基于用户适用度的开放数据质量提升研究*

王瑞云[1] 贾君枝[2]
(1. 山西大学经济与管理学院，太原 030006；2. 中国人民大学信息资源管理学院，北京 100872)

摘要：本文研究如何提高开放数据质量以更好地满足用户的应用需求。先分析用户需求匹配的行为过程，以北京开放数据门户网站的个体数据集为基本研究对象，选取浏览次数、下载次数和下载浏览比作为外部行为结果指标；然后分析外部指标与数据集的主题、元数据说明、及时性，以及数据表列数、行数等内在质量指标的可能的正相关关系；从相关分析中发现极端不符合正相关的异常数据集，联系这些数据集的用户选择情景深入讨论，提出针对这些异常数据集的质量提升建议。

关键词：开放数据质量；用户适用度；需求匹配；下载浏览比
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2018.12.003

当前大数据和“互联网+”等国家项目正致力于促进国家信息化发展，确保公民公平、公正、准确地获取到所需信息。2015年，国务院印发《促进大数据发展行动纲要》规划大数据发展目标，旨在2018年底前建成国家政府数据统一开放平台，2020年底前逐步实现信用、交通、医疗、卫生、就业、社保、地理、文化、教育、科技、资源、农业、环境、安监、金融、质量、统计、气象、海洋、企业登记监管等民生保障服务相关领域的政府数据集向社会开放[1]。政府和公用事业机构的开放数据是大数据的基本数据内容，因来源机构的信用担保可靠而具有很高的用户信任和应用价值。

国内开放数据发展起步晚于国外，在开放数据规模和质量上都存在一定的缺陷，亟需改进。根据互联网基金会发布的第四次《开放数据晴雨表》的评价，中国在115个国家/机构中排名71位。在开放数据评价的15个主题数据集中，我国只有人口普查细节数据和公共交通时间表两项排名在前69位，其他项单项排名都在70位以后；而且已提供的开放数据只达到该机构的最低要求，即数据集存在和可在线获取，没有达到整体可用和提供数据关键元素链接等更高要求[2]。

国外学者[3-7]的研究重视开放数据与用户需求的匹配，利用开放数据定量研究方法结合专家知识解决现实中的具体问题，如某一地区的人口下降和社区公共服务的可持续性，以及空气污染治理、疾病传播和控制等。国内学者[8-9]首先研究分析发达国家开放数据的经验以供借鉴，还有一些学者[10-13]采用各种定量研究方法进行国内开放数据的质量评价和质量提高研究。国内的定量研究通常采用问卷调查方法，对开放数据门户网站的整体质量进行评价。评价指标采用通用网站评价指标（如网站的下载速度等），不反映开放数据网站的重要特征。由于初期用户对开放数据使用很少，很多被调查的用户前期没有浏览和下载过开放数据[14]，回收的调查问卷准确性受到用户对开放数据认知的限制，所以现阶段有必要根据开放数据用户使用行为方面的特点，研究开放数据质量和质量提升。北京开放数据门户网站是国内开放数据各方面指标较好的网站之一，有大量的用户进行浏览和下载，本文后续部分以北京开放数据门户网站数据集作为案例数据来源。

1 研究的概念框架构建

1.1 用户适用度的用户行为表示

用户适用度是开放数据满足用户应用各方面需求

*本研究得到国家社会科学基金重点项目“基于关联数据的中文名称规范档语义描述及数据聚合研究”（编号：15ATQ004）资助。

的综合指标。开放数据的根本目的是利用[15]，开放数据集的用户适用度指标可以通过用户需求匹配和选择数据集的行为显式地表示出来。基于用户需求匹配和选择数据集的行为全过程见图1。首先用户面对开放数据门户网站的海量数据集，根据网站推广和导航进行初选，得到初步匹配需求的所有数据集集合。集合中的数据集都获得了用户浏览点击行为，该行为参数包括浏览的对象、浏览时间。本文只简单化选取当前时点数据集的累计浏览次数指标。其次，用户通过浏览数据集的内容说明和元数据，进一步精确地判断该数据集和自身需求的匹配程度。用户根据元数据详细说明来判断该数据集是否为所需内容，数据集的更新频率、最新更新时间影响用户对数据的及时性需求；数据集发布更新主体的可信程度影响用户对数据的可靠性需求；数据格式、数据集行列数等也是影响用户判定的质量指标。经过综合需求匹配阶段的精确判断，用户决定是否下载数据集提供的数据资源，符合用户精确需求匹配的数据集选入精确匹配数据集集合，并得到用户下载点击行为，行为参数具体包括下载对象和下载时间，本文表示为当前时点数据集的累计下载次数指标。用户开放数据的利用还包括手机端的APP用户关注的新型用户行为，表现为用户关注数的指标。最后，用户对下载到本地的数据资源进行处理，可能做出质量评价、问题反馈、提出进一步需求等一系列行为，作为门户网站未来提高数据集的质量参考。

1.2 用户适用度的数据集内在质量

数据的内在质量是用户选择的内因和基础，而行为统计为数据表示内在质量的需求匹配结果。用户适用度概念是由Vetrò等[3]提出，由于低质量的开放数据集增加了用户的再利用成本，从而不能满足用户显式和隐含的需要；并提出基于用户适用度的质量量度定义，包括从数据集到单元格不同粒度对象的9个质量量度定义，即创建更新可溯源性、及时性、过期延迟时间、数据单元和行的完整性、数据单元和数据集的标准符合性、单元粒度的易理解性和单元粒度的准确性。从上述九方面达到用户选择利用的要求，能降低用户的整体使用成本，提升数据集可靠性、及时性和准确性，从而大幅降低开放数据集的总体利用成本，整体上提高数据的用户适用度[5]，使数据集得到增值性的利用和再利用。

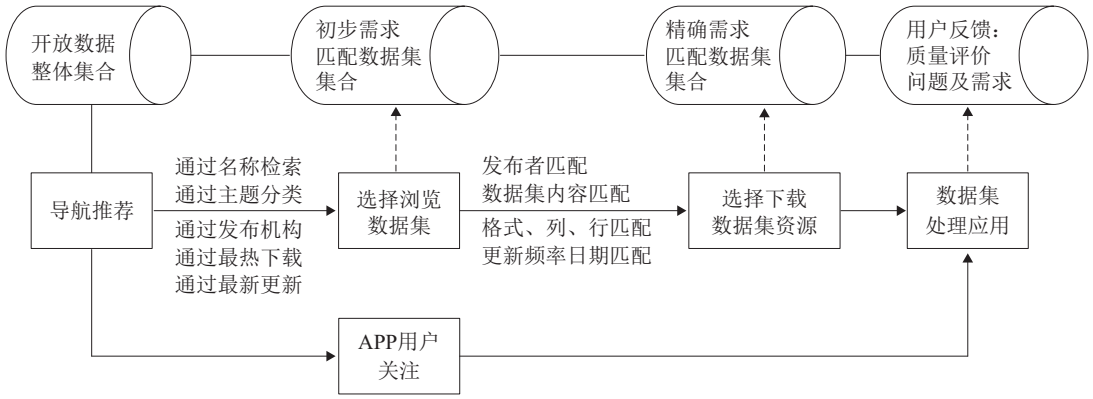

图1 用户需求匹配和选择数据集的行为全过程

另一个数据质量内在标准是关联开放数据质量的五星标准[16]，主要基于开放数据的发布格式和符合标准的程度，最低标准是存在任何格式用户可获取的开放数据，但是这些数据可能是图片格式，不方便用户的机器编辑处理。二星和三星的数据集分别是.xls和.csv格式的表格数据，这两个级别的数据集可以导入数据库；三星与二星的数据集相比，其优势体现在表格数据集不局限于微软的Excel数据（.xls格式）。四星的数据集符合W3C的开放标准，数据采用RDF表示，并且可以通过SPARQL查询获取。五星开放数据实现数据到其他提供方数据的关联。四星和五星的高质量数据方便用户集成多来源的开放数据，实现开放数据门户的互操作。国内的开放数据总体达到三星标准，提供.xls和.csv格式的表格数据，还有少部分的word文件和pdf图片文件。三星标准的数据集质量限制用户在多网站来源的数据集之间的互操作，提高用户的处理成本。而word和pdf图片格式的数据资源需要用户付出更高的处理成本，甚至需要安装专门软件处理数据，用户的利用成本更高。

1.3 基于用户适用度的数据质量框架

构建基于用户适用度的数据质量框架可以分为数据集内在质量指标和用户行为的外在质量指标。内在质量指标包括数据集的内容主题、数据集的元数据说明、数据集的及时性、数据列表现出的数据属性丰富度、数据行（多个表的总行数）表示出的数据规模5方面指标；用户行为的外在质量指标包括浏览次数、下载次数和用户关注数3个方面基本指标，以及计算出的下载浏览比、时段下载浏览比等分析性二级指标。

2 数据集内在质量与用户行为的关系

以北京开放数据门户网站作为实例研究对象，利用网络爬虫工具从门户网站的用户互动信息、数据集的主题导航、主题数据集的下载/浏览排行、数据集的热门下载等统计信息中获取数据集质量和用户行为数据，对该开放数据门户每个数据集的用户选择行为和数据集内在质量的关系进行分析，旨在为基于用户适用度的数据集质量提升奠定基础。

2.1 下载浏览关注与数据集内在质量的关系

2.1.1 下载次数与主题数据集个数的相关关系

门户网站共提供20个主题的1 023个数据集，由于网站数据集个数较多，为方便用户选择适合自身需求的数据集提供主题导航，通过主题数据集个数和主题内容两个属性向用户展示数据。门户网站给出按主题分类的数据集个数如表1所示，可以看出，不同主题的数据集分布差异明显。根据一般常识和开放数据提供者的考虑，提出假设S1。

S1：各主题的数据集个数与用户下载浏览次数正相关。

表1 按数据集个数排序的数据集主题情况

编 号	主题分类	数据集个数	编 号	主题分类	数据集个数
1	经济建设	298	11	交通服务	34
2	文体娱乐	124	12	财税金融	27
3	教育科研	81	13	农业农村	25
4	企业服务	75	14	劳动就业	19
5	社会保障	71	15	信用服务	15
6	政府机构与社会团体	65	16	餐饮美食	8
7	环境与资源保护	60	17	生活安全	8
8	医疗健康	51	18	消费购物	7
9	生活服务	47	19	宗教信仰	7
10	旅游住宿	40	20	房屋住宅	4

本文样本的获取时间为2018年10月10日，下载排名前30的数据集信息见表2[17]。由于“下载次数”比“浏览次数”更能体现开放数据集用户适用度的行为结果，故选取下载次数为首要因素排序。其中的6~8列在后文研究中使用。对浏览次数和下载次数按照主题分类汇总统计见图2。由于浏览次数远大于下载次数，为了图形显示清晰，图2中对浏览次数除以10。

下载量最多的数据集主题集中在教育科研、交通服务、旅游住宿、企业服务。教育科研主题占据下载次数排名第1和第2，该主题在下载次数前30的数据集个数为8，总下载次数12 087，远大于其他主题的数据集；但表1中该主题的数据集个数为81，排名第3，远少于第1主题的数据集个数298，所以教育科研主题是不支持假设S1正相关关系的一个异常。不支持假设S1最大的异常是表1中提供数据集个数最多主题的经济建设，在表2中下载量前30的数据集中没有出现。具体到经济建设主题内部，该主题按下载次数排名的数据集信息见表3。该主题下载次数排名前2的数据集在总体排名分别为111和136，其他的都在总体排名260以后。

上述两种异常否定了基于提供者和一般常识的假设S1。第一个异常的数据集主题是当前用户重点关注教育科研主题的外在表现，主题内容对浏览下载次数的影响远超过假设S1的正相关影响。第二个异常更需要开放数据门户管理者思考，经济主题的数据集

表2 按下载次数排序前30的数据集

编 号	数据集	浏览次数	下载次数	主 题	列 数	行 数	最新更新时间
1	小学	34 322	2 830	教育科研	4	1 217	2012/10/29
2	中学	32 160	2 620	教育科研	4	643	2012/10/29
3	土地用途分区	30 605	2 490	企业服务	2	30 946	2012/10/29
4	轨道交通线路	25 247	1 547	交通服务	2	20	2012/07/19
5	教育部直属院校	14 121	927	教育科研	8	25	2018/04/26
6	民办高校及独立学院	14 242	927	教育科研	8	16	2018/04/26
7	国务院委办属院校	10 639	902	教育科研	8	13	2018/04/26
8	市属高校	10 417	902	教育科研	8	41	2018/04/26
9	星级饭店	12 171	872	旅游住宿	5	783	2016/12/01
10	路况直播信息	8 759	709	交通服务	6	604	2018/04/22
11	主干路	5 966	656	交通服务	4	359	2012/07/19
12	中职	5 345	624	教育科研	4	126	2012/10/29
13	公路气象数据	8 324	602	交通服务	18	39	2018/03/31
14	三级医院	5 034	583	医疗健康	3	50	2012/09/10
15	备案停车场（位）	4 824	583	交通服务	9	6 134	2014/07/19
16	机场班车线路	5 272	577	交通服务	3	9	2012/07/19
17	森林公园	5 009	570	旅游住宿	5	31	2018/06/29
18	宾馆旅店	5 045	561	旅游住宿	3	6 198	2017/01/24
19	幼儿园	4 692	555	教育科研	4	1 165	2012/10/29
20	金叶级绿色饭店	4 905	546	旅游住宿	5	147	2014/07/19
21	北京地区博物馆	4 897	545	旅游住宿	5	161	2018/10/10
22	轨道交通站点	4 481	503	交通服务	2	298	2012/07/19
23	车管所	4 659	491	交通服务	7	8	2017/01/20
24	快速路	4 077	476	交通服务	4	80	2012/07/19
25	生态林管护面积分布	4 244	469	环境与资源保护	2	16	2013/05/05
26	超市	3 397	442	消费购物	10	1 941	2014/06/17
27	省道	3 998	442	交通服务	5	443	2012/07/19
28	银叶级绿色饭店	3 927	440	旅游住宿	5	129	2014/07/19
29	二级医院	3 962	439	医疗健康	3	108	2012/07/19
30	养老机构	3 548	434	社会保障	5	470	2018/04/13

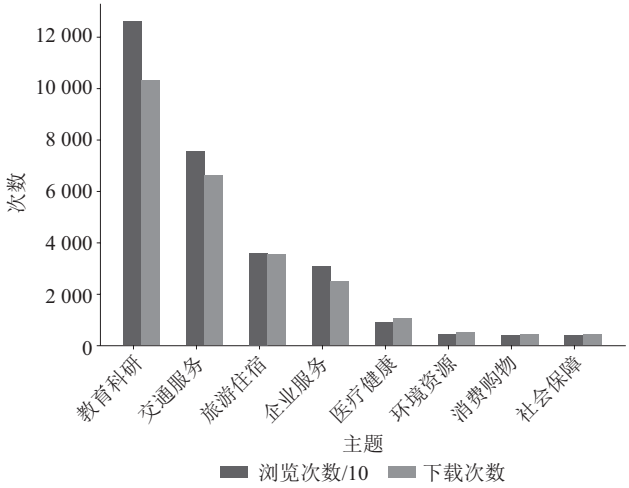

图2 主题分类的下载次数和浏览次数汇总（前8项）

提供的数据集个数很多，但是并没有被用户浏览和下载，网站需要对该主题的数据集增大推广力度，更好地满足用户需求，使该主题的数据集更多地被用户浏览下载。

2.1.2 下载次数与数据集及时性的正相关及异常

本文后续将研究6个正相关关系，分别是下载次数与数据及时性、数据表列数、数据表行数的3个相关关系，以及下载浏览比与数据及时性、数据表列数、数据表行数的3个相关关系。为准确地判定各数据集的各对指标的正相关关系是否成立，下面分别根据每个正相

表3 经济建设主题按下载次数排序前5的数据集

编 号	总体排名	数据集名称	浏览次数	下载次数	最新更新
1	111	北京市示范应用新能源小客车生产企业备案信息	2 461	238	2017/12/22
2		北京市示范应用新能源小客车产品备案信息	1 841	218	2017/12/22
3	136	北京市中小企业公共服务平台名单	1 040	143	2017/03/09
4		北京市软件产品检测机构认可名单	1 014	120	2017/05/31
5		北京市小企业创业基地名单	881	114	2017/03/09

关判断的两个指标，对表2的数据集进行聚类。本文的6个相关关系共涉及5个指标，分别为下载次数、下载浏览比、及时性、列数、行数；应用这5个指标对数据集进行聚类。聚类算法采用最小化组内距离、最大化组间距离的原则，分组参数设为5，编写程序计算。上述5个指标对表2的30个数据集的聚类分组结果见表4。

表4 数据集的下载次数、下载浏览比、及时性、列数、行数聚类分组赋值结果

分组赋值	下载次数各分组数据集	下载浏览比各分组数据集	及时性各分组数据集	列数各分组数据集	行数各分组数据集
5	1，2，3	12，15，19	5，6，7，8，10，13，17，18，21，23，30	5，6，7，8，13，15，26	1，3，15，18，19，26
4	4	7，8，14，16，21，23，24，26，30	9	9，10，17，20，21，23，27，28，30	2，9，10，11，22，27，30
3	5，6，7，8，9，10	10，22，25，27，28，29	15	1，2，11，12，19，24	20，21，28，29
2	11，12，13，14，15，16，17，18，19	9，11，13，17，18，20	20，26，28	14，16，18，29	12，14，24
1	20，21，22，23，24，25，26，27，28，29，30	1，2，3，4，5，6	1，2，3，4，11，12，14，16，19，22，24，25，27，29	3，4，22，25	4，5，6，7，8，13，16，17，23，25

表4第1列是5个分组对应的分值，第2~6列是按指标的聚类结果分组内的数据集编号。对5组分别按5级量级赋值，同一组内的数据集赋同一值，如第2列第1个分组“1，2，3”，表示1、2、3号数据集按下载次数分在一组，分值为5（5最好，1最差）。

指标及时性需要元数据给出固有的更新频率，在最新更新时间基础上分析。先按公式（1）计算数据集的延迟度。

延迟度=（研究时点—最新更新时间）÷更新周期 （1）

其中研究时点、最新更新时间的单位为年。由于门户网站数据集的元数据中没有提供更新频率，本文假设更新频率为1次/年。再根据公式（2）计算及时性。指标聚类采用的是最后计算出的及时性值，聚类结果见表4的第4列。

及时性=max（延迟度）—延迟度 （2）

利用表4“下载次数各分组数据集”和“及时性各分组数据集”的结果，对30个数据集的对应值进行成对比较，基本支持正相关15个数据集；找到极端不支持正相关的6个数据集，属于异常数据集，不符合及时性高数据集的下载次数高的正相关常识。这6个异常数据集为“小学”“中学”“土地用途区分”“北京地区博物馆”“车管所”和“养老机构”。其中有重要参考价值的是前3个数据集，分别为“小学”“中学”“土地用途分区”，这3个数据集下载次数最大，及时性反向最差，是极端负相关异常；该异常说明与这些数据集主题相关的社会问题得到大量用户关注，故下载和浏览次数最高。网站尤其需要解决异常数据集的及时性问题，及时更新数据集，更好地满足大量用户的数据及时性需求，避免严重挫伤大量用户的积极性。而另外的3个异常数据集为“北京地区博物馆”“车管所”和“养老机构”，及时性最高，下载数却排在表2的最后组，但是只是相对表2前面的20个数据集最低，放在全部数据集中下载数不低，可以排除该异常。

2.1.3 下载浏览次数与数据集行列数的正相关及异常

数据集的列数反映数据属性的丰富程度，行数反

映数据集的规模。数据集的列数和行数越多，说明数据集的质量越高，可以得到更高的用户下载浏览次数，一般列数、行数与下载浏览次数具有正相关关系。

利用表2“列数”和“行数”两列的数据聚类分组结果（见表4的“列数各分组数据集”和“行数各分组数据集”），将其分别与表4“下载次数各分组数据集”列的数据成对比较，分析两组正相关关系。结果表明，基本支持列数与下载浏览次数正相关的数据集有16个，正相关性不显著。支持行数与下载浏览次数基本正相关的数据集个数有13个，正相关同样不显著。

列数与下载浏览次数正相关的极端异常为“土地用途分区”数据集，下载次数最高极端反向对应了列数最少值，该异常需要对数据集的列进行深入分析，对于用户亟需的重点数据用2列是否足够表达实际数据的属性，能否满足用户的应用需求。行数正相关的极端异常为“轨道交通线路”数据集，下载次数较高反向对应了行数最小值，对此异常进行深入分析，该数据集为用户重点浏览下载的数据集，但只有16行数据，是否能满足用户的数据要求，是否需要细化数据粒度。

2.1.4 下载次数与及时性、列数、行数的正相关异常总结

综合下载次数与及时性、列数和行数的正相关的极端异常，需要提醒开放数据管理者注意共有的异常数据集（即用户下载次数最高的“小学”“中学”和“土地用途区分”数据集），更需要抓住用户需求迫切的契机，提高这些数据集的及时性，提高“土地用途区分”数据集的列丰富性和“轨道交通线路”数据集的行数。

2.2 用户下载浏览比与开放数据集内在质量的关系

2.2.1 用户下载浏览比的含义与计算

用户的下载浏览比反映用户在浏览数据集内容选择下载数据集链接数据资源的概率，代表用户根据数据集的元数据详细说明与自身需求进一步匹配选择的概率。表2中30个数据集的浏览次数和下载次数的分布见图3，图4中给出二者的3种方案的线性拟合，包括所有点的直线拟合、高区的直线拟合、低区的直线拟合。

分析数据集实际语义，下载次数小于浏览次数，下载次数与浏览次数正相关。下载次数y和浏览次数x函数关系如公式（3）所示。

$$y=a+bx \quad (x\geqslant 0,\ y\geqslant 0,\ 0\leqslant b\leqslant 1) \qquad (3)$$

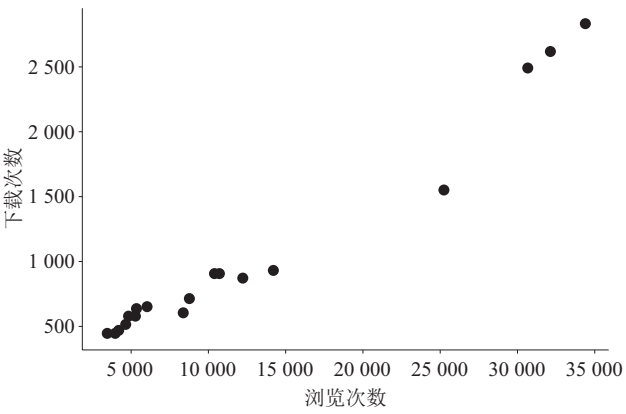

图3 浏览次数和下载次数散点分布

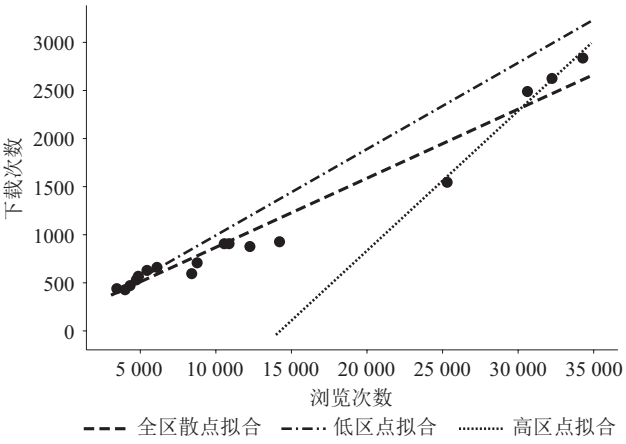

图4 整体拟合直线和高低区分别拟合直线

由于门户网站初期用户下载需要用户注册登录，而浏览不需要登录，所以某一段时间数据集的浏览次数增长，而下载次数为0，这时间点情景为$x>0$、$y=0$；该实际情景下，直线与x的交点$x\geqslant 0$，则与y交点处$y\leqslant 0$，即要求公式（3）中的参数$a\leqslant 0$。拟合结果如图4中的全部点、低区和高区的3条拟合直线，其参数（a，b）分为（155，0.071），（106，0.090），（−2 057，0.145）。其中两条拟合直线的参数$a>0$，严重违反实际情形；只有高区的拟合直线$a=-2\ 057$，不显著违背实际情况，其下载浏览比值为0.145。以上分析说明每个数据集的浏览下载拟合直线有显著差别，不能用同一条直线拟合。所以本文后续对每个数据集计算下载浏览比。

本文用两种方法计算下载浏览比。方法1：在公式（3）参数$a=0$时计算每个数据集的全局平均下载浏览

比。方法2：根据公式（4）计算在最近参考时间段（2018年8月23日—10月13日）的下载浏览比b_1[18]，这两种下载浏览比计算结果如图5所示。

$$b_1 = \frac{y_2 - y_1}{x_2 - x_1} \quad (4)$$

方法1的下载浏览比集中在0.07~0.14；方法2的下载浏览比中有8个数据集超过0.20，最高是幼儿园数据集达0.45。经济建设主题的数据集下载排名在100后，但方法2计算的下载浏览比较高，为0.35。

比较图5下载浏览比和图3的下载次数，可以看出，虽然前6个数据集浏览次数和下载次数都很高，但是下载浏览比很低。这说明门户网站前6个数据集虽然被大量的用户浏览，但是其中大部分用户进一步根据元数据判断数据资源与自身需求匹配时，没有选择下载数据集资源，数据集其他方面的质量可能无法满足用户需求。后面24个数据集用方法2计算的最近区间平均下载浏览比高于全局平均的下载浏览比，说明后面数据集的下载次数有加速发展的趋势。综合上述分析，两种方法下载比加权综合得到最终下载浏览比，用于数据集根据下载浏览指标的聚类，聚类结果见表4的第3列下载浏览比各分组的数据集。

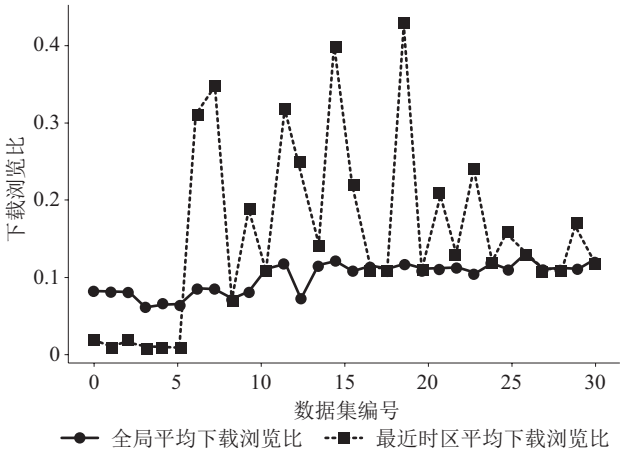

图5 数据集的平均下载浏览比和最近时段的下载浏览比

2.2.2 用户下载浏览比与数据及时性的正相关及异常

用表4中的下载浏览比和数据及时性的聚类分组值进行成对比较，验证这2个指标的正相关性。支持正相关的数据集有13个，正相关关系不成立。从本次正相关验证得到一个新的解释：2.1.2节的5个极端异常在本次不再是异常，而变成支持正相关的数据集，“小学”“中学”“土地用途区分”和“轨道交通线路”4个数据集的2个属性分类都在最低组，支持正相关，该结论可以部分解释2.1.2中下载次数与及时性正相关的极端异常，这4个数据集的下载浏览比指标低，表明下载次数的相对速度有降低的趋势，更是提醒管理人员尽快提高这些数据集的及时性，才有可能扭转下载次数下降的趋势。

本次相关验证在解释消除已有异常的同时，验证结果还发现了新的负相关的极端异常，异常数据集是表2的“中职”数据集和“幼儿园”数据集，这两个数据集的下载浏览比最好，表明它们有很好的下载应用趋势，但是及时性最差。这两个数据集在下载次数和及时性正相关验证中，没有表现出明显的相关异常。所以需要提醒网站管理者重视这两个隐藏的异常数据集的及时性质量提升。另外，不太极端的负相关异常还包括“三级医院”“机场班车线路”“快速路”数据集，也应该得到网站管理者注意，提早安排数据更新。

2.2.3 下载浏览比与数据集行列的正相关及异常

对表4中的下载浏览比与数据集的列数和行数的聚类数据进行成对比较，分别计算两个相关关系。支持下载浏览比与列数正相关的数据集有10个，正相关不能成立。异常数据集为“教育部直属高校”“民办高校及独立学院”数据集，下载浏览比最差，但数据集的列数在最好组；分析其原因是：①虽然教育科研主题是热门主题，但用户的关注热点在主题内部更加细分，这两个数据集正在逐步退出用户热门数据；②列数虽然多，但列内容不能匹配用户需要。

支持下载浏览比与行数正相关的数据集有13个，正相关不成立。但是在两个指标最高和最低两端组内，正相关表现比较显著。如下载浏览比最好的“备案停车场（位）”“幼儿园”数据集，正相关行数在最好组；下载浏览比最差的“轨道交通线路”“教育部直属高校”“民办高校及独立学院”数据集，正相关行数在最差组，该部分正相关也部分说明“教育部直属高校”和“民办高校及独立学院”数据集的列相关异常，可能是内在质量行数上存在缺陷；发现的行数正相关异常数据集是“小学”和“土地用途区分”数据集，下载浏览比最差，但数据表行数在最好组。说明数据规模对下载浏览比的影响远小于数据集主题内容的影响。

2.2.4 下载浏览与内在指标的正相关及异常小结

数据集的及时性、数据表列数和行数与下载浏览比的正相关关系都不能得到显著支持。但是2.1.2节下载次数与及时性正相关的极端异常在本节的下载浏览比与及时性的正相关得到部分解释，并且发现不太外显的2.1.2节没有发现的新隐含异常“中职”“幼儿园”数据集，需要提醒网站管理者注意这些隐含的异常。列相关的异常数据集中的“教育部直属高校”和“民办高校及独立学院”数据集，既可能是用户关注热门的细分和分支热门的转变，也可能有行数指标差的影响因素。行相关的异常数据集为“小学”和“土地用途区分”两个数据集，也可以从其下载比和及时性的同为最低正相关得到解释，这两个数据集行数虽然很多，但是及时性最差，所以下载浏览比最差。

2.3 下载浏览比与用户总体适用度的关系

下载浏览比可以显式地反映数据集与用户需求的匹配选择情况，在很大程度上，可以反映数据集的用户适用度质量，所以本文前面研究下载浏览比（下载次数）与其他质量指标的正相关关系，试图通过提高相关的质量指标来提高下载浏览比或下载次数，以期最终提高数据集的用户适用度。

下载浏览比过低表示数据集的质量有待提高的方面，但是并不能只限于提高下载浏览比。下载行为是用户根据数据集详细页面上元数据和数据说明，判断数据集的内容主题是否与需求的内容匹配；数据集的及时性是否符合用户要求，以及数据集的列数和行数与数据的属性丰富度和数据规模需求的匹配度。数据集元数据的准确说明为用户下载选择提供正确的依据，避免用户下载不适用数据的后期处理成本，对数据集的总体利用成本的降低和数据集的总体适用度有积极的作用。

因此，数据集的总体适用度质量需要在准确详细的数据集元数据基础上，保证数据集质量提升是建立在对总体成本有效降低的基础上，再提高重点数据集的相关指标质量进而提高下载次数和长期的下载浏览比。

3 研究结论及展望

本文基于用户利用开放数据的行为过程研究开放数据的用户适用度质量，研究对象涉及最微观的单个数据集和主题分类，通过研究下载次数、下载浏览比与数据集的及时性、列数和行数的正相关关系，发现极端不符合正相关关系异常数据集，深入分析异常数据集的应用情景，针对异常数据集，提出质量提升建议。

影响数据集下载次数和下载浏览比的最重要因素是数据集的主题内容和细分主题，门户网站应该根据用户的需求，发布更多热门主题的数据集，对数据集的主题分类尽量划分到热门主题，使数据集得到高的浏览次数和下载次数。

对于在多对正相关研究中发现的异常数据集，分析具体应用情景提出的建议应及时反馈给开放数据管理者。积极推进管理者利用相关关系改进热门重要异常数据集的质量缺陷。对于热门主题相关的异常数据集更为重要，重点提高异常数据集的及时性，长远提高异常数据集的下载浏览比；再进一步提高数据集列数，丰富数据集的属性信息，并且提高数据集的行数，从更细的粒度，提供规模更大的、更精准的数据，从而为用户提供更高的利用价值。最终不仅要提高数据集的当前下载浏览次数，更从长远发展的角度提高数据集的下载浏览比，提高开放数据的整体适用度。

另外数据集还应该保证元数据说明的准确性，提高下载次数和下载浏览比的工作应该在不增加后期应用成本的基础上进行，防止用户因下载不适用的数据集而浪费大量的后期处理成本。

本文研究的局限在于研究案例的开放数据还处在发展的初级阶段，无法获取多个阶段的用户行为数据比较，以及用户的行为数据还缺少后期应用成本数据；下一步研究将跟踪国内开放数据的发展，从更加系统的动态演变的角度关注开放数据的质量提升，同时关注关联数据技术在国内开放数据中的应用发展，提高开放数据机器处理方面的质量，更好地发掘海量开放数据的潜在价值。

参考文献

[1] 国务院关于印发促进大数据发展行动纲要的通知[EB/OL]. （2015-09-05）[2018-10-30]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[2] The Open Data Barometer[EB/OL]. [2018-05-11]. http://www.opendatabarometer.org.

[3] VETRÒ A, CANOVA L, TORCHIANO M, et al. Open data

quality measurement framework: Definition and application to open government data [J]. Government Information Quarterly, 2016, 33 (2): 325-337.

[4] ZAVERI A, RULA A, MAURINO A, et al. Quality assessment for linked data: A survey [J]. Semantic Web, 2016, 7 (1): 63-93.

[5] RUIJER E, GRIMMELIKHUIJSEN S, HOGAN M, et al. Connecting societal issues, users and data. Scenario-based design of open data platforms [J]. Government Information Quarterly, 2017, 34 (3): 470-480.

[6] THORSBY J, STOWERS G N L, WOLSLEGEL K, et al. Understanding the content and features of open data portals in American cities [J]. Government Information Quarterly, 2017, 34 (1): 53-61.

[7] WU C, KAO S C, SHIH C H, et al. Open data mining for Taiwan's Dengue Epidemic [J]. Acta Tropica, 2018, 183 (7): 1-7.

[8] 马海群，蒲攀. 国内外开放数据政策研究现状分析及我国研究动向研判 [J]. 中国图书馆学报，2015，41 (5)：76-86.

[9] 汪庆怡，高洁. 面向用户服务的美国政府开放数据研究及启示——以美国Data.gov网站为例 [J]. 情报杂志，2016，35 (7)：145-150.

[10] 姜鑫，马海群. 开放政府数据评估方法与实践研究——基于《全球开放数据晴雨表报告》的解读 [J]. 现代情报，2016，36 (9)：22-26.

[11] 张晓娟，孙成，向锦鹏. 基于开放数据晴雨表的我国政府数据开放提升路径分析 [J]. 图书情报知识，2017 (6)：60-72.

[12] 王今，马海群. 政府开放数据质量的用户满意度评价研究 [J]. 现代情报，2016，36 (9)：4-9.

[13] 马海群，唐守利. 基于结构方程的政府开放数据网站服务质量评价研究 [J]. 现代情报，2016，36 (9)：10-15，33.

[14] 北京政务数据资源网. 轨道交通线路 [EB/OL]. (2012-07-19) [2018-10-30]. ttp://www.bjdata.gov.cn/zyml/ajg/sjtw/6809.htm.

[15] W3C.Linked Data Glossary [EB/OL]. [2018-03-25]. https://www.w3.org/TR/ld-glossary/#x5-star-linked-open-data.

[16] 郑磊. 开放政府数据研究：概念辨析、关键因素及其互动关系 [J]. 中国行政管理，2015 (11)：13-18.

[17] 北京政务数据资源网. 最热下载 [EB/OL]. [2018-10-10]. http://www.bjdata.gov.cn/syzrxz/index.htm.

[18] 北京政务数据资源网. 数据一按主题 [EB/OL]. [2018-08-23]. http://www.bjdata.gov.cn/zyml/azt/jjjs/index.htm.

作者简介

王瑞云，女，1969年生，博士，讲师，研究方向：知识组织与管理，E-mail：ruiyunwang@163.com。
贾君枝，女，1972年生，博士，教授，博士生导师，研究方向：信息组织。

The Research of Improving Open Data Quality Based on Fitness for Use

WANG RuiYun[1] JIA JunZhi[2]
(1. School of Economics and Management, Shanxi University, Taiyuan 030006, China;
2. School of Information Resources Management, Renmin University of China, Beijing 100872, China)

Abstract: This paper aims at improving open data quality on fitness for use. Exploring behaviors of users' demand matching, using bjdata.gov.cn as study case, selecting downloads, reviews of every dataset and their express computing as output indices, we research the possibility of positive relationship of between above indices and other dataset's inner quality indices which containing content theme, metadata, timeliness, columns and rows of data resource table. More importantly we find out many exceptional datasets that don't extremely confirm to the positive relationship, and discuss further those datasets on their user selection context. Finally we suggest on quality improving for those exception datasets.

Keywords: Open Data Quality; Fitness for Use; Demand Matching; Ratio between Downloads and Reviews

（收稿日期：2018-11-15）