的系统参数的相关性检验, 并综合讨论两种评价方法的优势, 指出结合面向交互评价的观察角度以及面向系统评价的数据收集方法可以形成一种全面且经济的评价方法。

关键词: 信息检索系统系统评价; 面向系统方法; 面向交互方法; 系统交互

中图分类号: G356.8

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2015.02.012

Evaluation Method's Discussion for Information Retrieval System Based on UES

Zhuang Mengdie

(School of Information Management, Nanjing University, Nanjing 210093)

Abstract: This paper indicates the envisaging challenge and main limitation in its research according to collect and summarize the main stream approaches to information retrieval system and its current evaluation parameters. by correlation checkout form both of parameters based on UES model and elicited for system log, and integration to advantages on both, indicates to form a roundly and ecumenical evaluation method with interactive observation and system-based

Keywords: information retrieval system, system evaluation, system-based approach, interactive approach, interactions with system

1 引言

对信息检索系统评价的研究可以促进不同检索系统之间的比较, 并且加快信息资源的整合与深度利用^[1]。这是研究人员一直对它保持热情的原因。虽然各个信息系统存在很大的差异(例如, 信息类别, 信息资源类型等), 但主流的评价方法可以覆盖所有的类型。这些评价方法分为两种: 面向系统的评价方法和面向交互的评价方法。

其中, 面向系统的评价方法主要采用实验的方式来评价检索的算法。一系列实验开始于上世纪60年代初。第一个试验是在克兰菲尔德大学(Cranfield University)的航空学院。该学院研究人员编写了针对摘要索引搜索的评价方法。为了判断实验结果与用户查询之间的相关性, 研究人员设计了用于静态文档群的测试集。这种使用测试集的实验方法注重于检索系统在检索有关信息时的反馈速度和准确性, 它现在仍然被广泛使用在不同的信息检索系统评价中。研究人员研究的

作者简介: 庄梦蝶(1990-), 女, 南京大学硕士研究生, 主要研究方向: 信息检索, 信息系统评价。

基金项目: 2014年国家社科基金青年项目“大数据环境下移动互联网用户行为研究”(14CTQ021)。

收稿日期: 2014年4月28日。

问题是：这个检索算法有没有检索出与查询相关的文件？在这种评价方法中，用户被看成是一个抽象的概念^[2]，而用户的行为以及感受并没有被当成实验数据来收集。虽然长期以来信息检索系统都是从系统的角度来评价的，但搜索的准确度以及反馈速度等只是检索过程的一部分。当用户使用信息检索系统时，他们都参与了这个检索过程。检索系统的性能最终体现在用户是否能够使用该系统得到想要的信息。所以真正适用的信息检索系统在拥有优良的算法之外，还应该能帮助用户的信息认知并提供舒适的用户体验。

面向交互的评价方法是在1970年提出的。该评价方法是将用户的交互表现作为一个评价参数^[3]。它主要采用访谈和实验来观察系统和用户之间的交互，并从中提取数据来衡量用户在使用信息检索系统时的满意程度。研究问题也与传统的面向系统方法有所不同：用户是否能够使用系统检索到令他满意的文件？这类评价方法早期只有定性的描述。随着评价方法的改进，在通过分析访谈和实验收集的数据后，用户的参与度和满意度水平可以按照程度的深浅量化^[4]。但相比面向系统的评价方法，面向交互的评价方法研究的时间短，也没有形成标准的研究范式^[5-6]。大部分研究参数的设计只针对特定的一个信息检索系统，因而无法进行系统与系统之间的对比。

下面对这两种评价方法作进一步的剖析。

2 面向系统的信息检索系统评价方法

就像早期的克兰菲尔德测试一样，定量实验被广泛使用于面向系统的信息检索评价中，以便评价检索算法的性能。研究人员假设用户的查询和系统返回结果之间存在稳定以及独立的二元关系。Moghadasi在他的研究中将面向系统的评价方法描述为一个通过“量化系统检索出的文件组与假设的相关文件组之间的相似度，以观察检索系统的功效”的评价方式^[7]。这些定量实验通常使用一些测试群^[8]。这些测试群包含了设定的用户查询以及与其相关的文件检索结果^[9-11]。实验时，研究人员会在检索系统上测试一系列的查询。根据返回的文件列表与测试集的假设结果来计算检索算法的评估参数。再结合查询时间等基本数据进一步评估这一检索系统。表1列举了面向系统评价方法的常用评估参数^[12]。

在全球范围内，有一系列的工作室主持并鼓励面向系统的评价方法研究，例如TREC和CLEF。TREC是文本检索会议（Text Retrieval Conference）的缩写，由美国国家标准与技术研究院（NIST）和美国国防部联合赞助。它创立于1992年并在每年举办一系列的课题研讨会。TREC长期支持大数据信息系统检索算法评价的课题，同时支持着有关医药、信息行为以及知识库加速等项目。近年来，TREC专注于将实验数

表1 面向系统评价方法常用参数

参数	描述
Recall	Recall也称作召回率，用查询到的相关文件数除以数据库中相关文件的总数来计算
Precision	Precision也称作准确率，用查询到的相关文件数除以查询到的文件总数来计算
F-Measure	F-Measure对Recall和Precision两个值做加权调和平均。 $F = 2 * (Precision * Recall) / (Precision + Recall)$
Average Precision (AP)	AP同样是结合了Precision和Recall的参数。它表示的是Precision-Recall曲线下的面积
Mean Average Precision (MAP)	MAP是用来评价一组测试集。它将每个测试集的AP求平均值
Geometric Average Precision(GMAP)	GMAP也是用来评价一组测试集。它将每个测试集的AP求算术平均值
Precision at n	n表示前n个查询结果。Precision at n用前n个结果中相关的文件数除以n来计算。因为用户倾向于查看排名靠前的结果，所以这个参数更符合实际情况
Mean Reciprocal Rank(MRR)	MRR是对一组测试集各自的搜索排名的倒数求平均值。它适用于相关文件总数少，Precision值很高的系统

据, 临床纪录与公众检索资源整合的系统评价。这个项目研究的是如何将现实生活中个体的医疗情况和各种实验或者临床数据相关连并整合成一个开放的数据库。这将有助于医护人员通过实验数据及以往的病例快速做出合适的诊断, 同时公众也可以检索这些数据来加强自身的健康知识。但针对这类系统的评价方法面临一个挑战: 用户群对信息的需求有差异。例如在这个项目中 Lupu 和他的同事指出: 普通用户和专家对于化学资料的看法与需求都存在很大差距^[13]。普通用户希望数据库能涵盖主要医药常识且简单易懂, 而专家们则希望得到全面的医药信息并包含专业而深入的解释。这使得用户对查询与文件之间的关联集合有不同的要求。这与面向系统评价方法假设的查询与文件存在稳定且独立的二元关系相矛盾。除此之外, 另一个机构 CLEF (Conference and Labs of the Evaluation Forum) 注重多信息类型的系统评价。近来由 CLEF 举办的主要项目是 CHiC。CHiC 旨在针对数字化的文化遗产数据集开发评价系统。然而这个数据集同时包含了文本数据以及图片数据。在自然情况中, 用户对文本排序的敏感度较图像更高。例如, 用户倾向于按排序结果顺序浏览文本文件, 但浏览图片文件的顺序则相对随机。对文本检索上有效的评价参数不能反映图像检索的实际使用情况。

然而, 面向系统的评价方法受到参数设计的局限, 面临以下两项挑战。

(1) 面向系统评价方法假设查询与文件间的关联是稳定且独立的二元关系, 这与真实的用户需求相冲突。不同用户群体对信息需求的不同导致了他们对查询与文件的关联度认知存在差异。例如在 TREC 医疗信息资源整合评价项目中, 适用于普通用户的检索算法不能满足专家的信息需求。而面向系统评价方法使用的评估参数都是基于对检索结果与查询语句之间存在稳定且独立的二元关系的假设。这违背了真实的用户需求, 因而其评价结果不能充分反应检索系统在真实使用时的性能。新的信息检索系统评价应当增加评价的维度, 多方面地对系统进行评价。

(2) 面向系统评价方法的传统评价参数集中在对算法的分析上, 忽视了用户对结果主观选择性, 只针对算法的评价结果不能反映用户的实际使用情况。用户是参与信息检索系统应用的重要组成部分。他们在面对检索结果时并不是一定按照系统的返回结果顺序浏览, 而是主观选择认为相关联的检索结果。例如在 CLEF 的 CHiC 项目中, 人们对文本结果排序的敏感度较图像结果更高。对于文本数据, 用户倾向于选择排序靠前的结果。而对于图像数据, 用户的这一倾向并不明显。应用在文本数据检索算法上相当有效的评价参数不能适用于图像数据检索算法。新的信息检索系统评价方法应当在不同的数据集上保持稳定性, 并加深对用户的信息寻求模式的理解, 充分反映检索的真实结果。

3 面向交互的信息检索系统评价

虽然信息检索系统在过去很长的一段时间里都以面向系统的角度进行评价, 但是近几十年来研究人员提出了面向交互的评价方法。在传统的信息检索评价, 也就是基于系统的评价中, 用户没有被直接研究。而从系统交互的观点看, 信息系统返回给用户的信息以及用户在使用信息检索系统时的体验会影响用户对这个信息资源的理解, 并进一步作用于用户的决策表现。Cool 和 Belkin 认为: “由于信息检索的研究人员认识到信息查询过程中固有的交互性, 面向交互的研究已经得到重视并将会加强。^[14]” 面向交互的信息检索评价是一门独立的学科, 它结合了信息检索研究、信息行为研究以及人机交互研究等多个领域知识^[12], 研究的是人与信息系统之间的交互以及用户在使用信息系统的感受。其中的评价数据收集方法与社会科学学科 (例如心理学, 教育学) 的方法类似, 主要采用访谈和实验来观察用户使用系统的行为与感受。首先, 研究人员会选取一个信息检索系统作为实验工具。然后, 针对这个系统设计一套具体的任务, 并要求用户在实验中完全遵照任务的指示。用户与信息系统交互的感受会在任务刚结束后以系统日志文件、问卷

调查、访谈报告和其他自我报告等形式记录下来^[15]。不同的评价方法会自行定义完全不同的评价参数,但大部分参数可以归类到主要的几个参数集。表2列举了在面向交互的评价方法中常用的参数集。

然而,面向交互的信息检索评价方法也面临以下3项挑战。

一是,面向交互的信息检索评价结果过于总结性,无法给出明确的指导结论来帮助系统改进。因为这类方法使用的数据大部分是非直接的观察物,例如用户完成任务的程度、用户对系统的主观满意度等。这些参数包含大量不易观察且不易准确量化的用户心理活动,用户的主观表述和研究人员对表述的理解也会使结果存在偏差。这些原因使得相应的评价结果具有局限性。同时,用户的个体差异也严重影响他们的主观感受。例如,用户的教育程度,用户的年龄等个人背景等。这也使得评价结果非常笼统,只能解释检索过程中出现的各类现象,而无法给出明确结论来帮助系统改进。新的评价方法应包含可以准确量化的评价参数,并淡化个体差异对评价结果的影响。

二是,面向交互的信息检索评价结果在应用于多类系统间的横向比较时存在局限性。现存的所有面向交互的评价方法都是根据某一信息检索系统设计,这使得评价方法非常有针对性,但也使研究人员无法用一种评价方法对多个信息系统做可靠的横向对比。例如UES (User Engagement Scale)^[15]就是一种用以分析用户印象为主的多维

度评价方法。这个评价方法包含了多个维度:系统和内容的美观程度、系统的新颖性和实用性、用户参与使用过程的积极程度以及用户对此体验整体评价^[16]。然而,研究人员发现,即使这个评价方法在多种类型的系统上都具有良好的信度与效度^[17],但在不同的系统中(如文字检索系统^[16])维度内的少许问项非常不稳定。若要达到完全一致的效果需要对个别问项进行修正或者删除。新的评价方法应该增强其应用的广泛性,使之适用于系统间的横向比较。

三是,面向交互的信息检索评价方法需要投入大量的时间成本与人工成本。面向交互的评价方法主要采用问卷调查和访谈等数据收集形式,这需要大量人力等参与。而这些数据收集需要在用户使用系统之后,并有时需要若干次跟踪采访,这使得实验过程相对于面向系统的评价方法需要更长的时间以及更多的人力投入。Toms、Freund和Li描述交互式信息检索评价的研究“研究费时,还需要大量的人力。这使参与者和研究者都会感到十分劳累”^[18]。新的评价方法应该开发新的数据收集方式以降低成本,同时增加非主观参数的个数以减少个体差异对结果的影响,并且提出标准的参数计算体系以便对比不同系统的性能。

4 基于UES模型的案例研究

面向系统的评价方法只集中在对算法的分析上,从而忽视了用户对检索结果评价的影响。同时,面向系统评价方法假设查询与文件间的关

表2 面向交互的主要评价参数集

参数	描述
实验背景	包含了检索系统的背景以及参与者的背景。系统的背景包括检索类型、数据集的类型。参与者的背景包括年龄、性别、对主题的熟悉度、地点等
交互参数	参数集由从系统日志中提取的变量构成。与面向系统的评价方法参数不同,Interaction参数完全描述用户与系统的交互过程。例如用户浏览页面的数量、检索语句的长度、检索语句的数量、检索的时长等
使用性能	数据集由实验的直接结果构成,描述系统对用户完成任务起到的作用。例如任务的完成数量、任务的完成质量、完成任务消耗的时间
易用性	表示用户通过使用系统检索信息后主观上对系统的评价。这涉及系统的界面设计的便捷度、用户体验的愉悦度

联是稳定且独立的二元关系，这与真实的用户需求相冲突。基于交互的信息检索评价方法虽然将用户的作用计入评价，但因为参数使用大量主观的数据，这使得结论具有局限性，不能直接且明确的帮助系统改进。而评价方法的设计步骤也限制了系统间的横向对比。除此之外，基于交互的数据收集方式耗时费力，不利于大规模评价的展开。本案例选取 UES 模型^[15]对两种评价参数之间的相关性进行研究。UES 模型（User Engagement Scale）是 2008 年 O’ Brien 和 Toms 提出的一个用来量化用户对使用科技产品的态度的模型。UES 包含 6 个维度，共 31 个评价指标参数。与其他面向交互的评价方法相比，UES 是第一个将参与度当作是用户的心理感受来研究的模型。同时认真推敲了其维度与维度之间的关系，从 400 多个问项中得出了稳定的六维结构并筛选出 31 个评价指标参数。近 10 年来，通过在网上购物、文本检索、图像检索等不同检索环境下的反复实验，UES 模型的这一结构均显得比较稳固。本案例选取 UES 模型是因为它评价指标参数清晰，问项结构稳定，适合与外部参数做相关性检验。同时，纯粹以用户心理为评价角度，适合与系统自动记录的数据进行对比。各个维度内的问项表 3 显示了它的 6 个属性。

此案例所使用的 UES 数据和系统数据是在 2013 年的 CHiC 电子文化遗产项目中采集的^[19]，测

试系统是图像信息检索系统。实验任务是让参与者随意的检索或浏览他们想要的图片，直到他们感觉无趣为止。而 157 名参与者在使用该系统进行检索后也完成了一份 UES 问卷。相关的系统日志则被自动记录在服务器里。UES 数据和系统日志数据在试验后被导入 SPSS 软件进行分析。

通过计算 15396 条系统日志记录数据生成 14 个系统参数，并通过探索性因子分析法（EFA）^[20]对生成参数进一步分析，筛选出 3 个能够代表用户交互总体情况的系统参数。表 4 展示了这 3 个系统参数。

将每一个用户对应的 3 个系统参数看做用户的系统属性，记为因子 1，它所表示的是在用户的整个检索过程中用户对这个检索系统感到有趣的程度。为了研究来自系统数据的属性与 UES 属性的关系，将因子 1 与 UES 模型的 6 个属性做 Pearson 相关性检验，结果如表 5 所示。

其中，美观度以及易用性都与因子 1 之间几乎没有关系（ $r < 0.1$ ）。这可能是因为实验环境是图像信息检索系统。由于图像本身的吸引力，用户比较不容易被界面的美观度所影响；而实验任务是让用户自由的检索，这会让他们觉得没有负担；同时系统与市面上主流的检索系统很相似，这使得用户能容易且熟练地使用实验中的检索系统。这些试验环境因素使得用户对系统交互界面的美观度和易用性感受几乎没有影响到他们的实

表 3 UES 的 6 个属性

UES 维度	描述
美观度	美观度主要描述用户对一个电脑程序界面外观的评价
实用性	实用性测量用户对此系统检索性能的评价以及用户是否会向他人推荐这个系统
积极程度	积极程度表示的是这个系统所吸引用户、引发用户兴趣以及娱乐用户的程度
专注力	专注力指的是用户在使用系统时注意力的集中程度
新颖性	新颖性表示的是用户觉得系统有着意想不到的、令人惊讶的以及不熟悉的特征的程度
易用性	易用性描述的是用户感觉进行和完成任务所需要花费的精力

表 4 系统参数描述

系统参数	描述
timelookatbookbag	用户浏览收藏夹的时长
saved	用户所保存到收藏夹的页面总数量
totaltime	用户在感到无聊之前浏览页面的总时长

表5 UES属性与因子1的相关性检验

UES属性	因子1
美观度	0.097
实用性	0.277
专注力	0.354*
积极程度	0.383*
新颖性	0.393*
易用性	0.072

实际检索行为。

而值得注意的是,专注力、积极程度和新颖性与因子1的相关性被归为中等偏强(>0.35)。这说明用户在这3个方面的感受可以通过系统参数表现出来,也意味着表示用户整体行为的系统参数很有可能贡献于这3组UES属性。专注力和积极程度这两个维度本身是描述用户在使用系统时个人的投入状态,这与系统记录下来的用户行为有着良好的正相关。而就新颖性来说,它与美观度、易用性等同样是描述用户对一个系统的感受,但它却可以直接影响到用户的行为。因为这个检索环境与主流搜索引擎十分相似,使用户感觉到“新颖的”其实还是检索的结果。当用户对搜索结果感到新颖的时候,会延长花费在检索上的动作和时间以便更好的探究。专注力、积极程度和新颖性这3个参数可以被系统参数表示。具体可以替代到UES维度中的系统参数还需要通过更多横向实验进一步细化。

5 信息检索系统评价方法探讨

面向交互的评价方法因为关注的纯粹是用户和系统的交互,不存在对关联度的假设,同时考虑到用户在信息检索过程中的主导作用。而面向系统的评价方法所使用的系统日志数据属于系统自动纪录的客观数据,一方面不需要额外的人力来采集数据,增强了评价方法的经济性。另一方面交互的信息是用户的行为表现,从而降低了主观表述与表述理解带来的影响。对于这两种评价方法,一味地单打独斗是不可行的。而若将两种评价方法相结合可以部分程度上抵消这些局限。

从上述案例可见,表示整体用户行为的因子1与专注力、积极程度以及新颖性有着较为显著的关联。这样的结果进一步确认了面向系统的评价方法的参数和面向交互的评价方法的参数之间存在相关性,而将其中一类的评价参数带入另一类评价方法中是合理并可行的。

若从系统数据中提取与交互相关的数据以整合成新的面向交互的评价参数,并以面向交互的评价算法为指导,形成新的可以横向对比系统表现的参数计算体系。这将解决两类评价方法的局限,并更加全面、经济地评价信息检索系统。这将需要以下两个步骤。

(1)研究并提取系统日志中包含的用户交互的相关数据。系统日志自动记录了用户与系统交互的时间与动作。Toms和她的同事通过反复关联用户参数与系统参数^[21],得出这两类数据存在很强的内在联系。从这些原始数据中,可以提取出与交互相关的二级数据。例如用户查询的语句、用户停留在某一个特定模块的时间、用户查看其检索结果的时间以及用户使用自主查询或者系统导航功能的次数等。这些数据间接反映了用户在使用该信息检索系统时的感受。譬如当用户对一个结果感兴趣,他会较久地停留在这个检索结果上。这类数据虽然来源于采集简单的系统日志数据,但与以往的面向系统的评价参数不同,它直接以数字的形式表现了用户在使用系统时的感受。使用代表用户交互的系统参数的优势在于系统日志数据记录的是用户使用检索系统的实际情况,这将用户需求的间接表现计入评价参数,增强了评价方法的全面性。

(2)将表现用户交互的系统数据与已有的面向交互的信息系统评价参数相关联。通过信息行为研究以及人机交互研究,可以将从系统日志中提取出的相关数据与已有的面向交互的评价参数相关联^[21]。以研究这些系统参数分别属于哪一个面向交互的评价参数集;是否存在一个参数计算体系,能用多个系统数据来量化面向交互的评价参数;这个参数计算体系是否在不同类型的系统上可信且有效。这个方法将有3个优势,一是因

数据来源于系统自动记录的日志文件,这直接降低了传统的交互数据收集方法所需的人工以及时间成本。二是若存在评价范式,它可用来衡量不同系统之间的差异。这也为完善标准化的信息检索系统评价模型打下基础。三是经横向比较之后的评价结果可以为信息检索系统提供更明确的指导。

相比仅面向系统的评价方法,这个新体系减少了传统的面向系统方法中数据采集过程的时间成本和人力成本,增强了经济性。同时采用用户的行为数据来表示他的主观体验,也降低了主观表述与表述理解带来的影响。相比仅面向交互的评价方法,这个新体系考虑到了用户在信息检索过程中的主导作用,与检索系统的实际使用结果更为接近。

6 结论

综合面向系统的评价方法与面向交互的评价方法所使用的实验方法与评价参数的讨论,信息检索系统评价研究面临着挑战:一方面,面向系统的评价方法假设查询与文件间的关联是稳定且独立的二元关系,这与真实的用户需求相冲突。同时因为只关注检索算法,忽视了用户对结果的主观影响;另一方面,面向交互的评价方法使用大量主观的数据,因而结论往往只能解释现象,不能直接且明确地帮助系统改进。而针对单一系统设计的评价方法也限制了系统间的横向对比。同时数据收集方式耗时费力,不利于大规模评价的展开。

案例实验证实用户的整体行为与 UES 模型的维度有着较为显著的关联,说明使用一类的评价参数带入另一类评价方法的思路是合理并可行的。实验仍留下了许多可研究的角度,像是如何在交互式信息检索系统评价的实验中植入更加标准化的任务,例如给予用户一定的时间限制等。进一步改进此类的实验能够允许这个新思路在不同系统中的横向对比以及允许提取不同的系统参数用以分析。

两种单独的信息检索评价方法若一直互相

孤立,无法突破已有的局限。新的研究方法应当结合两种研究方式的优点:从便于收集的系统中数据中提取与交互相关的数据以整合成新的评价参数,并结合这些参数和现有的面向交互评价模型提出全面的参数计算体系。这有望完全以系统数据取代传统耗时耗人力的问卷、采访数据以便降低实验的成本,并且全面考虑用户在信息检索中的作用。

参考文献

- [1] Smeaton A, Over P, Kraaij W. Evaluation Campaigns and TRECVID [C]//Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. New York: ACM Press, 2006: 321-330.
- [2] Mandl T. Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance [J]. Informatica, 2008, 32(1): 27.
- [3] Salton G. Evaluation Problems in Interactive Information Retrieval[J]. Information Storage and Retrieval, 1970(6): 29-44.
- [4] Fidel R. Qualitative Methods in Information Retrieval Research[J]. Library and Information Science Research, 1993(15): 219.
- [5] Borlund P. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems [J]. Information Research, 2003(8): 152.
- [6] Jarvelin K, Kekalainen J. Cumulated Gain-Based Evaluation of IR Techniques[J]. ACM Transactions on Information Systems (TOIS), 2002(20): 422-446.
- [7] Moghadasi S, Ravana S, Raman S. Low-cost Evaluation Techniques for Information Retrieval Systems: A Review[J]. Journal of Informetrics, 2013,7(2): 301-312.
- [8] Sanderson M, Scholer F, Turpin A. Relatively Relevant: Assessor Shift in Document Judgements[C]//17th Australasian Document Computing Symposium. Australia: RMIT Press, 2010: 60-67.
- [9] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. New York: ACM Press, 1999.
- [10] Clarke C, Kolla M, Cormack G, et al. Novelty and Diversity in Information Retrieval Evaluation[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Informa-

- tion Retrieval. New York, 2008: 659–666.
- [11] Melucci M, Baeza-Yates R. Advanced Topics in Information Retrieval[M]. Berlin: Springer, 2011.
- [12] Kelly D. Methods for Evaluating Interactive Information Retrieval Systems with Users[J]. Foundations and Trends in Information Retrieval, 2009, 3(1/2): 1–224.
- [13] Lupu M, Huang J, Zhu J, et al. TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation [C]//REC, ACM SIGIR Forum, 2009, 43(2): 63–70.
- [14] Cool C, Belkin N J. Interactive Information Retrieval: History and Background[M]. London: Facet Publishing, 2011: 1–14.
- [15] O’ Brien H L, Toms E G. What Is User Engagement? A Conceptual Framework for Defining User Engagement with Technology[J]. Journal of the American Society for Information Science and Technology, 2008, 59(6): 938–955.
- [16] O’ Brien H L, Toms E G. Examining the Generalizability of the User Engagement Scale (UES) in Exploratory Search[J]. Information Processing & Management, 2013, 49(5): 1092–1107.
- [17] O’ Brien H L, Toms E G. The Development and Evaluation of a Survey to Measure User Engagement in E-commerce Environments[J]. Journal of the American Society for Information Science and Technology, 2010, 61(1): 50–69.
- [18] Toms E G, Freund L, Li C. WiIRE: The Web Interactive Information Retrieval Experimentation System Prototype[J]. Information Processing & Management, 2004, 40(4): 655–675.
- [19] Hall M, Villa R, Rutter S, et al. Sheffield Submission to the CHiC Interactive Task: Exploring Digital Cultural Heritage[C]//4th International Conference of the CLEF Initiative, Valencia, Spain, 2013: 192–211.
- [20] Fabrigar L R, Wegener D T. Understanding Statistics: Exploratory Factor Analysis[M]. New York: Oxford University Press, 2012.
- [21] Toms E G, O’ Brien H L, Kopak R, et al. Searching for Relevance in the Relevance of Search [M]//Context: Nature, Impact, and Role. Berlin: Springer Berlin Heidelberg, 2005: 57–78.

(上接第72页)

- 研究[D].北京:中国科学院研究生院科技政策与管理科学研究所,2008: 17.
- [2] 胡明铭, 黄菊芳. 同行评议研究综述[J]. 中国科学基金,2005(4): 251–253.
- [3] 吴述尧. 科学进步和同行评议[J]. 科技导报,1997(4): 23–25.
- [4] 李芬, 朱紫阳, 丁枝秀. 关于中国同行评议研究状况的分析报告: 基于CNKI核心期刊的文献研究[J]. 中国科学基金, 2009(3): 177–182.
- [5] Department of Health and Human Services.2011 NIH Director’ s Pioneer Award Program (DP1)[EB/OL]. [2011–02–20].<http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-10-008.html>.
- [6] Howard Hughes Medical Institute.2008 HHMI investigator competition[EB/OL].[2009–02–01]. <http://www.hhmi.org/research/competitions/investigator2008>.
- [7] 陈涛, 钱万强. 让科学家在生命医学研究之刃自由舞蹈[J]. 中国基础科学,2010(3): 39–44.
- [8] 林强. 谈科学研究工作中的同行评议制度[J]. 研究发展与管理,1992(1): 57–59.
- [9] 王平, 宋子良, 刘爱玲. 省级同行评议专家选择: 理论与实现[J]. 科技管理研究, 1997(4): 38–40.
- [10] 陈进寿. 从人际关系谈同行评议制的改进[J]. 中国科学基金,2002(3): 56–58.
- [11] 陈喜乐, 李腾达, 刘伟榕. 定性评价方法的极限与超越[J]. 发展研究,2014(5): 16–19.