

科学数据用户相关性指标研究^{*}

高飞¹, 石蕾², 王健¹, 张贵兰¹, 刘建平¹

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 国家科技基础条件平台中心, 北京 100862)

摘要: 通过对科学数据用户相关性判断行为的研究, 探索科学数据相关性判断过程中所使用的指标及模式, 加深对科学数据相关性判定机制的理解, 为设计开发智能化数据搜索引擎提供算法和理论支撑。使用出声思考和事后访谈两种方法对用户相关性判断行为进行研究。数据在真实的环境中收集, 使用摄像机记录实验过程, 采用扎根理论对访谈数据进行编码分析。在定性研究的基础上设计问卷, 开展大样本问卷调查。科学数据用户相关性判断过程使用的标准可以分为数据本体性和数据可用性两类, 共9个指标。与其他信息类型相比, 科学数据检索有更强的目的性, 相关性判断过程不能缺少对数据主题性、质量和权威性的评估, 其他指标在具体情境中, 只有受到相关信息需求刺激才会调用。

关键词: 信息检索; 科学数据; 用户相关性; 标准

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2017.11.005

科学数据指通过科技活动或其他方式所获取的反映客观世界本质、特征、变化规律等的原始基本数据, 根据不同科技活动需要进行系统加工整理的各类数据集, 及用于支撑科研活动的科学数据集^[1]。任何一项科研活动都离不开对数据资源的分析、实验、综合等活动, 然而随着数据规模的扩大和数据共享工作的进行, 科学数据获取效率依旧很低。造成这种现象的直接原因是没有高效的科学数据检索工具, 而更深层次的原因是对科学数据相关性判断机制理解不足, 特别是基础的科学数据相关性标准不健全及指标使用研究薄弱。针对此问题, 本文采用实证的研究方式, 探索科学数据用户相关性判断过程中的使用指标。

1 相关研究

1.1 科学数据元数据标准研究

科学数据是科技工作者在研究工作中观测、收集或加工处理后得到的反映世界客观本质的数据。数据

生产的目的是支持科研活动开展。通过文献调研发现, 现有科学数据标准多基于元数据, 其目的是进行科学数据存储管理和实现数据共享。国外的元数据标准发展比较完善, 知名的元数据标准有都柏林核心元数据、生命科学元数据、地球科学元数据、社会科学元数据等^[2]。都柏林核心元数据不是专门的科学数据标准, 是最早制定的关于网络信息资源的元数据标准, 很多类型的信息资源元数据标准都参考都柏林核心元数据制定。在地球科学元数据标准中比较常用的有美国联邦地理数据委员会元数据标准和国际标准化组织地理信息技术委员会标准两类, 美国联邦地理数据委员会元数据标准主要包括标识信息、数据质量信息、空间数据组织信息、空间参照系统信息、实体和属性信息、分发信息和元数据参考信息, 国际标准化组织地理信息技术委员会标准主要包括标识信息、数据质量信息、数据志信息、空间数据表示信息、参照系统信息、要素分类信息、分发信息、元数据参考信息。生命科学元数据主要包括达尔文核心元数据, 其标准可分为标识信息、事件信息、生物发现信息、生物分类信息、地质环境信息、

^{*} 本研究得到中国农业科学院科技创新工程项目 (编号: CAAS-ASTIP-2016-AII) 资助。

位置信息、位置资源关系信息和参考信息。社会科学类元数据标准可分为文档描述信息、研究描述信息、数据文件描述、变量描述信息、其他相关材料和描述信息。其他标准分类限于篇幅不逐一赘述。国内元数据标准研究起步较晚,但发展迅速。如中国科学院核心元数据标准、中国国家图书馆资源信息核心元数据标准、医药卫生科学数据元数据标准等。科学数据标准的研究初衷并不是为某一科学领域制定,但从后续发展看,学者更倾向制定适用于某个学科领域的元数据标准。所以科学数据元数据标准越来越多,但并没有很好地解决数据共享问题。

1.2 用户相关性指标研究

用户相关性指标研究主要围绕信息对象、用户认知和情境三要素间的复杂关系展开^[3],通过不断增加信息对象类型和情境类型的方法来探索更多的相关性标准及其使用模式。早期主要针对学术论文等学术信息类型,之后逐渐扩展到新闻、音乐、视频、医学影像等类型。相关性指标使用模式研究的数量相对较少,主要思想是借助标准使用频率评估其在不同情境下的重要性,进而估算其使用概率。Park认为影响用户相关性判断的因素有用户认知背景、主题相关背景和信息使用动机三类^[4];Cool等探索除主题相关外其他对文档相关性判断存在影响的因素,分别为主题、内容或信息、格式、表述、价值取向以及用户自身等^[5];Barry认为影响用户文档选择的因素主要有七类,包括文献内容、用户研究背景和经历、用户个人偏好、与其他信息和资源的关系、文献来源、文献本身的情况,以及用户个人情况^[6];Wang等建立了用户文档选择认知模型,并提出11项相关性标准,具体为主题性、研究方向和水平、研究领域、新颖性、期望质量、实时性、阅读时间、可获取性、特殊需求、权威性、来源^[7];Markkula等根据记者对图片的选择行为提出7个图像的相关性标准,具体为主题性、技术性、文本影响、视觉效果、花费、时效性、个人情感^[8];Rieh认为用户主要根据主题性、信息质量和认知权威性对网页信息进行判断^[9];Balatsoukas等使用眼动仪探索用户在网络环境中如何进行相关性判断,用户使用最多的标准依次为主题性、范围、使用者的背景和质量^[10]。

综合已有的研究,用户相关性标准研究在信息对象和标准数量上有了比较重大的进展,但一些重要信息类型(如科学数据)的研究较少。现有基于元数据的

科学数据标准的目的仅是对数据进行科学地存储管理,并没有注重用户查询数据时判断决策的主体地位。因此,虽然数据共享的规模日益扩大,但用户获取数据的效率很低。本文针对此问题,进行科学数据用户相关性标准研究。

2 实验方案

2.1 情景实验

选取被试人员36名,其专业分别为生物化学与分子生物学、微生物学、农业遥感、农业经济等。其中,硕士研究生32名,博士研究生4名;男生14名,女生22名。被试人员都参与了科研项目,对于科学数据有较强需求,符合实验条件。

被试人员根据当前研究任务和真实需求自拟任务进行网络数据检索。主试人员提供可以自由检索的环境,保证被试人员能够进行出声思考,不受外界影响。被试人员的检索时间在30分钟以内,实验结束分3种情况:(1)查询到相关数据,并打开浏览;(2)查询到相关数据,并进行下载;(3)对多个检索结果进行评估后,没有得到相关数据。使用录像机记录检索过程和出声思维数据。

检索实验反映被试人员当前阶段的真实需求。该任务通过20—30分钟的检索会话完成,反映被试人员短时记忆中信息的加工情况。检索实验完成后立刻进行访谈,访谈时让被试人员观看自己检索数据的视频。访谈主要围绕需要哪些数据,数据来源,是否找到相关数据,从哪些线索判断数据相关/不相关,为什么这些线索可以判断数据相关/不相关。访谈全程录像,以便于后续数据分析。

2.2 问卷调查

情景实验结束后,根据分析结果设计用户相关性指标重要程度测量问卷,并进行问卷调查。通过网络问卷的形式,对问卷进行有偿发放与回收。调查目标对象是硕士及以上学历研究生和数据使用频率较多的本科生,这类对象具有一定的数据科研基础,学习或工作中经常使用科学数据,能够根据自身经验对标准的重要程度作出判断。共发放问卷669份,回收669份,其中有效问卷544份。回收率100%,有效率81%。

2.3 数据编码与分析

出声思考和访谈转录为文本数据,基于扎根理论进行编码分析,重点是对文本中出现的相关性判断指标进行编码。这些指标是用户头脑存储的知识,是用户评估数据价值的工具性概念或逻辑关系判断根据,在文本中体现为“与我研究相关”“最新的”“有名的”“因为它比较准确,所以”等概念性语句。将标志性文本编码,并根据其意义命名(在编码过程中不断进行修改),分别为“主题性”“时效性”“权威性”“准确性”,由此得到指标集合。编码出现不一致的情况时,对原材料进行备忘记录与上下文分析,重新编码。编码表在编码过程中不断修改,最终的编码如图1所示。编码由第一作者完成,并由其他3位编码员进行信度检验,编码间的信度分别为80%、81%和79%,均在Krippendorff认可的探索研究可靠性结论的概率范围内(67%—80%)^[11]。编码实例如表1所示。

1 被试	3.10 其他
1.1 任务	4 数据作用
1.2 数据需求	4.1 提供新想法
1.3 数据类型	4.2 支撑研究结论
2 检索方式	4.3 丰富文章,使成果更加容易被认可
2.1 简单检索	4.4 提供技术方法指导
2.2 高级检索	4.5 保留以备后期研究
3 标准	4.6 作为研究对象
3.1 主题性	4.7 其他
3.2 可获取性	5 判断
3.3 质量	5.1 相关
3.4 专业需求	5.2 不相关
3.5 新颖性	5.3 部分相关
3.6 权威性	6 满意度
3.7 便利性	6.1 满意
3.8 可理解性	6.2 比较满意
3.9 时效性	6.3 不满意

图1 编码表

表1 编码实例

文本实例	指标	作用	判断	满意度
IP: 主要是看了一下到底是干什么的,觉得这里面都和我研究的有点偏题了 ^{3.1} ,可能和我的研究有点关系,但是不是特别相关 ^{5.2}	主题性	-	部分相关	-
IP: 因为在NCBI里面搜索出来的基本上就是比较权威的 ^{3.6} ,也是能下载的 ^{3.2}	权威性、可获取性	-	-	-
IP: 不太相关 ^{5.2} ,可以看下。这篇的主题也是可以多了解一下的 ^{3.5}	新颖性	-	部分相关	-
IP: 既然完成任务的,对这些数据还是比较满意的 ^{6.2}	-	-	-	比较满意
IP: 这样来讲,它们支撑起整个文章的骨架 ^{4.2} ,但是想有血有肉的话还需要更多的数据来补充 ^{4.3} ,论文啊,研究啊,才更有意义一些	-	支撑文章、丰富内容	-	-

3 结果与发现

3.1 科学数据用户相关性指标

对文本数据编码、概念归类后得到9个科学数据用户相关性指标定义(见表2)。

从表3可见,主题性是使用频率最多的标准,92.66%的数据进行了主题性判断。主题性应该满足空间位置相符、时间覆盖度相符、内容与研究主题相符的条件。

专业需求指数数据能够满足用户研究领域的特殊需要,如数据发挥的作用、数据的特定格式要求等。专业需求使用频率仅次于主题性,44.04%的数据对其进行评估。

时效性指数数据在时间上满足研究需求。用户对时

效性的评估主要包括两个方面:一是数据的时间覆盖度符合研究需求;二是数据是实时的,能够不断更新。通常数据满足一个方面就能达到用户要求。17名用户使用这条标准,44.95%的数据对时效性进行评估。用户通过数据起止时间、数据更新时间等评估数据的时效性。

质量指数数据的优劣,包括数据的完整性、准确性、可信程度。质量是用数据相关性判断的重要指标,33.94%的数据对该标准进行评估。在科技工作中,数据质量决定研究结果的可靠性,因此在选择数据时,用户会对数据质量进行仔细筛查。

权威性指数数据在专业领域内的认可程度。14.68%的数据权威性进行评估。用户的一次相关性判断不会对所有数据标准进行加工,通常是根据具体情境选择

表 2 相关性指标定义

标 准	定 义	文本实例
主题性	数据与研究主题相符	我要研究2010年焦作市土地利用情况,当然要查询2010年焦作市统计年鉴了
		这个数据与我的研究内容正好一致
专业需求	数据能够满足用户研究领域的特殊需要	这个基因的功能正好是我需要的
		数据描述对基因序列的功能作了描述,不是我想要的
时效性	数据在时间上满足研究需求	研究小麦生育期,我一般选择1981—2010年的数据
		看更新时间啊,我们需要的数据一般都是比较新的
质量	数据的优劣	根据数据来源、数据格式、引用情况,判断数据质量
		这个数据缺失值,质量不高
权威性	数据在专业领域内的认可程度	这个数据平台是国家建立的,上面的数据都是比较权威的
		可以看数据发布机构,比较知名的机构,数据权威性强
可获取性	外界条件不影响用户通过各种方式获得数据	会关注下载链接,看是否能够得到数据
		一般不会使用这里面的数据,没有下载权限
新颖性	没有接触过或者不熟悉的数据	这个数据有认证标志,都是研究比较透彻了,不是我想要的
		以前没有见过这个数据所以想看看
便利性	用户得到数据和使用数据的难易程度	数据太多了,我就看前几条,这样比较方便
		我一般选择云量少的数据,后期处理起来比较简单
可理解性	数据是否容易被理解和接受	选择清晰度较高的图片,看起来比较容易

使用标准。当需要判断的数据量过大时,用户可使用权威性指标对数据进行筛选。一般是从发布数据的机构、数据所在平台、发表数据的期刊、数据发布者等方面来判断数据的权威性。

可获取性指外界条件不影响用户通过各种方式获取数据。可获取性的使用频率较低,11.01%的数据对该标准进行评估。可获取性影响用户对数据的使用,通常用户在初步判断数据相关后才关注该标准。但也有部分用户在相关性判断过程中先关注数据的获取途径,在确定数据可下载之后,才会进行下一步判断。

新颖性指未接触过或者不熟悉的数据。新颖性指标使用频率较低,仅有5.50%的数据对新颖性进行评估。这说明科学数据用户在选择数据资源时,通常已经明确对数据的需求,数据选择的目性强。

便利性指用户得到数据和使用数据的难易程度。在相关性判断过程中,用户更倾向于选取检索结果排列靠前的数据,认为这些数据更容易得到,同时用户也会考虑数据获取后的使用情况,他们更希望获取的数据容易使用。

可理解性指数据是否容易被理解和接受。只有2个用户在检索过程中涉及这条标准。在一个简短的检索会话中,用户对数据的可理解性关注较少。造成这种现

表 3 用户相关性指标使用情况

标 准	数据/条	数据百分比	频 率	百分比	用户人数/个
主题性	101	92.66%	221	43.25%	33
专业需求	48	44.04%	162	31.70%	21
时效性	49	44.95%	51	9.98%	17
质量	37	33.94%	37	7.24%	20
权威性	16	14.68%	16	3.13%	12
可获取性	12	11.01%	12	2.35%	10
新颖性	6	5.50%	6	1.17%	5
便利性	4	3.67%	4	0.78%	4
可理解性	2	1.83%	2	0.39%	2

象的原因有两个:一是用户熟悉检索平台使用,对所呈现的数据格式比较了解,不会刻意关注;二是数据呈现的内容更加直接,用户可以在获取数据后使用专业工具进行数据分析,对于不熟悉的检索环境,用户则希望系统提供的数据检索结果更容易理解。

3.2 相关性指标使用模式

研究发现,一次相关性判断不会对所有相关性指标进行加工;一条数据相关性判断常使用1—4个指标,

平均一次使用1.88个指标;使用2个指标的情况最多,占数据总量的71.56%;使用1个指标的情况占比22.02%;使用3—4个指标的情况较少,占比6.42%(见表4)。

主题性是使用最广泛的指标,所有的数据相关性判断均使用了该指标,其中22.02%的数据仅对主题性一个指标进行评估。可以认为,主题性是相关性判断的基础。在特定情况下,尤其是非相关性判断时,仅使用主题性指标就能得到判定结果。对主题性进行加工后,用户一般还会对质量、专业需求、时效性、便利性、可获取性进行评估。在2个指标的评估模式中,“主题性+专业需求”“主题性+质量”和“主题性+时效性”组合出现的频率较高,分别占数据总量的38.53%、14.68%和11.93%。在真实检索环境中,在判定数据主题相关后,数据的质量、专业需求或者时效性满足其一,基本就能得到相关结果。数据可获取性也是比较重要的判断指标,关系着用户是否能得到数据实体,但需要和其他指标一起使用。用户对数据进行相关性判断时,头脑对指标的加工速度是非常迅速的。用户一次加工的指标不会超过4个。只有遇到特殊信息需求时,才会对相应指标进行加工。

表4 相关性指标使用模式

标准使用模式	数据量/条	百分比	人数/个
主题性	24	22.02%	17
主题性+专业需求	42	38.53%	12
主题性+质量	16	14.68%	6
主题性+时效性	13	11.93%	9
主题性+便利性	5	4.59%	4
主题性+可获取性	2	1.83%	1
主题性+时效性+质量	2	1.83%	1
主题性+便利性+可获取性	2	1.83%	2
时效性+主题性+便利性	1	0.92%	1
时效性+主题性+可获取性	1	0.92%	1
主题性+权威性+时效性+可获取性	1	0.92%	1

3.3 科学数据用户相关性指标因子分析

用户根据其相关性判断时标准使用情况,对9个标准的重要性作出判断。评价标准为非常重要5分,比较重要4分,一般3分,不太重要2分,不重要1分。对量表信度进行 α 信度系数检验, $Cronbach's\alpha$ 为0.882 ($\alpha>0.5$), 信度系数表现良好,量表可信度高。

3.3.1 指标重要程度

对用户而言,数据质量是最重要的指标(均值为4.18),其次是数据的主题性(均值为4.12)、权威性(均值为4.06)。相对而言,数据新颖性、可理解性和便利性重要程度一般,均值分别为3.49、3.48和3.48(见表5)。在科研工作中数据质量决定研究的成败,因而在数据选择时质量评估占据最多的权重。主题性的重要程度仅次于质量,只有符合研究主题的数据,用户才有进一步探索的兴趣。数据权威性评分也较高,位列第三,说明数据在研究领域里获得认可的程度对用户判断有重要影响。时效性均值为3.84,略低于质量、主题性和权威性指标。数据的时效性较易评估,用户通常会关注数据的发布时间和更新时间,期望得到实时的数据以及符合其研究需求的数据。可获取性均值为3.71,用户比较关注能否得到具体的数据资源,并不介意获取过程需要付出的费用。用户对数据新颖性要求不高,但如果数据能对用户带来一定启发,用户判断数据是相关的可能性增大。用户认为数据可理解性的重要程度一般,但若有多项数据可供选择,用户会选择更容易理解的数据。专业需求是比较重要的指标(均值为3.82),在科研项目进行过程中,用户对数据的专业性要求会逐步提高。便利性在用户感知评分中均值为3.48,相较于其他指标,用户认为数据的便利性不太重要。通常用户在数据查询时对数据获取的便利性要求不高,在数据筛选使用过程时,才会对该指标进行评估。

表5 指标重要程度

指 标	均 值	标准差
主题性	4.12	1.179
可获取性	3.71	1.154
新颖性	3.49	1.191
时效性	3.84	1.173
质量	4.18	1.015
权威性	4.06	1.089
专业需求	3.82	1.210
可理解性	3.48	1.150
便利性	3.48	1.227

3.3.2 因子分析

对标准量表进行KMO和Bartlett的检验, KMO值为0.899 ($KMO>0.5$), Bartlett球形度检验显著性为0

($p<0.01$), 各指标间关系良好, 适合进行因子分析。由图2可知, 曲线在第二个拐点处趋于平稳, 提取两个因子比较合理。两个因子累计贡献率为62.64%。在社会科学研究中, 累计贡献率大于60%即满足数据要求。

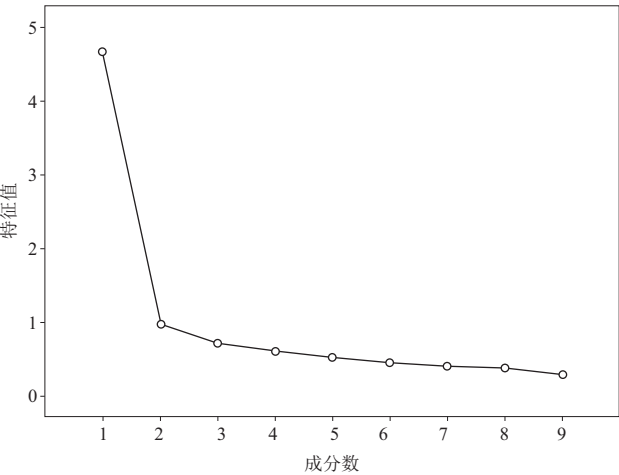


图 2 碎石图

第一公因子方差贡献率为35.21%, 主题性、时效性、新颖性、质量和权威性有较大的因子负荷值(见表6)。这些指标可以共同反映用户在相关性判断时对数据本体的需求, 命名为数据本体性。第二公因子方差贡献率为27.43%, 可获取性、便利性、专业需求和可理解性有较大的因子负荷值。这些指标共同反应了用户数据获取与使用阶段的需求, 命名为数据可用性。研究发现, 用户在进行相关性判断时, 两个因子多数情况下同时发挥作用; 但在信息不足或者用户想快速找到相关数据时, 只对数据本体性进行判断也能满足用户需求。如用户可在判断过程中使用“主题性”“主题性+质量”“主题性+时效性”等判定模式。数据可用性对用户相关性判断也有重要影响, 但需结合数据本体性同时进行加工才能发挥作用, 单独对数据可用性进行评价无法得到数据是否相关的判定结果。

4 总结

综合两个实验不难发现, 用户相关性判断过程中常用的指标为主题性、时效性、质量、权威性和可获取性。而在用户指标的重要性感知中, 比较重要的指标有主题性、质量和权威性。这表明在用户认知中, 相关性判断过程不能缺少对数据主题性、质量和权威性的评估, 其他指标在具体情境中, 受到相关信息需求刺激才会调用。用户的相关性判断过程是在自我认知的控制

表 6 旋转成分矩阵

标 准	因子1负荷值	因子2负荷值
主题性	0.608	0.418
可获取性	0.473	0.505
新颖性	0.650	0.313
时效性	0.767	0.266
质量	0.831	0.217
权威性	0.814	0.199
专业需求	0.329	0.679
可理解性	0.278	0.783
便利性	0.159	0.841

下, 对感知线索有意识加工和无意识加工共同进行的。用户在数据检索过程中, 进行一次相关性判断平均使用1.88个指标, 最多不会超过4个指标。智能搜索引擎的设计不需要添加过多筛选条件, 可以提供用户自定义设置搜索模式的功能。研究发现, 数据主题相关是用户相关性判断的先决条件。在此基础上, 用户更重视数据的质量、时效性与权威性。此外, 用户也比较注重数据的专业需求, 数据能否满足用户的专业需求通常可以决定数据的相关性, 在数据搜索引擎开发的过程中应该重视这个指标。

与其他信息类型相比, 科学数据检索具有更强的目的性。Barry认为新颖性是文档选择的重要评估指标, 用户会因为好奇想了解新的知识而选择一个文档; 科学数据用户则检索目标十分明确, 更需要对其研究有支持效果的数据^[6-7]。Wang等研究发现, 用户对超过半数的文档评估只用了1个指标^[7], 但在数据选择时使用的指标更多, 相关性判断过程中71.56%的数据使用了2个指标。科学数据用户需要能表达数据权威性的指标信息, 如数据被引用率, 目前这些信息并不完善, 用户会通过发表文章的影响因子判断数据的权威性。

与科学数据元数据标准相比, 用户相关性指标更注重用户判断的主体地位。用户在数据相关性判断过程中从数据的本体性和可用性对数据进行评价。在判断时能够加工的信息量有限, 信息量过大可能会造成用户选择困难。科学数据元数据标准的目的在于能够合理完整地描述信息资源, 在进行数据存储时保证信息资源不会丢失, 所以元数据标准是对数据资源本身的描述。用户相关性指标同样注重数据本身, 如数据的主题性、质量、时效性。同时, 用户也需要数据的使用信息, 如数据是否能获得、是否能满足专业需求等。

科学数据作为一种总量与经济效益不断提升的信息载体,以其为对象开展相关性研究具有重要意义。本文通过实证研究探讨科学数据用户相关性判断过程中使用的指标,以便于能更好地理解科学数据用户相关性判断机制,为设计开发智能搜索引擎提供理论和算法基础。

研究的被试人员根据方便取样原则,均为来自农业领域的研究生,其在研究工作中有数据需求,也对数据有一定程度了解,符合本研究的被试人员要求。但相对广泛的科学数据用户而言,研究被试人员类型比较单一。因此,后续研究需要扩大被试人员范围,增加专家用户和普通用户。情景实验选取被试人员的数量虽达到要求,但就每个学科的被试人员数量仍相对较少,被试人员在检索中使用的平台也具有一定的局限性,在后续研究中不仅要增加每个学科的被试人数,还要增加学科的范围,以便于获取更加全面和可靠的数据和实验结果。

参考文献

- [1] 中华人民共和国科学技术部.科学数据共享概念与术语_概念_1.2[EB/OL]. (2010-08-14)[2017-04-19].<http://www.docin.com/p-149852222.html>.
- [2] 赵华,王健.国内外科学数据元数据标准及内容分析[J].情报探索,2015, 208(2):21-24.
- [3] 王健,周国民,王剑,等.认知导向信息需求研究综述[J].图书情报工作,2013,57(10):136-141.
- [4] PARK T K.The nature of relevance in information retrieval:an empirical study[J].Library Quarterly,1993,63(3):318-351.
- [5] COOL C,BELKIN N J,FRIEDER O,et al.Characteristics of text affecting relevance judgments[J].Automotive News,1993,17(4):77-84.
- [6] BARRY L C.User-defined relevance criteria:an exploratory study[J]. Journal of the American Society for Information Science and Technology, 1994,45(3):149-159.
- [7] WANG P,SOERGER D.A cognitive model of document use during a research project.study I.document selection[J].Journal of the American Society for Information Science,1998,49(2):115-133.
- [8] MARKKULA M,SORMUNEN E.End-user searching challenges indexing practices in the digital newspaper photo archive[J]. Information Retrieval,2000,1(4):259-285.
- [9] RIEH S Y.Judgment of information quality and cognitive authority in the web[J].Journal of the Association Society for Information Science and Technology,2002,53(2):145-161.
- [10] BALATSOUKAS P, RUTHVEN I.An eye-tracking approach to the analysis of relevance judgments on the web:the case of Google search engine[J].Journal of the American Society for Information Science and Technology,2012,63(9):1728-1746.
- [11] KRIPPENDORFF K H.Content analysis:an introduction to its methodology[M].Beverly Hills:CA:Sage,2004.

作者简介

高飞,女,1988年生,博士研究生,研究方向:信息检索,E-mail: gaofei20120214@163.com。

石蕾,女,1982年生,硕士,副研究员,研究方向:科技管理。

王健,男,1971年生,博士,研究员,研究方向:信息处理。

张贵兰,女,1993年生,博士研究生,研究方向:信息检索。

刘建平,男,1989年生,博士研究生,研究方向:信息检索。

An Exploratory Research on Scientific Data User Relevance Criteria

GAO Fei¹, SHI Lei², WANG Jian¹, ZHANG GuiLan¹, LIU JianPing¹

(1.Agricultural Information Institute of CAAS, Beijing 100081, China; 2.National Technology Foundation Platform Center, Beijing 100862, China)

Abstract: The paper aimed at exploring the scientific user relevance criteria by studying the relevance judgment behavior, to deepen the understanding of user relevance judgment mechanism, and to provide algorithm and theoretical basis for designing and developing intelligent data search engine. The study was conducted by two methods, which is think aloud and interview. The data were collected in real environment, and the whole process was recorded by camera. The analysis of verbal reports was based on Grounded Theory. On the basis of qualitative research, a questionnaire was designed to carry out a large sample of questionnaires. The scientific user relevance criteria can be divided into two categories: data ontology and data availability, a total of 9 criteria. Compared with other information types, scientific data retrieval was more purposeful so it was necessary to evaluate the topicality, quality and authority of the data in relevance judgment process. Other criteria would be called only when stimulated by the relevant information needs in specific circumstances.

Keywords: Information Retrieval; Scientific Data; User Relevance; Criteria

(收稿日期: 2017-09-25)