



图书馆大数据与大数据图书馆

Big Data in Library and Big Data Library

2017@ 贵阳

CONTENTS

01 背景与现状

02 设计与实现

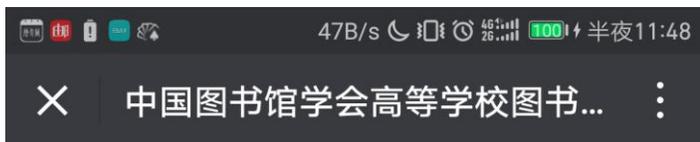
03 应用示例

04 问题及方向

背景与现状

1

背景与现状



大学之问：图书馆该去哪儿？ | 案例

2017-03-22 麦可思 谌超

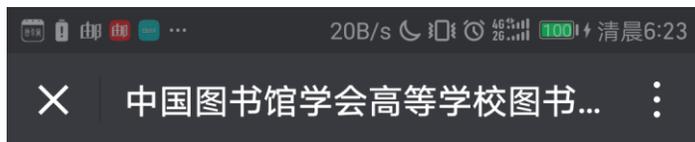
中国图书馆学会高等学校图书馆分会

1. 就业蓝皮书	2. 新生	3. 学生事务	4. 教学教务	5. 招生宣传
6. 就业	7. 专业建设	8. 质量年报	9. 战略规划	

麦可思研究专注高教管理，订阅[麦研图书馆](#)，可查看更多主题文章哦！



“学术图书馆死了。”这是美国阿尔弗莱特大学图书馆主管布莱恩·沙利文在展望学术图书馆的未来时发出的感叹。尽管悲观至极，却道出了图书馆发展的现实——馆藏量不再是图书馆唯一引以为傲的资本，以人为



学术图书馆之死

“学术图书馆死了。”美国阿尔弗莱特大学图书馆主管布莱恩·沙利文在名为《2050年学术图书馆尸检报告》一文中发出如此感叹。“死因是图书馆建筑逐渐演变成机房、学习空间和信息技术的聚集地。”

尽管沙利文的话略带调侃，但不可否认的是我们正见证着大学图书馆的这一改变。从所谓的柏林大脑（柏林自由大学的语言学系图书馆）到伍斯特大学屡获殊荣的金色蜂巢图书馆，从昆士兰大学伊普斯维奇图书馆森林式的内在到阿伯丁大学（邓肯莱斯图书馆七层高的通透设计，我们看到的是建筑形态充分体现了当今图书馆的使用功能，面对图书馆的认知也构建了我们的



数字时代来了，大学图书馆死了？ | 案例

原创 2016-01-24 麦可思 麦可思研究

关注麦可思研究，订阅高教管理宝典

阅读推荐：校园规划

日前，教育部印发修订版《普通高等学校图书馆规程》指出，高校图书馆的建设和发展应与学校的建设和发展相适应，其水平是学校总体水平的重要标志。而事实上，在数字化时代的当下，大学图书馆的生存发展却并非易事。

当 斯科特拿到老师第一节课发出的参考书单

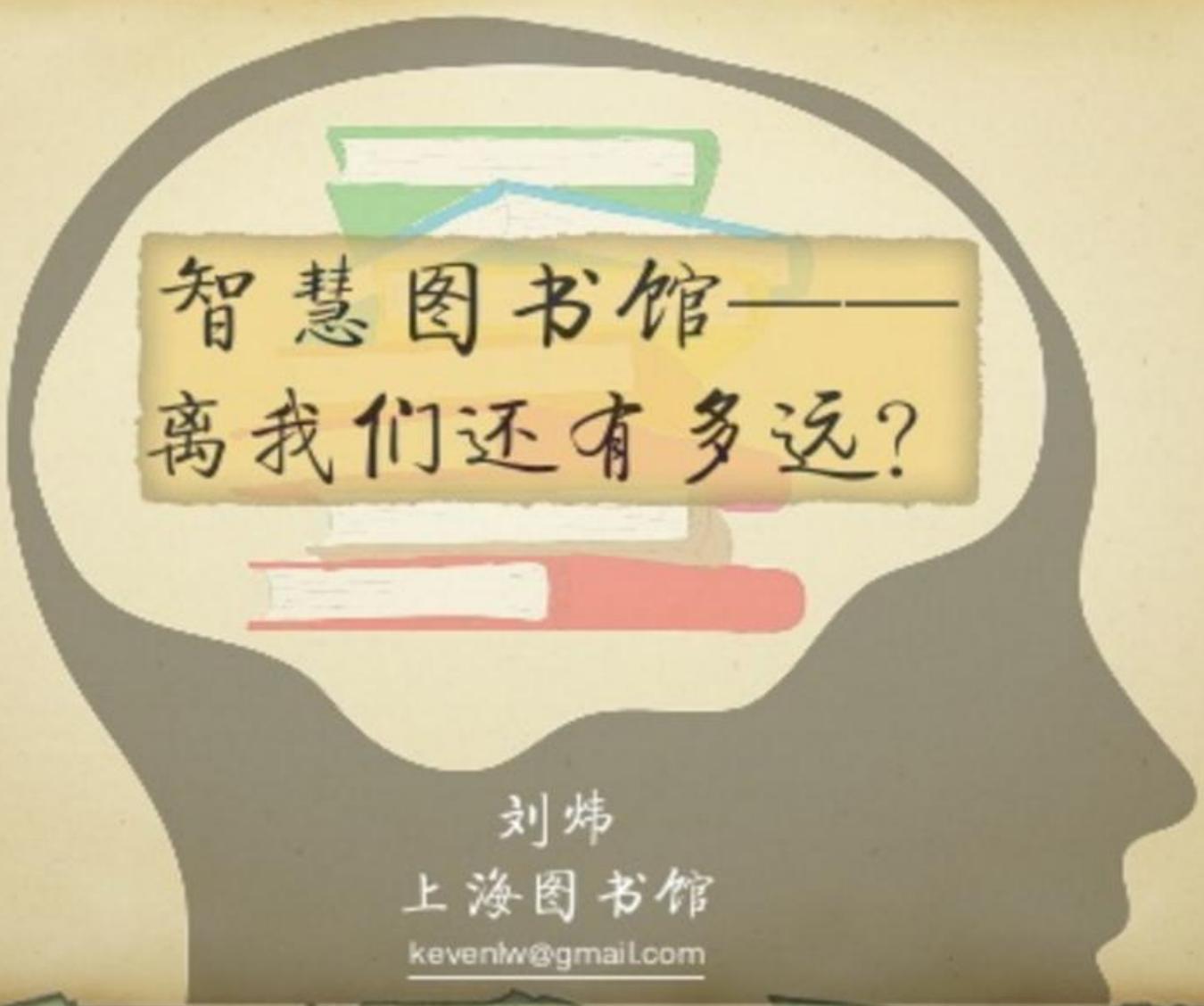
- 空间
 - 实体 Physical
 - 资源
 - 数字 Digital
 - 服务
 - 混合 Hybrid
 - 技术
 - 透明 Transparent
-
- 越来越多的人通过网络获取资源
 - 越来越多的人来到图书馆获取空间

背景与现状



- Dec. 2011
- Digital library → Smart library → Knowledge Center

背景与现状



智慧图书馆——
离我们还有多远？

刘炜

上海图书馆

keventw@gmail.com

背景与现状

Welcome to the IKCEST



International Knowledge Centre
for Engineering Sciences and Technology
under the Auspices of UNESCO
联合国教科文组织国际工程科技知识中心

Scholar

News

Technology

Dataset

Other >>



Home

News

Symposium

Training Workshop

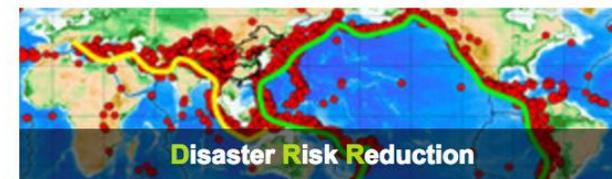
Collaboration

Assisting UNESCO

Technology

About IKCEST

Knowledge Service Sectors



Disaster Risk Reduction



Engineering Education



Silk Road Sciences and Technology



MOOC Online Education

SRST

Free online courses from the best universities and institutions of the Belt and Road

背景与现状



背景与现状

Everything **HOLLIS** **Articles**

Keywords anywhere	contains	<input type="text"/>	AND
Title	contains	<input type="text"/>	AND
Author / Creator	contains	<input type="text"/>	AND
Subject	contains	<input type="text"/>	AND

Resource Type:

Language:

Publication Date:

Start Date:

End Date:

Location:

Find more articles

-  Harvard users: [sign in to find more articles](#)
-  Alumni: [find resources for accessing articles](#)

Help with HOLLIS+

背景与现状



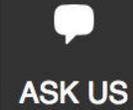
Search

Hours & locations

Borrow & request

Research support

About us



Search

Articles, e-books, & more

by

Keyword

for

ex: carbon nanotubes



[Go to BartonPlus advanced search](#)

It's coming in June! Try our new search early »

Hours & locations



Barker Library

1pm-6pm today, 24/7 Study

📍 10-500 📞 617-253-0968



Dewey Library

1pm-6pm today, 24/7 Study

📍 E53-100 📞 617-253-5676

NEWS

[Try our new beta search tool](#)



FEATURED STORY

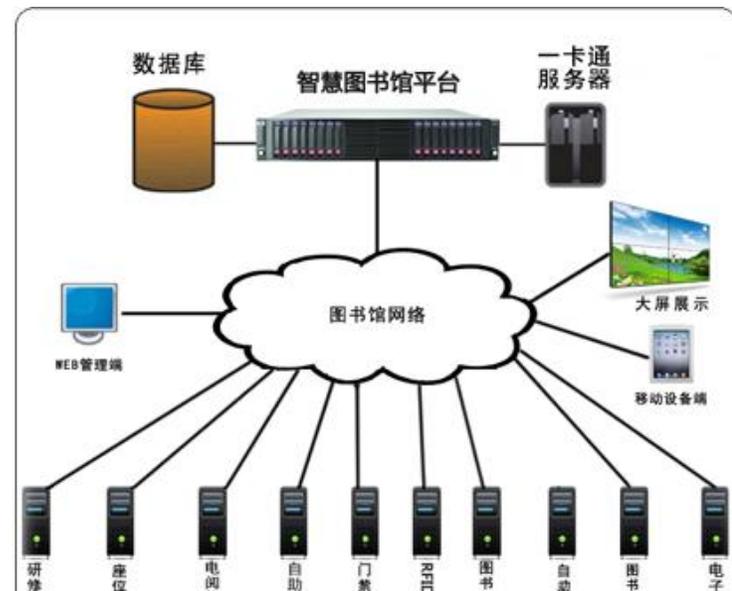
Classroom Earth: Drones over Death Valley

All news & events

背景与现状



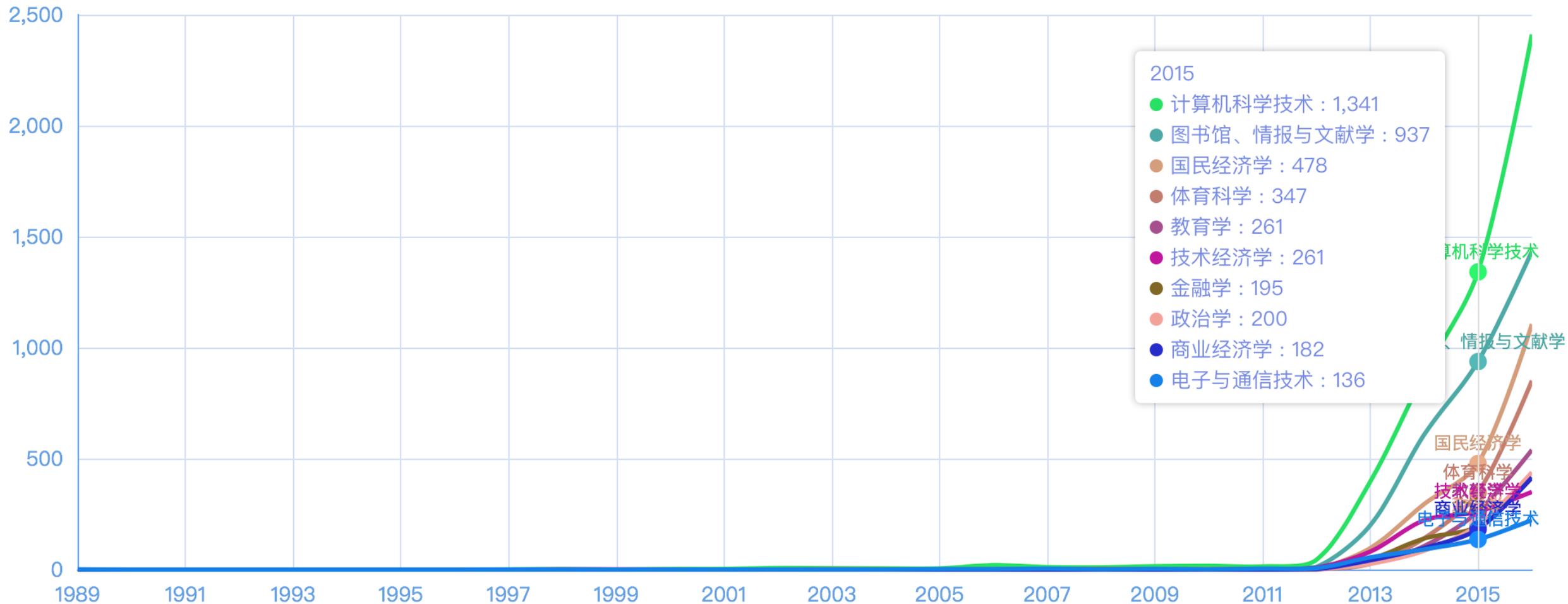
- 图书馆运行状态和数据的图形化展示
- 电子资源绩效评估
- 为学科发展提供数据支撑
- 为图书馆服务提供数据支撑
- 自动化系统数据的深度挖掘与利用



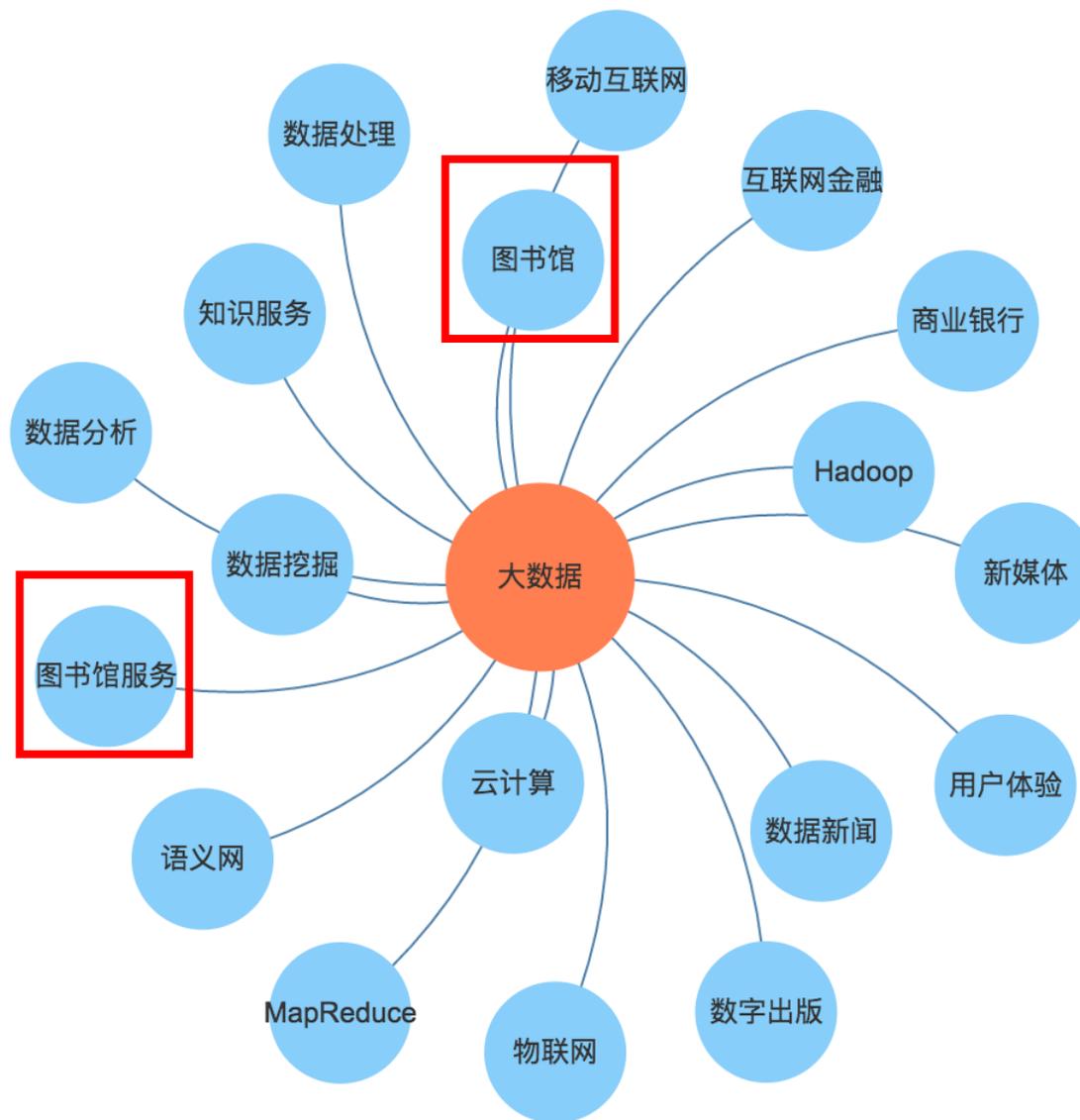
背景与现状

大数据历史热度

● 计算机科学技术 ● 图书馆、情报与文献学 ● 国民经济学 ● 体育科学 ● 教育学 ● 技术经济学 ● 金融学 ● 政治学 ● 商业经济学 ● 电子与通信技术



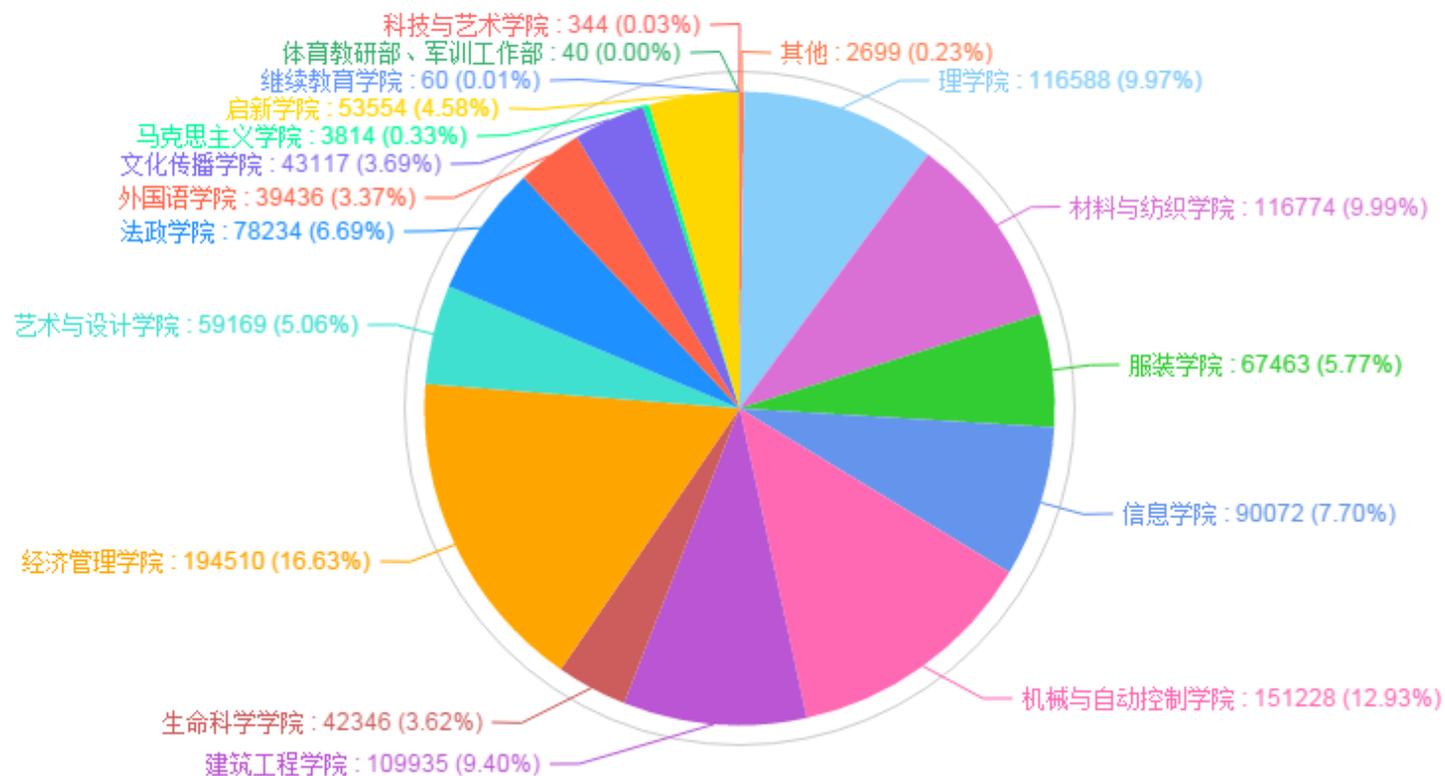
背景与现状



背景与现状

- 其他
- 理学院
- 材料与纺织学院
- 服装学院
- 信息学院
- 机械与自动控制学院
- 建筑工程学院
- 生命科学学院
- 经济管理学院
- 艺术与设计学院
- 法政学院
- 外国语学院
- 文化传播学院
- 马克思主义学院
- 启新学院
- 继续教育学院
- 体育教研部、军训工作部
- 科技与艺术学院

进馆量学院分布 (2015年)

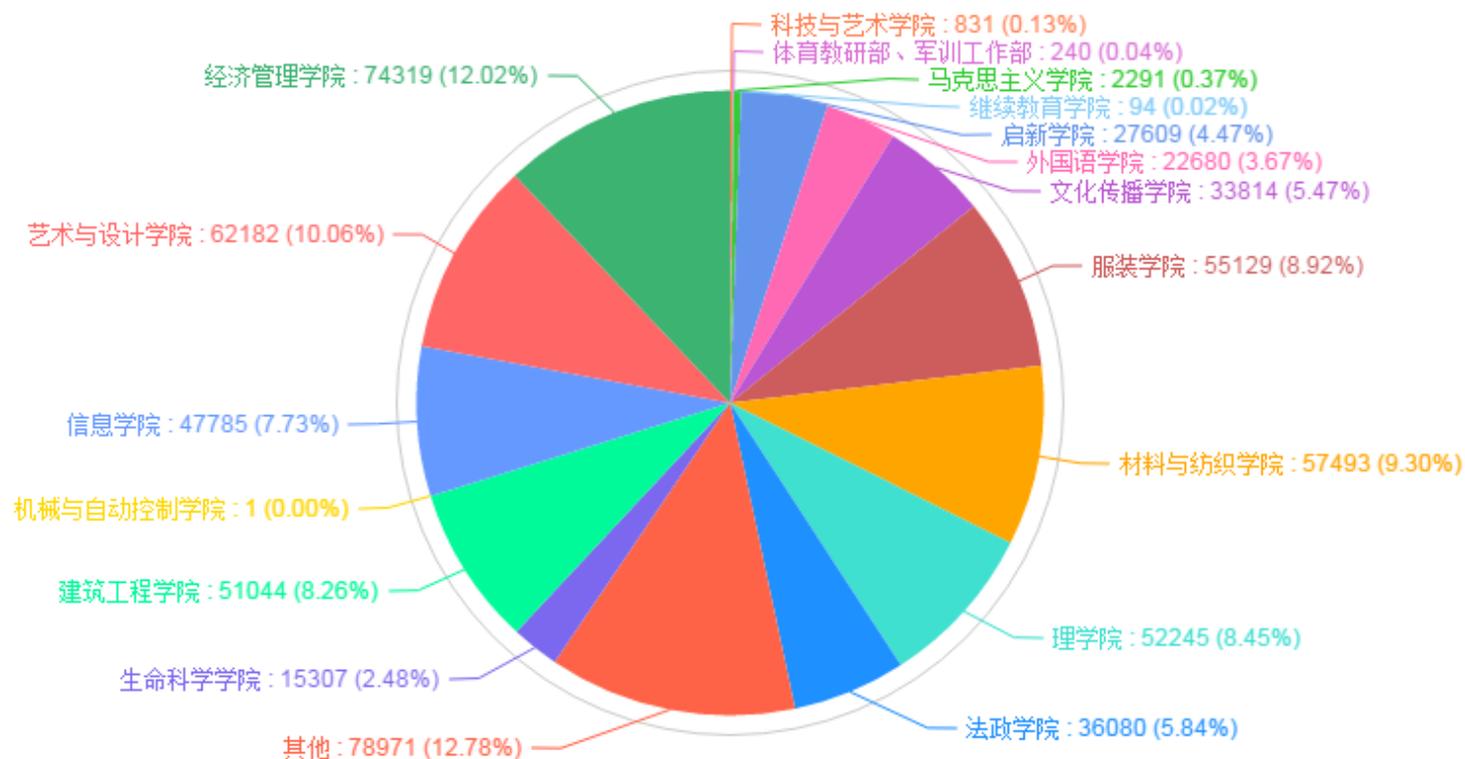


背景与现状

- 科技与艺术学院
- 继续教育学院
- 体育教研部、军训工作部
- 马克思主义学院
- 启新学院
- 外国语学院
- 文化传播学院
- 服装学院
- 材料与纺织学院
- 理学院
- 法政学院
- 其他
- 生命科学学院
- 建筑工程学院
- 机械与自动控制学院
- 信息学院
- 艺术与设计学院
- 经济管理学院

各学院借还量分布

(近三年)



背景与现状

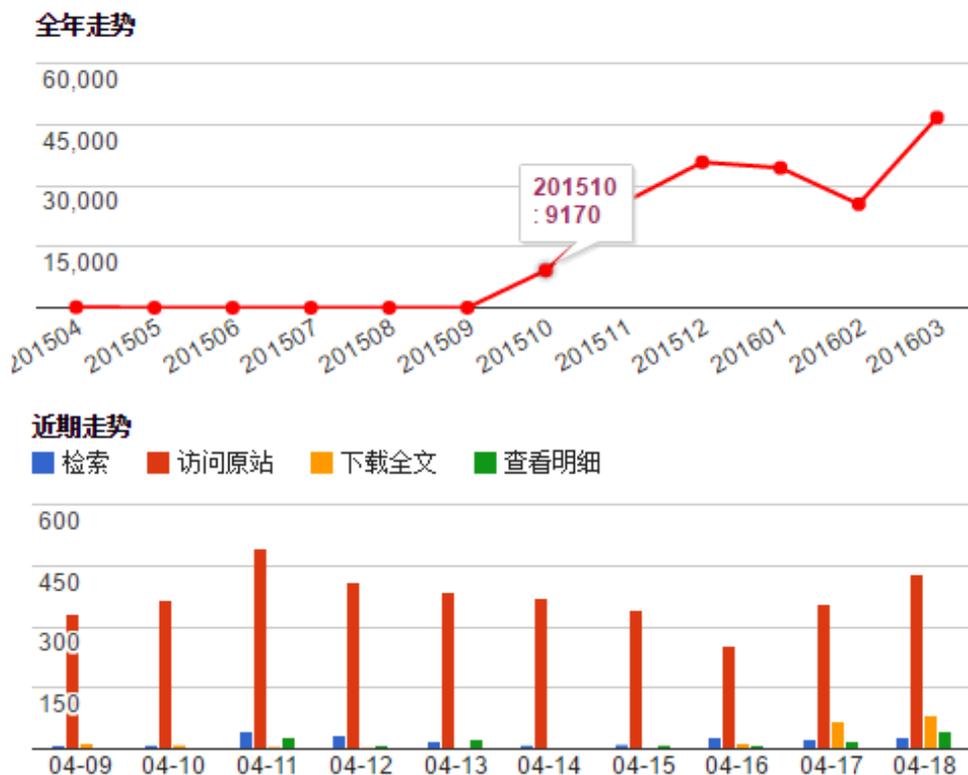
资源导航 | 聚搜索 | 联邦搜索 | 统计排行

资源库名称	访问次数					操作
	总数	检索	访问原站	全文下载	资源明细	
中国知网(CNKI)资源总库	208647	1071	73682	133174	720	统计图表
北大法宝	14773	0	14734	39	0	统计图表
超星数字图书馆	3530	240	3259	5	26	统计图表
Westlaw	3351	51	3205	93	2	统计图表
万方数据知识服务平台	2289	595	944	702	48	统计图表
月旦法学知识库	2287	110	1861	307	9	统计图表
读秀中文学术搜索	2214	633	1297	223	61	统计图表
北大法意网	1902	0	1902	0	0	统计图表
HeinOnline	1766	70	847	838	11	统计图表
LexisNexis专业法律数据库	1737	0	1614	123	0	统计图表
人大复印资料	1394	494	821	0	79	统计图表
JSTOR回溯资源数据库	1227	140	517	553	17	统计图表
西南政法大学学位论文库[本地]	1162	282	551	289	40	统计图表
维普知识发现系统	756	85	421	232	18	统计图表
全国报刊索引-晚清与民国期刊全文数据库	751	421	139	171	20	统计图表
台湾华艺科学库、学位论文数据库	728	0	612	116	0	统计图表
慧科新闻(报纸)数据库	692	0	691	1	0	统计图表
法律家·中国法学多用途教学案例系统	593	0	593	0	0	统计图表

我要咨询



中国知网(CNKI)资源总库统计图



背景与现状



➤ 匹配：了解读者需求



➤ 关联：挖掘读者需求



➤ 创造：预测读者需求

- 2016 Top trends in Academic Libraries

- 1. 研究数据服务(RDS-Research Data Services)
- 2. 数字学术交流(Digital Scholarship)
- 3. 馆藏评估趋势 (Collection assessment trends)
- 4. ILS系统与内容提供商的合并(ILS and content provider /fulfillment mergers)
- 5. 学习支持(Evidence of learning: Student success, learning analytics, credentialing)
- 6. 高等教育信息素养框架新方向(Critical information literacy in the Framework)
- 7. 替代计量(Altmetrics)
- 8. 新创馆员职位(Emerging staff positions)
- 9. 开放教育资源(Open Educational Resources-OER)

设计与实现

2

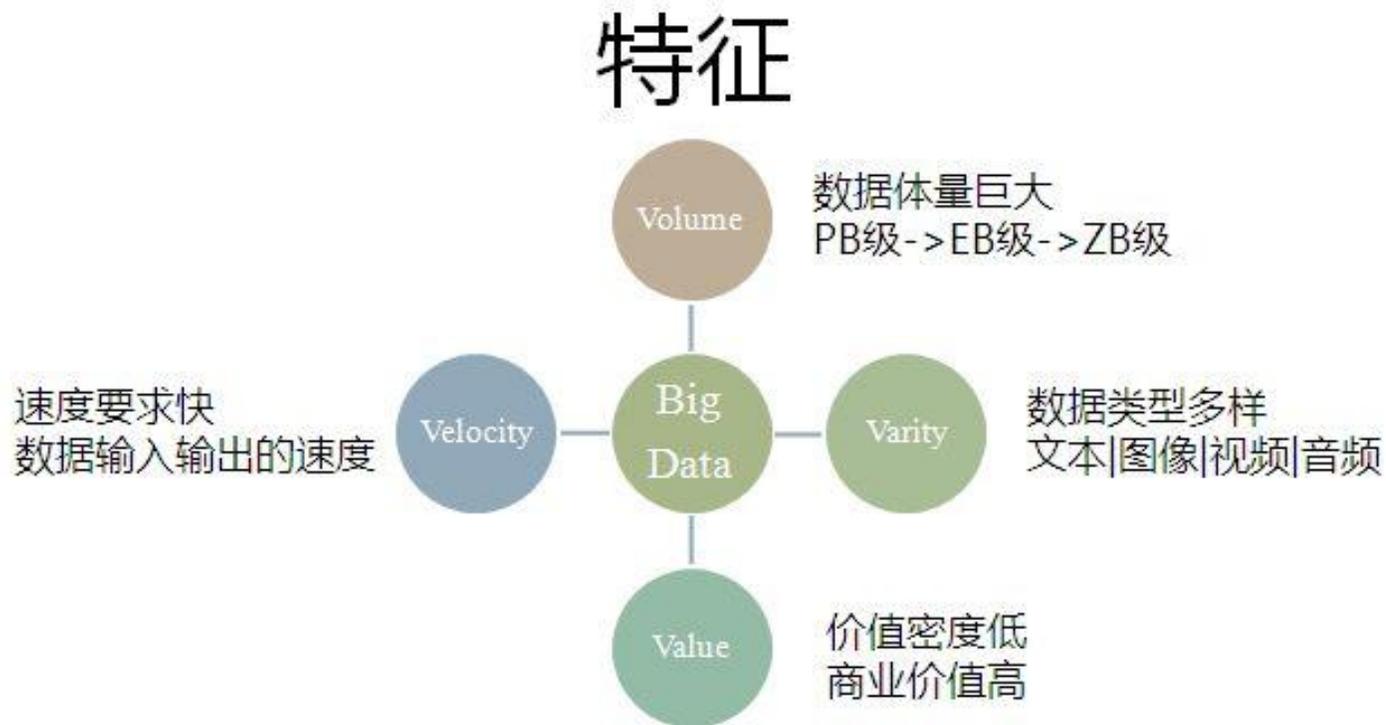
设计与实现

- 核心挑战之一是处理海量的非结构化数据。这种新数据呈现与大海类似的性质，所以称之为**数据海**。



- 应用需求特点

- 海量数据
- 多源数据
- 速度要求快
- 隐含的用户需求



◆ 大数据平台 Hadoop

- HDFS（Hadoop Distributed File System）：
是一个分布式文件系统，用来存储海量数据。特点是：高可靠性、高扩展性和高吞吐；
- Map/Reduce：是一种编程模型，用于大规模数据集的并行计算框架和思想。特点是：易于编程、高容错性和高扩展性；
- 擅长存储大量的半结构化的数据集，数据可以随机存放，一个磁盘的失败不会带来数据丢失。
- 擅长分布式计算——快速地跨多台机器处理大型数据集
- 高可靠性、高可拓展性、高容错性和高效性



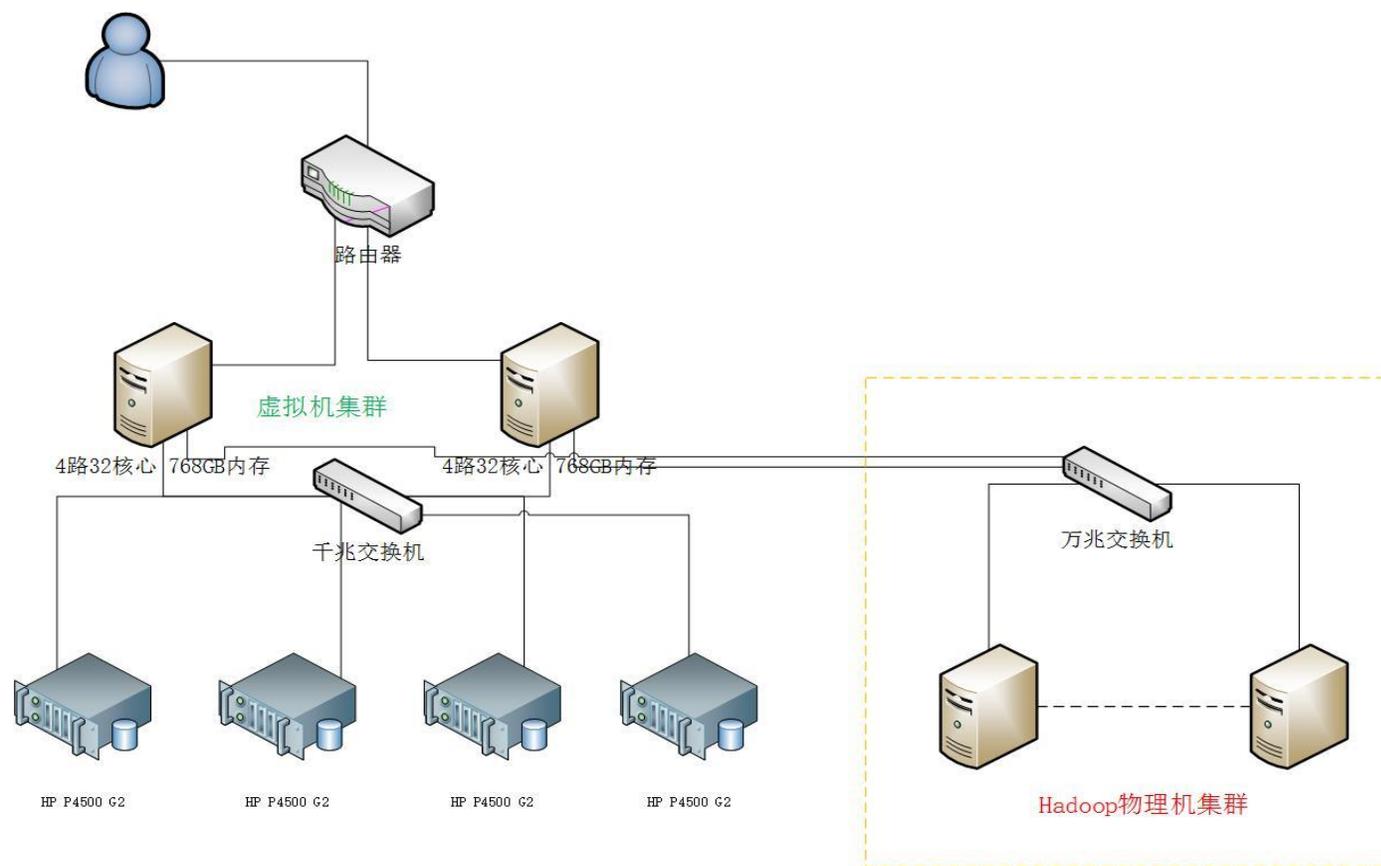
虚拟机集群

1. 时间响应要求不高的计算任务：
如ES集群索引计算。
2. 常规应用：如WEB、爬虫等。
3. 数据仓储：如原始元数据及计算出的结果数据。
4. 辅助物理集群：如部署Hadoop的NameNode节点。

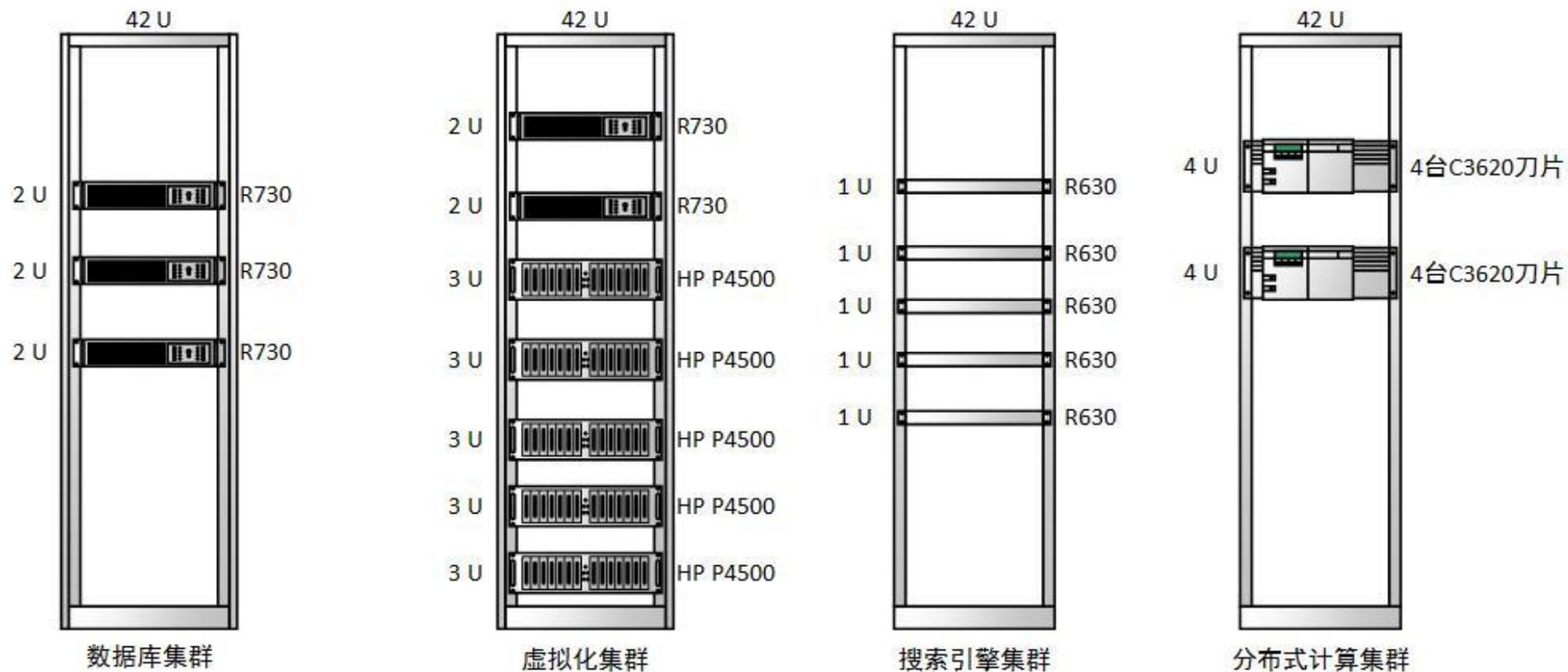
物理机集群

1. Hadoop物理机集群：Slave计算节点，目前物理节点有3个，虚拟节点有4个。

Cloud: Physical & Virtual



设计与实现



数据库集群: Mysql Cluster

虚拟化集群: 2台R730 750GB内存&5台HP P4500存储

搜索引擎集群: 5台Elastic Search集群

分布式计算集群: 8台 Spark集群

- Internet Archive : 6000TB , 20TB/月
- CADAL项目 : 2 , 760 , 000册
- SUMMON : 800 , 000 , 000条
- Scopus: 27,000,000
- 中文期刊篇名 : 64 , 873 , 000条
- 中国学者SCI收录 : 2 , 100 , 000篇
- 中国专利 : 6 , 910 , 000件
- 作者数据 : 7 , 520 , 000
- 机构数据 : 2 , 750 , 000

应用示例

3

应用示例_数据清洗

◆ 问题描述

- 存量数据4000万条 (sci_old)
- 新增数据290万条 (sci_new)
- 要求从新增数据 (sci_new) 中剔除原来已经存在的数据 (sci_old)，把剩余数据添加到存量数据表中。

◆ 问题分析

- 计算方法：将sci_old表中的每行数据和sci_new的所有数据进行关键字字符串的匹配；
- 计算量：需要进行 $290\text{万} \times 4000\text{万} = 1160000$ 亿次匹配，而每次匹配需要对 title、author、year等多个字段进行字符串匹配；
- 分析结论：字符串的匹配是一个相对较大计算量的计算，大约速度 400万次/秒；也就是全部计算完成需要8055个小时完成。

应用示例_数据清洗

◆ 实现思路

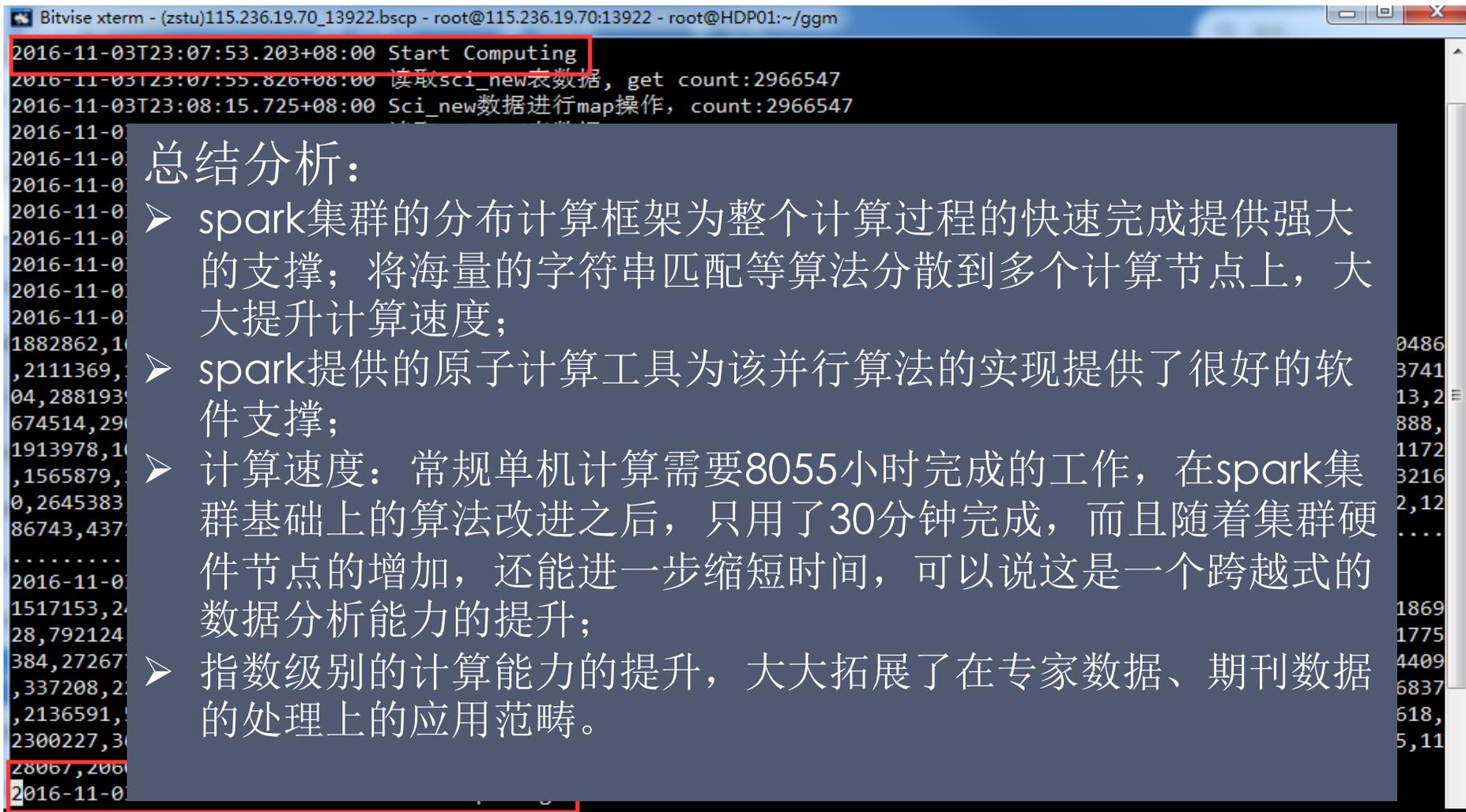
- 改进算法：按特征值进行分组后处理，避免两表乘积造成的海量计算；
- 采用spark集群进行分布式的并行计算；

◆ 实现过程

采用scala语言进行实现并运行在在spark集群上

- 从title字段中取特征值
- 以特征值为key，将sci_old表和sci_new表进行合并，总得到50014433行数据；
- 分组，总得到37913516组。
- 按组进行过滤，剩下同时包含sci_old和sci_new数据的组 148万，对这些组内数据进行复杂字符串的匹配；

应用示例_数据清洗



Bitvise xterm - (zstu)115.236.19.70_13922.bscp - root@115.236.19.70:13922 - root@HDP01:~/ggm

```
2016-11-03T23:07:53.203+08:00 Start Computing
2016-11-03T23:07:55.826+08:00 读取sci_new表数据, get count:2966547
2016-11-03T23:08:15.725+08:00 Sci_new数据进行map操作, count:2966547
```

2016-11-0
2016-11-0
2016-11-0
2016-11-0
2016-11-0
2016-11-0
2016-11-0
2016-11-0
1882862, 1
, 2111369,
04, 288193
674514, 29
1913978, 1
, 1565879,
0, 2645383
86743, 437
.....
2016-11-0
1517153, 2
28, 792124
384, 27267
, 337208, 2
, 2136591,
2300227, 3
28067, 206
2016-11-0

0486
3741
13, 2
888,
1172
3216
2, 12
....
1869
1775
4409
6837
618,
5, 11

总结分析:

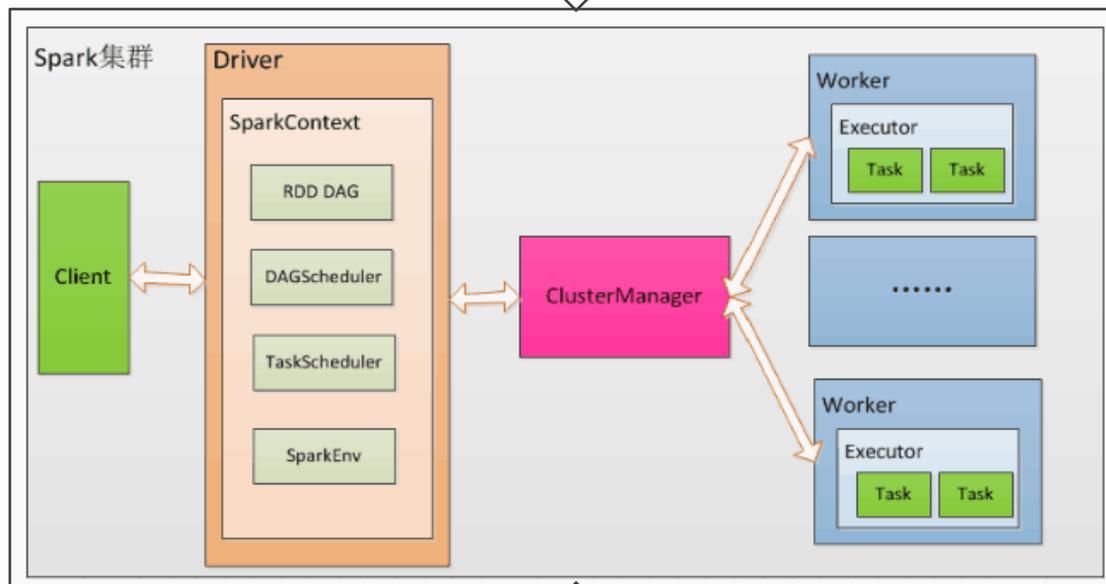
- spark集群的分布计算框架为整个计算过程的快速完成提供强大的支撑; 将海量的字符串匹配等算法分散到多个计算节点上, 大大提升计算速度;
- spark提供的原子计算工具为该并行算法的实现提供了很好的软件支撑;
- 计算速度: 常规单机计算需要8055小时完成的工作, 在spark集群基础上的算法改进之后, 只用了30分钟完成, 而且随着集群硬件节点的增加, 还能进一步缩短时间, 可以说这是一个跨越式的数据分析能力的提升;
- 指数级别的计算能力的提升, 大大拓展了在专家数据、期刊数据的处理上的应用范畴。

应用示例_书目查重

成员馆提交待查重数据



数据格式检查



CADAL基准数据库

应用示例_书目查重

CADAL原始数据存在字段数据不全，录入错误及重复数据问题。通过与国图数据（149万条）、读秀（292万条）、南大（118万条）、北大（96万条）、清华（127万条）及人大（148万条）数据的相似度计算对CADAL的数据进行了补全和纠错，确保基准比较数据准确。

• 采用文本相似度计算方式 —— 提高查重准确度

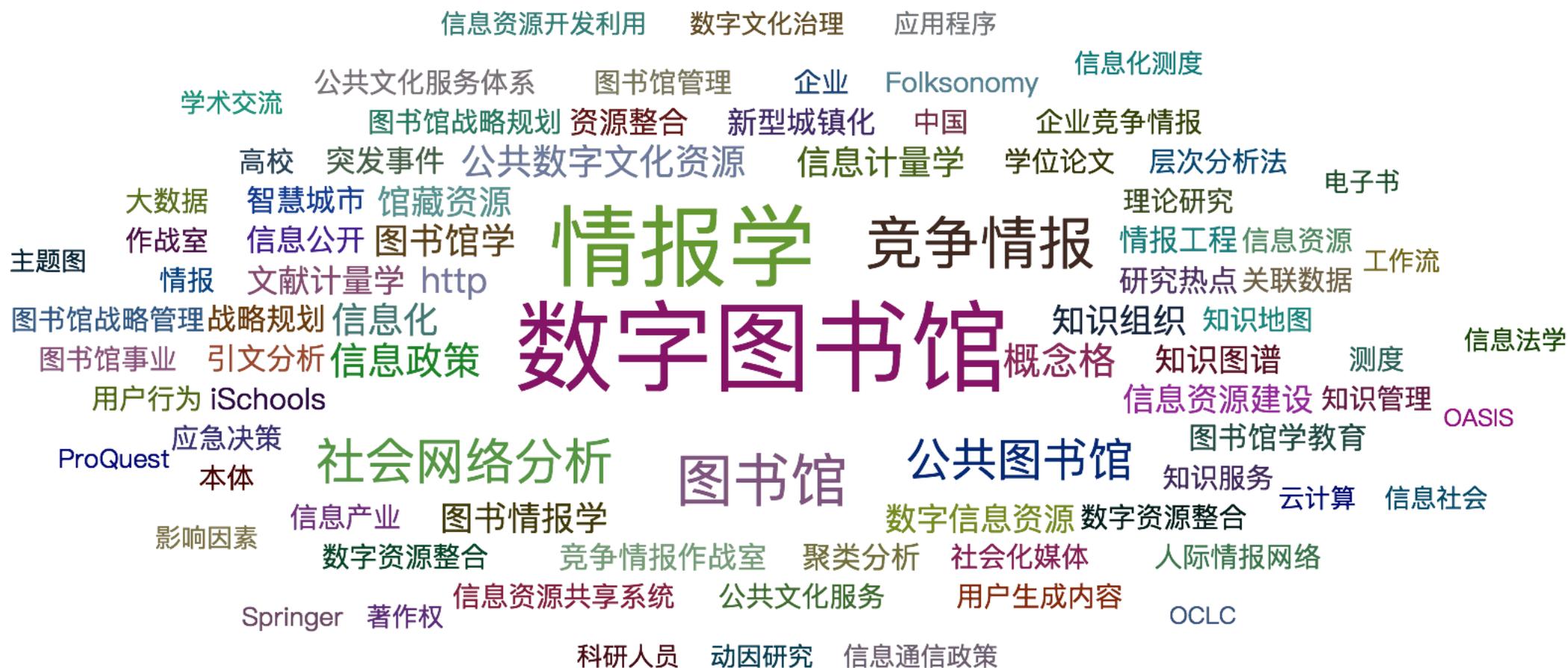
1. 提取每条记录的指纹：包括字段内容及相应权重，权重可以根据不同的情况作调整。
2. 利用相似度算法比较指纹的匹配程度，生成一个0—1的匹配度
两个指纹间的相似度可用下面的公式表示

$$S(P, T) = \frac{|\{p_i | (p_i, t_j) \in R_s\}| + |\{t_j | (p_i, t_j) \in R_s\}|}{|P| + |T|} \times 100$$

采用多种算法（如：DiceCoefficient、LevenshteinDistance、GST等）对指纹进行相似度匹配，结合权重信息，给出最终相似度。根据相似度值确定哪些为相同，哪些需人工确认。

应用示例

图书情报核心期刊关键词top100-云图



应用示例

图书情报核心期刊TOP10作者列表

姓名	文章数量
邱均平	246
王知津	163
毕强	149
朱庆华	143
本刊讯	140
柯平	135
肖希明	130
马海群	130
郑建明	127
李纲	102

大图、中图、情报学报TOP10作者列表

姓名	文章数量
邱均平	45
苏新宁	30
武夷山	29
侯汉清	28
化柏林	26
叶鹰	26
李纲	26
马费成	26
张玉峰	24
叶继元	22

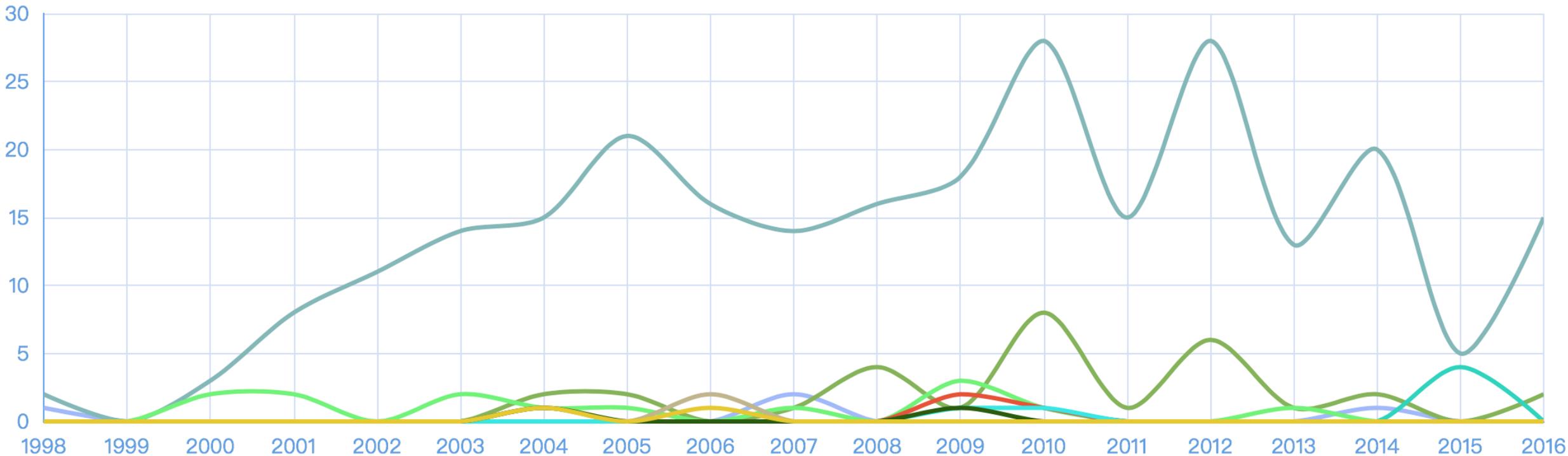
邱均平专家成果关键词云图



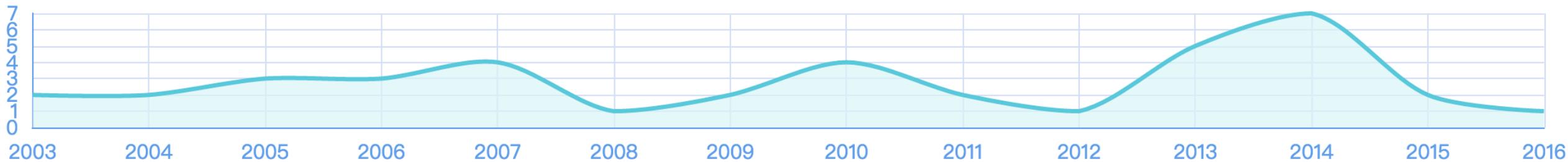
应用示例

大学评价历史热度

● 体育科学 ● 科学学与科技管理 ● 教育学 ● 图书馆、情报与文献学 ● 未知学科 ● 国民经济学 ● 管理学 ● 商业经济学 ● 工业经济学 ● 系统学



邱均平在关键词(大学评价)历年成果折线图

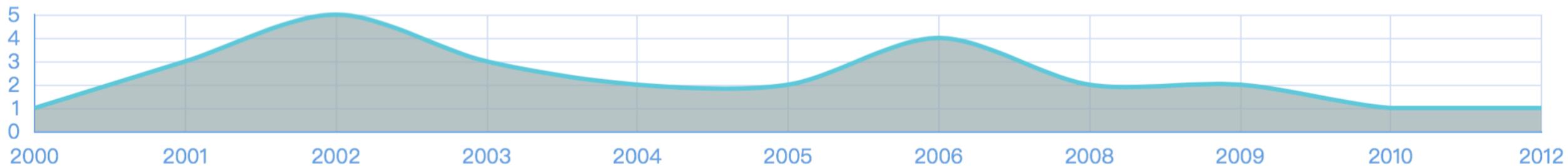


武书连专家成果关键词云图

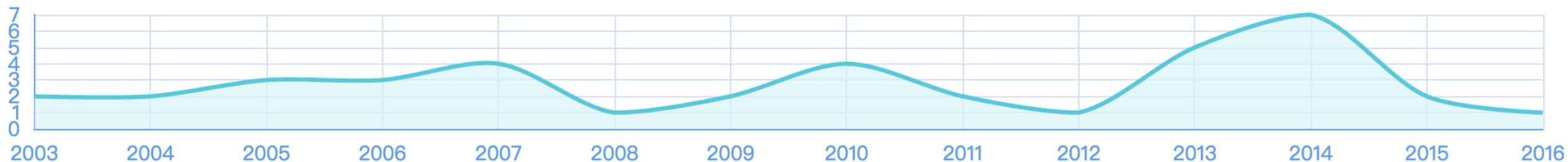


应用示例

武书连在关键词(大学评价)历年成果折线图



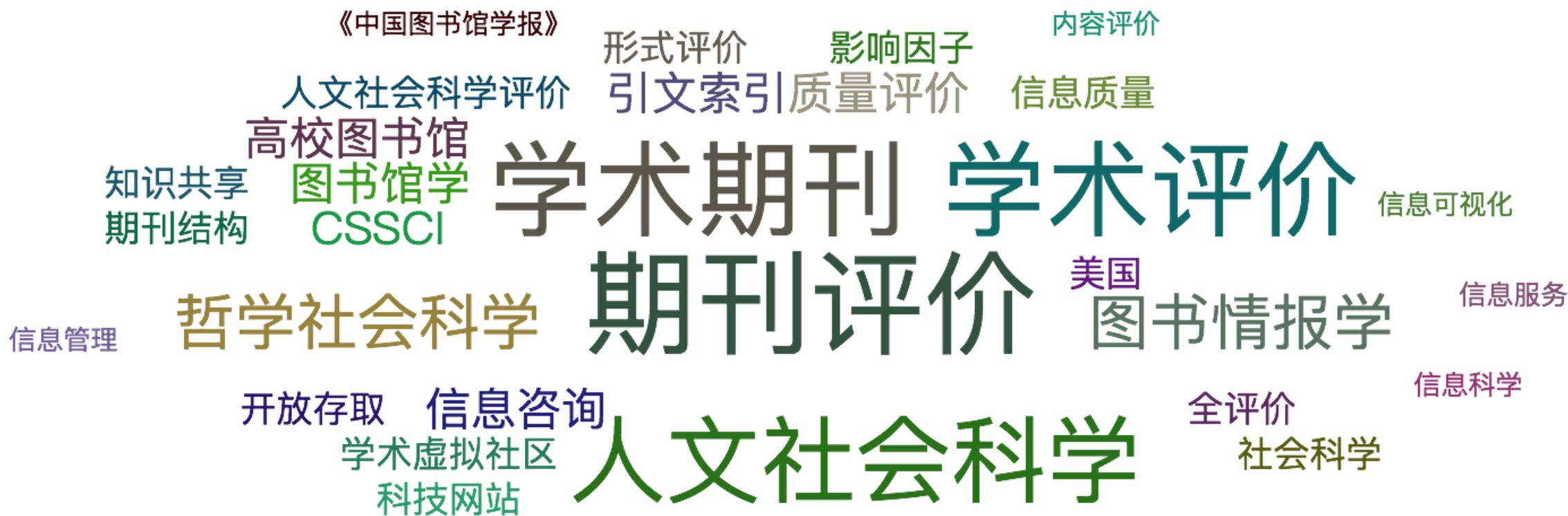
邱均平在关键词(大学评价)历年成果折线图



王知津专家成果关键词云图

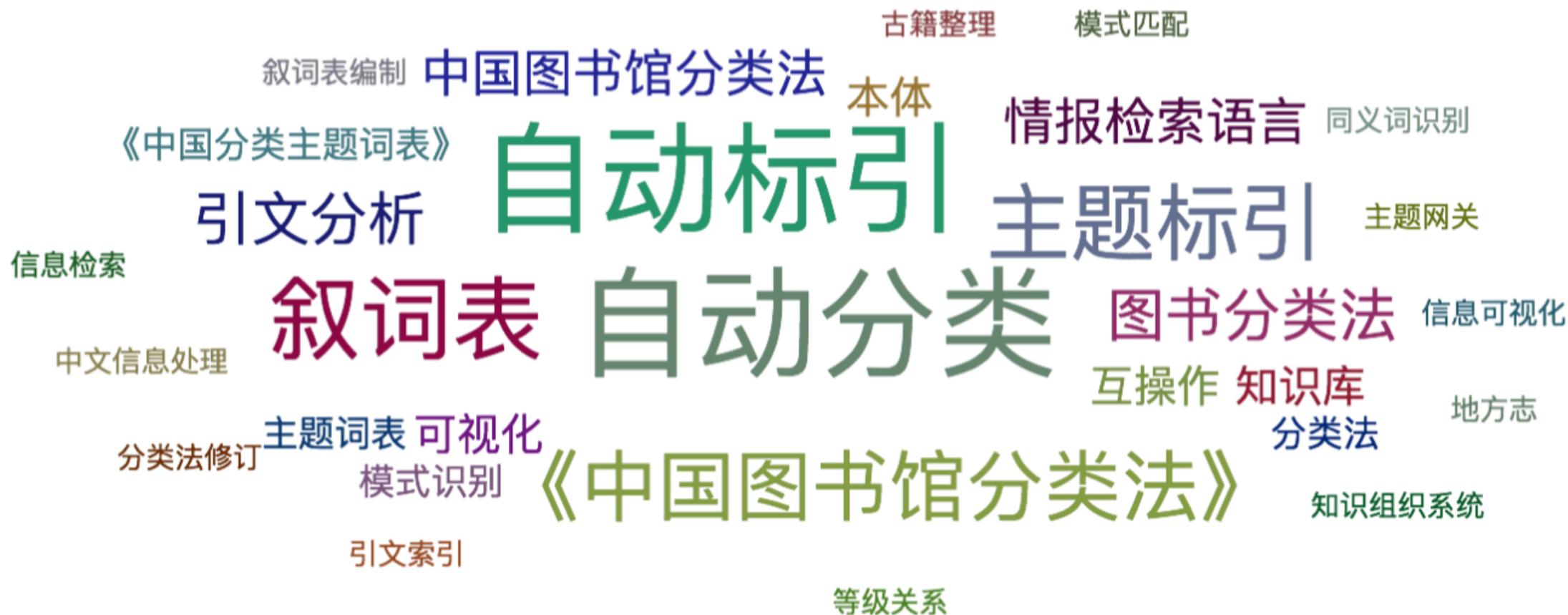


叶继元专家成果关键词云图



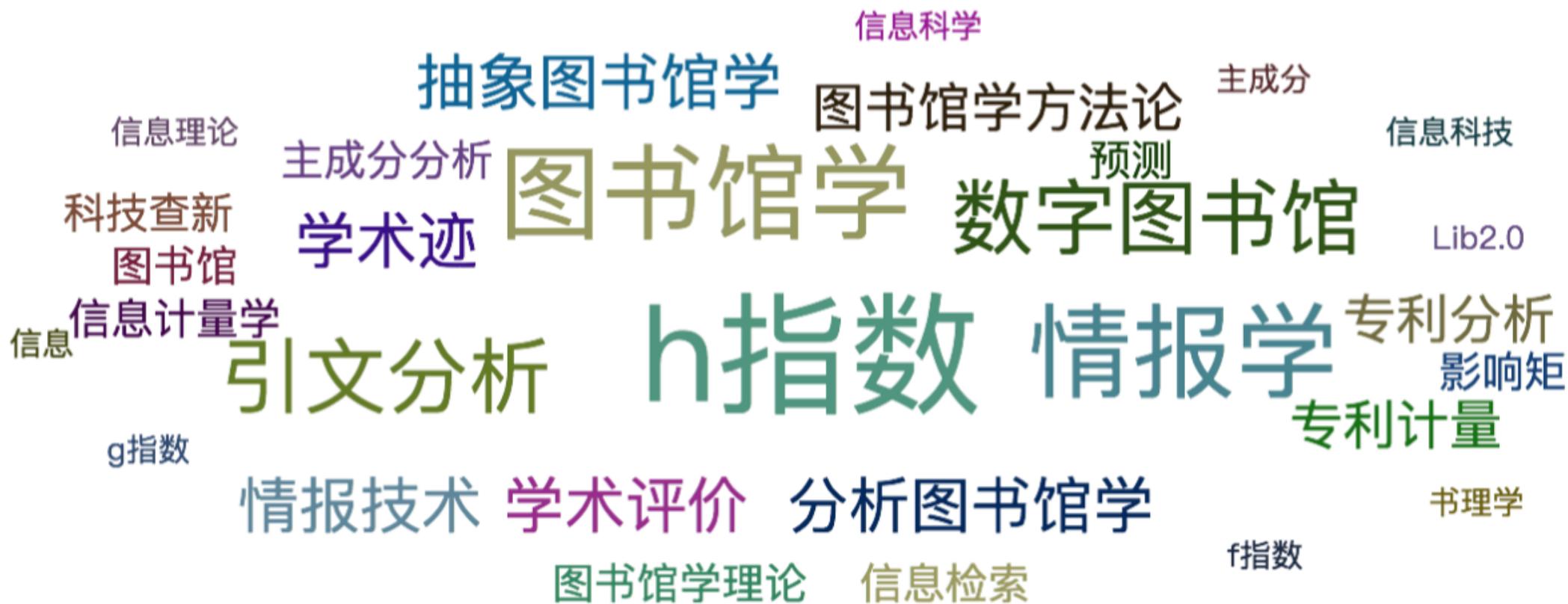
应用示例

侯汉清专家成果关键词云图



应用示例

叶鹰专家成果关键词云图



应用示例



问题及方向

4

问题与展望

- 数据来源 Data Source
- 数据获取 Data Acquisition
- 数据清洗 data cleaning
 - 同名问题
 - 机构规范
 - 学科映射
 -
- 数据挖掘 Data Mining
 - 学者认定 Scholar ID
 - 合作者 **co - author**
 - 工作经历 personal resume
 - 学术图谱 Knowledge Map
- 可视化工具

问题与回答

搜索 cers_unionresources_v41 (114361086 个文档) 的文档, 查询条件:

must estype.creatorStandard.raw term 邱均平

搜索 返回格式: Table 显示数量: 50 显示查询语句

查询 5 个分片中用的 5 个, 826 命中, 耗时 0.003 秒

index	_type	_id	_score	id
cers_unionresources_v41	estype	94FC314A-4D0E-E711-8B70-0050569B7A51	12.8890295	94FC314A-4D0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	4C70913D-4C0E-E711-8B70-0050569B7A51	12.8890295	4C70913D-4C0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	95822F6A-4F0E-E711-8B70-0050569B7A51	12.8890295	95822F6A-4F0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	69AB345D-4F0E-E711-8B70-0050569B7A51	12.8890295	69AB345D-4F0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	FBB90C4E-520E-E711-8B70-0050569B7A51	12.8890295	FBB90C4E-520E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	FB57AC6B-570E-E711-8B70-0050569B7A51	12.8890295	FB57AC6B-570E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	D28B1298-570E-E711-8B70-0050569B7A51	12.8890295	D28B1298-570E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	05B59DC3-560E-E711-8B70-0050569B7A51	12.8890295	05B59DC3-560E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	1B271E53-520E-E711-8B70-0050569B7A51	12.8890295	1B271E53-520E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	1B0B3206-590E-E711-8B70-0050569B7A51	12.8890295	1B0B3206-590E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	9A9FC729-590E-E711-8B70-0050569B7A51	12.8890295	9A9FC729-590E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	BACA2824-600E-E711-8B70-0050569B7A51	12.8890295	BACA2824-600E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	DD5D3B4E-5A0E-E711-8B70-0050569B7A51	12.8890295	DD5D3B4E-5A0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	E34EAC75-590E-E711-8B70-0050569B7A51	12.8890295	E34EAC75-590E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	DD74211F-6F31-4FDF-837A-411C650BE00C	12.8890295	DD74211F-6F31-4FDF-837A-411C650BE00C
cers_unionresources_v41	estype	7580DD48-4E0E-E711-8B70-0050569B7A51	12.8578615	7580DD48-4E0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	0E8084D8-4D0E-E711-8B70-0050569B7A51	12.8578615	0E8084D8-4D0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	16D1C612-4C0E-E711-8B70-0050569B7A51	12.8578615	16D1C612-4C0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	BCC535A9-4E0E-E711-8B70-0050569B7A51	12.8578615	BCC535A9-4E0E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	DAF6AEE7-520E-E711-8B70-0050569B7A51	12.8578615	DAF6AEE7-520E-E711-8B70-0050569B7A51
cers_unionresources_v41	estype	0D1C9C48-4F0E-E711-8B70-0050569B7A51	12.8578615	0D1C9C48-4F0E-E711-8B70-0050569B7A51

"key": "武汉大学信息管理学院",
"doc_count": 180

评价研究中心",

研究中心",

学院",

研究中心",

息学院",

中心",

管理系",

系",

院信息学系",

问题与展望

邱均平专家成果关键词云图



邱均平专家成果关键词云图



数据可视化

邱均平成果关键词历年分布图



THANK YOU