基于Splunk的图书馆数据库资源用户访问行为研究
---以北京工业大学图书馆为例

雷东升

北京工业大学图书馆

2017年6月8日

目录

引　言

- 图书馆资源保障体系建设重点转向电子资源
- 访问日志数据记录了用户访问的详细信息
- 研究用户访问行为改进服务质量、提高数据库资源利用率
- 北京工业大学图书馆数据库资源丰富、完备。

Splunk的功能

- Splunk是面向大数据的机器数据引擎。
- 使用 Splunk 可收集、索引和利用所有应用程序、服务器和设备（物理、虚拟和云中）生成的快速移动型计算机数据
- Splunk是一个托管的日志文件管理工具，它的主要功能包括：日志聚合功能、搜索功能、提取、对结果进行分组，联合，拆分和格式化、可视化功能、电子邮件提醒功能

Splunk的功能

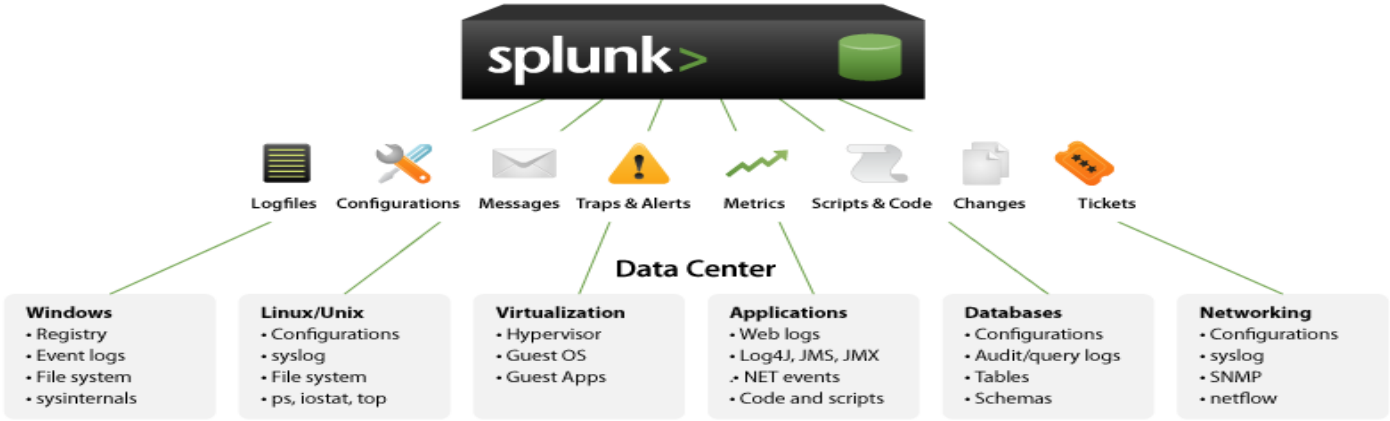


图 索引任何数据

Splunk的功能

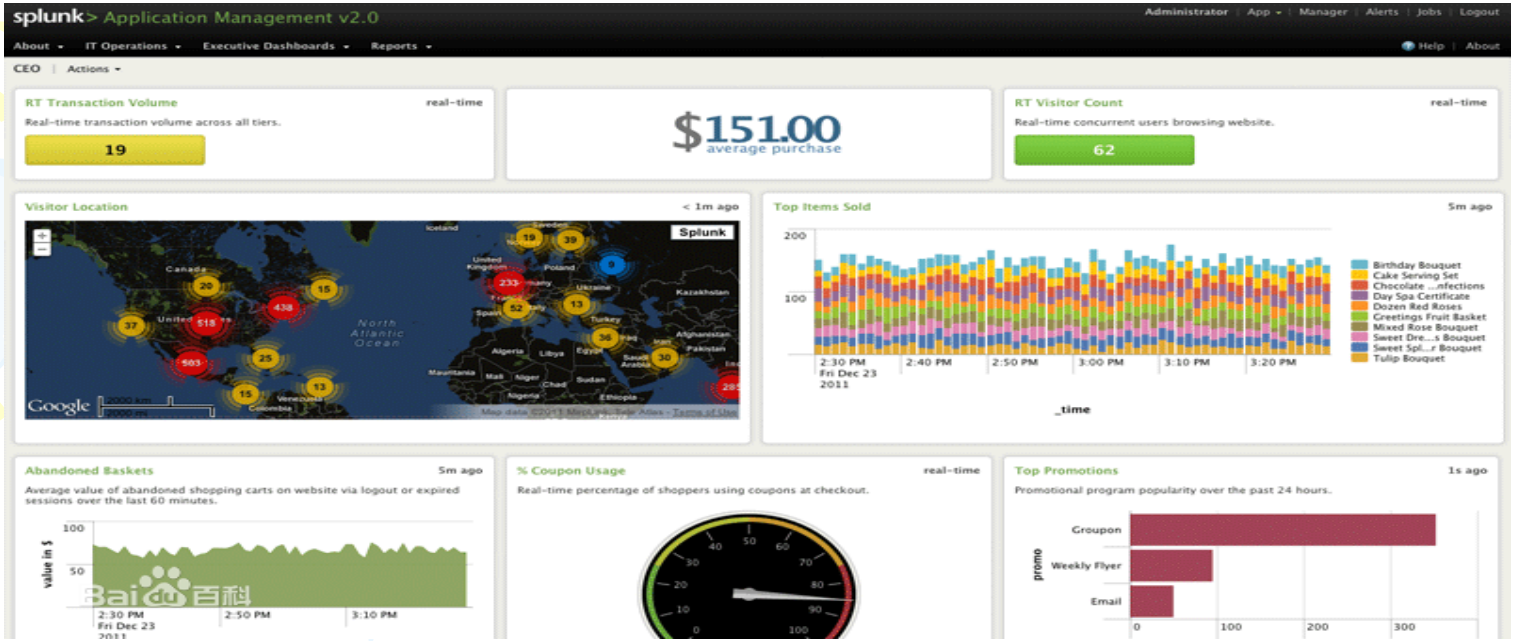


图 自定义仪表板和视图

研究思路

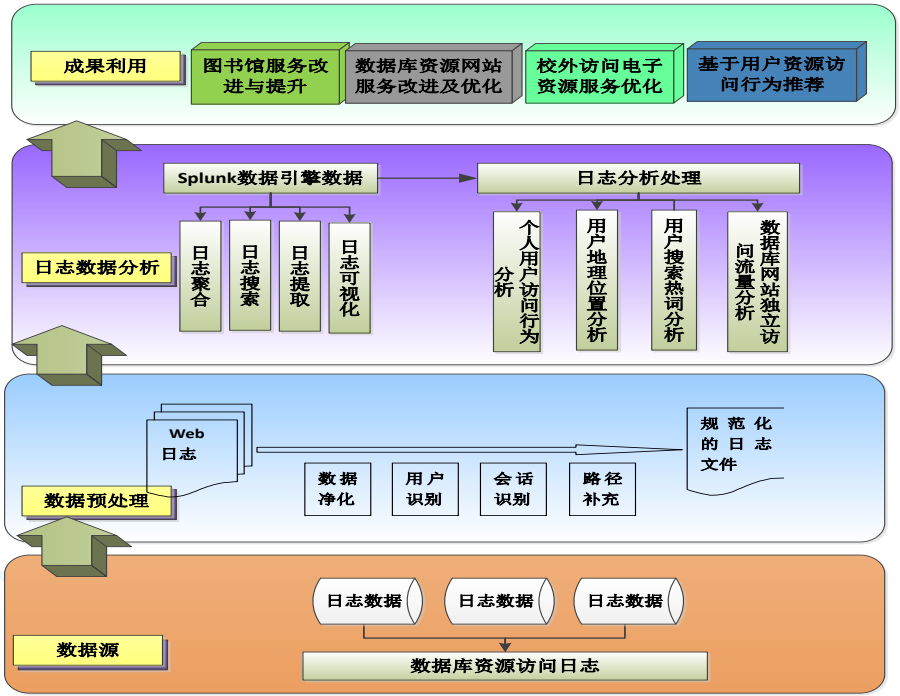


图　基于Splunk的图书馆数据库资源用户访问行为模型

日志数据的预处理及导入

- 数据库资源访问日志属于Web日志，数据是半结构化的
- 北京工业大学图书馆数据库资源是通过EZproxy代理服务器来实现访问的，EZproxy代理服务器日志记录了用户访问数据资源的全部信息。用户访问图书馆数据库产生的Web日志数据量庞大，每天生成一个文本型文件，文件大小100M-500M Byte。
- 本文使用2015年6月1日至7日北京工业大学图书馆数据库资源访问日志作为数据来源，避免了国内法定长假对分析普遍意义上的用户访问行为可能造成的偏差。

日志数据的预处理及导入

- 为保证所有文件的日志记录可以被同时搜索，给所有日志文件创建共同索引Library_data。“ezproxy20150601.log”文件记录了2015年6月1日北京工业大学图书馆数据库的详细使用信息，其他文件依次记录后6天的图书馆数据库访问信息，将6月1日至6月7日的Web日志文件添加到Library_data索引中。Splunk自动确定数据源类型为组合型访问日志文件（Combined Access Log File）。
- Splunk中的事件通常称为记录或者数据行，每一个事件都有一个时间戳。2015年6月1日至6月7日，北京工业大学图书馆数据库资源访问日志文件共记录了1,743,096条事件。将数据导入Splunk后，可以对日志进行分析。

用户访问行为可视化分析
-----网站访问量分析

- 网站访问量即指网站流量（traffic），是描述访问网站的用户数量以及用户所浏览的网页数量的指标 。衡量标准有两个：①访问数IP（Internet Protocol）；②综合浏览量PV（Page View）。测量时常以日为标准，即根据日独立IP或PV来计算。
- 在两种衡量标准中，访问数指独立IP数，规定在00:00-24:00（即一整天）内同一IP地址只被计算一次。综合浏览量指PV数，指页面点击量，用户每次刷新即被统计一次。

用户访问行为可视化分析

-----网站访问量分析 PV

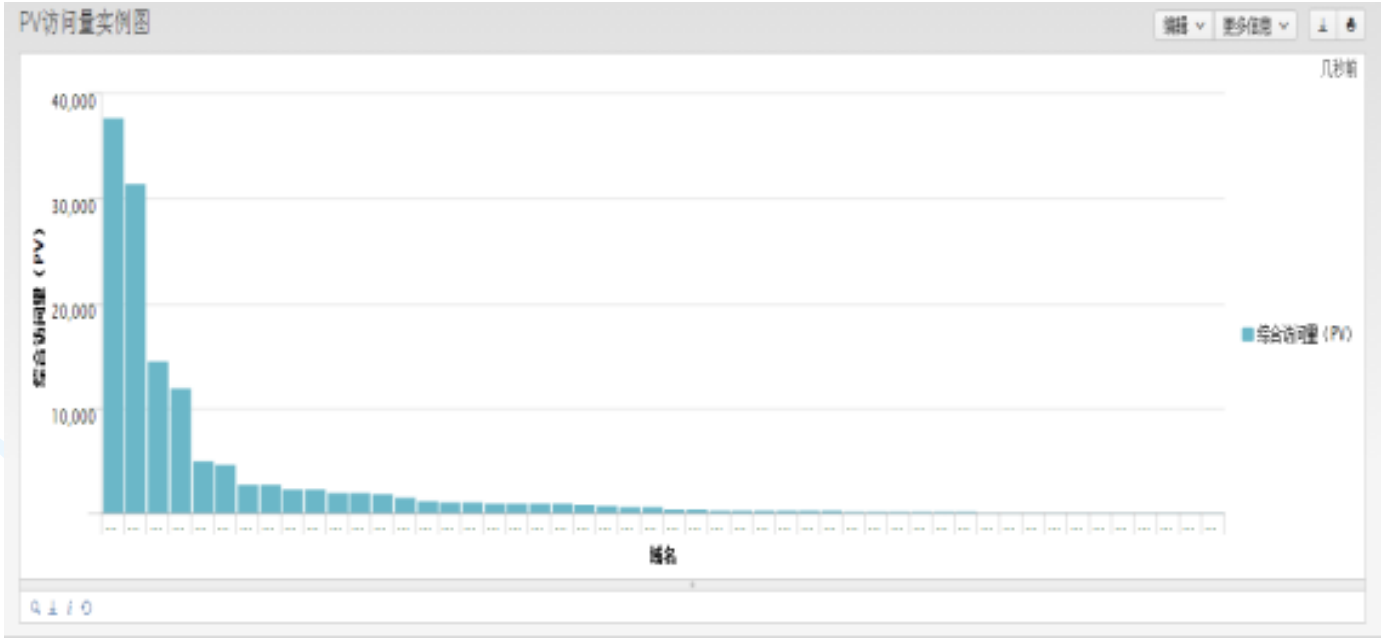
2015年6月1日的Web日志记录所包含的全部网站的PV访问量，使用Splunk软件对日志记录所包含的全部网站的PV访问量进行处理。

域名	综合访问量（PV）
http://epub.cnki.net:80	51358
http://www.cnki.net:80	50946
http://piccache.cnki.net:80	25544
http://acad.cnki.net:80	16926
http://adp.cnki.net:80	7708
https://libproxy.bjut.edu.cn:443	6610
http://gongjushu.cnki.net:80	5476
http://images.webofknowledge.com:80	5064
http://docdownload.cnki.net:80	4898
http://cdmd.d.cnki.net:80	4142
http://ac.els-cdn.com:80	3934
http://libproxy.bjut.edu.cn:80	3404
http://pdf.d.cnki.net:80	3348
http://caj.d.cnki.net:80	3312
http://ieeexplore.ieee.org:80	3028
http://caj.tmp.d.cnki.net:80	2822
https://scifinder.cas.org:443	2490

图网站综合浏览量（PV）统计图

用户访问行为可视化分析

-----网站访问量分析 PV



图网站综合浏览量（PV）柱状图

用户访问行为可视化分析

-----网站访问量分析 PV

2015年6月1日网站综合浏览量（PV）排名

排名	域名及端口	事件数（单位：个）
1	http://epub.cnki.net:80	51358
2	http://www.cnki.net:80	50946
3	http://piccache.cnki.net:80	25544
4	http://acad.cnki.net:80	16926
5	http://adp.cnki.net:80	7708
6	https://libproxy.bjut.edu.cn:443	6610
7	http://gongjushu.cnki.net:80	5476
8	http://images.webofknowledge.com:80	5064
9	http://docdownload.cnki.net:80	4898
10	http://cdmd.d.cnki.net:80	4142
11	http://ac.els-cdn.com:80	3934
12	http://libproxy.bjut.edu.cn:80	3404
13	http://pdf.d.cnki.net:80	3348
14	http://caj.d.cnki.net:80	3312
15	http://ieeexplore.ieee.org:80	3028
16	http://caj.tmp.d.cnki.net:80	2822
17	https://scifinder.cas.org:443	2490
18	http://staticieeexplore.ieee.org:80	2000
19	http://kreader.cnki.net:80	1776
20	http://apps.webofknowledge.com:80	1584

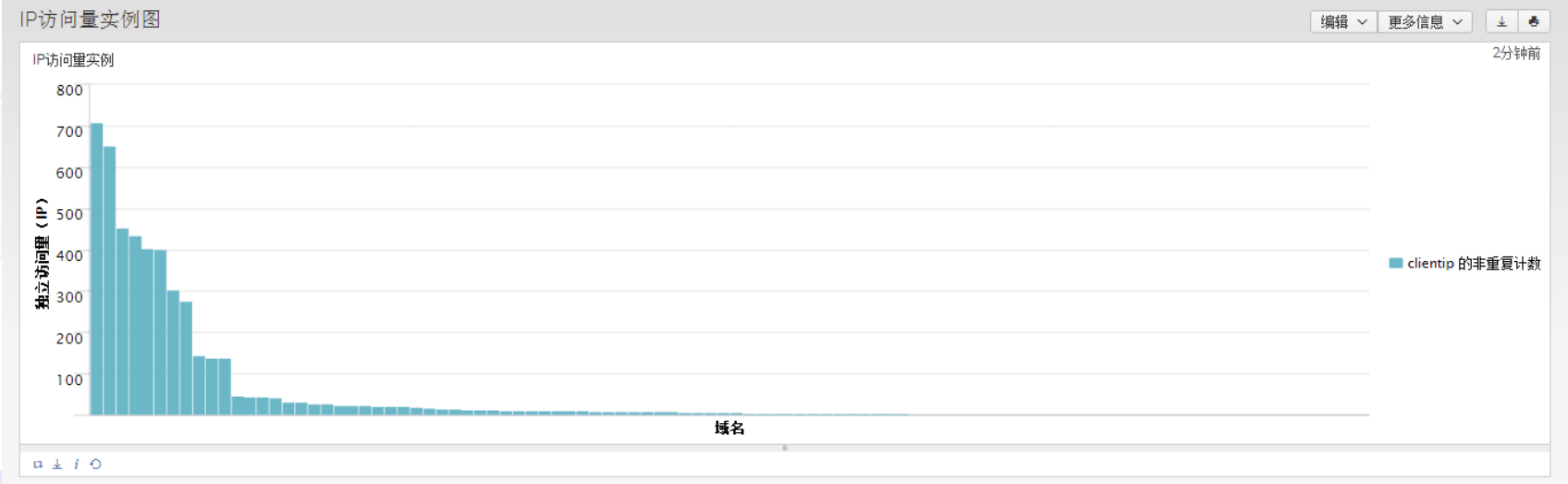
结合图表及统计结果，中国知网是学校师生访问的主要数据库资源，以ISI Web of knowledge为检索平台的数据库Web of Science访问量仅次于中国知网。

用户访问行为可视化分析

-----网站访问量分析 PV

通过对网站PV访问量进行排名，可知学校师生高频使用的数据库依次为：中国知网（CNKI）、Web of Science、IEL（IEEE/IET全文数据库）数据库、CA(SciFinder Scholar)网络数据库、“读秀”知识库、万方数据库、ScienceDirect Online（SDOL）全文数据库等。

用户访问行为可视化分析
-----网站访问量分析　　IP访问量



2015年6月1日网站独立访问量（IP）柱状图

用户访问行为可视化分析

-----网站访问量分析　IP访问量

2015年6月1日网站IP访问量排名

排名	域名及端口	独立用户个数（单位：个）
1	https://libproxy.bjut.edu.cn:443	709
2	http://libproxy.bjut.edu.cn:80	651
3	http://epub.cnki.net:80	453
4	http://piccache.cnki.net:80	435
5	http://acad.cnki.net:80	403
6	http://www.cnki.net:80	401
7	http://docdownload.cnki.net:80	303
8	http://adp.cnki.net:80	276
9	http://cdmd.d.cnki.net:80	145
10	http://caj.d.cnki.net:80	137
11	http://pdf.d.cnki.net:80	137
12	http://caj.tmp.d.cnki.net:80	45
13	http://apps.webofknowledge.com:80	44
14	http://pcs.webofknowledge.com:80	43
15	http://images.webofknowledge.com:80	41
16	http://customimages.webofknowledge.com:80	32
17	http://www.sciencedirect.com:80	31
18	http://pdf.tmp.d.cnki.net:80	28
19	http://cdn.els-cdn.com:80	27
20	http://172.24.10.20:8088	24

用户访问行为可视化分析
-----网站访问量分析　IP访问量

中国知网（CNKI）和Web of Science是北京工业大学图书馆数据库中使用频率最高的数据库。其他数据库如：ScienceDirect Online（SDOL）全文数据库、万方数据库、CA（SciFinder Scholar）网络数据库、“读秀”知识库也是师生获取电子资源的主要来源。除综合性数据库外，北京工业大学师生经常访问的数据库更多涉及计算机科学、电子学、化学、材料等学科领域。

搜索热词可视化结果及分析 1

在众多数据库中选择中国知网并对其相应字段加以分析。使用Splunk筛选出仅包含用户在中国知网上检索的事件。截取前300个搜索结果，删除不可识别的结果，最终得出排名前50位的关键词。这些关键词就是在2015年6月1日至7日北京工业大学用户访问中国知网数据库中的搜索热词。使用IBM的Many Eyes网站绘制词云。

搜索热词可视化结果及分析 2

在词云中，搜索热度由字号体现出来。搜索热词中"北京工业大学"被搜索的频率最高，"中国"、"北京"、" 美国"等表示地理位置的词汇也出现在词云中。分析认为，图书馆数据库访问者更希望获取本校在各研究领域所产出的文献，因此在中国知网上进行检索时键入"北京工业大学"作为关键词。在学术研究中，需要参考学习国外的科研成果与研究技术，所以在关键词中出现"美国"、"德国"的一类代表国家名称的词汇。"半导体激光器"、"丙烯酸甲酯"、"城市形象"、"钢结构"等专业学术名词也出现在热词词云中，而"可穿戴设备"、"智能家居"、"经济全球化"等词汇是当前的研究热点。

用户地理分布可视化结果及分析1

合法用户访问高校图书馆电子资源的途径主要有2种：①在校园内（包括在图书馆内），即校园网用户；②在校园外，非校园网用户。据美国B. Franklin等人的调查资料显示，美国读者使用图书馆数据库资源时通过第一种途径占85.31%；第二种途经占14.69%。这一结果表明，校外访问图书馆数据库资源已经成为图书馆合法用户的重要访问渠道，这种访问方式的增长趋势日益明显。

用户地理分布可视化结果及分析2

使用Splunk的Baidu Map应用程序获取网站访问者的地理信息。该应用程序提供的geoip命令可以获取Web日志记录中的IP地址，并能够绘制包含这些IP地址的世界地图。geoip命令将事件中的IP地址与对应地理位置联系起来，将一系列的地理信息添加到搜索结果中，这些信息包括经度、纬度、国家、城市等。搜索事件的地理可视化有助于迅速发现和定位潜在问题或故障的具体位置，减少平均的故障恢复时间。

用户地理分布可视化结果及分析3

图访问者世界地理分布图

图访问者全国地理分布图

个体用户访问行为结果及分析

图书馆管理人员希望了解每个用户访问目的，获取这些用户的访问行为。通过分析用户行为，以发现不同用户对图书馆电子资源的不同偏好。学校图书馆可以这些依据加强图书馆服务建设，增设用户感兴趣的电子资源，放弃利用率较低的电子资源，实现对资金的合理分配。

Splunk通过用户地理位置信息可以定位到具体用户，针对用户在Web日志中产生的某些记录进行深入分析。根据字段获取用户访问时间、地点、访问目的等信息，挖掘用户所关注的学术领域、作者及相关文献。

个体用户访问行为结果及分析1

用户信息表

字段	描述	值
user	用户名	07XXX
clientip	用户IP	96.244.252.113
clientip_city	城市	Crofton
clientip_country_code	国家代码	US
clientip_country_name	国家名称	United States
clientip_latitude	纬度	39.0142
clientip_longitude	经度	-76.6794
clientip_region_name	地区名	MD
date_year	年份	2015

用户访问网站及事件数

域名及端口	事件数（单位：个）
http://acad.cnki.net:80	155
http://adp.cnki.net:80	47
http://caj.d.cnki.net:80	18
http://caj.tmp.d.cnki.net:80	1
http://cdmd.d.cnki.net:80	8
http://docdownload.cnki.net:80	30
http://epub.cnki.net:80	454
http://pdf.d.cnki.net:80	9
http://piccache.cnki.net:80	71
http://www.cnki.net:80	305

个体用户访问行为结果及分析2

该用户仅访问了中国知网，并对搜索的文献进行下载。进一步分析该用户的搜索关键词，发现该用户对关键词“稻田”、“免耕”、“水稻”、“产量”、“长期试验”进行检索，并通过全文搜索对“不同施肥制度对黄泥田土壤酶活性及养分的影响”、“土壤表层管理对部分土壤化学性质的影响”进行检索。在对作者的分析中发现有“徐阳春”、“高亚军”、“郝晓晖”、“张扬珠”、“黄东迈”等作者姓名。根据以上信息分析，可以得知该用户需要关于“土壤”方面的文献，涉及到化学等学科，检索的作者均为该领域学者。

用户访问行为分析1

根据网站访问量和用户个体访问行为分析，北京工业大学师生对中国知网（CNKI）的数据库资源使用频率占据绝对优势。中国知网、Web of Science、万方数据库是被用户访问的频繁数据库，其它专业数据库资源网站的访问量呈指数下降，说明全校师生对其他一些专业数据库的使用较少。作为理工科院校，学校师生更偏向于访问涉及计算机科学、电子学、化学、材料等学科领域的数据库资源。

用户访问行为分析2

根据用户地理分布可视化结果可知，北京工业大学图书馆数据库存在远程访问用户，这些用户分布在全国乃至世界范围内。这说明北京工业大学图书馆开通了校外访问服务，为外出交流、讲学、出国访学的用户使用图书馆的数据库资源提供了校外访问途径。

数据库资源建设的建议

通过用户对数据库资源访问行为分析，对图书馆数据库资源建设提出如下可行性建议。

- 减少或放弃购买使用效率低或访问有效性低的数据库。
- 围绕学校学科建设重点和用户感兴趣领域引进相关数据库。用不同的分析方法挖掘用户感兴趣的学科研究领域，结合学校学科建设重点引进相关领域数据库资源，合理利用经费。
- 改进数据库资源远程访问技术。由于图书馆数据库存在远程用户，需要通过提高远程访问技术为校外访问建立更为安全、稳定的访问服务。

总 结

以北京工业大学图书馆数据库访问日志作为研究对象，以数据库用户访问行为作为研究点，应用大数据分析软件Splunk研究分析图书馆用户访问数据库资源的行为。

通过对用户访问行为的分析，找到全校师生经常访问的数据库资源，挖掘出一段时间内被用户高频检索的词汇，讨论了校外访问对图书馆数据库资源的使用情况，对个体用户的兴趣方向、涉及学术领域及所关注的文献作者进行挖掘分析。为学校数据库资源建设提出合理化建议。

谢 谢！