社会化新媒体与非结构化大数据分析

Research on New Social Media and Unstructured Big Data Analysis

张华平 博士 副教授

大数据搜索与挖掘实验室

kevinzhang@bit.edu.cn

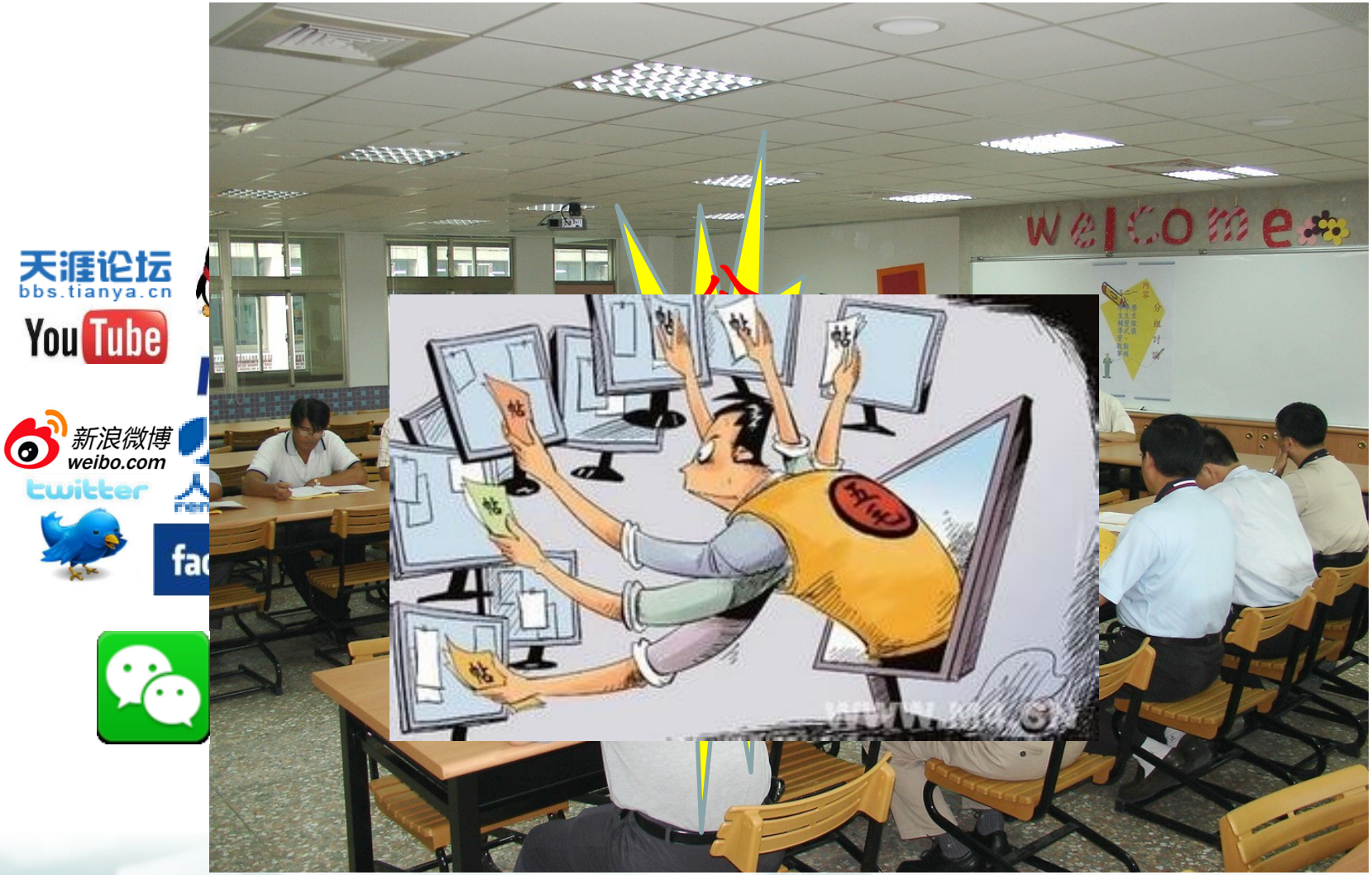
www.nlpir.org

中国数据分析师行业峰会

2015.9

大数据论坛 BigdataBBS.com

Gov1.0遭遇Web3.0



大数据论坛
BigdataBBS.com

从棱镜手机监控看大数据洞察力...

纲要

非结构
大数据

I 社会化新媒体

II 非结构化大数据

III 大数据搜索与挖掘关键技术

IV 新媒体分析实战案例

大数据论坛
BigdataBBS.com

新旧媒体冰火两重天

应用 网易首页 登录

2013年多家时尚杂志倒闭，2014年呢？

2014-01-16 14:00:24 来源:时代周报(广州) 有0人参与 分享到

时尚平媒华丽逆袭，谁找到了新模式？

吴黛西

2007年岁末，设计师Marc Jacobs在洛克菲勒的Rainbow Room召开了年度假日聚会，并邀请了800名客人。主要客人包括来自Vogue、W和Harper's Bazaar等杂志的贵宾。衣香鬓影、喧歌闹舞、摇摇摆摆，红男绿女如飞蛾般穿梭，脱衣舞女郎、无上装帅哥，以及一场撒向贵宾的钱雨。这是时尚杂志编辑们最好的时光。葡萄美酒夜光杯，从纽约到伦敦，从巴黎到北京，穿越普拉达的华服，俯视蒂凡尼的珠宝……

时尚杂志的不景气之年

好景不常，2010年，英美就开始逐渐停掉一批出版物。殿堂级时尚杂志《i-D》也宣布改为两月一刊，去年《新闻周刊》索性停止纸质杂志出版，英国倒闭98家出版社，国外时尚平媒已风声鹤唳。随即，这场雪灾就蔓延到国内。2013年，财讯传媒集团旗下《东方壹周》和《他生活hislife》两本时尚刊物相继停刊。时尚传媒集团欲停止旗下三本刊物《Good Housekeeping时尚生活》、《COSMOPOLITAN Bride时尚新娘》和《Men's Health 时尚健康（男士）》的出版。2014开年伊始，南方都市报旗下唯一一本时尚刊物《风尚周报》也因营收艰难传将停刊。

新旧媒体冰火两重天II

新旧媒体冰火两重天 III

近年央视广告收入显疲态 增速仅省级卫视一半

2014-06-04 11:21:07 来源: 21世纪经济报道(广州) 有2157人参与 分享到

上市公司相关品牌中标2014年央视广告的金额前十列表

品牌/上市公司简称	金额（万元）	行业
露露/承德露露	67207	饮料
美的/美的集团	38096	家电
格力/格力电器	27754	家电
老板电器/老板电器	22933	家电
青岛啤酒/青岛啤酒	20119	酒业
云南白药/云南白药	19769	医药
燕京饮料/燕京啤酒	16654	饮料
双汇/双汇发展	14396	食品
华菱汽车/华菱星马	12497	交通
一汽轿车/一汽轿车	11460	交通

数据来源：北京中视百纳国际广告有限公司

有研究机构数据表明，近年来，央视广告收入疲态已显。根据CTR数据预测，2013年电视媒体广告刊例增长11.3%，其中，**央视增长14.7%，省级卫视增长28%；** 2012年央视广告收入为269.76亿元，同比增长低于15%，而百度广告收入为222.46亿元，同比增长53.5%。

有广告行业专家对21世纪经济报道表示，由于互联网、移动端广告给电视媒体带来的冲击，除了少数省级媒体外，电视媒体的广告收入增长基本都在放缓。他指

传统媒体之殇

↗ 生产《渠道《品牌《社群

- 写书的不如卖书的；卖书的不如卖粉丝的

亚马逊 amazon.cn

当当网 dangdang.com

传统媒体之殇

- 平面纸媒传统造成信息孤岛基因
  - 历史内容挖掘的价值增值；
  - 世界为我所用：
  - 今日头条： 我们不生产新闻,我们只是新闻的搬运工
- 内容为王 vs. 用户为王
- 小数据精英 vs. 大数据技术

大数据论坛 BigdataBBS.com

传统媒体之殇

福建新闻 NEWS.ONFJ.COM

福建在线 | 新闻 | 社会 | 汽车 | 健康 | 财经 | 经济 | 出国 | 法律 | 福建 | 福州

当前的位置：福建在线 > 新闻中心 > 福建新闻 > 社会新闻 >正文

1 原始报道

厦门女子玩触屏磨坏手指软骨 或致永久性致残

来源: 台海网-海峡导报 时间: 2012-06-25 10:26 编辑: 陈华春

福建在线www.onfj.com6月25日讯(海峡导报记者 陈洋钦 通讯

着苹果手机和平板电脑的流行，一些“苹果病”也随之而来。比较

市民吴小姐最近被确诊患上了一种 “新型”的 “苹果病”———

ONFJ.CO

sina新浪福建 新浪福建 > 城事 > 社会 > 正文

2012欧lx35 “3·5”行动火热进行中

0%购置税优惠 0:50分期付款或0利息

2 一次转发

厦门女子玩触屏磨坏手指软骨 或致

sina新浪数码 新浪数码 > 笔记本 > 正文

4 二次加工后的转发

女子玩平板手指永久性致残 专家说活该

3.中关村在线2天后的二次所谓“原创”

女子玩平板手指永久性致残 专家说：P

类型：[原创] 作者：中关村在线 关晓萌 日期：2012-06-27 05:59:00

原文地址：http://pad.zol.com.cn/303/3030356.html

中关村在线消息：当年郑智的无影脚飞踢西塞小腿致残事件，如今还历历在目，但毕竟足球场上断胳膊断腿都是再正常不过的事情。如果“永久性致残”事件发生在一个平板电脑玩家的身上，还是个女玩家，似乎就不那么正常了。

06月27日 08:31 中关村在线 微博

件，如今还历历在目，但毕竟足球场上断

残”事件发生在一个平板电脑玩家的身上

观小车被恶意刮花 原是老人跌倒失手划伤

车刚开半个月 临近报废年限突然自燃

车看到警察就掉头 原来暗藏3把枪

假制服回原公司盗窃作案四起被判六个月

網易 数码 网易 > 数码频道 > 正文

5各大网站跟风转发

请输入关键词 新闻 搜索

女子玩平板手指永久性致残 专家说：P

2012-06-27 06:03:45 来源: 中关村在线网站(北京) 有15人参与 手机看新闻 转发到微博 0

新浪微博 beta 首页 广场 微群 应用 游戏 搜索微博、找人 ICTCLAS张华... 手机 找人 消息 帐

举报处理大厅 > 不实信息

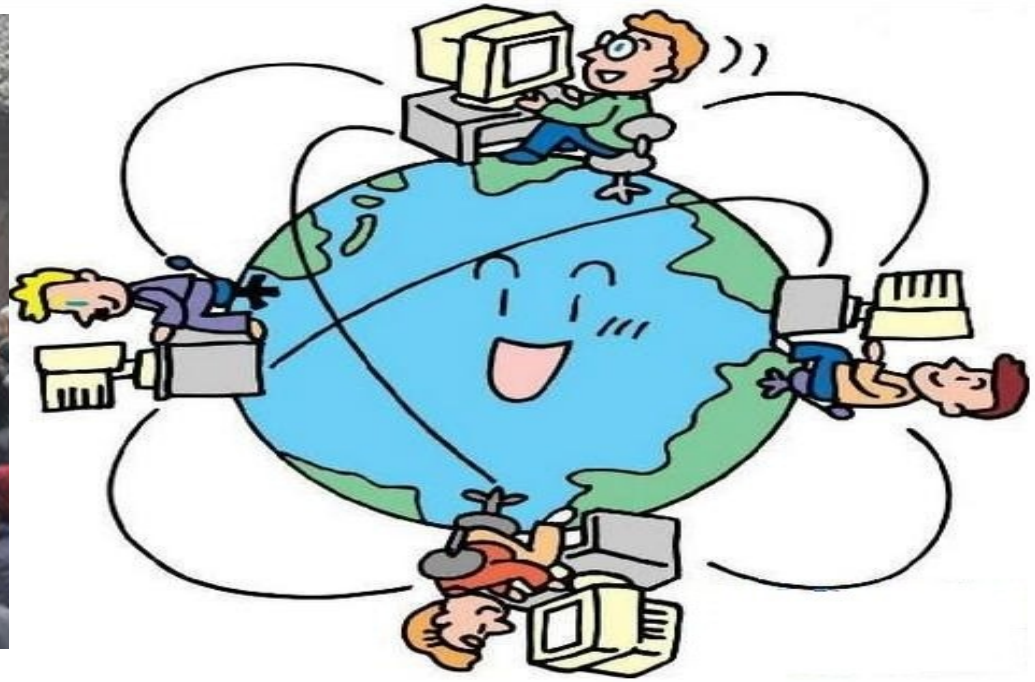
6.微博不实举报：错打五十大板:)

女子玩平板手指永久致残

受理编号: K1CaJ6wpc6q8i 状态: 举证阶段 > 判定阶段 > 结果公示 距离判定结束时间:23时55分57秒 访问次数: 287

社会化媒体

- 社会化媒体（社交媒体）运用易涉入和传播的沟通技术并以社会化交流为目的的媒体。特点：社会关系+传媒

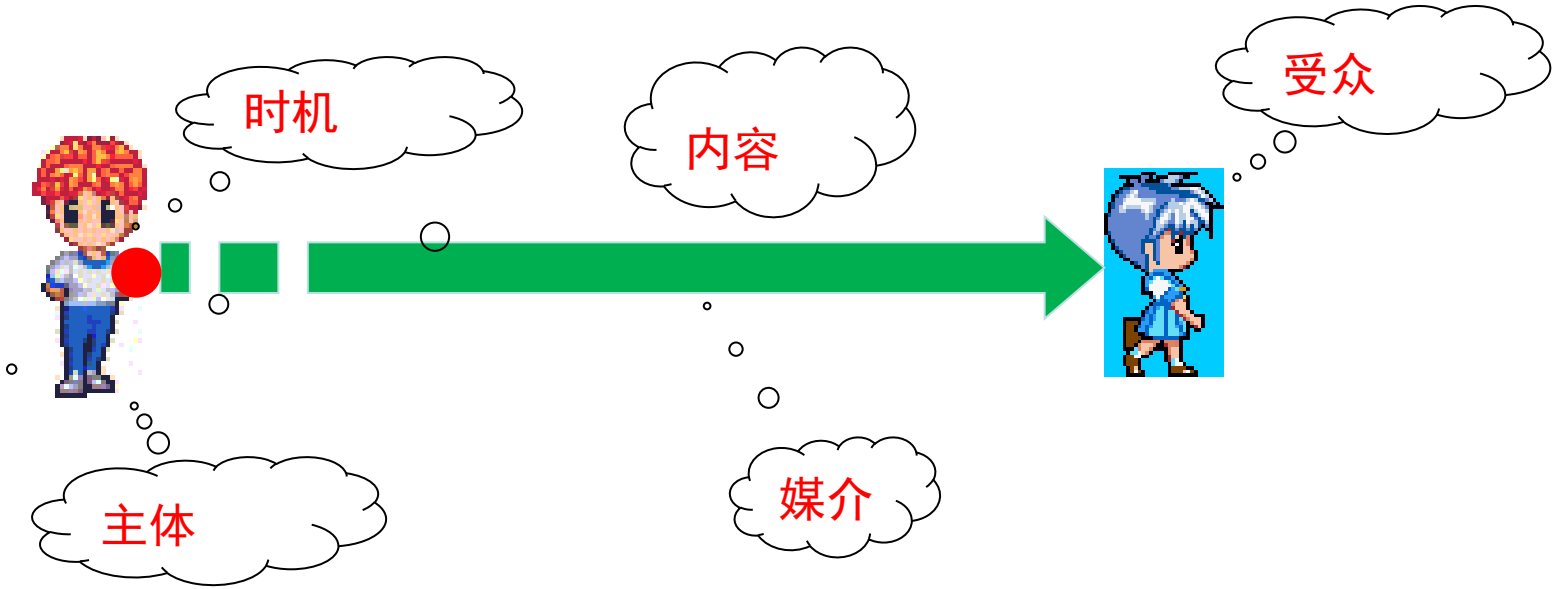


大数据论坛
BigdataBBS.com

社会化媒体发展历程

BigdataBBS.com

传播要素解析



大数据论坛
BigdataBBS.com

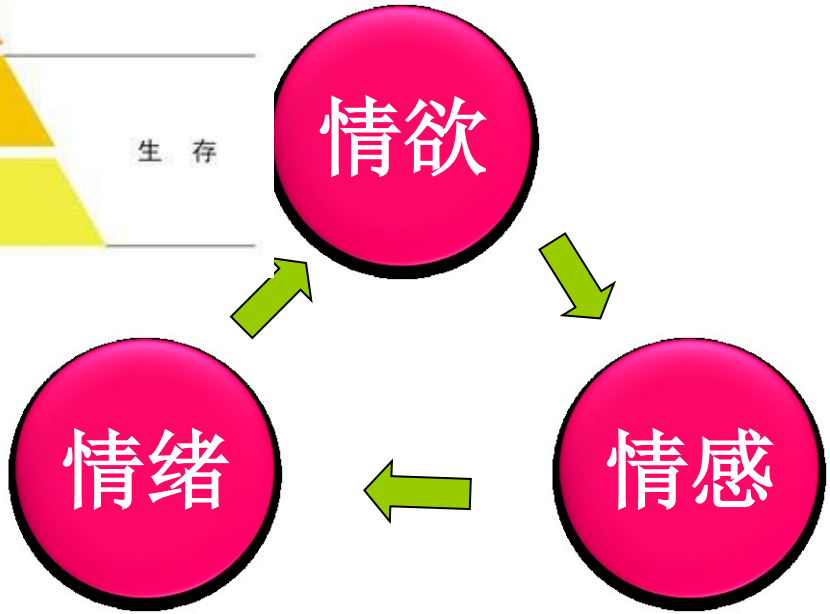
传统媒体 vs. 新媒体

传媒时代	报纸/电视	互联网1.0	新媒体
内容	正式	半正式	非正式
传播方式	一对多广播，无反馈的；	少对多浏览，弱化社交	多对多，社交型，
主体	授权机构，少数	大部分网民	几乎所有人
受众	被动接受，参与感弱	主动获取，部分参与	主动推送，收发全参与
生产过程	先审后发	先发后审	即发少审
时机/速度	24-72小时	1-2小时	即时，快且影响面广
代表	人民日报，CCTV	新浪新闻，博客，	微博,微信，facebook
场景	政府宣传，传教	小范围演讲互动	对等交流

BigdataBBS.com

社会媒体传播实战技巧：内容

马斯洛需求层次理论



情与代入感

艾克曼：喜、怒、哀、
仇官仇富情绪

大数据论坛
BigdataBBS.com

社会媒体传播实战技巧：内容

1. 代入感原则，拒绝自嗨

社会媒体传播实战技巧：内容

↗2.拉家常讲故事，拒绝高大全空洞说教

熬夜看球

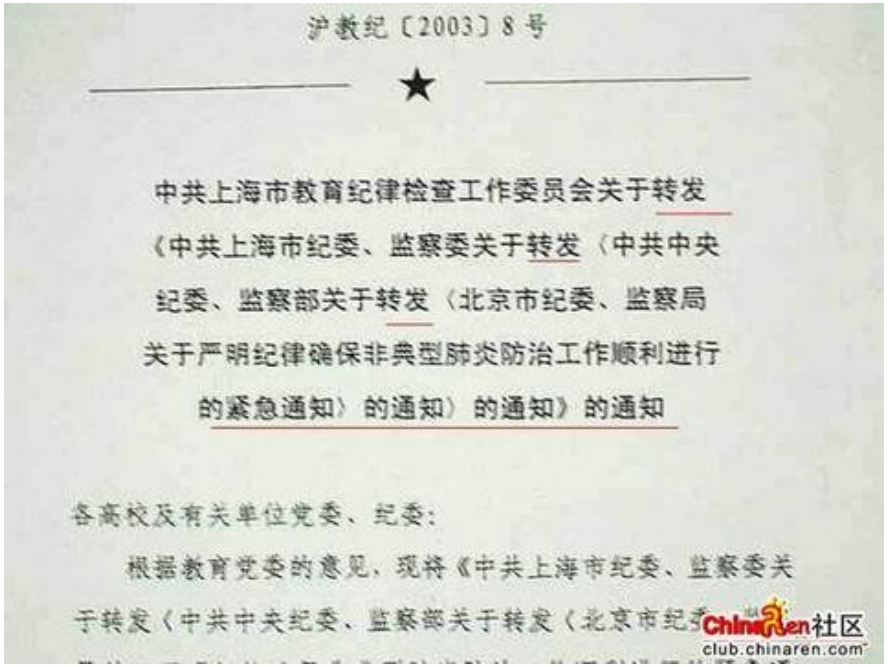
社会媒体传播实战技巧：内容

↗3. 角色个性化拒绝平庸：幽默风趣，借题发挥

社会媒体传播实战技巧：内容

- 主题创意（幽默、借用，故事性）、好记易传播（3-4音节）

打土豪，分田地
微笑局长
表叔
房姐
我爸是李刚
土豪，我们交朋友吧
光盘计划



沪教纪〔2003〕8号

★

中共上海市教育纪律检查工作委员会关于转发《中共上海市纪委、监察委关于转发〈中共中央纪委、监察部关于转发〈北京市纪委、监察局关于严明纪律确保非典型肺炎防治工作顺利进行的紧急通知〉的通知〉的通知》的通知

各高校及有关单位党委、纪委：

根据教育党委的意见，现将《中共上海市纪委、监察委关于转发〈中共中央纪委、监察部关于转发〈北京市纪委

ChinaRen社区 club.chinaren.com

大数据论坛 BigdataBBS.com

纲要

非结构
大数据

I 社会化新媒体

II 非结构化大数据

III 大数据搜索与挖掘关键技术

IV 新媒体分析实战案例

大数据论坛
BigdataBBS.com

什么是大数据

↗ 我们的见解

- 大数据是指 、多源异构、实时 然语言处理、信息 知识，转化为智慧
- 是一场新的 革命。（全量分析 地认识世界的混杂 断）

论坛
BigdataBBS.com

杨达才启示：1+1>>2才是大数据

大数据论坛
BigdataBBS.com

近半世纪来的三次革命

计算机时代

20世纪
70年代

计算方式的革命

互联网时代

20世纪
90年代

信息传播方式的革命

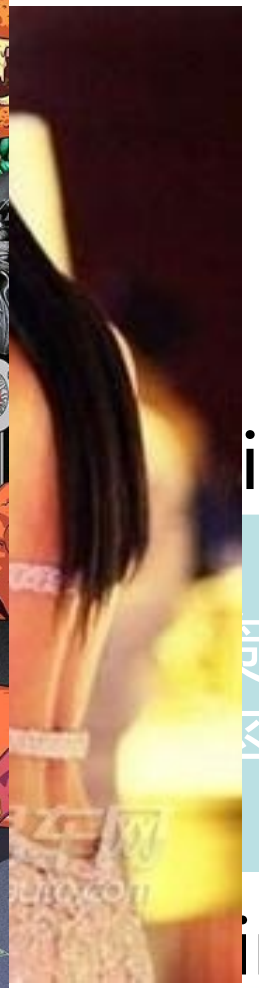
大数据时代

21世纪10
年代

决策方式的革命

大数据论坛
BigdataBBS.com

大数据颠覆决策模式



ion

ime

数据论坛
DataBBS.com

大数据时代的特征

大数据论坛
BigdataBBS.com

大数据应对之道:知著、见微、晓意

大数据论坛
BigdataBBS.com

不“晓意”的尴尬

不“晓意”的尴尬

- Who was Chinese President in 2012?
- Hu.
- Who?

Hu Jintao

- Who is Chinese President?
- Xi.
- She?

Xi Jinping

大数据论坛 BigdataBBS.com

机器理解自然语言？

KFC店里现在都挂着新的海报，上面写了这样一句话……

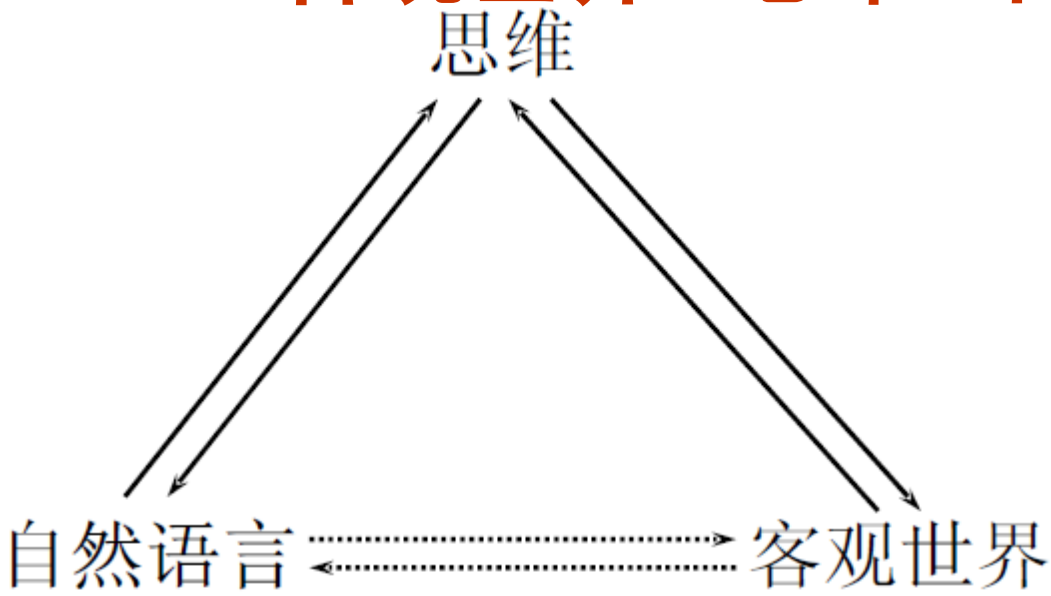
右面的鸡才是最好的

WE DO CHICKEN RIGHT

我们是做炸鸡的，我们做鸡的，我们行鸡道，我们当鸡的，我们一定要把鸡打成右派!!!

大数据论坛
BigdataBBS.com

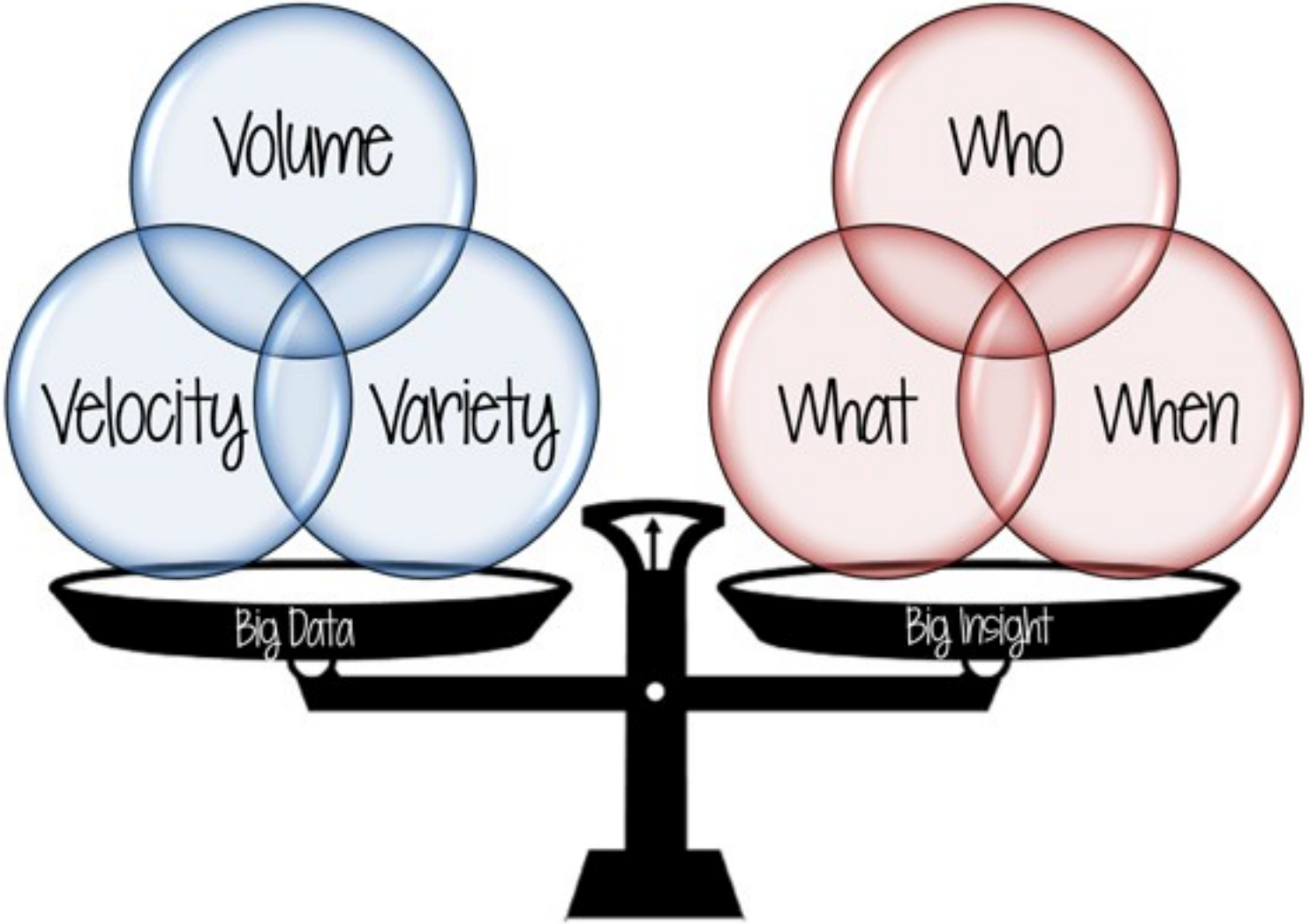
客观世界->思维->自然语言



- 衰减效应：
  - 思维最多只能反映80%的客观世界；
  - 自然语言只能反映80%的思维：词不达意，答非所问；
  - 听众最多只能听懂80%；
  - 听懂的部分只有80%能反映到思维中；
  - 分析客观世界的最多只能利用80%。

大数据论坛 BigdataBBS.com

大数据更大意义上是非结构化内容理解



大数据论坛
BigdataBBS.com

纲要

非结构
大数据

I 社会化新媒体

II 非结构化大数据

III 大数据搜索与挖掘关键技术

IV 新媒体分析实战案例

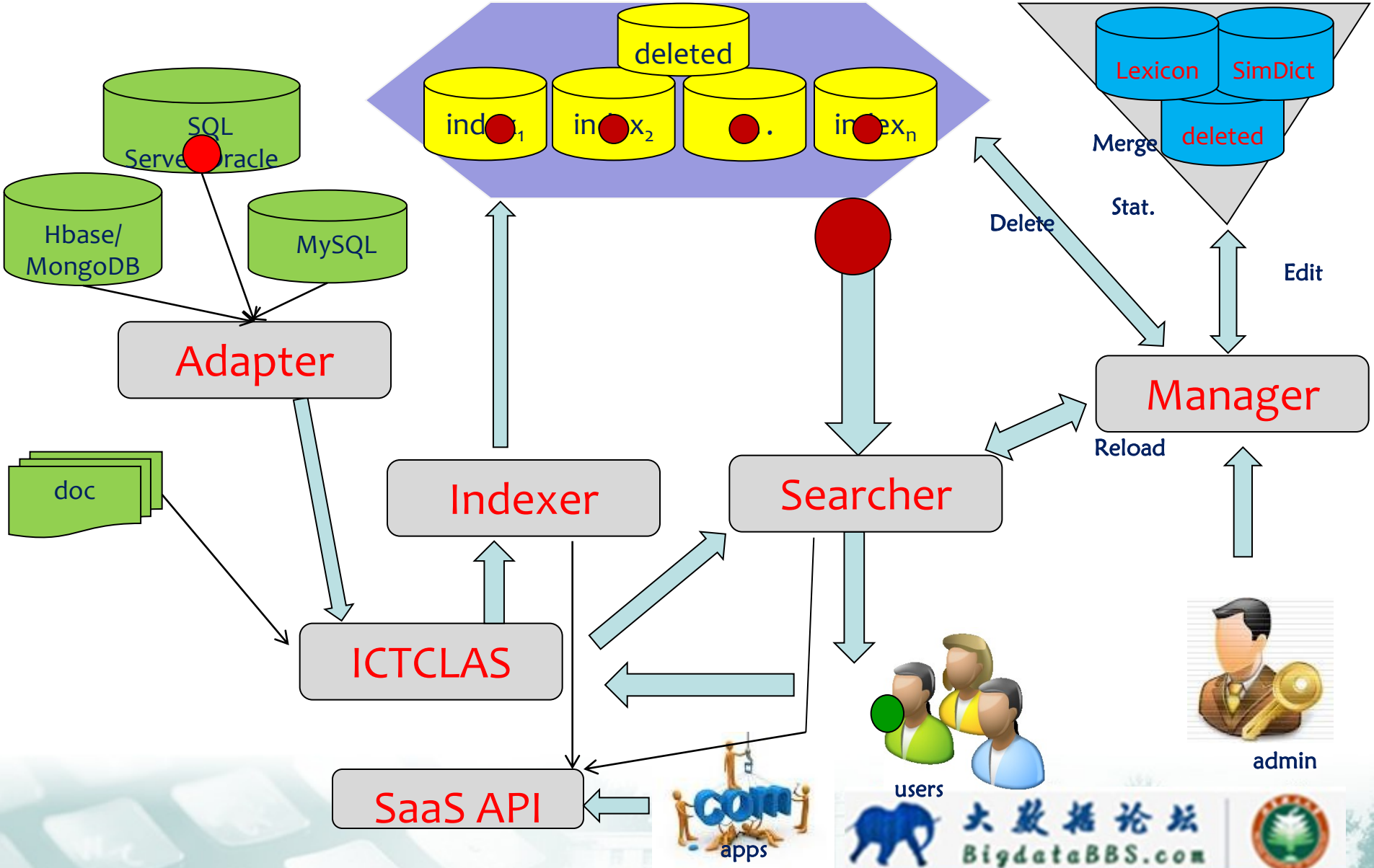
大数据论坛
BigdataBBS.com

JZSearch大数据精准搜索

- **搜索基本功能：**
  - 多字段关联搜索、指定字段排序、精确搜索与模糊搜索
- **搜索特色功能：**
  - 内嵌正负面情感等极性分析、语义联想搜索、临近搜索、搜索结果去重；
  - 内嵌了ICTCLAS智能分词系统；
  - 数据库实时同步：数据库增删改10秒内即可同步到搜索；
- **搜索维护功能：**
  - 单点故障容错；支持增量索引；自动备份与恢复机制；自动缓存机制；自动优化机制；搜索屏蔽与恢复；

大数据论坛
BigdataBBS.com

JZSearch Big Data Search Engine Architecture



JZSearch内部语法示例

- [FIELD] * [NEAR] 尚福林 ##负面JZSearch## 12
- [FIELD] price [RANG] 1.0 9.0 [FIELD] name [AND] 牛奶儿童
- [FIELD] name [PREF] 张
  - **姓名字段name必须以“张”作为前缀开头**
- [FIELD] id [PREC] 123
  - **字段id必须以“123”精准匹配，如“1234”或者“0123”均不作为匹配结果；**
- [field] content [complex] 统计局||中国统计局||CPI 骗人||砖家 10
  - **要求content字段内，同时出现{统计局||中国统计局||CPI }{骗人||砖家},且两者之间的距离必须在10个词内。**

大数据论坛 BigdataBBS.com

搜索结果自动分组统计

标准 图书馆 www.bzsb.info

加入收藏 设为首页 帮助 关于我们

网站▼ | 登录 | 注册 | 返回首页

标准分类

- 植物检疫、病虫害防治(18)
- 瓜果、蔬菜种植与产品(14)
- 农药管理与使用方法(12)
- 豆类、薯类作物与产品(10)
- 食品卫生(8)
- 农机具(7)
- 基础标准与通用方法(5)
- 农林技术(5)
- 种籽与育种(5)
- 粮食加工与制品(5)
- 经济作物综合(4)
- 畜禽饲料与添加剂(4)
- 植物保护综合(2)
- 一般有机化工原料(2)
- 基础标准与通用方法(2)
- 标志、包装、运输、贮存(2)
- 食品加工与制品综合(2)
- 标准化、质量管理(1)
- 粮食、饲料作物综合(1)
- 标准化、质量管理(1)
- 蔬菜加工与制品(1)
- 调味品(1)
- 蔬菜罐头(1)
- 烟草制品(1)

标准号 关键词: 土豆 检索 ☑ 按国内标准 ☐ 按国外标准

分类检索 高级检索

首页 < 1 2 3 4 5 6 7 8 > 尾页

对搜索结果自动分类统计

自动搜索数据库里面的内容，数据库更新10秒内，会直接反映在搜索结果里。

标题

标准号	名称	中标分类名称
DB13/T 396.7-1999	旱地地膜覆盖栽培技术规程	种籽与育种
DB13/T 398.7-1999	旱地地膜覆盖栽培技术规程	农林技术
DB13/T 865-2007	马铃薯有机栽培技术规程	瓜果、蔬菜种植与产品
DB1302/T 156-2001	无公害马铃薯生产技术规程	瓜果、蔬菜种植与产品
DB1304/T 071-2001	无公害农产品生产技术规程 马铃薯	农林技术
DB1304/T 102-2001	马铃薯脱毒种薯繁育技术规程	农林技术
DB1304/T 136-2004	棉花与土豆（马铃薯）间套复种生产技术规程	瓜果、蔬菜种植与产品
DB1306/T 11-2005	无公害马铃薯生产技术规程	瓜果、蔬菜种植与产品
DB1308/T 007-1999	马铃薯茎尖脱毒技术规程	豆类、薯类作物与产品
DB1308/T 008-1999	脱毒马铃薯微型种薯生产技术规程	豆类、薯类作物与产品
DB1308/T 009-1999	脱毒马铃薯基础种薯生产技术规程	豆类、薯类作物与产品
DB1308/T 010-1999	脱毒马铃薯合格种薯生产技术规程	豆类、薯类作物与产品
DB1308/T 011-1999	马铃薯脱毒种薯质量标准	豆类、薯类作物与产品
DB1308/T 012-1999	马铃薯商品薯（块茎）质量标准	豆类、薯类作物与产品
DB1308/T 088-2005	食用马铃薯淀粉	豆类、薯类作物与产品

BigdataBBS.com

我们的工作：JZSearch精准搜索引擎

少数民族语言精准搜索引擎

搜索

偏移量	长度	文件路径	标题	全文
0	11532	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--خەلق نوری	正文
57433	11532	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--خەلق نوری	正文
69320	10674	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--شىنجاڭ خەۋەرلىرى--خەلق نوری	正文
80102	9776	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--تىبەت خەۋەرلەر--خەلق نوری	正文
89986	10393	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--دۇنيا خەۋەرلىرى--خەلق نوری	正文
101320	4984	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--مۇشەت ئاسراش--خەلق نوری	正文
106304	3690	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--دىن--خەلق نوری	正文
109994	4304	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--قىزىق نۇقتا--خەلق نوری	正文
114298	4037	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--شەخسىلەر--خەلق نوری	正文
118335	8526	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网维吾尔文--سۈرەتلىك خەۋەرلەر--خەلق نوری	正文
127751	4385	F:/data/MultiLang/Corpus/20100426/000@20100426	人民网--خەلق نوری--amp;nbsp; سەھىيە amp;nbsp; مەجىسۇس	正文

下一页 尾页 共有 1380 条记录，当前第1/92 页

باش بەت قىلىڭىبەت ساقلاڭتەھرىر خەت ساندۇقىنالاقىلىشىڭباش بەتمەركەز人民网维吾尔文--خەلق نوری
خەۋەرلىرىخەلقنارا خەۋەرلەرشىنجاڭ #x0D; #x0D; #x0A; روھىمەملىكەت
خەۋەرلىرىئىقتىسادجەمئىيەتمانارىپتەنتەربىيەمىللەت ۋە مەدەنىيەتمۇھىت ئاسراشدىنقىزىق نۇقتاشەخسلەرسۈرەتلىك
خەۋەرلەرماتېرىيالمەخسۇس سەھىيەزەھەر ۋە ئەيدىزىبىلگى خەۋەرلەر مەملىكەتلىك ئەرلەر ۋاسكىتبول كەسپىي بىرلەشمە
مۇسابىقىسىنىڭ نومۇرىي ھەل قىلغۇچ مۇسابىقىسىدە گۇاڭدۇڭ كوماندىسى شىنجاڭ كوماندىسىنى يەڭدى شىنجاڭ
تۈرپاندا كەم كۆرۈلىدىغان بورانلىق ھاۋا رايى كۆرۈلدى گەنسۇ جاڭيېدا قۇم بورانلىق ھاۋارايى كۆرۈلدى 1-پەسىلدە سىرتتىن
شىنجاڭغا سېلىنغان مەبلەغ نوخشاش ۋاقىتتىكىدىن 4 يېرىم ھەسسە ئاشتى روسىيەنىڭ تۈنجى تورگۈمدىكى ياردەم
بۇيۇملىرى يېتىپ كەلدى شىنجاڭ قۇمۇلدا قۇم بورانلىق ھاۋا رايى كۆرۈلدى ئالدىن مەلۇمات: ئادىل ھوشۇر خەلق تورىدا

大数据论坛 BigdataBBS.com

NLPIR大数据搜索与挖掘技术开发平台

- NLPIR网络搜索与挖掘共享开发平台，针对语言信息内容处理的全技术链条的共享开发平台。15年专业研究与工程积累，提供应用软件及各平台下的二次开发包，非商用永久免费。www.nlpir.org下载。

自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform

- 核心功能包括：
  - 搜索类：全文精准检索；
  - 语言类：新词发现，分词标注，统计分析与术语翻译；关键词提取；
  - 文档类：文本聚类及热点分析；分类过滤；自动摘要；文档去重；情感分析

大数据论坛
BigdataBBS.com

NLPIR大数据搜索与挖掘技术开发平台

大数据搜索与挖掘系统(试用版)

新词发现 | 语料库分词 | 词频统计及翻译 | 聚类 | 分类 | 正负面分析 | 摘要及关键词提取 | 文档去重 | HTML正文提取 | 全文检索 | 编码转换

新词存放地址： C:\Users\pc\Desktop\新建文件夹 (2)\新词.txt ... 编辑 导入用户词典

语料源所在路径： ...

分词结果存放路径： ... 语料库分词

☑ 手工输入语料: 开始分词

北杜市（日本），2009年10月16日 探访日本大型太阳能电池试验场 10月15日，在日本山梨县北杜市，一名日本技术人员正在介绍北杜试验场。北杜市的大规模太阳光发电研究所北杜试验场，是日本最主要的太阳能电池测试场所之一，这里对来自中国在内的十几个国家的太阳能电池生产厂商的产品进行对比试验，也为日本研究新型太阳能电池提供参考依据。 新华社记者刘华摄

分词结果提示

北杜市/n （/wkz 日本/nsf ）/wky ，/wd 2009年/t 10月/t 16日/t 探访/v 日本/nsf 大型/b 太阳能/n 电池/n 试验场/n 10月/t 15日/t ，/wd 在/p 日本/nsf 山/n 梨/n 县/n 北杜市/n ，/wd 一/m 名/q 日本/nsf 技术/n 人员/n 正/d 在/p 介绍/v 北杜/n 试验场/n 。/wj 北杜市/n 的/ude1 大规模/b 太/b 阳光/n 发电/vn 研究所/n 北杜/n 试验场/n ，/wd 是/vshi 日本/nsf 最/d 主要/b 的/ude1 太阳能/n 电池/n 测试/vn 场所/n 之一/rz ，/wd 这里/rzs 对/p 来自/v 中国/ns 在内/u 的/ude1 十几/m 个/q 国家/n 的/ude1 太阳能/n 电池/n 生产/vn 厂商/n 的/ude1 产品/n 进行/vx 对比/vn 试验/vn ，/wd 也/d 为/v 日本/nsf 研究/v 新型/b 太阳能/n 电池/n 提供/v 参考/vn 依据/n 。/wj 新华社/nt 记者/n 刘华/nr 摄/vg

关于 退出

NLPIR之关键词提取

BigdataBBS.com

NLPIR之大数据聚类


```
<?xml version="1.0" encoding="gb2312" standalone="yes" ?>
- <LJCluster-Result>
  <clusnum>200</clusnum>
  + <clus id="0">
  + <clus id="1">
  - <clus id="2">
    <feature>甲型 流感疫苗 流感病毒 流感确诊病例 流感疫情 流感病例 重症病例 接种疫苗 季节性流感</feature>
    - <docs num="150">
      <doc>多国H1N1流感疫情呈严重态势 美国疫苗供不应求</doc>
      <doc>卫生部就甲型H1N1流感疫情防控及疫苗预防接种工作答问</doc>
      <doc>甲型流感疫情在多国呈现严重态势</doc>
      <doc>中国报告4例接种甲流疫苗后死亡病例(实录)</doc>
      <doc>日本拟撤销对孕妇接种流感疫苗限制</doc>
      <doc>卫生部召开甲型流感防控疫苗接种通气会</doc>
      <doc>全球甲型流感疫苗接种进入高潮</doc>
      <doc>沪上近7万人已接种疫苗 天气转冷甲流危险性将累加</doc>
      <doc>对鸡蛋过敏者不宜接种甲流疫苗</doc>
      <doc>钟南山：甲流疫苗副作用小 学校应优先接种</doc>
      <doc>甲流疫苗答问公布:对鸡蛋过敏者不宜接种疫苗</doc>
      <doc>撩开甲型H1N1流感神秘的面纱:寻踪 变异 应对</doc>
      <doc>世卫公布甲型H1N1病毒调查：新病毒 人类无免疫</doc>
      <doc>福建、山东发生聚集性甲型Ｈ１Ｎ１流感疫情</doc>
      <doc>北京首现甲流死亡病例：北航军训团1名新生死亡</doc>
      <doc>流感疫苗对孕妇及胎儿免疫率达90%</doc>
      <doc>甲流可控可防可治</doc>
      <doc>首批甲流疫苗已运抵澳门 高危族23日起优先接种</doc>
      <doc>新疆报告1例甲流死亡病例 3类人群禁止接种疫苗</doc>
      <doc>法专家称接种甲型流感疫苗有助长期防疫</doc>
      <doc>接种疫苗仍是最有效手段</doc>
      <doc>美国甲流疫苗严重短缺 感染人数以百万计</doc>
      <doc>专家提醒:注射甲流疫苗后也应注重个人防护</doc>
      <doc>欧洲卫生专家称：甲流可能导致欧洲4万人丧命 </doc>
      <doc>成都严禁以防控甲型流感为由推销其它疫苗</doc>
```

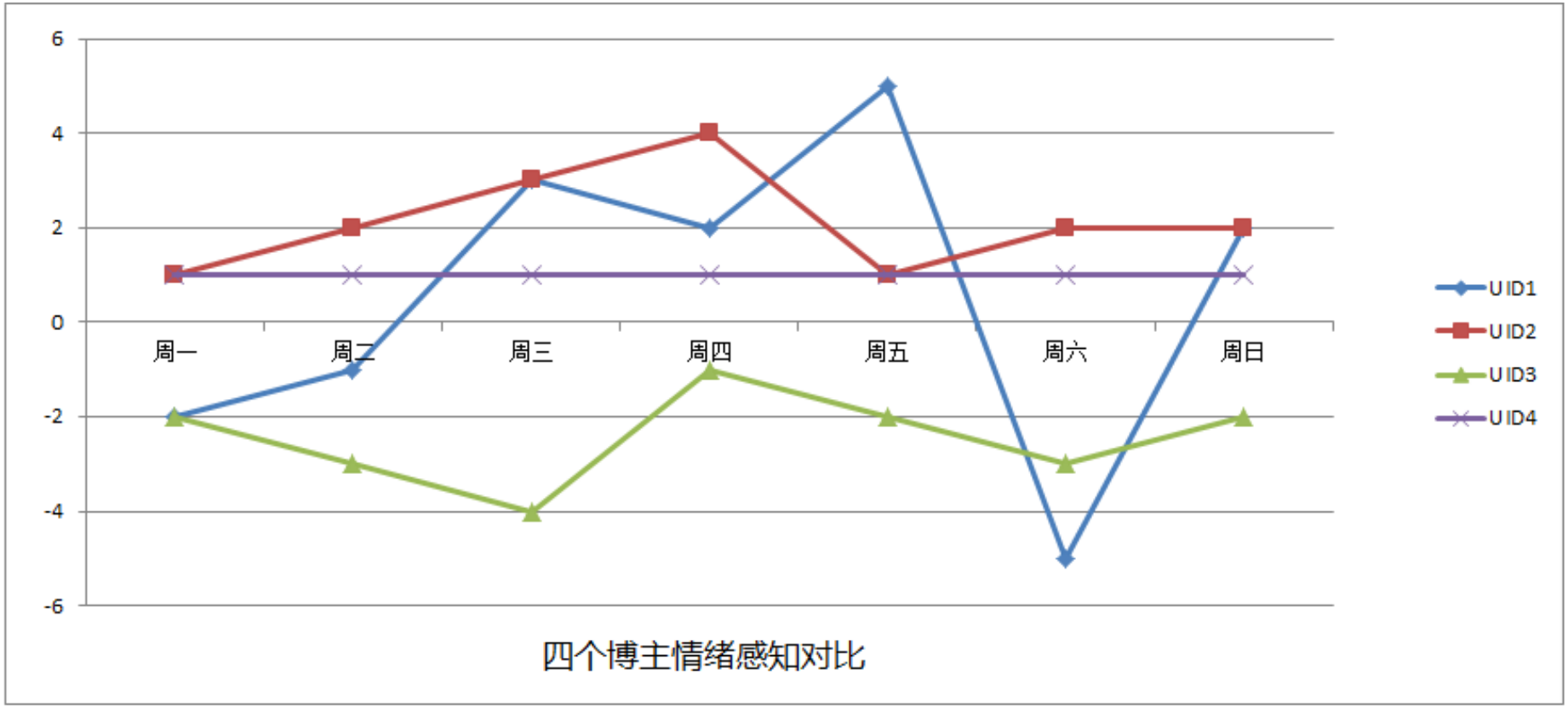

大数据论坛 BigdataBBS.com

大数据应用案例：用户画像

利用每日新增的网络访问日志数据500G，采用大数据挖掘技术，提供商情精准营销增值服务。

大数据论坛
BigdataBBS.com

微博博主情绪感知



大数据论坛
BigdataBBS.com

价值观挖掘：一言一行背后的机理

id	昵称	身份	pid		仁爱向量	传统向量	刺激向量	安全向量	成就向量	普世向量	权力向量	自我定向向量	遵从向量
116	孙健追求arete	国际关系学院	####	7	10	5	6	3	4	9	2	1	8
18	账号异常		####	2	10	4	5	6	3	8	7	1	9
39	IT疯云		####	2	10	1	4	7	5	8	9	6	3
53	全球热门新闻搜罗	新闻机构	####	1	10	2	4	8	6	7	9	3	5
40	海天5	基督教伯特利教	####	1	10	2	5	4	9	6	8	3	7
83	城管不好干		####	1	10	2	3	7	8	6	4	5	9
74	陈晓发	广东工业大学	####	8	10	1	3	4	9	6	5	7	2
2	喻国明	中国人民大学新	####	1	10	3	2	8	6	5	9	4	7
65	ICTCLAS张华平博士	张华平博士的微	####	1	10	3	4	8	6	5	7	9	2
5	小c是纯洁的猴子	西南石油大学二	####	9	10	2	4	6	8	5	1	3	7
68	林芙比	香港大学（200	####	6	10	2	9	3	8	5	1	4	7
103	小珍_QQ596859972	中医养生投资顾	####	2	10	1	7	8	9	5	4	3	6
95	我不叫成韦华	上海师范大学	####	5	10	3	6	1	3	4	9	8	7
19	李承鹏	记者、评论员	####	8	10	2	6	9	3	4	7	1	5
121	雙低青年	司法腐败严重则	####	5	10	1	8	6	3	4	7	2	9
78	Queen奇琦琪		####	6	10	4	9	1	5	3	2	8	8
30	东营日报社徐艺菡	东营日报社 东	####	5	10	2	6	9	8	3	1	4	7
92	Valder_	广州市商贸职业	####	4	10	1	8	6	7	3	9	5	2
62	青菜小玩子		####	6	10	3	4	9	8	2	7	1	5
70	mama咪吖	华南农业大学	####	8	10	7	9	2	5	1	3	6	4
26	何兵	中国政法大学教	####	1	9	4	2	8	6	10	5	3	7

大数据论坛
BigdataBBS.com

黄金眼Web大数据搜索与挖掘云平台

改革 ○ 关键词 ○ 人物 ○ 地点 ○ 机构 ○ 媒体 ◉ 作者 分析

· 事业单位养老改革再推进

当前用户：admin　　首页　设置　修改密码　注销

财经 CAIJING

新媒体数据资源管理中心
大数据智能搜索与挖掘平台

智能搜索：　搜索

财经首页　财经搜索　财经导航　财经分类　财经聚类　平台挖掘　我的工作区

《财经》十五年

每一个词都是一个故事
一段历史

· 地产商“基金”冲动

纲要

非结构
大数据

I 社会化新媒体

II 非结构化大数据

III 大数据搜索与挖掘关键技术

IV 新媒体分析实战案例

大数据论坛
BigdataBBS.com

大数据[illegible]解读

- 2008.5.20台湾地区[illegible]

  我们55；台湾47；民主2[illegible]全球11；社会11；中华民[illegible]平8；未来7；历史7；经济7[illegible]5；宪政5；国家5；价值5；[illegible]

- 2012马英九就职演说

  我们80；台湾42；两岸21；国际19；经济19；文化18；产业16；未来14；社会13；中华民国12；发展12；民主11；人民11；国家10；自由10；成长10；竞争力9；和平9；政府8；人才8；关系8；人8；创造8；开放8；安全7；我国7；推动7；环境7；机会7；司法7；支柱7；幸福7；合作7；大陆6；宪法6；全球6；政策6；世界6；公民6；改革6；投资6；民众6；价值6；

华商论坛 BBS.HSW.Cn
BigdataBBS.com

大数据挖掘：甄嬛体？汪峰体？

有一位网友统计了汪老师在大陆发行的9张专辑共117首歌曲的歌词

同一词语在一首歌出现只算一次。形容词，名次和动词的前十名分别是（词语后面的数字是出现的次数）：

	形容词		名词		动词
0	孤独：34	0	生命：50	0	爱：54
1	自由：17	1	路：37	1	碎：37
2	迷惘：16	2	夜：29	2	哭：35
3	坚强：13	3	天空：24	3	死：27
4	绝望：8	4	孩子：23	4	飞：26
5	青春：7	5	雨：21	5	梦想：14
6	迷茫：6	6	石头：9	6	祈祷：10
7	光明：6	7	鸟：9	7	离去：10
9	理想：6	8	瞬间：8	8	再见：9
9	荒谬：5	9	桥：5	9	埋：6

再来一个，比如某人的生日19820307：**自由，桥，再见，迷惘，生命，死，孤独，鸟**

润色一下：

站在通向自由的桥上

再见了，迷惘的生命

犹如死亡般的孤独

将不在桎梏这只小鸟

大数据论坛 BigdataBBS.com

风 = [地点]{将}有[级
[地点]{将}有[级
[地点]有[级数]
[地点]{将}有[级
[地点]风力[级数

雨雪 = [地点]{将}有

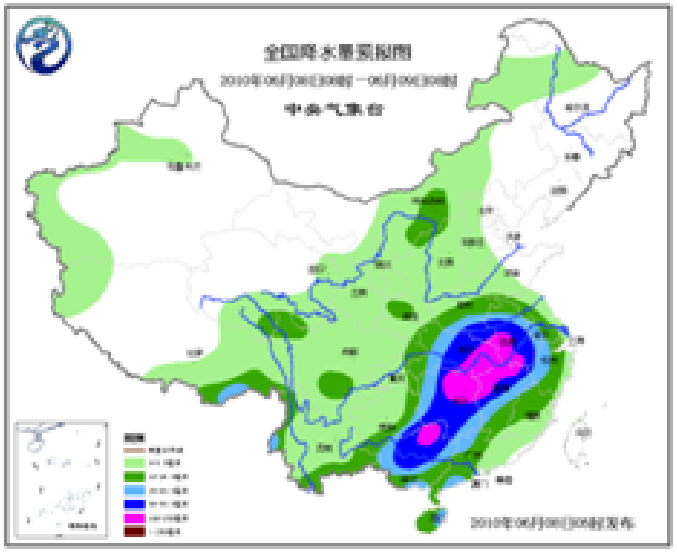
气温 = {时间}[地点]
{时间}[地点]
{时间}[地点]
{时间}[地点]

能见度 = {时间}[地点
{时间}[地点
{时间}[地点
{时间}[地点

时间 = 今天白天到夜
今天下午到夜
今天白天,
今天夜间,
今天早晨,
今天早晨到上
[DAY]日开始,
[DAY]日起,
[DAY]日早晨和

未来三天，青藏高原中东部及其以东大部地区自西向东、自北向南将有一次明显降雨过程，其中，湖北大部、河南南部、安徽、江苏西南部、浙江西部、福建大部、江西、湖南、广西、广东中西部、海南等地有大到暴雨，局地有大暴雨。另外，西藏东南部、云南西北部和东南部的部分地区有大雨。上述部分地区并伴有短时雷雨大风或冰雹等强对流天气。

具体预报如下：



8日08时至9日08时，新疆西部、西藏东部和西部的部分地区有小到中雨（雪）或雨夹雪，西北地区中东部、内蒙古中西部和东北部、黑龙江西部、山西、河北西部、黄淮大部、江淮、江汉、江南、华南大部以及西南地区等地有小到中雨或阵雨，其中，**西藏东南部、云南西北部、河南南部、湖北大部、安徽大部、江苏西南部、浙江西部、江西中北部、湖南大部、贵州东南部、广西大部、广东西部、海南东部等地有大到暴雨，其中，湖南东北部、江西北部、湖北东部、安徽西南部、广西东北部等地的部分地区有大暴雨**；上述部分地区还将伴有短时雷雨大风或冰雹等强对流天气。新疆西部和北部、甘肃西部、宁夏、内蒙古中西部等地有4~6级风，北部海有5~6级偏南风。

风或冰雹等强对流天气}

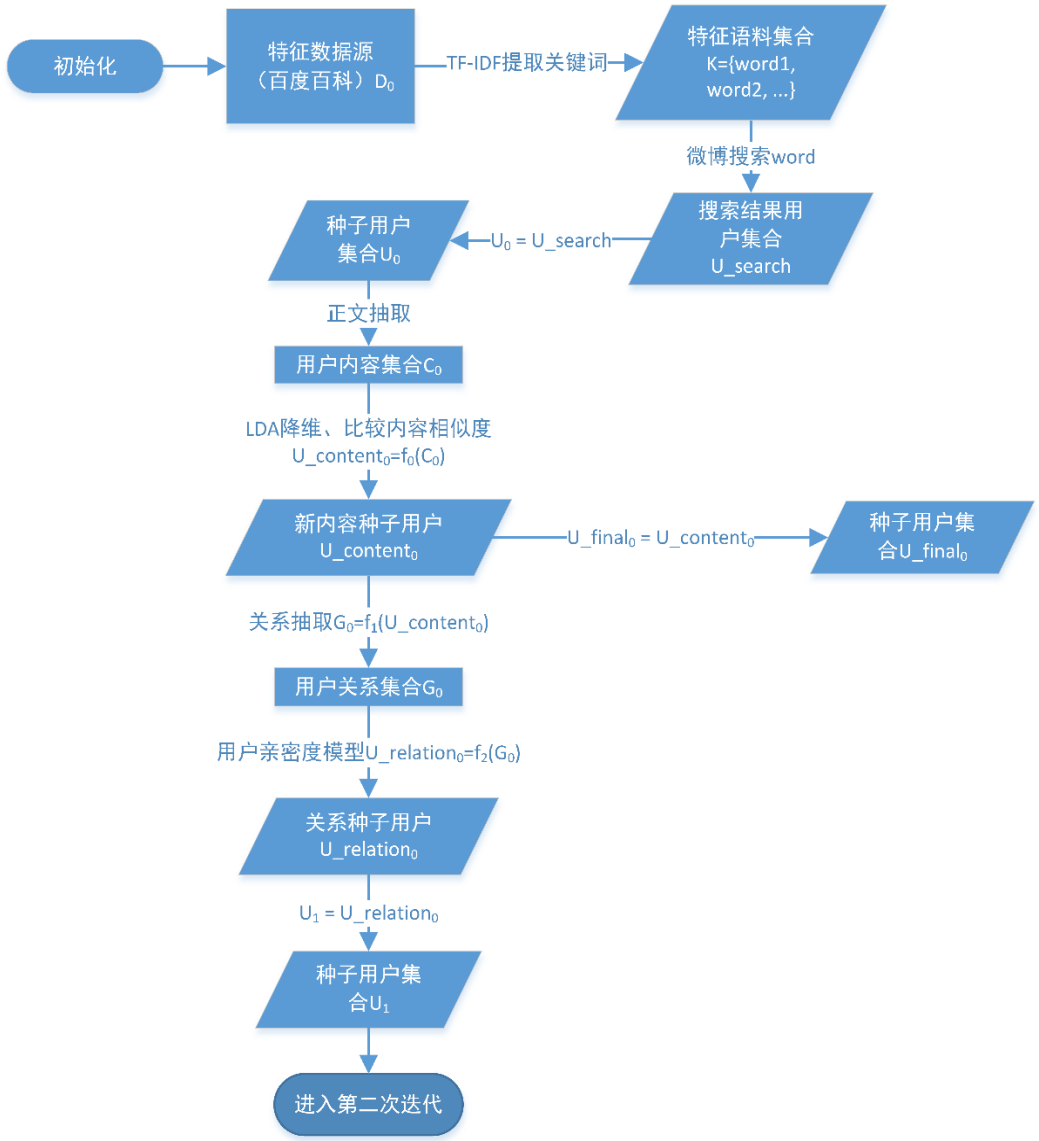
失独老人发现项目

本文选择对“失独”这一主题进行实验分析。因为家中唯一的子女不幸离世，这样的家庭被称为“失独家庭”。家中的老人即被称为“失独老人”。

通过在微博平台上寻找与“失独”相关的群体，合理地检验模型的有效性，并且结合微博文本分析方法和关系分析方法对这一特定群体进行案例分析，从数据的角度对案例进行分析解释。

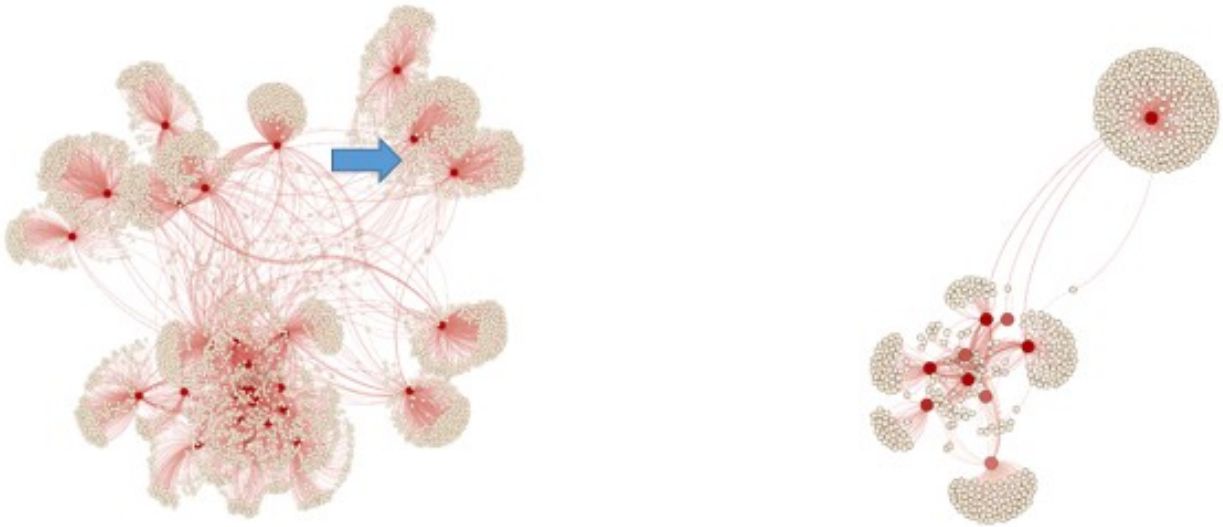
大数据论坛
BigdataBBS.com

失独老人发现算法流程



大数据论坛 BigdataBBS.com

失独老人算法迭代演变过程



语义种子用户的关系网络的迭代演变过程

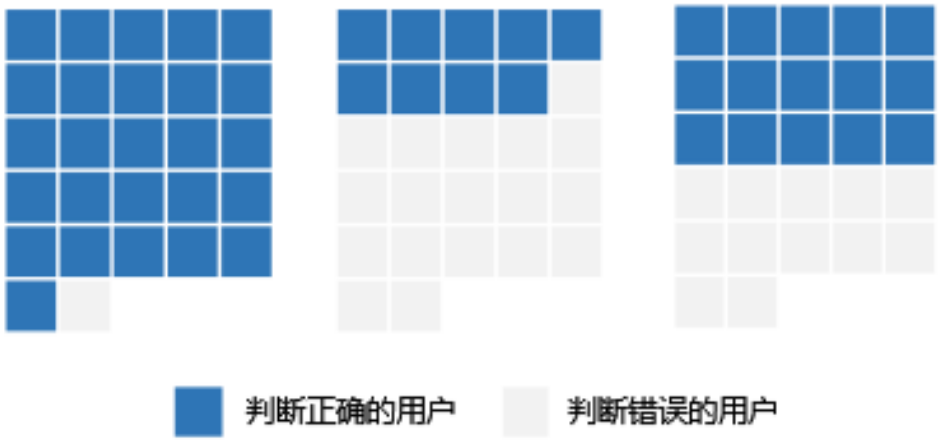
用户关系网络描述表

	第一次迭代	第二次迭代	第三次迭代	第四次迭代
节点数量	227580	12410	5490	637
边数量	296111	18536	7655	885

大数据论坛
BigdataBBS.com

失独老人发现算法评测

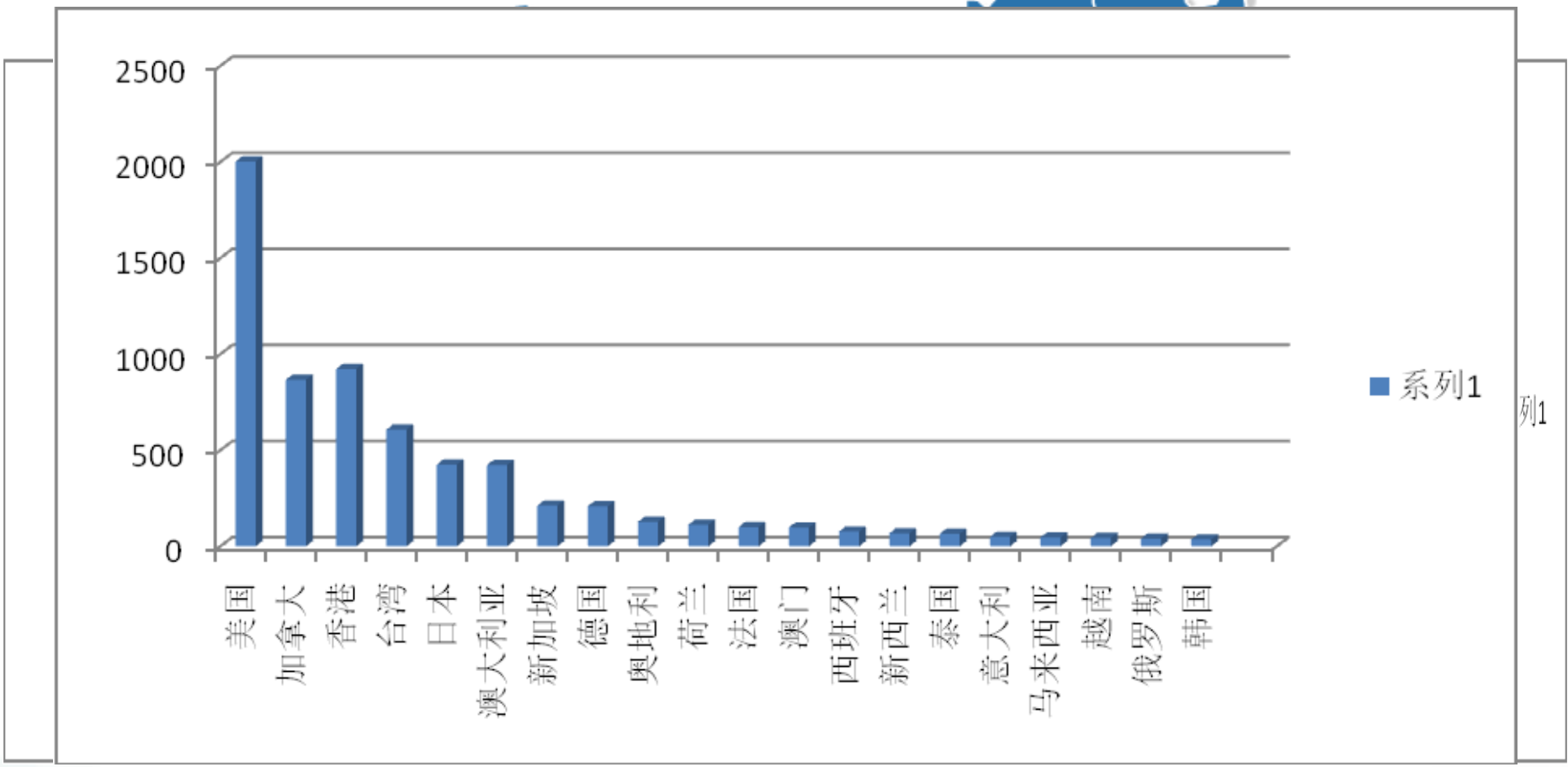
	特定群体发现模型（I）	搜索发现算法（II）	改进的搜索发现算法（III）
正确的用户数量	26	9	15
正确用户占比	0.96	0.33	0.56



大数据论坛
BigdataBBS.com

“张灵甫”事件的新媒体传播分析

“张灵甫”事件的新媒体传播分析



大数据论坛
BigdataBBS.com

所有参与者的观点分析

数据论坛
gdataBBS.com

草根的观点分析

Copyright 2015

大v的观点分析

张灵甫
鉴定费
技术少将
历史
没有
公知
颜值
厚葬
傻逼
抗战老兵
脑残
抗日
将领
铠甲
革命烈士
抗战
英烈祠
英雄
国民党
汪记国民党
并非抗日名将
抗日
名将
杀妻
戏子
历史学者
内战
问题
进攻解放区
尸还魂
少将
共产党
旅级
媒体
职务军衔
解放军
战场
国防部
少将旅长
高级将领
马谡
真相
部队
评论
师长
果粉
文章
纪念
参加
遗骨
发现
大陆
团支书
长沙会战
常德会战
争议
回复
军队
文艺少将
民国
羊圈
军事
就是
父亲
儿子
基层军官
山东媒体
战斗
人民
应该
上海
纪念章
重量级勋章
政府
新闻
正面战场
解放军
农户
军人
进入
抗日系统
军衔
将领
作战
洛阳铲
认为
抹黑
跳梁
大捷
朋友
作者
可能
需要
村民
战争
评价
抗日军
抗日
抗日
抗日

媒体的观点分析

Copyright 2015

tagxedo.co

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
微信：drkevinzhang
公众号：大数据千人会
欢迎大家访问大数据搜索与挖掘实验室官网：http://www.nlpir.org

大数据论坛
BigdataBBS.com