



图书馆大数据与大数据图书馆

Big Data in Library and Big Data Library

黄晨 浙江大学图书馆

2017@ 湖州

CONTENTS

01 背景与现状 BACKGROUND

02 设计与实现 TECHNICAL ISSUE

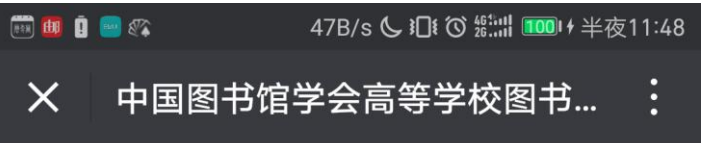
03 案例分享 CASE STUDY

04 合作展望 COOPERATION

背景与现状 BACKGROUND

1

背景与现状



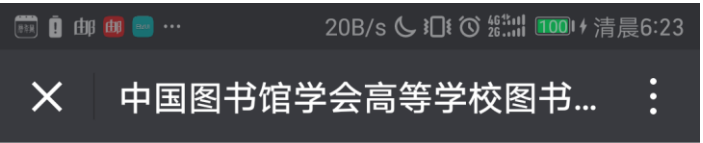
大学之问：图书馆该去哪儿？ | 案例

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，而面对图书馆的认知也构建了我们



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

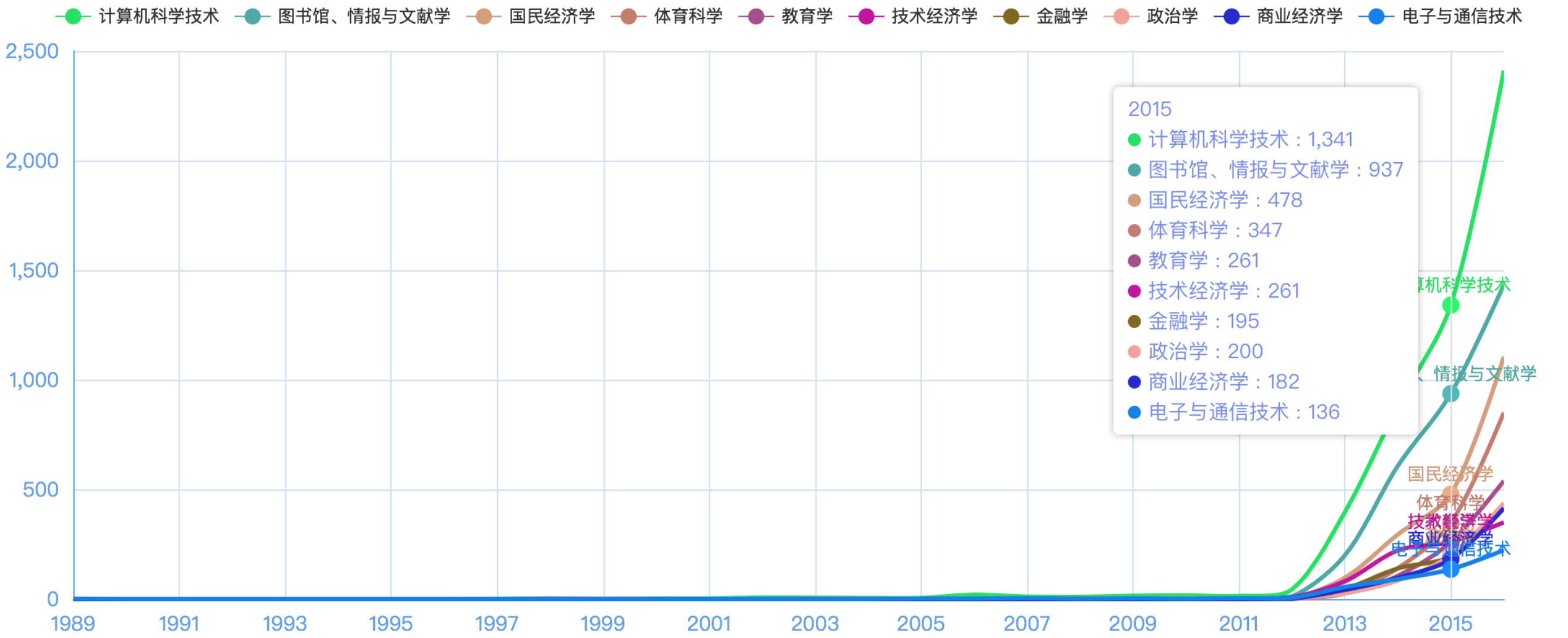
关注麦可思研究，订阅高教管理宝典
阅读推荐：校园规划

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

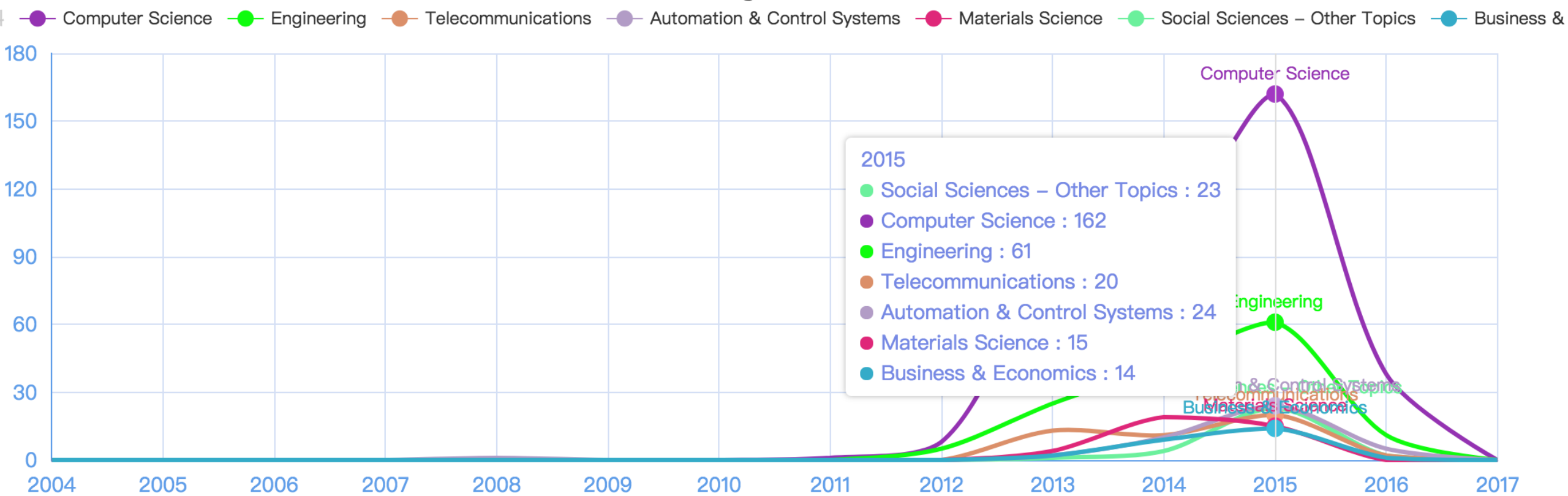
背景与现状

大数据历史热度

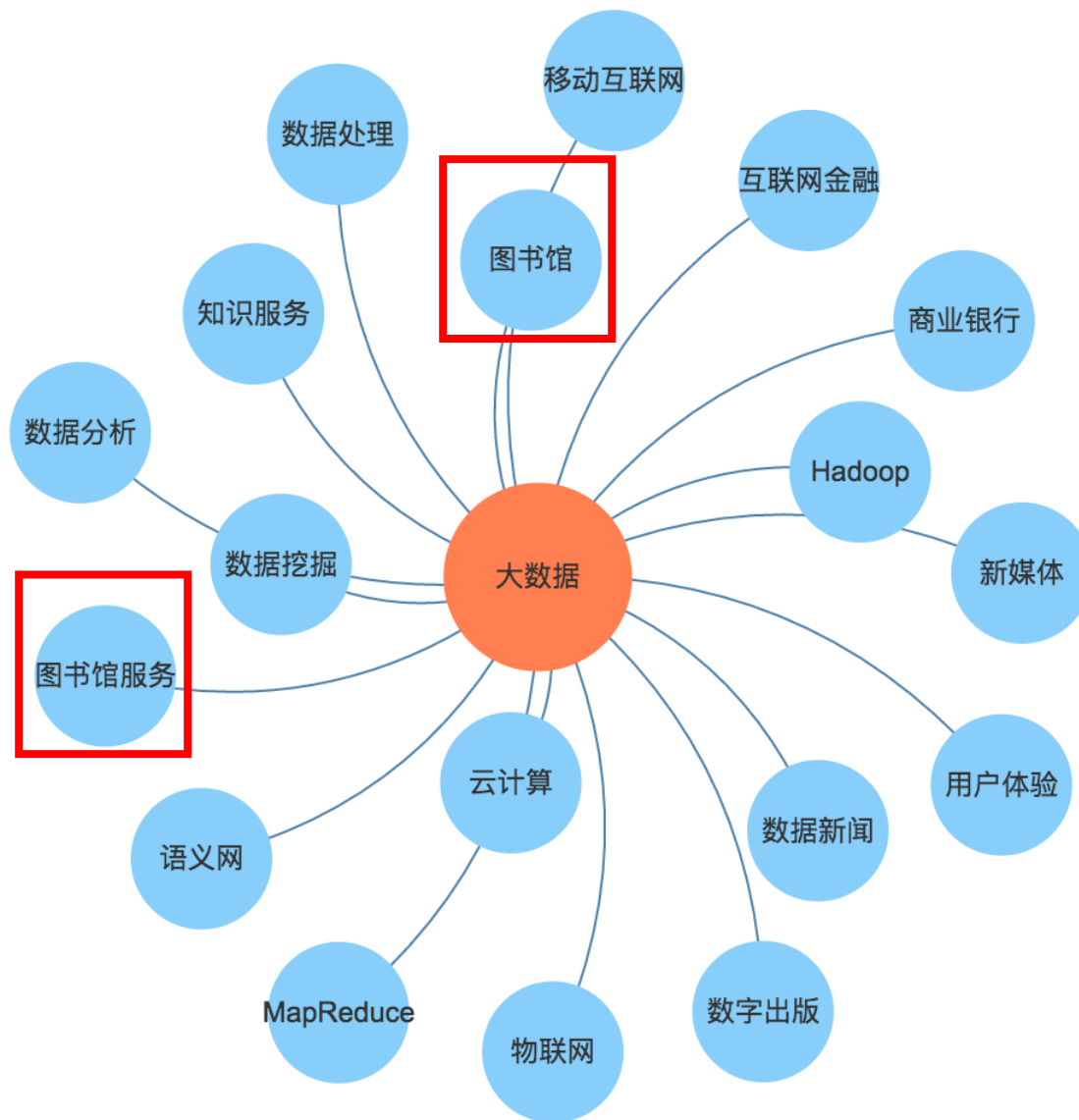


背景与现状

大数据(big data)历史热度



背景与现状



背景与现状



背景与现状

Keywords anywhere	contains		AND	Resource Type:	Any resource type
Title	contains		AND	Language:	Any language
Author / Creator	contains		AND	Publication Date:	Any year
Subject	contains		AND	Start Date:	Day Month Year
				End Date:	Day Month Year
				Location:	All locations (HOLLIS)

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



红楼梦



学术搜索

找到约 216,000 条结果（用时0.03秒）

我的著作引用情况

文章

我的图书馆

时间不限

2017以来

2016以来

2013以来

自定义范围...

[引用] **红楼梦**评论

王国维 - 2005 - 上海古籍出版社

被引用次数: 424 相关文章 引用 保存 更多

[引用] **红楼梦**人物论

王昆仑 - 2004 - 北京出版社

被引用次数: 284 相关文章 引用 保存 更多

[引用] **红楼梦**的两个世界

余英时 - 2006 - 上海社会科学院出版社

被引用次数: 136 相关文章 引用 保存 更多

《**红楼梦**》第22回回末考察——与梅节兄讨论

蔡义江 - 红楼梦学刊, 2001 - cqvip.com

. 我的维普: 帐户余额; 充值记录; 下载记录; 我的收藏. 购物车; 充值; 客服. 首页 | 期刊大全 | 论文选题 | 论文检测 | 期刊投稿 | 在线出版 | 优先出版 | 维普官方天猫店 | 云发表. 您的位置: 网站首页 >

《中文科技期刊数据库》> 人文社科 > 文学 > 中国文学 > 摘要 《**红楼梦**》第22回回末考察——

被引用次数: 82 相关文章 引用 保存 更多

不限语言

中文网页

简体中文网页

[引用] **红楼梦**资料汇编

朱一玄 - 2001 - 南开大学出版社

被引用次数: 150 相关文章 引用 保存

☒ 包括专利

☒ 包含引用

难忘的历程——《**红楼梦**》英译事业的描写性研究

陈宏薇, 江帆 - 中国翻译, 2003 - cqvip.com

背景与现状



红楼梦

检索

高级检索

☐ 在限定条件下重新检索 ☐ 在结果中检索 ☒ 重新检索

精炼检索

取消所选分面

- ☐ 只检索学术文章
- ☐ 只检索被引资源
- ☐ 不包含报纸文章
- ☐ 只检索馆藏电子资源

内容类型

分布图

- ☒ 全部
- ☐ 图书(8015)
- ☐ 期刊(53590)
- ☐ 报纸(19310)
- ☐ + 学位论文(2884)
- ☐ 会议论文(906)
- ☐ 专利(109)

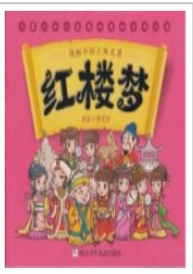
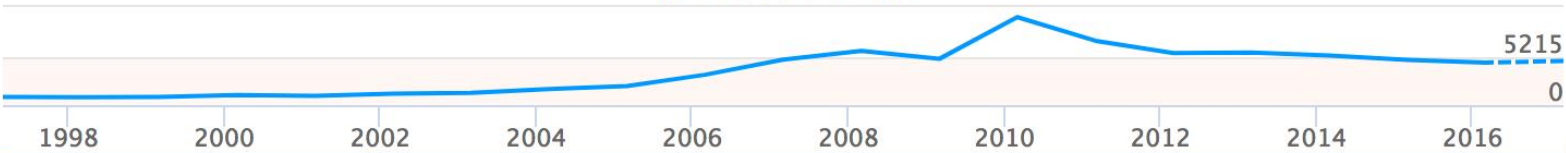
确定

检索 红楼梦 返回 90,191 结果。总被引频次: 4926711次

排序: 默认排序



相关论著发文量趋势图



保存题录

[图书] 红楼梦 引证(3894)

收藏

分享到

作者: 曹雪芹著

出处: 杭州: 浙江少年儿童出版社 2013 183页

ISBN: 978-7-5342-7182-3

主题词: 漫画-连环画-作品-中国-现代

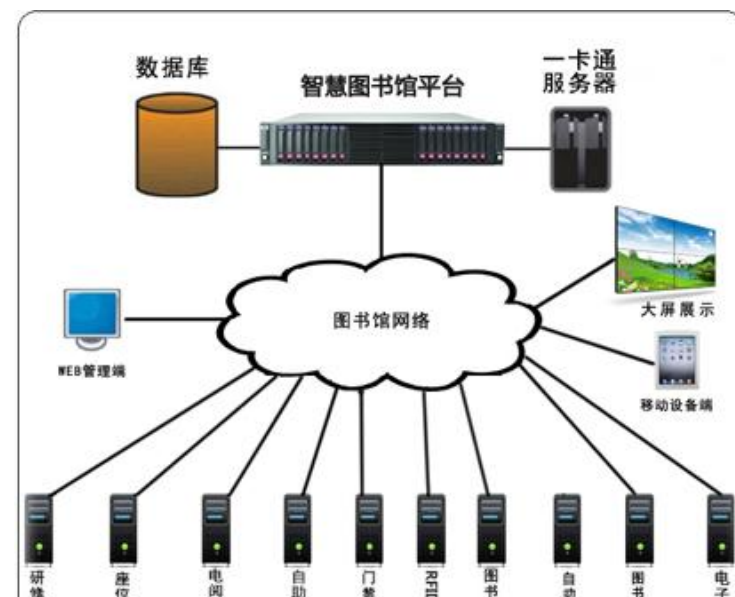
摘要: 《漫画中国古典名著: 红楼梦》精选了原著里最精彩、最生动的故事, 特邀国内一批优秀的插画家绘制精美的漫画, 为小朋友展开一次最奇妙的纸上阅读之旅, 让古典名著变得可亲、可读、可感、可赏: 做一场荣衰无常的红楼情梦...

获得途径: 图书试读 图书馆文献传递

背景与现状



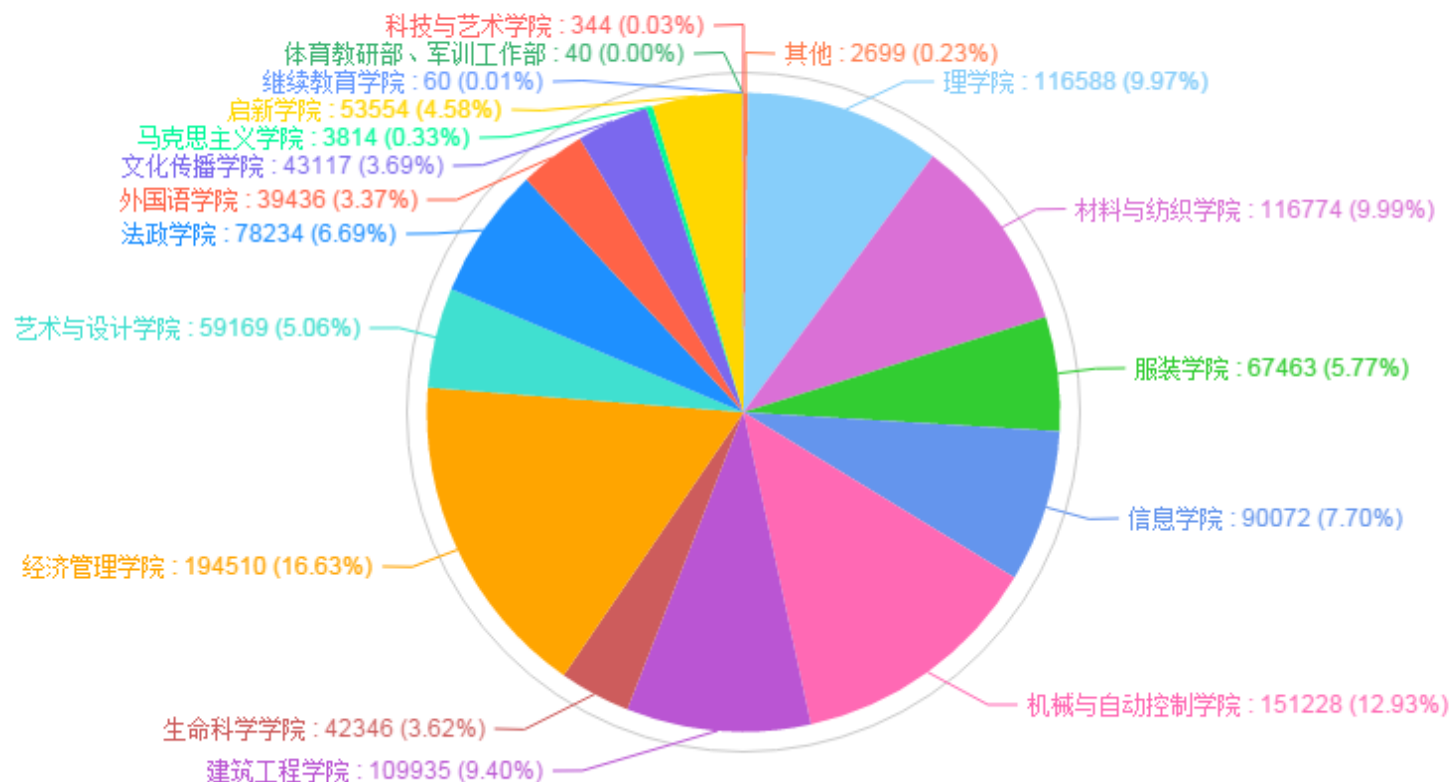
- 图书馆运行状态和数据的图形化展示
- 电子资源绩效评估
- 为学科发展提供数据支撑
- 为图书馆服务提供数据支撑
- 自动化系统数据的深度挖掘与利用



背景与现状

- 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

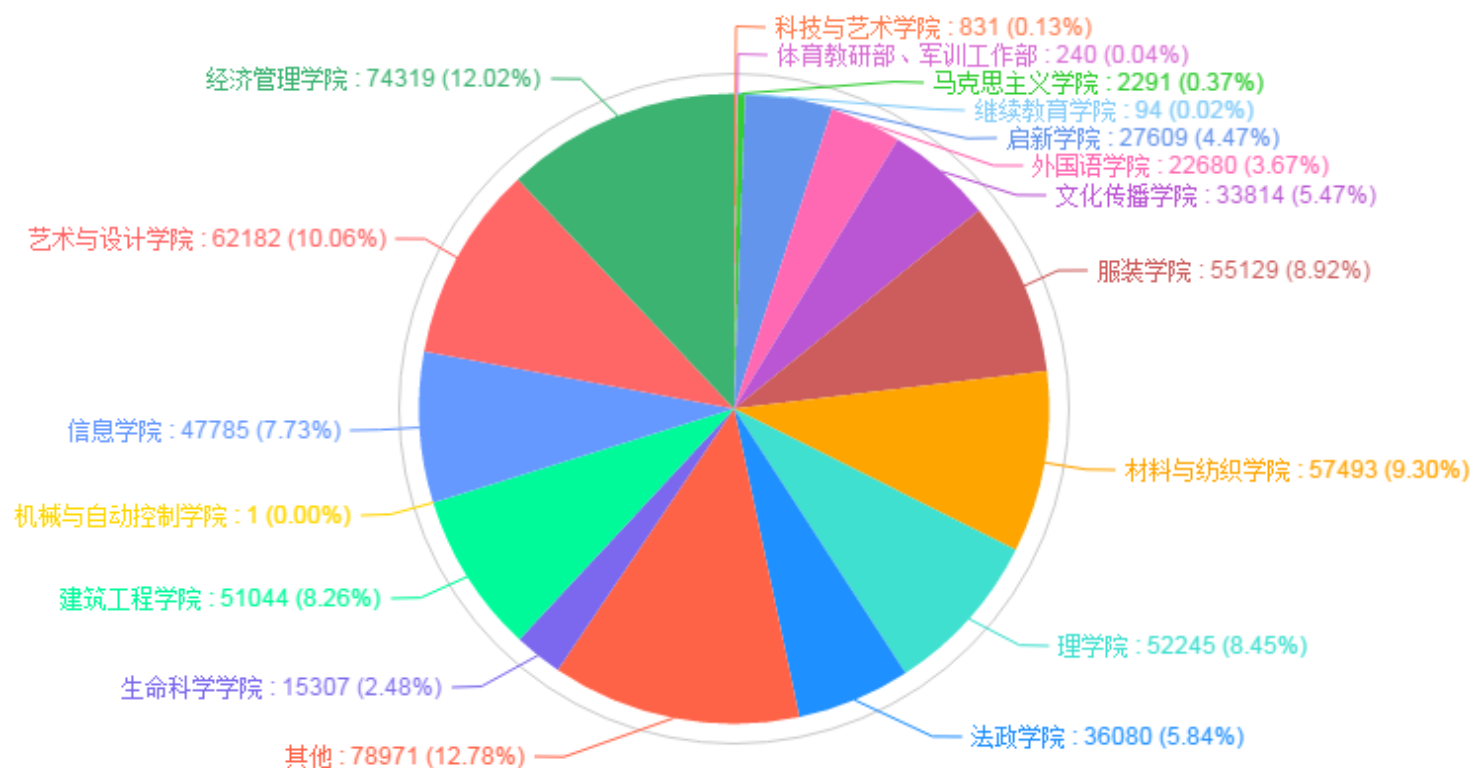
进馆量学院分布（2015年）



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

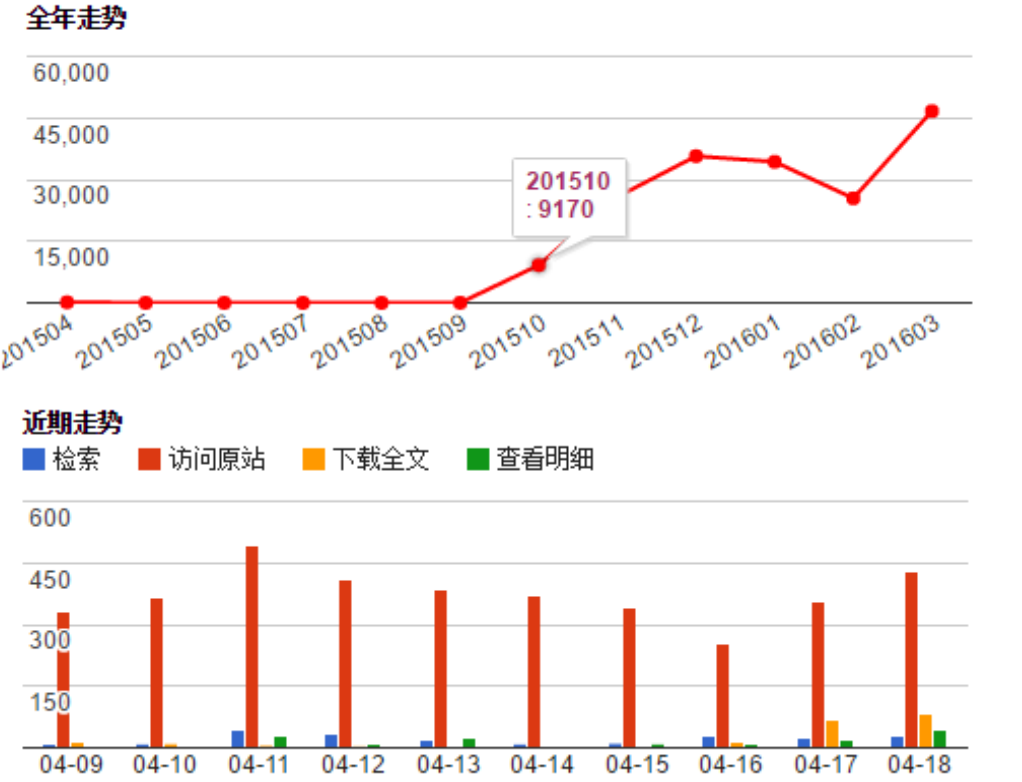
资源导航 | 聚搜索 | 联邦搜索 | 统计排行

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

SciVal

Configure, visualize and export information according to your personal needs through SciVal's integrated modular platform:



Overview

Get a high-level overview of the research performance of your Institution, other Institutions, Countries and Groups of Researchers.



Benchmarking

Compare and benchmark your Institution to other Institutions, Researchers and Groups of Researchers using a variety of metrics.



Collaboration

Explore the collaboration network of both your Institution and other Institutions.



Trends

Get the current scientific trends to determine a new research strategy, find collaboration opportunities and rising stars.

Don't have access to SciVal? Complete the consultation [request form](#) to be contacted by your local account manager.



[主页](#) [关于我们](#) [产品与服务](#) [新闻中心](#) [技术支持](#) [联系我们](#)

[汤森路透知识产权与科技](#) [科技信息产品与服务](#) [InCites](#)

InCites

[学术研究，出版与分析](#)

[知识产权解决方案](#)

[生命科学信息，服务与技术](#)

[科技信息产品与服务](#)

InCites数据库中集合了近30年来Web of Science核心合集七大索引数据库的数据，拥有多元化的指标和丰富的可视化效果，可以辅助科研管理人员更高效地制定战略决策。

基于Web of Science核心合集七大索引数据库30多年客观、权威的数据，InCites数据库中可以提供：

- 涵盖全球5000多所名称规范化的机构信息
- 囊括30多年所有文献的题录和指标信息
- 更丰富、更成熟的引文指标
- 包含了基于中国国务院学位委员会和教育部《学位授予 和人才培养学科目录（2011年）》的学科分类

您可以利用InCites数据库：

- 定位重点学科/ 优势学科，发展潜力学科，优化学科布局

INCITES平台产品

[Essential Science Indicators](#)

[Journal Citation Reports](#)

市场资料

[新一代InCites平台简介](#)

[InCites 数据库快速使用指南](#)

更多信息

[新一代InCites平台及InCites数据库培训](#)



New: Patent Market Tracker Mid-Year Report

Twice yearly, Innography publishes its Patent Market Tracker™ (PMT) report, an analysis of the key trends and market conditions driving the patent marketplace. The Mid-Year 2017 PMT update covers US patent purchases and sales for the first six months of 2017.

[Read the Report >>](#)

背景与现状

2017

浙 江 大 学 SCI论文十年回顾与展望

SCI publications of Zhejiang University during 2006-2015:
Retrospect and Prospect



浙江大学图书馆
浙江大学信息资源分析与利用研究中心

2017

浙 江 大 学 ESI学科排名分析及发展预测

Analysis and Prediction of ESI Discipline Ranking
in Zhejiang University



浙江大学图书馆
浙江大学信息资源分析与利用研究中心

2017

浙江大学学术论文竞争力分析

电气工程



2017

浙江大学学术论文竞争力分析

电力系统

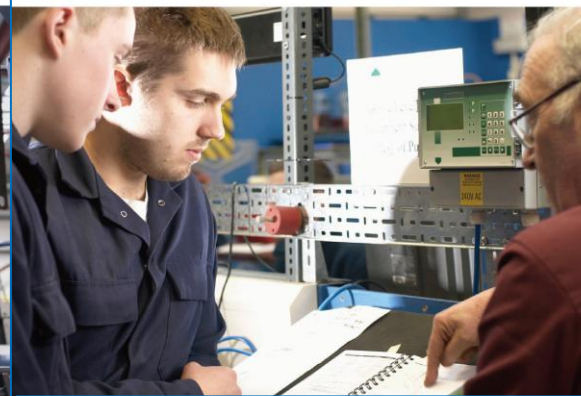


浙江大学图书馆
浙江大学信息资源分析与利用研究中心
2017年7月

2017

浙江大学学术论文竞争力分析

控制学科



2017

浙江大学学术论文竞争力分析

电机学科



浙江大学图书馆
浙江大学信息资源分析与利用研究中心
2017年7月

背景与现状



情报 服务 产品

1

【战略情报】

28份
战略咨询报告

2

【学科情报】

40+学科 100+份
专业分析报告

3

【人才评估】

8个学院
200+人次

4

【科研情报】

30+份
专利分析报告

- 图书馆大数据
 - 生产大数据
 - 分析大数据
- ⑩ 大数据图书馆
 - 大数据收藏
 - 大数据服务

设计与实现 TECHNICAL ISSUE

2

- 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**。

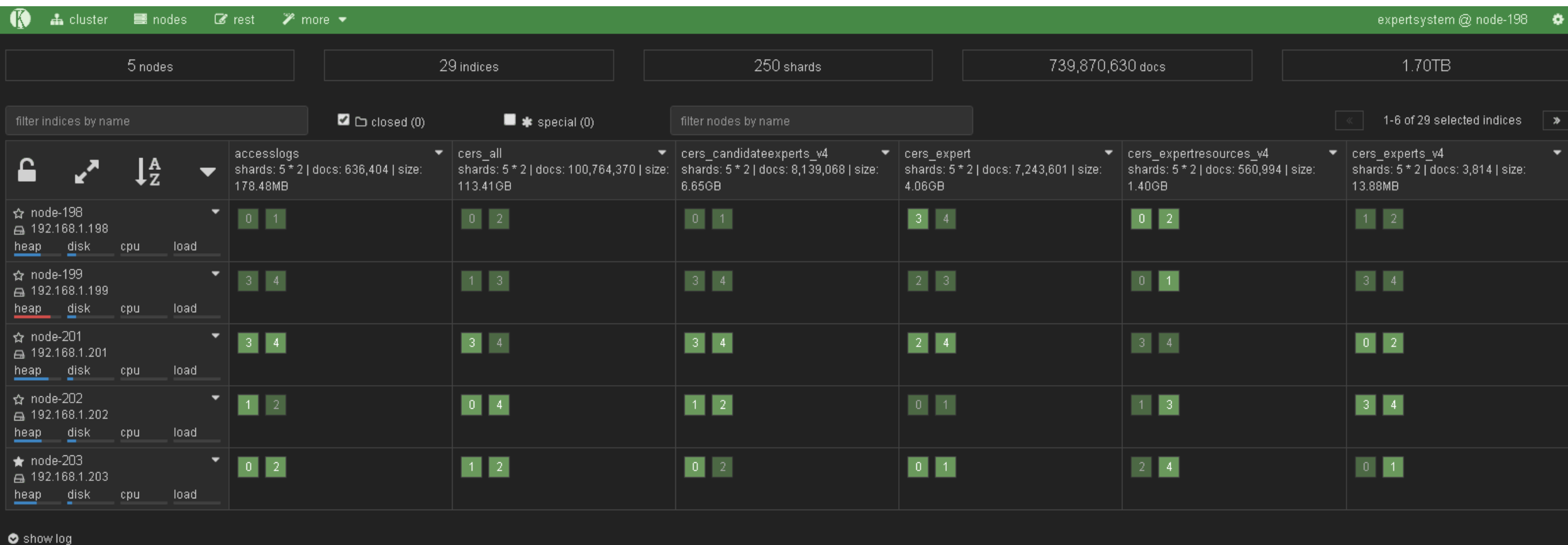


◆ 大数据平台 Hadoop

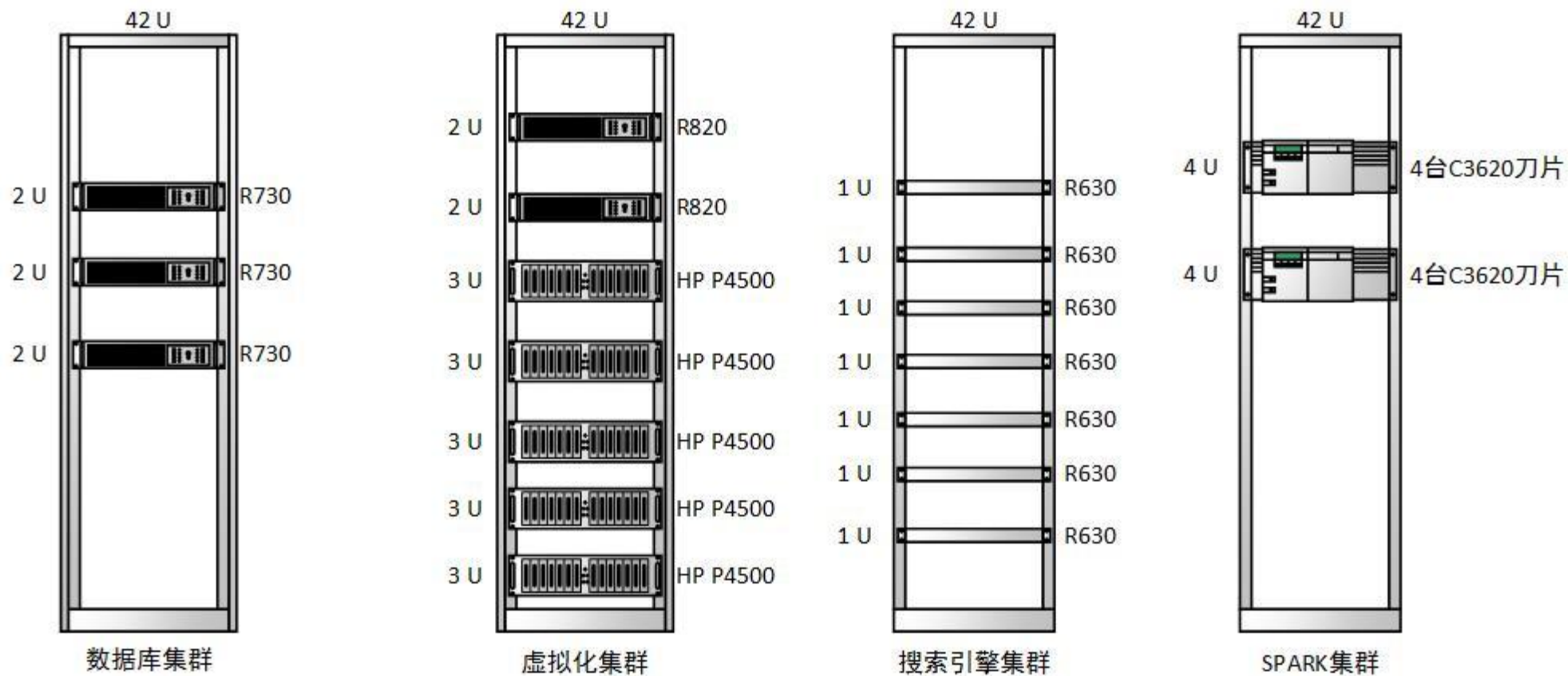
- HDFS（Hadoop Distributed File System）：
是一个分布式文件系统，用来存储海量数据。特点是：高可靠性、高扩展性和高吞吐；
- Map/Reduce：是一种编程模型，用于大规模数据集的并行计算框架和思想。特点是：易于编程、高容错性和高扩展性；
- 擅长存储大量的半结构化的数据集，数据可以随机存放，一个磁盘的失败不会带来数据丢失。
- 擅长分布式计算——快速地跨多台机器处理大型数据集
- 高可靠性、高可拓展性、高容错性和高效性



◆ 分布式搜索引擎 Elastic Search



设计与实现



数据库集群：3台R730 Mysql Cluster

虚拟化集群：2台R820 750GB内存&5台HP P4500存储

搜索引擎集群：6台Elastic Search集群

分布式计算集群：8台 Spark集群

功能效果

◆ 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。

功能效果

◆ 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50,014,433行数据；
- 分组，总得到37,913,516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；

功能效果

A screenshot of a terminal window titled "Bitvise xterm - (zstu)115.236.19.70_13922.bscp - root@115.236.19.70:13922 - root@HDP01:~/ggm". The terminal shows Spark job logs with timestamps and counts. Overlaid on the terminal is a semi-transparent blue box containing the text "总结分析:" followed by three bullet points. The first bullet point discusses the Spark distributed computing framework. The second bullet point mentions Spark's atomic computing tools. The third bullet point discusses the improvement in calculation speed, noting that a task that normally takes 8055 hours is completed in 30 minutes. The terminal text in the background includes "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", and "2016-11-03T23:08:15.725+08:00 Sci_new数据进行map操作, count:2966547".

功能效果

数据类型	条 数
中文期刊	73762997
专利	8064424
学位论文	1639787
会议论文	1351523
成果奖励	594749
项目成果	187307
图书	1460885
SCI	2076547
EI	1400282
SCOPUS	13815659

数据增量：2 万条/天

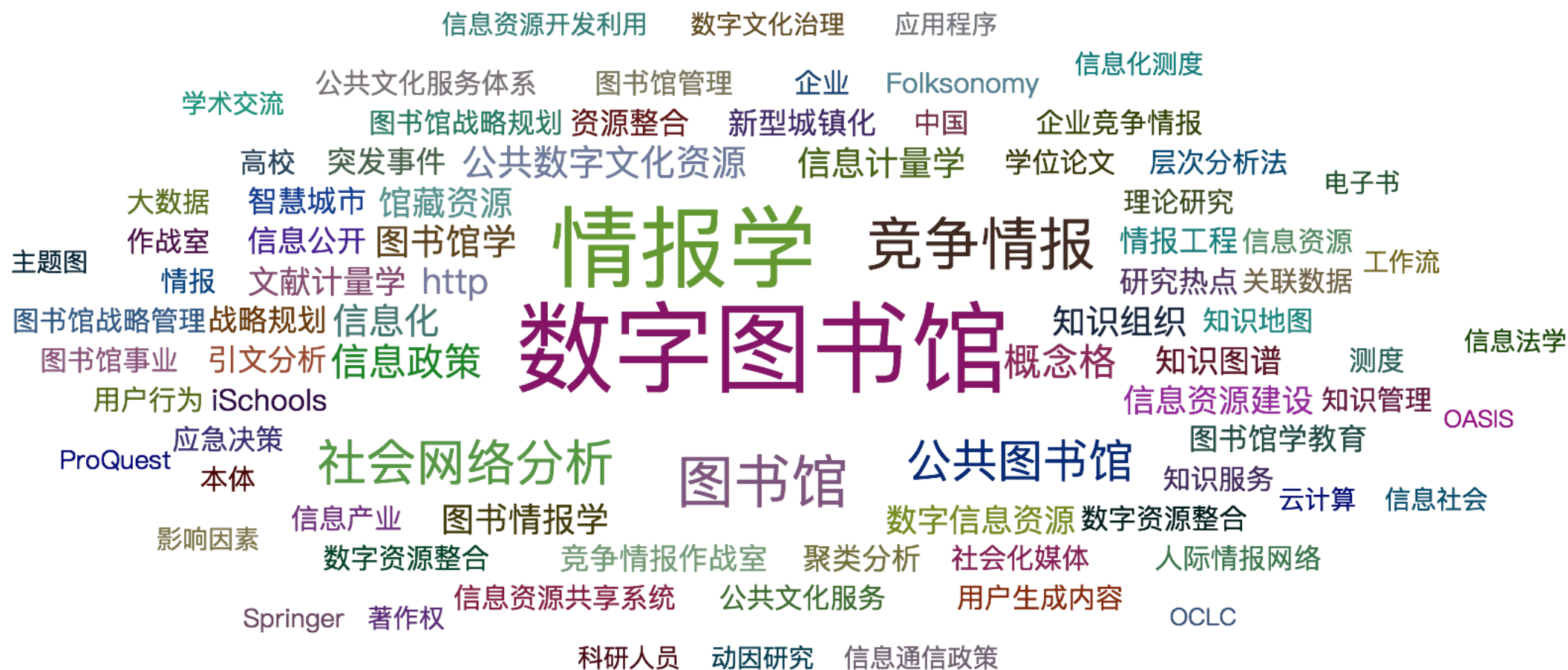
案例分享

CASE STUDY

3

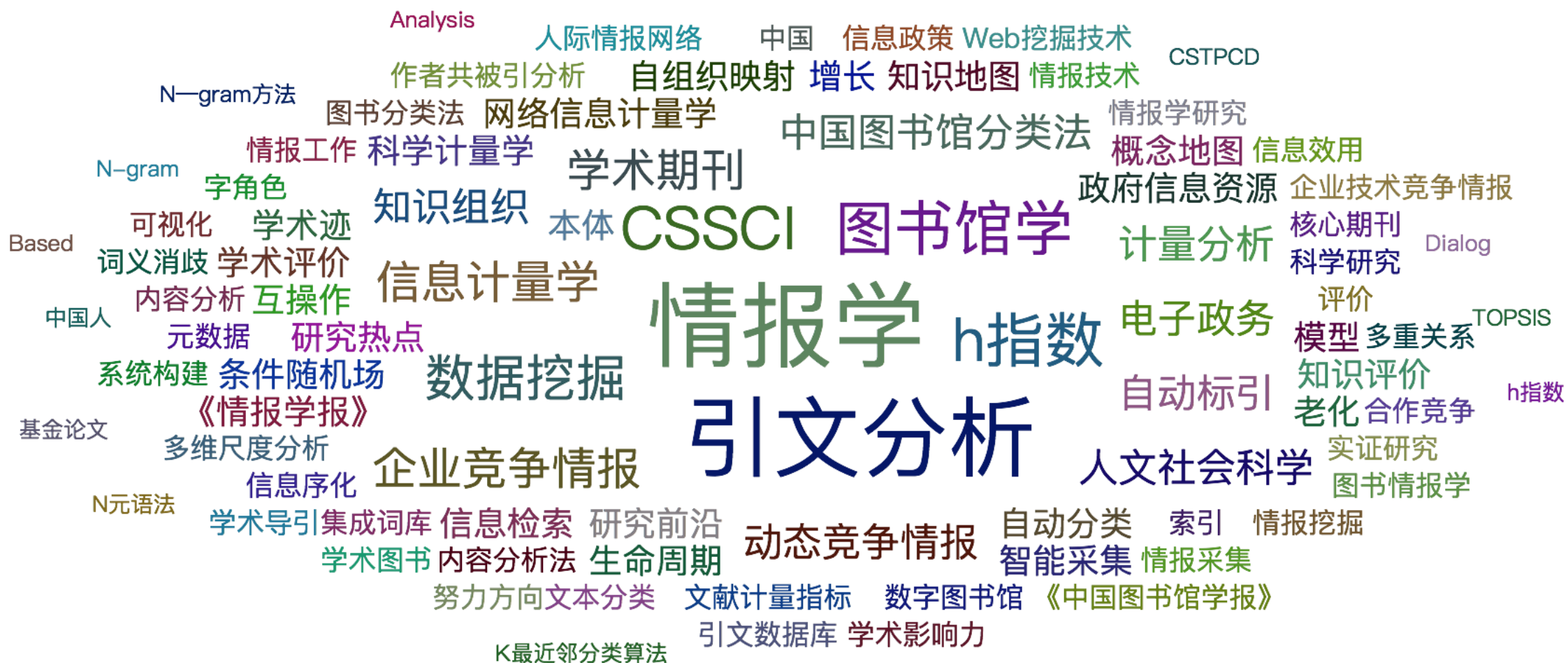
案例分享_期刊分析

图书情报核心期刊关键词top100-云图



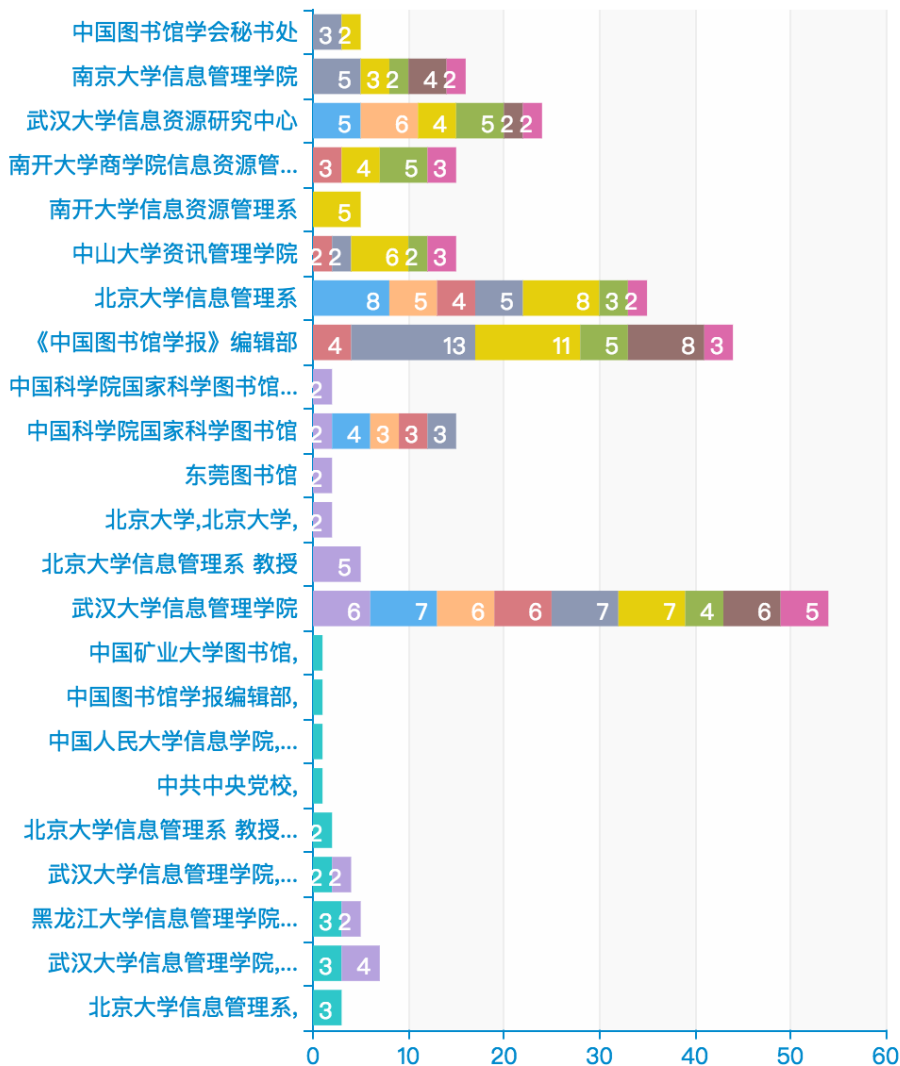
案例分享_期刊分析

大图、中图、情报学报关键词top100-云图

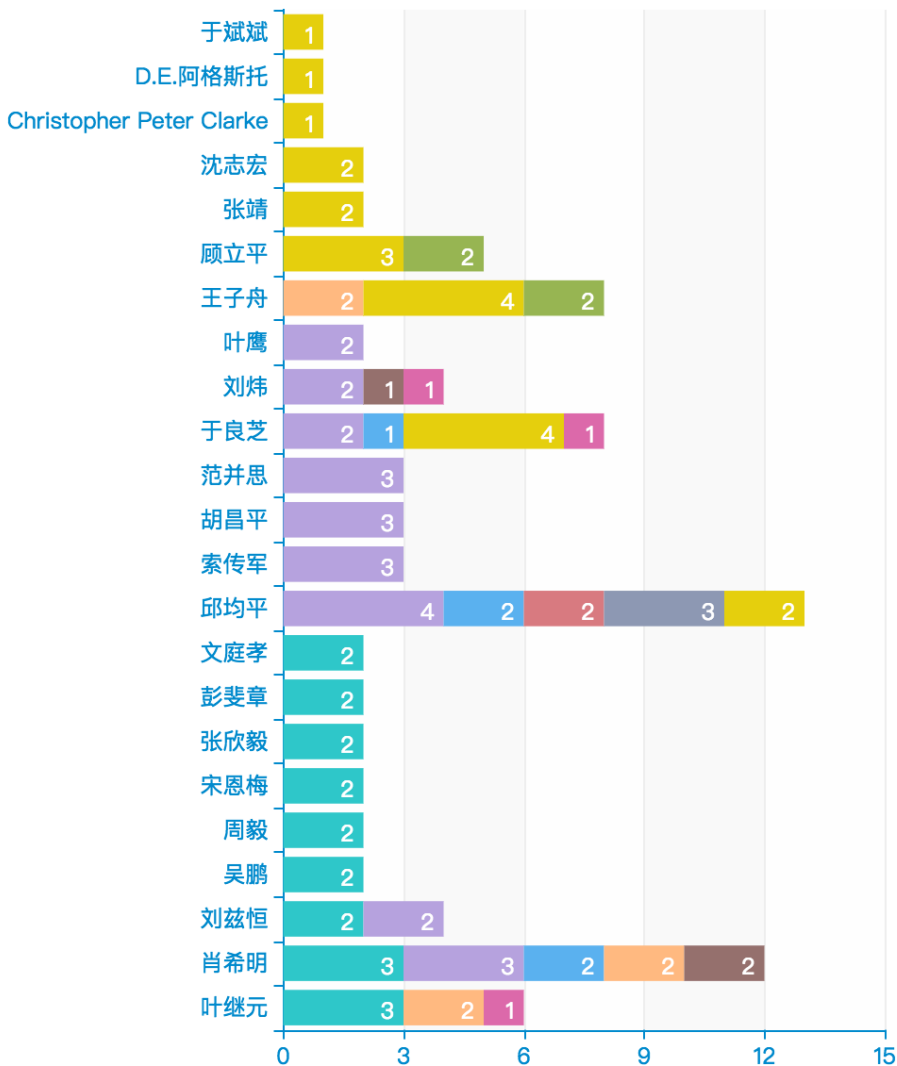
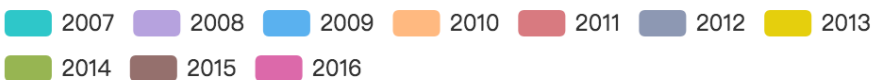


案例分享_期刊分析

本刊近年发文机构排行



本刊近年高发文作者排行



案例分享_中国工程科技专家库



登录 注册 移动端 历史更新



专家

Expert

从37859种中文期刊抽取了 8139066 位专家

关键词

keyword

抽取关键词约 4000万

请输入专家姓名（必填）

请输入机构

请输入学科

请输入关键词

开始搜索

最近热词

智能行 袁隆平 申长雨 陈竺 智能制造
机器人 石墨烯 碳纤维 大数据 袁隆平
申长雨 陈竺 智能制造 机器人 石墨烯
碳纤维 大数据袁隆平 申长雨 陈竺 智

数据更新

今日新增XXX位专家信息
系统已收录XX位专家信息
今日XX位专家信息被认领
总计XX位专家信息被认领

系统消息

2017年6月12日 与浙江大学合作
2017年6月12日 与浙江大学合作
2017年6月12日 与浙江大学合作
2017年6月12日 与浙江大学合作

案例分享_中国工程科技专家库



吴朝晖

认领

交通运输工程;计算机科学技术;信息与系统科学相关工程与技术;电子、通信与自动控制技术;经济学;基础医学;管理学;机械工程;纺织科学技术

浙江大学

个人简历

根据数据搜索表明吴朝晖（专家）可见的最早中文论文发表是1991年在计算机科学（期刊）上的知识库的一致性问题和检查方法一文，最早英文论文发表是2000年在BIOMEDICAL AND ENVIRONMENTAL SCIENCES（期刊）上的Oxidative stress before and after operation in patients with chronic cholecystitis containing gallstone一文。截至2016年在计算机科学、浙江大学学报（工学版）、计算机应用研究、FRONTIERS ON SEPARATION SCIENCE AND TECHNOLOGY、PLANT SCIENCE、PLANT BREEDING等6种期刊上共发文157篇，其研究领域主要集中在计算机科学技术、电子、通信与自动控制技术、经济学等方向。

关键词

ACT-R 驾驶行为建模 认知体系 倒排索引 语义相似度
Web服务 服务发现 服务组合 智能代理 黑板系统
分布式问题求解 协同决策 流程服务 回溯树
面向服务的体系架构 VDHA 网络 e-Science 协议
VDHA_Grid 中药复方配伍规律 关联规则发现 数据挖掘

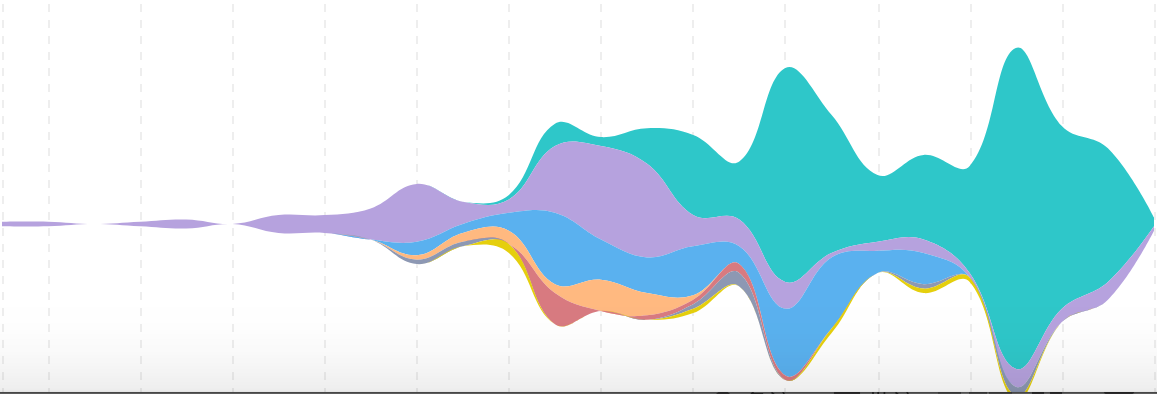
语义相似度关键词联想词

句子相似度 计算机应用 对等网络
文本分类 中文信息处理 语义Web服务 信息内容
概念相似度 文本聚类 查询扩展 模式匹配
HowNet 信息量 信息检索 义原 问答系统
概念格 聚类 语义距离 领域本体 语义网 叙词表
查全率 服务匹配 服务发现 主观判断 P2P

学术图谱

专家图谱

专利 期刊论文 项目成果 SCI论文 会议论文 成果奖励 图书



案例分享_中国工程科技专家库



语义相似度

搜索

相似关键词: 本体 语义距离 知网 领域本体 语义检索 语义相关度 查询扩展 信息检索 Web服务

此关键词早期提出专家

涵盖此关键词论文



余正涛

机构: 北京理工大学计算机科学与工程系,北...
学科: 计算机科学技术
期刊: 计算机工程与应用
论文: 基于问句语料库的受限领域自动应答...



张骞

机构: 东北大学计算机软件国家工程研究中...
学科: 计算机科学技术
期刊: 华中科技大学学报(自然科学版)
论文: SCORM学习资源的语义相似度度量 ...



陆

机构: 南京农业大学信息科技学院,南京农业...
学科: 图书馆、情报与文献学

基于问句语料库的受限领域自动应答系统

(31-33+89) 2003
余正涛,樊孝忠,宋丽哲
受限领域自动应答;问句语料库;语义标注;空间向量模型;语义相似度;问句相似度

CNKI

万方

SCORM学习资源的语义相似度度量

(297-299) 2003
张骞,张霞,刘积仁
导航;元数据;资源实体;资源本体;语义相似度

CNKI

万方

用于信息检索的同义词自动识别及其进展

(87-93) 2004
陆勇,侯汉清
同义词识别;语义相似度;词汇相似度算法;向量空间;共现分析

CNKI

万方

维普

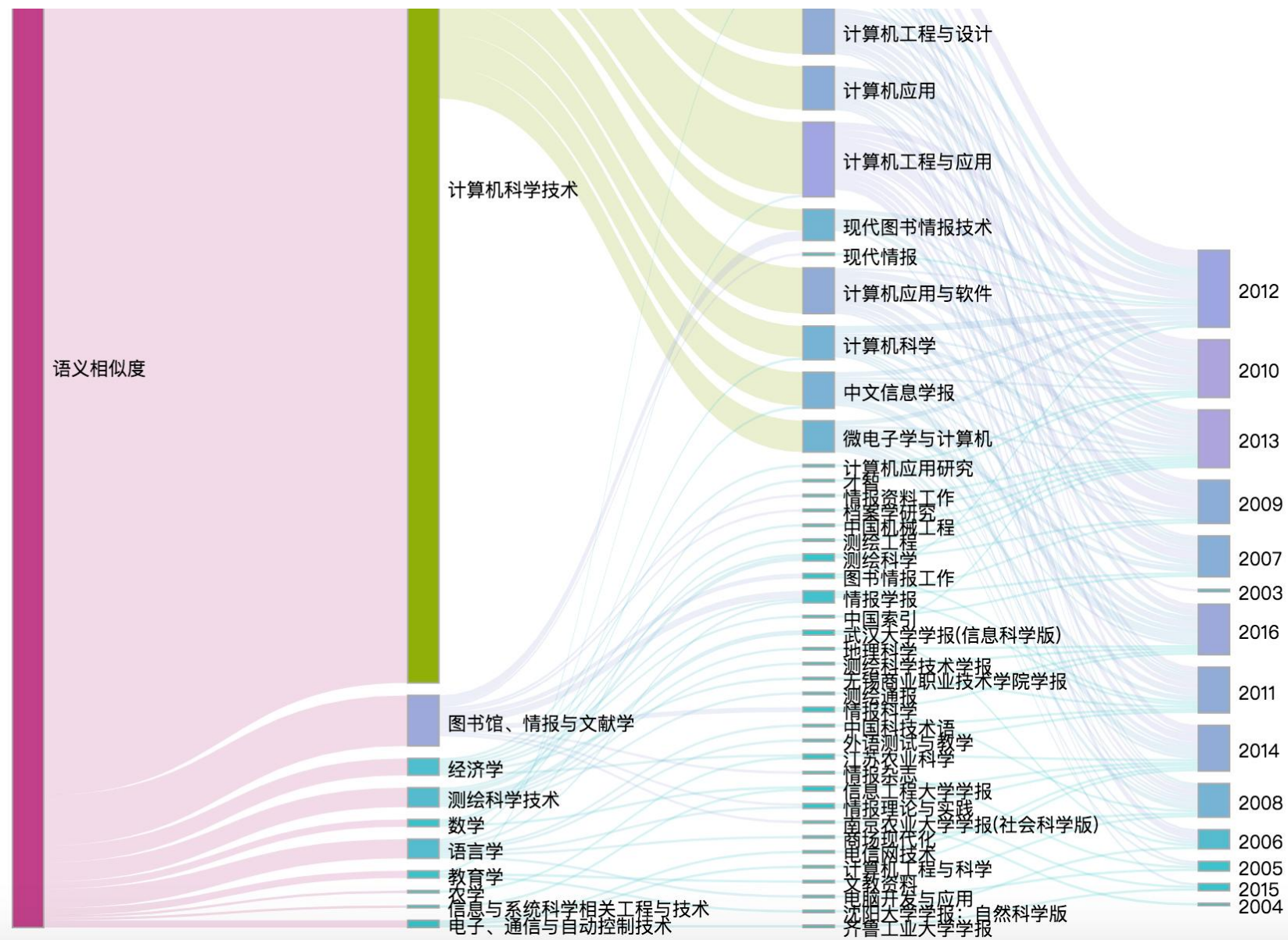
一种基于潜在语义分析的查询扩展算法

CNKI

万方

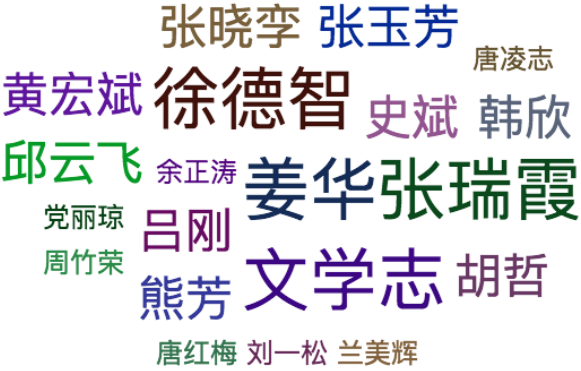
维普

案例分享_中国工程科技专家库



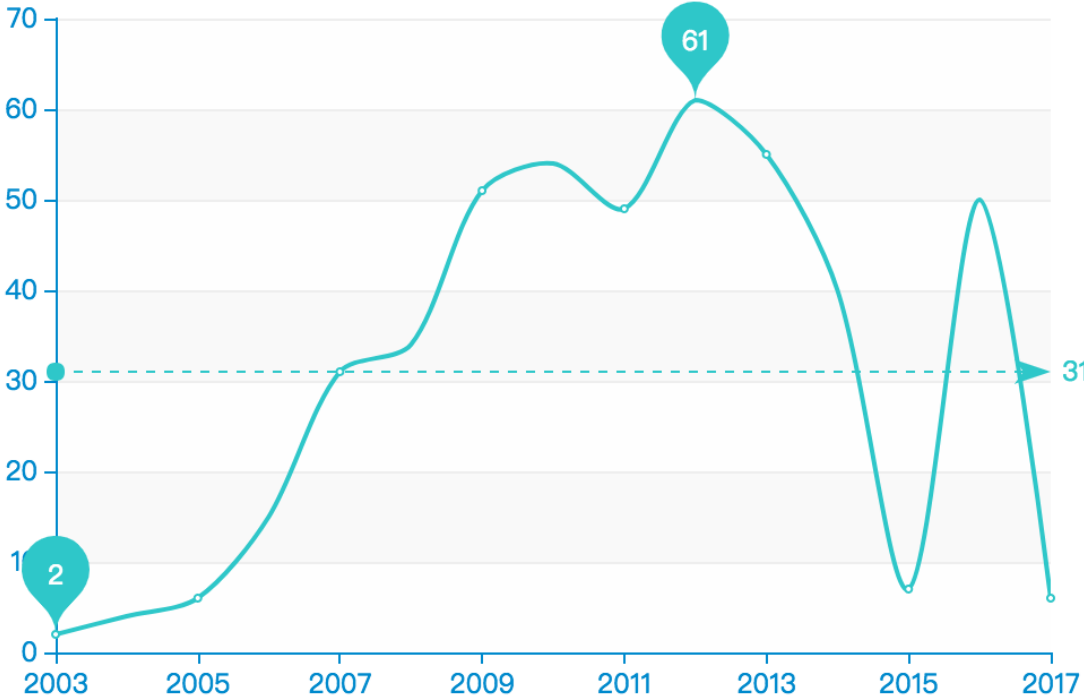
案例分享_中国工程科技专家库

围绕此关键词的专家网图



此关键词历史热度

语义相似度历史热度



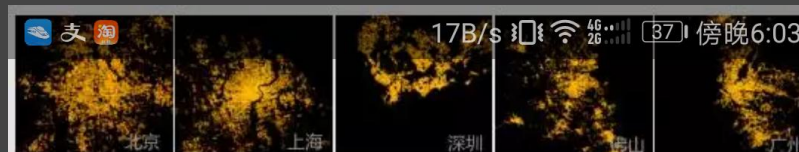
合作展望

COOPERATION

4

合作展望

- 数据获取 Data Acc
- 数据清洗 data cle
 - 同名问题
 - 机构规范
 - 学科映射
 -
- 数据挖掘 Data Mi
 - 学者认定 Scholar I
 - 合作者 co - author
 - 工作经历 personal
 - 学术图谱 Knowled
- 数据可视化
- 数据共享



THANK YOU