DRAS数字资源使用分析系统及解决方案

同方知网（北京）技术有限公司
高等教育分公司总经理　王峰

内容

I. 图书馆数据挖掘的意义

II. 《DRAS数字资源使用分析系统》介绍

III. 未来构想——以清华大学使用日志为例

图书馆“大数据”时代来了！

- 随着大数据时代的来临，图书馆也不可避免的受到了大数据信息浪潮的冲击。
- 科学研究的变化要求数字图书馆大数据的支撑。越来越多的学科领域完全建立在大量数据的基础上，比如系统生物学、宏生态学、基因组学等。
- 用户信息素养的变化，用户服务要求越来越个性化、学科化，要求图书馆从大量的数据中分析潜在的价值，从而决定着大数据时代的图书馆的发展水平及方向。

——陈传夫 《大数据时代的数字图书馆》

2012年中国图书馆年会

图书馆的“大数据”

数据挖掘的意义

- 1、实现针对不同读者的个性化服务
  - 以“资源”为核心——以“读者”为核心
  - 跟踪服务、精准服务、知识关联服务、宣传推广服务
- 2、提供研究动向以及研究热点的变化
  - 分析学科研究热点的动向
  - 掌握科研人员的研究进展
- 3、为资源采购部门提供资源评价的建议
  - 分析资源的使用情况，评估性价比
  - 收集资源访问历史，预测读者关注的热点
  - 评估资源建设的合理性

《DRAS数字资源使用分析系统》介绍

一、系统开发背景

近几年，随着图书馆购买数字资源的比例逐年上升，数字资源建设从高速增长期进入稳定增长期。

美国研究图书馆协会（ARL）统计显示：

美国研究型大学图书馆的数字资源经费占文献总经费的比率在2010-2011年度就已达到62.47%，超过了印本资源。

图书馆发展电子馆藏工作的重心已经从最初的资源引进转向了有效的数字资源使用评估。

一、系统开发背景

随之带来了一系列问题：

图书馆的资源配置是否合理，如何使用有限的经费选购日益增多的数字资源？

如何将数字资源与学科建设相结合，使引进的数字资源得到有效利用？

如何保证数字资源的合理合规使用？

这一系列的问题都涉及到对数字资源的使用评估。

一、系统开发背景

图书馆数字资源使用日志分析的难题：

- 目前各个数据库的统计标准和计量方法不同，导致不同数据服务商提供的统计数据无法比较；
- 缺乏在线行为统计的说明，导致统计数据难以理解；
- 数据库商提供的使用数据的真实性及有效性无从判定，统计数据无法真实反映读者需求的问题。

二、系统简介及网络架构

1、 系统简介

- 通过技术方案，采集机构范围内访问数字资源的底层非结构化的web日志，对数据进行清洗、会话识别及分析，挖掘读者使用数字资源的行为数据，为图书馆提供客观真实且遵循统一标准的数据库访问日志，便于图书馆对数据库的使用价值进行客观分析评估，科学合理选购数据库，从而为图书馆采购决策工作及挖掘读者使用需求提供支持。

二、系统简介及网络架构

2、系统设计目标

1）一站式统计图书馆采购的所有中外文数字资源的使用日志，掌握资源真实使用情况，为图书馆资源采购提供数据支撑。

2）监测读者恶意下载行为，保障数字资源的合理合规使用。

3）通过对读者检索行为、浏览行为、下载行为等数据的挖掘，把握读者真实需求，为图书馆开展个性化、学科化服务提供数据支撑。

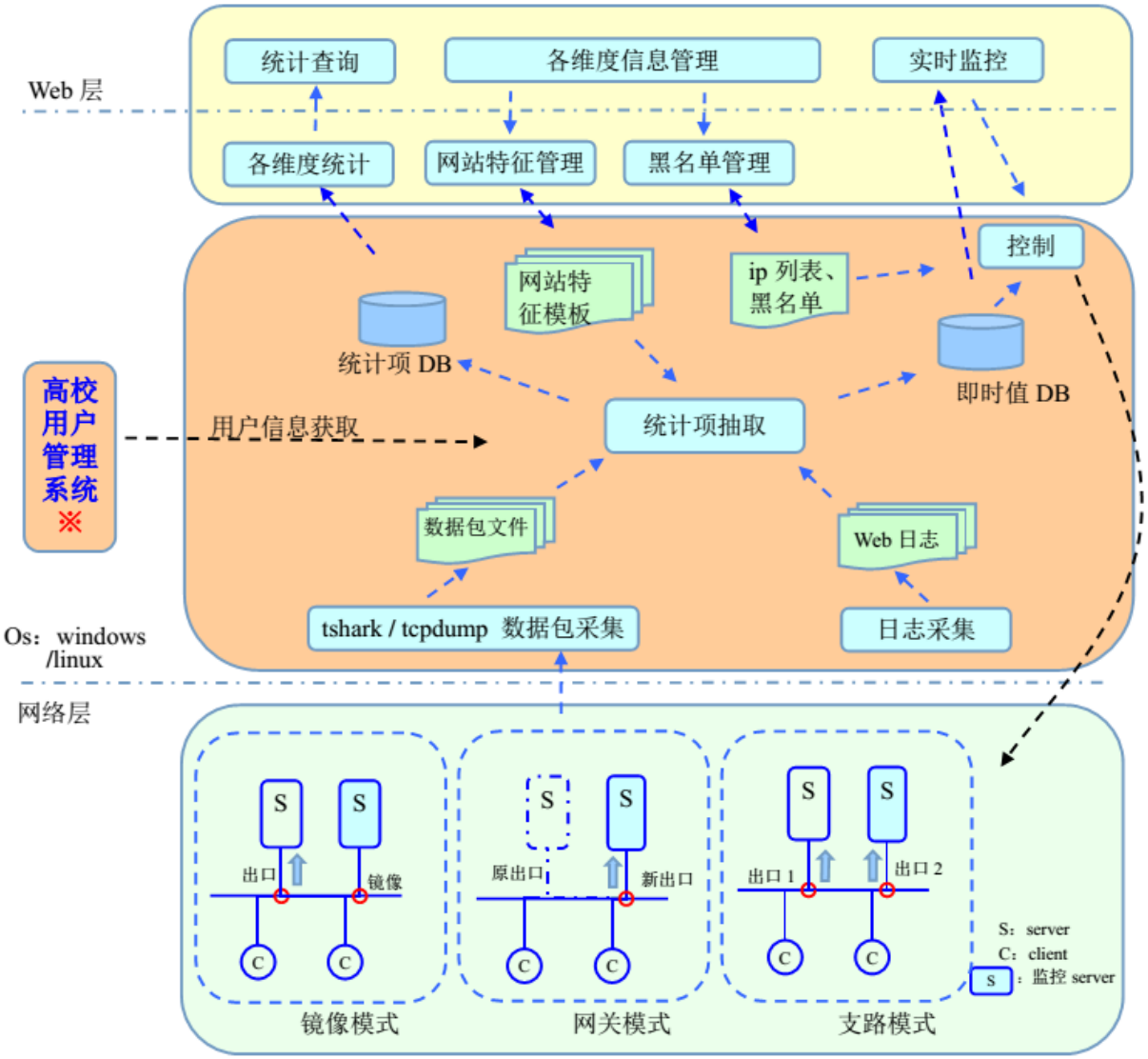
二、系统简介及网络架构

3、系统难点及解决方案

- **难点：**本系统中涉及到Web日志数据的统计和分析，需要从海量的互联网访问日志中采集出所有读者访问数据库的使用日志。
- **解决方案：**结合各高校的网络环境的现状，采用有针对性的几种实施方案，将数字资源的访问流量从海量的网络请求中分离出来，是本统计系统获取统计数据的基础。

二、系统简介及网络架构

整体架构



网络结构：结合高校既有网络结构，采用合理切入方式，分流出电子文献的访问请求，是数据采集的基础。

数据采集：利用 tshark/tcpdump 等抓包工具，收集数据包，存储为源数据文件。

日志采集：针对实时显示的如流量、访问次数等信息，通过 web 日志形式收集，便于快速分析，实时控制。

统计项抽取：基于源数据文件，通过分组、清洗、抽取等步骤，获得须统计项目，存入数据库。

实时控制：针对即时数据（访问频次、流量），根据阀值，判断不良操作行为，加以控制。

各维度统计：基于用户身份、请求资源、访问行为、时间等信息，提供丰富的统计方式，直观呈现统计结果。

网站特征管理：分析各资源网站的特征信息，形成规则模板，提供有效的管理方式，便于识别操作行为及资源类别。

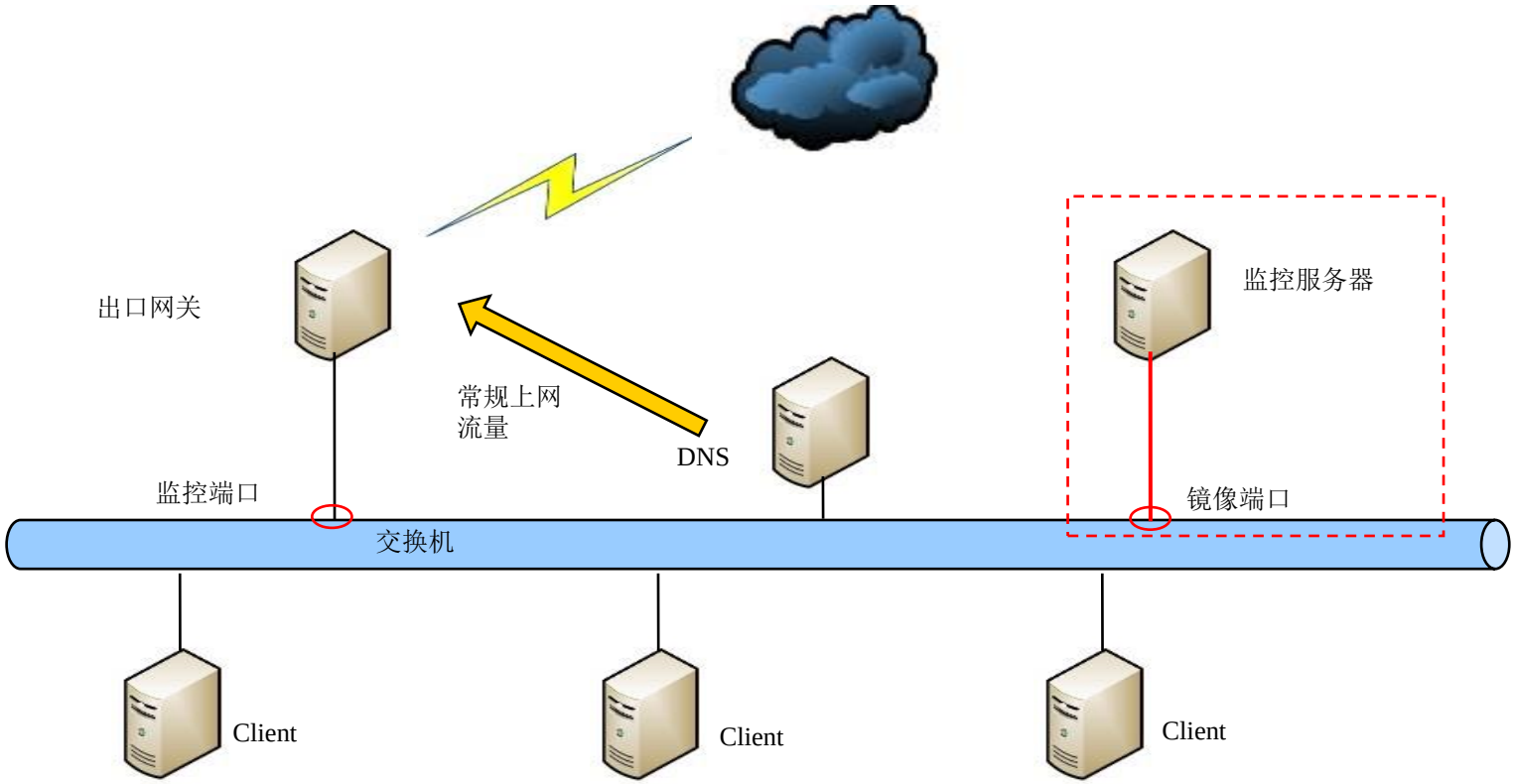
黑名单管理：通过数据分析或人工录入，对异常 ip、用户、行为等进行管理，便于有效控制。

※高校用户管理系统：各高校的用户管理系统同本统计系统数据互访，便于身份的识别。

二、系统简介及网络架构

a） 镜像模式

在交换机镜像端口，增设监控服务器，截获出口网关的数据包，作存储、统计。



二、系统简介及网络架构

a）镜像模式

优点：

1、不需要更改网络拓扑结构，工程量小

2、日志获取完整度高、丢包率低

缺点：

1、涉及部分隐私问题

解决办法：将服务器、系统操作权限交给高校网络中心统一管理

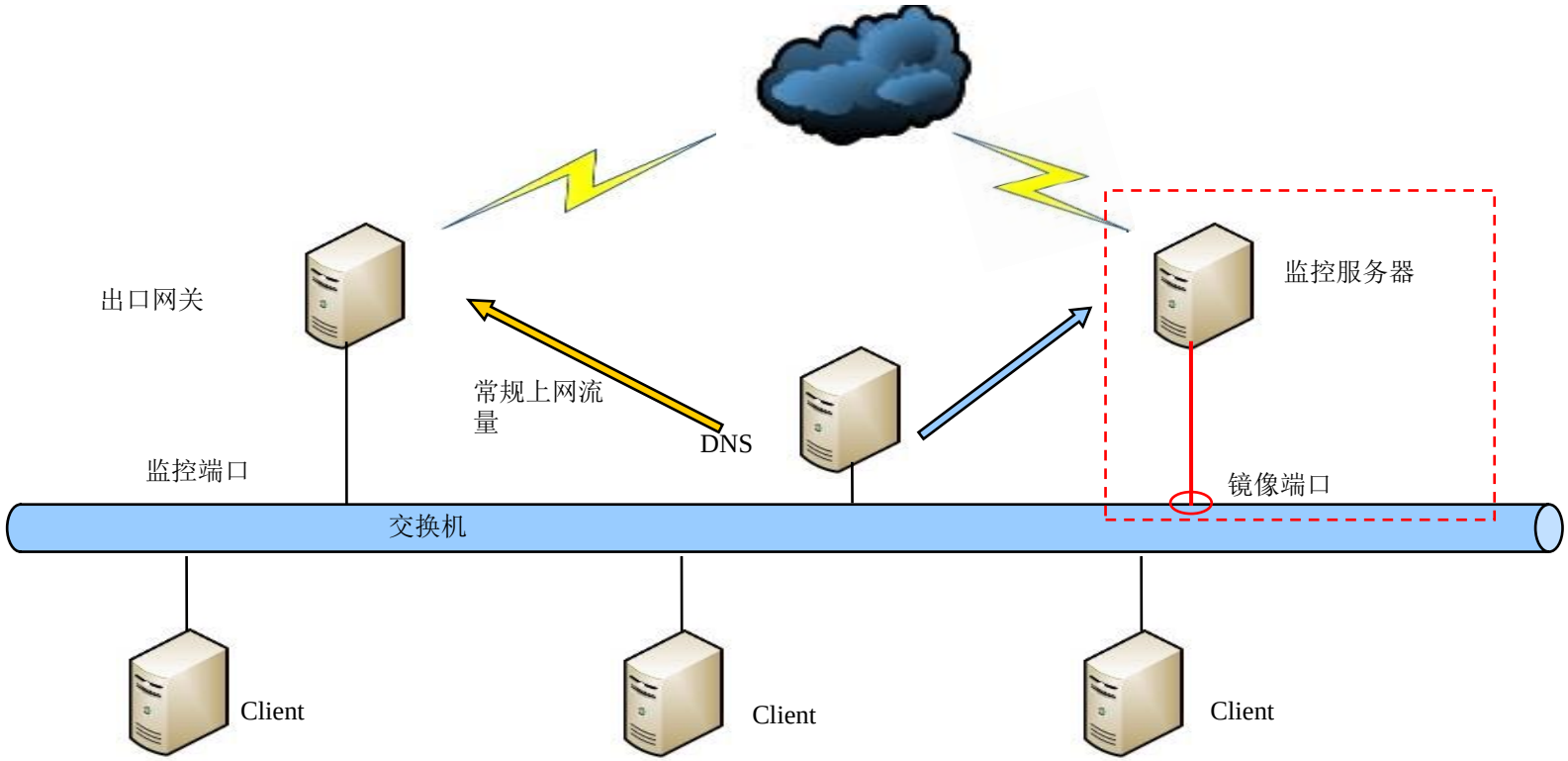
适合：

此模式适合无法改变任何架构，对用户访问可控性要求低时使用。

二、系统简介及网络架构

b） 支路模式（分离DNS）

增设网络出口，对目标文献服务器访问请求的发送、接收由监控服务器承担。



二、系统简介及网络架构

b） 支路模式（分离DNS）

优点：

1、隐私问题涉及少，只需监听数据库的使用日志

2、日志详细度高，有利于对使用数据深度分析

缺点：

1、需要较小程度更改网络拓扑结构，对IPT或DNS指向进行修改，存在DNS后续维护问题。

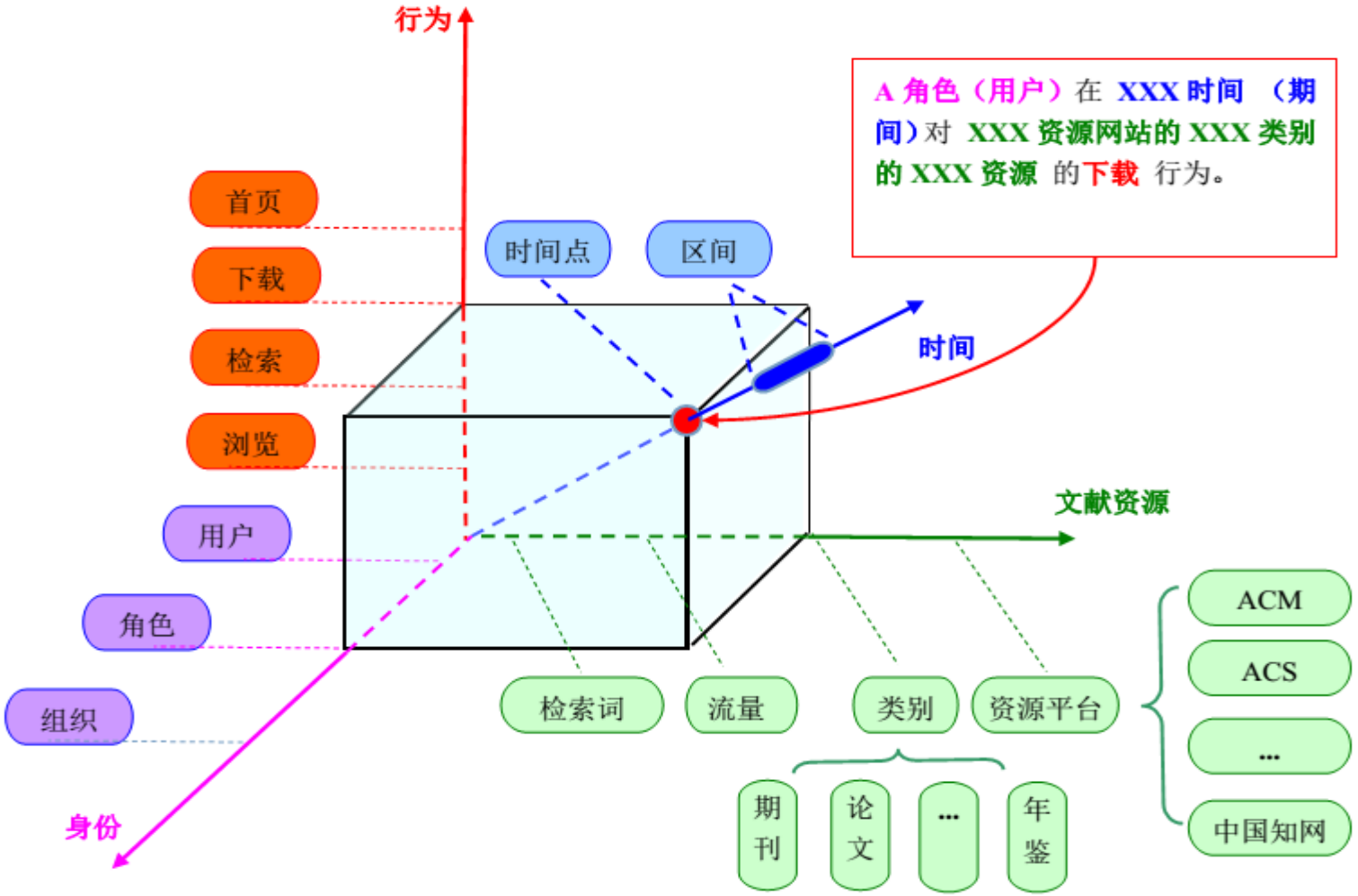
适合：

此模式适合可调整DNS或IPT时使用。

系统功能设计

- CNKI数字资源使用分析系统
  - 数据库管理
    - 数据库列表
    - 数据库域名及IP地址检测
    - 院系IP地址列表
  - web日志数据准备
    - web日志数据采集
    - 数据清洗
    - 用户识别
    - 会话识别
  - 时间维度统计分析
    - 数据库使用统计分析
    - 资源使用统计分析
    - 院系使用统计分析
  - 检索词统计分析
    - 数据库使用检索词分析
    - 院系使用检索词分析
  - 读者使用行为分析
    - 读者使用整体统计
    - 读者使用详情日志查询
  - 流量监控
    - 设置阀值
    - 实时监控
    - 违规日志查询

多维度的数据统计和分析



四、系统主要功能模块及页面展示

1、数据库管理模块

1） 数据库列表：管理图书馆已订购数据库的名称、域名、对应IP地址等信息。

2）数据库域名及IP地址检测：系统后台检测数据库的域名及IP地址是否发生变化并推送提示更新。当学校图书馆采购了新数据库时，可以到系统后台申请添加数据库的域名及IP地址等信息。

2）院系IP地址列表：登记院系的IP段地址信息。

四、系统主要功能模块及页面展示

编号	数据库名称	数据库类型	资源类别	语种	域名	数据库厂商	服务器IP	服务器地域
1	中国学术期刊网络出版总库	全文	期刊	中文	http://www.cnki.net	CNKI	211.151.247.92	北京市
2	中国博士学位论文全文数据库	全文	学位论文	中文	http://www.cnki.net	CNKI	211.151.247.92	北京市
3	中国优秀硕士学位论文全文数据库	全文	学位论文	中文	http://www.cnki.net	CNKI	211.151.247.92	北京市
4	万方学术期刊	全文	期刊	中文	http://g.wanfangdata.com.cn/	万方	121.194.12.3	北京市
5	ASM（美国微生物学会）	全文	期刊	外文	http://www.journals.asm.org/	ASM	171.66.122.153	美国加利福尼亚州斯坦福市
6	ACS（美国化学学会）	全文	期刊	外文	http://pubs.acs.org/	ACS	118.186.70.47/…	北京市
7	JoVE（可视化实验期刊数据库）	全文	视频	外文	http://www.jove.com/	JoVE	166.78.179.190	美国德克萨斯州圣安东尼奥市
8	Web of Science	题录	其他	外文	http://webofknowledge.com	汤森路透	167.68.24.3	美国
9	SpringerLink电子期刊	全文	期刊	外文	http://link.springer.com/	Springer	203.69.81.32/…	台北市
10	MyiLibrary电子图书	全文	图书	外文	http://lib.myilibrary.com/…	MyiLibrary	198.183.167.199	美国

四、系统主要功能模块及页面展示

2、日志统计分析模块

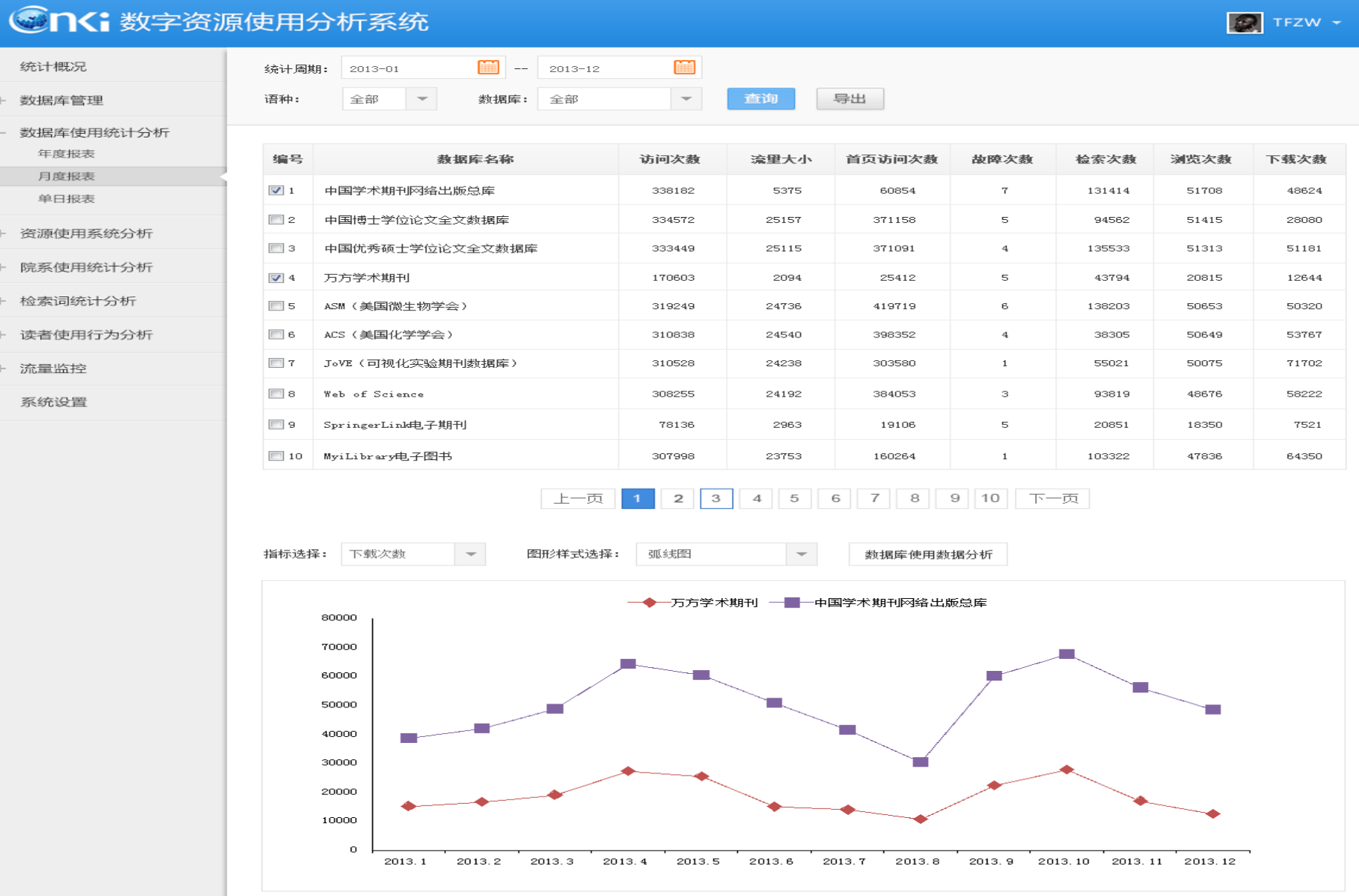
1）数据库访问维度：可以查看和分析各个数据库的使用数据及不同数据库在不同时间段的使用量。

2）资源访问维度：可以查看和分析各种资源的使用数据及不同资源在不同时间段的使用量。

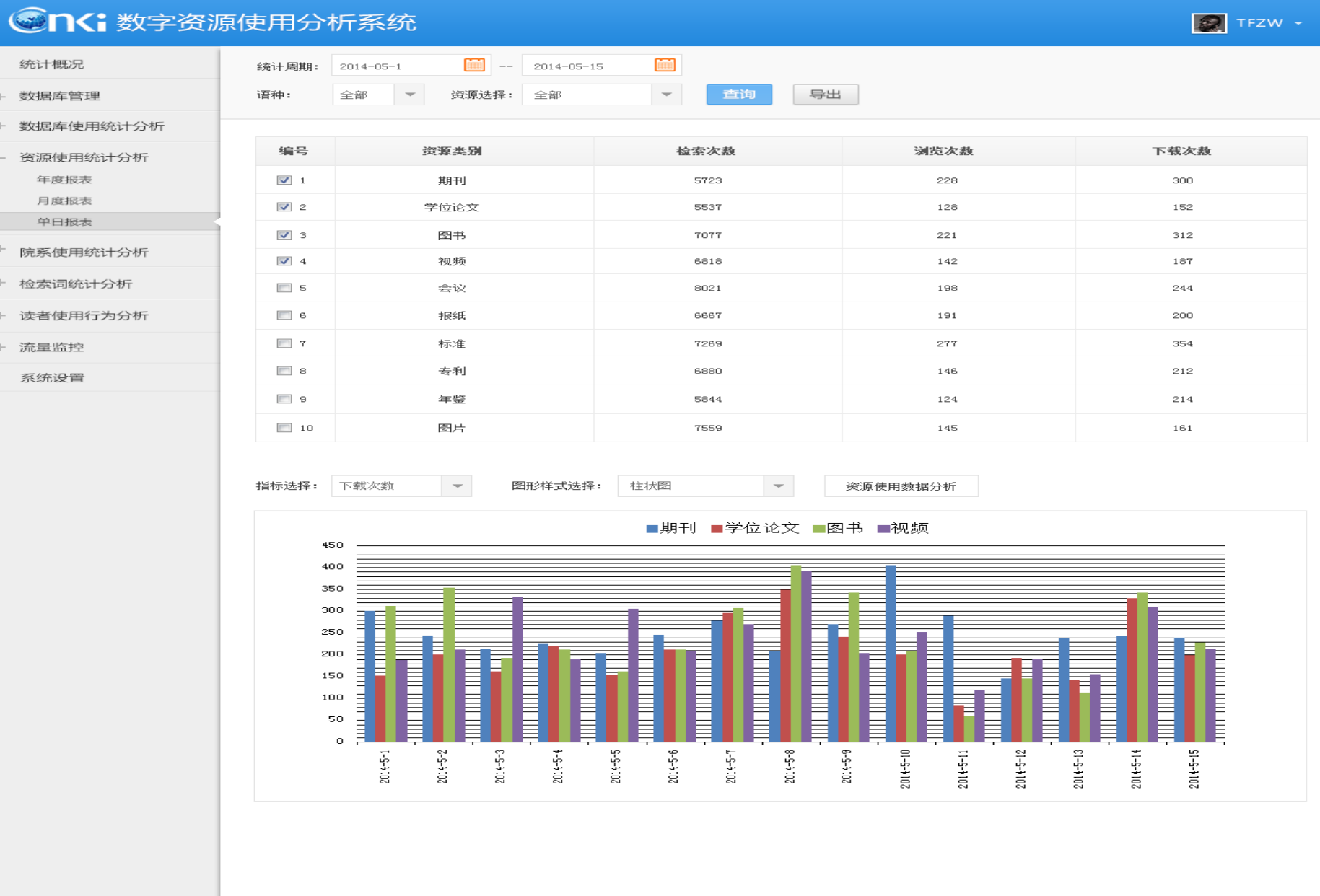
3）院系访问维度：可以查看和分析各个学院的使用数据及不同学院在不同时间段的使用量。

注：根据以上三种访问维度，按照年、月、日三个时间频度，提供使用数据的查询和分析（包括访问次数、首页访问次数、检索次数、下载次数、浏览次数、故障次数、流量大小）。

四、系统主要功能模块及页面展示



四、系统主要功能模块及页面展示



四、系统主要功能模块及页面展示

3、检索词统计分析模块

1） 数据库使用检索词分析：提供不同数据库在不同时间段的热点检索词排名及频次。

2） 院系使用检索词分析：提供不同学院在不同时间段的热点检索词排名及频次。

四、系统主要功能模块及页面展示

四、系统主要功能模块及页面展示

4、读者使用行为分析模块

1） 读者使用整体统计：根据读者的IP地址提供不同读者在不同时间段的访问次数、检索次数、浏览次数、下载次数等使用数据，并且可以选择多个IP的使用行为数据进行对比分析。

2） 读者使用详情日志查询：根据读者的IP地址提供不同读者在不同时间段对不同数据库的详细使用行为。

四、系统主要功能模块及页面展示

数字资源使用分析系统

TFZW

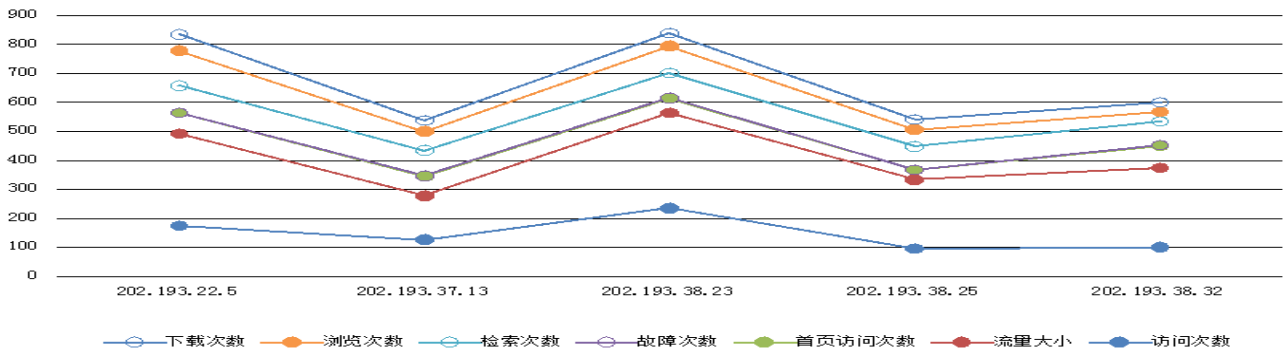
- 统计概况
- ＋ 数据库管理
- ＋ 数据库使用统计分析
- ＋ 资源使用统计分析
- ＋ 院系使用统计分析
- ＋ 检索词统计分析
- － 读者使用行为分析
  - 读者使用整体统计
  - 读者使用详情日志查询
- ＋ 流量监控
- 系统设置

统计周期： 2014-03-25 -- 2014-05-25 查询 导出

编号	用户IP	访问次数	流量大小	首页访问次数	故障次数	检索次数	浏览次数	下载次数
1	202.193.22.5	175	318	73	0	93	121	56
2	202.193.37.13	128	151	67	2	88	65	38
3	202.193.38.23	237	329	50	1	86	93	45
4	202.193.38.25	97	238	34	0	81	58	34
5	202.193.38.32	101	275	76	3	81	33	33
6	202.193.38.45	174	282	86	1	77	51	66
7	202.193.38.47	159	282	83	4	76	59	52
8	202.195.42.47	35	96	17	2	27	12	22
9	202.195.42.23	96	159	23	3	71	47	31
10	202.195.42.25	237	349	74	6	66	63	53

上一页 1 2 3 4 5 6 7 8 9 10 下一页

指标选择： 全部 图形样式选择： 弧线图 读者使用数据分析



四、系统主要功能模块及页面展示

5、流量监控模块

1） 设置阀值：针对不同的数据库设置不同的下载阀值、每个IP最大下载量或流量的上限等。

2） 实时监控：实时监控每个IP及学校整体的使用流量或下载次数，对异常行为进行报警，系统设置里可以选择预警时发短信给管理员。

3） 违规日志查询：支持查询历史违规使用日志。

四、系统主要功能模块及页面展示

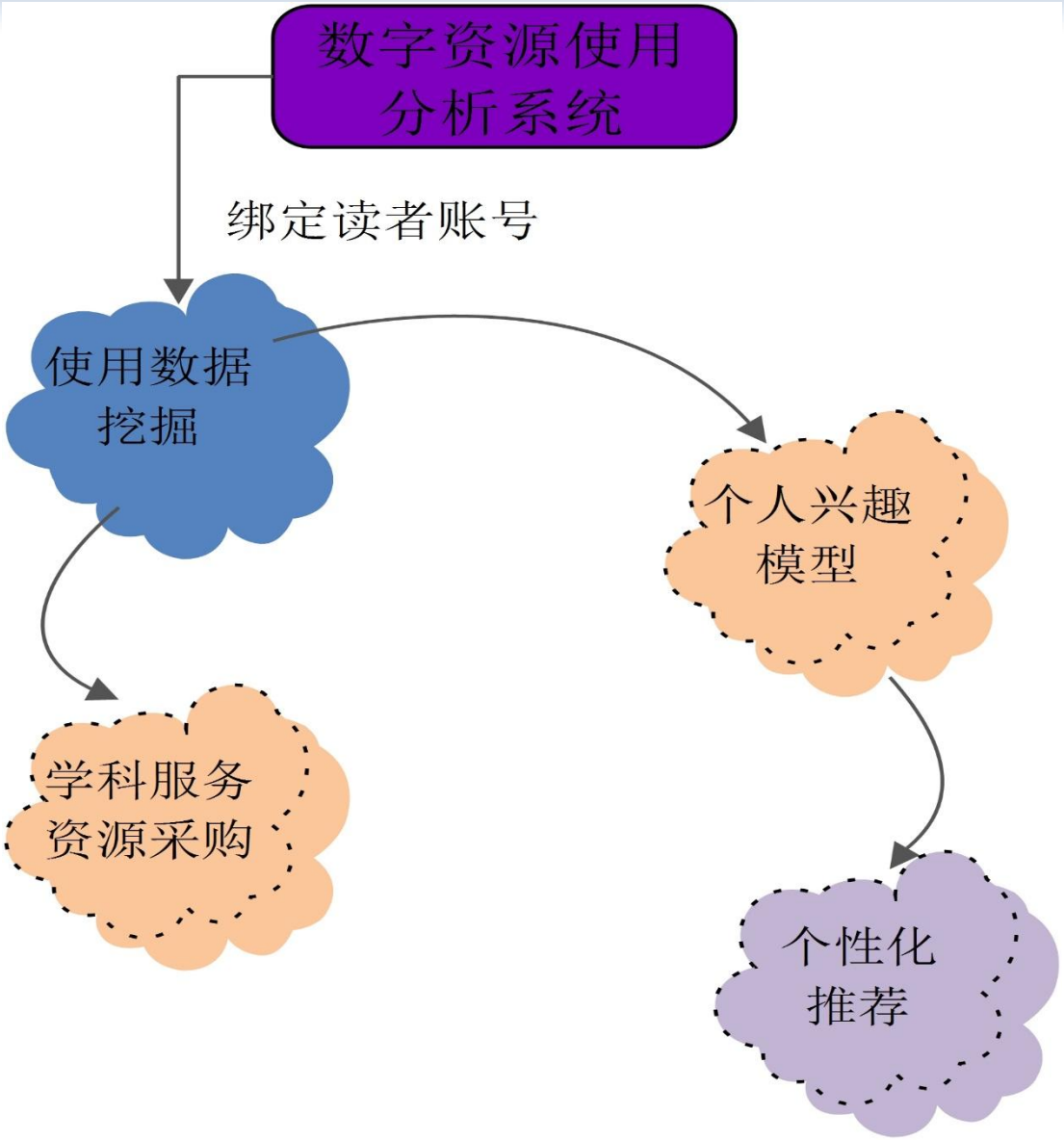
编号	用户IP	数据库名称	下载上限	实际下载	时间
1	202.193.38.45	SpringerLink电子期刊	100	112	2014-5-15
2	202.195.42.32	ASM（美国微生物学会）	200	208	2014-5-15
3	202.198.207.2	ASM（美国微生物学会）	200	317	2014-5-15
4	202.198.207.31	ACS(美国化学学会）	100	136	2014-5-15
5	202.195.42.25	SpringerLink电子期刊	100	135	2014-5-15
6	202.198.207.21	中国学术期刊网络出版总库	300	313	2014-5-15
7	202.193.38.32	万方学术期刊	200	205	2014-5-15
8	202.198.207.31	SpringerLink电子期刊	100	116	2014-5-15
9	202.193.38.45	SpringerLink电子期刊	100	109	2014-5-15
10	202.198.207.41	ACS（美国化学学会）	100	137	2014-5-15

下一步规划

- 未来构想：通过大数据和数据挖掘技术，对收集的读者行为数据进行处理和特征提取，准确的把握读者的行为特征和偏好，建立读者需求兴趣模型，从而构建数字图书馆个性化服务系统，为读者主动推送所需要的精准的各类型资源和服务。

即将发布，敬请期待！

下一步规划



案例：清华大学CNKI详细下载日志深入挖掘

- **分析对象：**《中国学术期刊网络出版总库》
- **分析行为：**下载、检索
- **时间范围：**2013年7月1日-12月31日
- **分析学科：**TOP 3学科——材料科学与工程、动力工程与工程热物理、管理科学与工程
- **挖掘目标：**热门文献、热门关注期刊、热门关注机构、热门关注作者、热门检索词

热门检索词 TOP 20

清华大学热门检索词

检索词	检索次数	检索词	检索次数
临床路径	5990	增塑剂	2521
综述	4014	测定	2503
北京	3927	人工智能	2464
中国	3631	设计	2265
美国	3461	城镇化	2240
食品包装	3261	规划	1996
经济研究	3196	城市	1977
a	2992	塑化剂	1912
检测	2962	中国社会科学	1879
the	2588	发展	1843

1、材料科学与工程

排序	热门关注机构	下载量
1	中南大学粉末冶金国家重点实验室	257
2	北京航空航天大学材料科学与工程学院	146
3	东北大学材料与冶金学院	123
4	北京航空航天大学航空科学与工程学院	111
5	北京科技大学材料科学与工程学院	95
6	西北工业大学凝固技术国家重点实验室	95
7	合肥工业大学材料科学与工程学院	93
8	西北工业大学	89
9	中南大学材料科学与工程学院	88
10	哈尔滨工业大学材料科学与工程学院	84
10	纳米材料国内外研究进展ⅰ——纳米材料的结构、特异效应与性能	17

2、动力工程与工程热物理

排序	热门关注机构	下载量
1	清华大学热能工程系	425
2	中国科学院工程热物理研究所	259
3	清华大学汽车安全与节能国家重点实验室	183
4	清华大学热能工程系热科学与动力工程教育部重点实验室	183
5	清华大学热科学与动力工程教育部重点实验室	163
6	上海交通大学机械与动力工程学院	159
7	西安交通大学动力工程多相流国家重点实验室	137
8	西安交通大学能源与动力工程学院	137
9	华中科技大学煤燃烧国家重点实验室	131
10	哈尔滨工业大学能源科学与工程学院	129

3、管理科学与工程

排序	热门关注机构	下载量
1	西安交通大学管理学院	105
2	清华大学经济管理学院	88
3	浙江大学管理学院	49
4	清华大学公共管理学院	43
5	华南理工大学工商管理学院	41
6	南开大学周恩来政府管理学院	33
7	东北财经大学工商管理学院	32
8	南京航空航天大学经济与管理学院	32
9	中国科学院心理研究所	28
10	东北大学工商管理学院	27
10	企业环境下个人获取知识和输出知识能力对组织公民行为影响的研究	11

感谢您的关注

同方知网（北京）技术有限公司