

基于知识图谱的 通用数据价值洞察平台

DataExa创始人 洪万福

定义：实体及其之间的关系图。

本质：由知识组成的语义网络。

目的：让用户能够更快更简单的发现新的信息和知识。





搜索引擎、问答系统



金融反欺诈



反恐、情报分析



精准营销

数据模型：更适合复杂、互联性、低结构化的数据。

信息检索：精确度、完整性、关联度、智能化程度更高，互操作和用户体验更好。

分析能力：基于关系和图的分析能力更强。

更好地表征和计算这个多维的世界



互联网、移动互联网



引擎



人机交互

语义分析

可视化构建

SNS分析

推荐引擎

图谱数据

机器学习

规则引擎

图数据库

并行计算



可视化适配



分析平台



行业数据资源

数据源

数据库

文本、语音

图片

视频

手段

NLP

规则引擎

机器学习

人工

知识类型

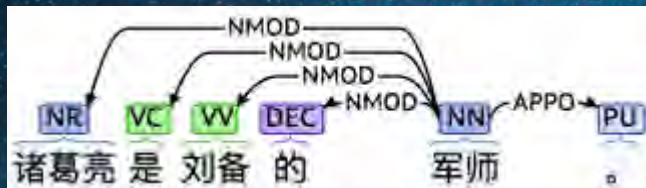
实体


属性


标签

关系

PERSON PERSON
诸葛亮 是 刘备 的 军师 。



 命名实体识别

 句法分析

 文本挖掘（主题模型等）

 关系提取

正则表达式

```
/^(\w)+(\.\w+)*@(\w)+((\.\w+)+)$/
```

规则引擎

```
if user.consumption.total>=200000 then  
user.type='VIP'
```

网页信息抽取

XPath : `//div[@class=summary]/p`

CSS : `("div h1 a", "innerHTML")`

混合

NR师从NR => {2}-老师->{1}

DataExa Insight 数据洞察平台 v2.1

DataExa Insight > 项目管理 > ml1b-rdd例子 > 编辑流程

交通拥堵预测

保存 运行 其他流程

- 管理操作
- 基本形状
- 数据加载
- Dataset数据操作
- RDD数据操作
- 数据持久化
- 数据类型处理
- 特征工程
- 神经网络
- 分类
 - 逻辑回归分类
 - 朴素贝叶斯分类
 - 决策树分类
 - 随机森林分类
 - 梯度提升决策树分类
 - 多层感知机分类
 - OneVsRest分类
- 回归分析
- 聚类
- 推荐引擎
- 模型评估



```
graph TD; Start((开始)) --> Merge(( )); FileRDD[文件(RDD)] --> Merge; Merge --> FileLibsvm[文件(libsvm)]; FileLibsvm --> Split[Split]; Split --> Train[随机森林分类模型训练]; Train --> Predict[随机森林分类预测];
```

节点信息

名称: 随机森林分类模型训练

描述:

参数

rdd: \${output}_1[2]

classes: 5

trees: 5

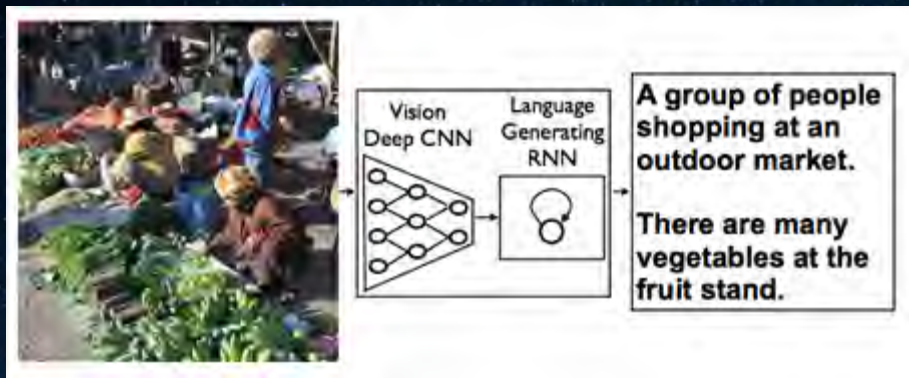
featureSubsetStrategy: auto

maxDepth: 5

maxBins: 8

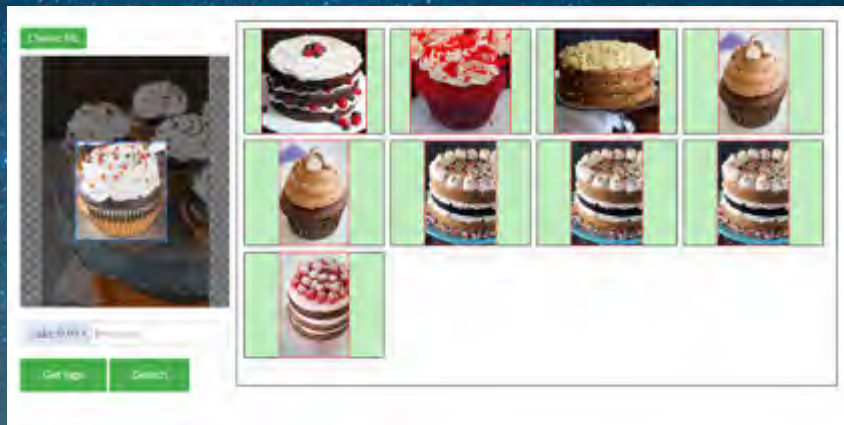
seed: 10

设计 代码 评估



图片内容解析

图片相似度



WordNet管理

分词名称: 查询

词性: 词类:

描述:

最佳线性无偏预测(best linear unbiased prediction, 简称BLUP), 又音译为“布拉普”[1], 是统计学上用于线性混合模型对随机效应进行预测的一种方法。最佳线性无偏预测由C.R. Henderson提出。随机效应的最佳线性无偏估计 (BLUP) 等同于固定效应的最佳线性无偏估计 (best linear unbiased estimates, BLUE) (参见高斯-马尔可夫定理)。因为对固定效应使用估计一词, 而对随机效应使用预测, 这两个术语基本是等同的。BLUP被大量使用于动物育种。

近义词

最佳线性无偏估计 布拉普 BLUP	名称: <input type="text" value="最佳线性无偏估计"/> 权重: <input type="text" value="1"/>
保存	新增
删除	

反义词

相关词

线性混合模型 随机效应	名称: <input type="text" value="线性混合模型"/>
----------------	---

同义词
近义词
反义词
词向量
词距离
情感度

.....



表达式设置

增加/删除	类型	类型名称	默认值	默认表达式
+	属性	title	未知标题	<code>{title}</code>
-	属性	summary	未知摘要	<code>{div[@class=summary]}</code>
-	属性	body	位置正文	<code>{div[@class=body]}</code>

取消 确定



实体对齐

实体距离

上下文特征选取

冲突解决

图模型

向量空间模型

排序学习方法

四维分析 {
实体
关系
时间
空间

图分析 {
关系挖掘
路径提取
最短路径分析
关联度分析
出入度分析
社区发现
强连通分量
.....

自然语言

王菲的前夫的女儿

内置函数

relationship(诸葛亮).depth(2).limit(20)

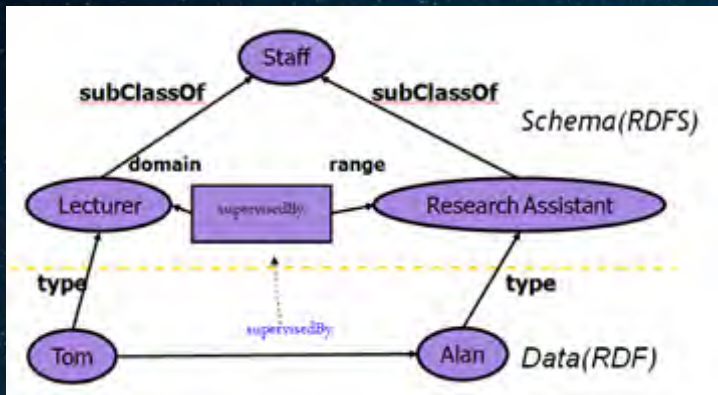
图查询

```
m=[:]  
g.v(1).out('likes').in('likes').out('likes').groupCount(m)  
m.sort{-it.value}
```

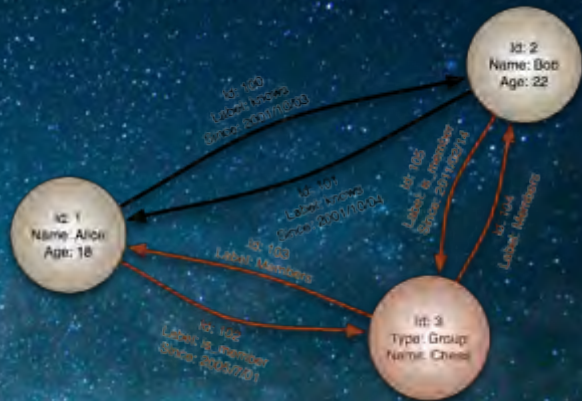
推理器

2016年A股主板发生的并购案

RDF



PGM





Neo4j



Titan



Jena



机器学习

Representation Learning

数学运算

order[2015-2016].sum

规则引擎

```
if user.consumption.total >= 200000 then  
  user.type = 'VIP'
```

自定义

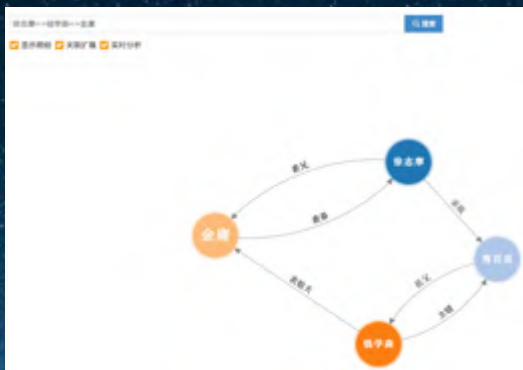
亲戚关系推理器

根据图谱查询结果进行可视化展示的样式适配

实体检索



关系搜索



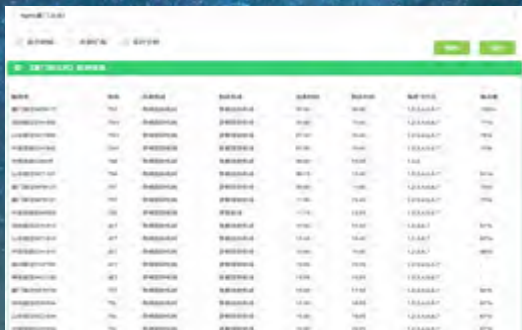
范围搜索



精准问答



列表结果



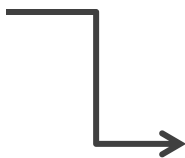
名称	类型	所属公司	发行日期	发行地区	发行语言	发行数量
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	中国大陆	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	中国香港	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	中国台湾	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	马来西亚	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	新加坡	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	泰国	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	印度尼西亚	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	菲律宾	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	越南	普通话	1000000
仙剑奇侠传三	电视剧	上海仙剑网络科技有限公司	2009年6月20日	柬埔寨	普通话	1000000

图表展示



实用的智能语义计算平台

DataExa-Sati是一个高效实用的语义计算平台，基于自然语言处理、图存储计算、问答系统等技术，提供一套成熟的行业知识图谱构建体系，深度挖掘海量非结构、半结构化数据背后隐藏的价值。



- 基于知识图谱的语义分析；
- 模糊语义识别，智能关联相关内容；
- 支持语境化的人机对话；
- 内置丰富的查询函数，简化复杂的查询逻辑；
- 查询结果输出和可视化适配，智能匹配用户最终
- 丰富的数据展示，可视化手段

特性



支持多源异构知识提取；



支持中文和英文；



基于图的存储和计算机制；



面向领域的词网和图谱自定义管理；



模糊语义识别，智能关联相关内容；



内置丰富的查询函数，简化复杂的查询逻辑；



快速分析数据实体在时间、空间下的关联性和因果性；



丰富的数据展示和可视化手段；

应用层

语义搜索

人机对话

分析研判

行业图谱

知识图谱层

知识融合

知识表征

搜索查询

逻辑推理

图谱管理

实体管理

关系管理

系统管理

数据管理

集体管理

算法管理

模型管理

智能组件层

DataExa-NLP

DataExa-Insight

DataExa-Taqforge

DataExa-Octopus

DataExa-RE

DataExa-Zion

DataExa-SNA

DataExa-SatiCore

基础服务层

Hadoop

Spark

Tensorflow

DataSet

图数据库

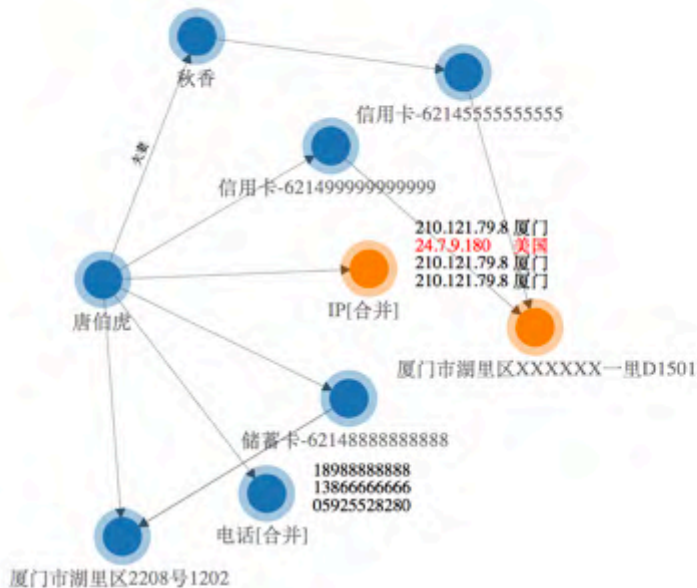
图计算

GPU

CPU

关键节点时间轴

- 现在**
- 2014/5/8 购买保险
ORD2014050810287
- 2014/2/20 [流失风险] 大额转账:
ORD201402207
- 2013/10/15 信用卡额度提升
ORD2013052112235
- 2013/5/21 购买理财产品
ORD201305212235
- 2013/3/2 申请信用卡 621499999999888
- 2012/8/5 申请房贷 ORD20208051211
- 2012/2/10 申请储蓄卡 621488888888888
- 过去**



异常分析[2016.6.1 - 2016.6.30]

- 信息 [厦门市湖里区XXXXXX一里D1501] 被他人使用
- 手机银行登录地址异动
- 总体评分下降12分 [960 -> 948]

用户画像

姓名: 唐伯虎
证件: 3505XXXXXXXXXXXXXX

90后 已婚

产品信息

储蓄卡: 6214888888888
信用卡: 621499999999999
理财产品: ORD201305212235
保险: ORD201405081028

消费行为-银联



商户	金额
厦门市老知青餐饮	¥199.99
厦门市佳丽海鲜海鲜餐厅	¥586.45
福州三七巷周大福珠宝	¥4890.00
北京京友快捷酒店	¥289.95

消费行为-电商



商户	金额	商品
J.com	¥19.99	大米

导出 打印 快照 | 标注 备注 关联 | 视图 统计 显示比例 | 帮助

实体



相关

搜索相关...

关系

其他

- 父亲
- 丈夫
- 第四个儿子
- 岳父
- 侄女

路径分析 最短路径 关系挖掘 关联度 出入度 过滤 排列 高级搜索

本-拉登



属性

照片



中文名

奥萨马·本·穆罕默德·本·阿瓦德·本·拉登

职业

政治“阿尔-伊达”大本营创始人,政治基地组织首领

星座

白羊座

别名

乌萨马·本·拉丹

宗教

伊斯兰教

出生地

沙特阿拉伯吉达市

籍贯

沙特阿拉伯

标签

世界罪犯

人物

伊斯兰教人物

企业级的可视化机器（深度）学习平台

DataExa-Insight是一个简单易用的大规模机器学习平台，通过集成行业成熟的机器学习框架，提供可视化建模、配置化参数、流程化操作、模板化任务等功能，大大降低客户进行数据价值洞察的成本。

- 关注业务全局，避免价值孤岛；
- 建立全景业务图谱，深度挖掘潜在价值；
- 无代码、灵活的可视化建模和分析流程设计；
- 强大的分布式智能计算引擎；
- 包括自然语言、推理引擎、机器学习等；
- 丰富的数据展示、可视化手段。

特性



支持多源异构数据集成；并行化分布式的数据处理组件，支持清理、替换、组合、采样、去重、拆分等数据预处理操作；



支持Spark、Tensorflow多个机器学习深度学习（或混合）平台；



超细粒度的配置化算子；提供多个分类、聚类、回归、主题模型、推荐算法，同时支持前沿的深度学习、在线学习、贝叶斯推荐等算法；



可视化和模板化的机器学习操作，降低使用难度；提供在线模拟功能；



支持YARN、Mesos、GPU等计算集群；



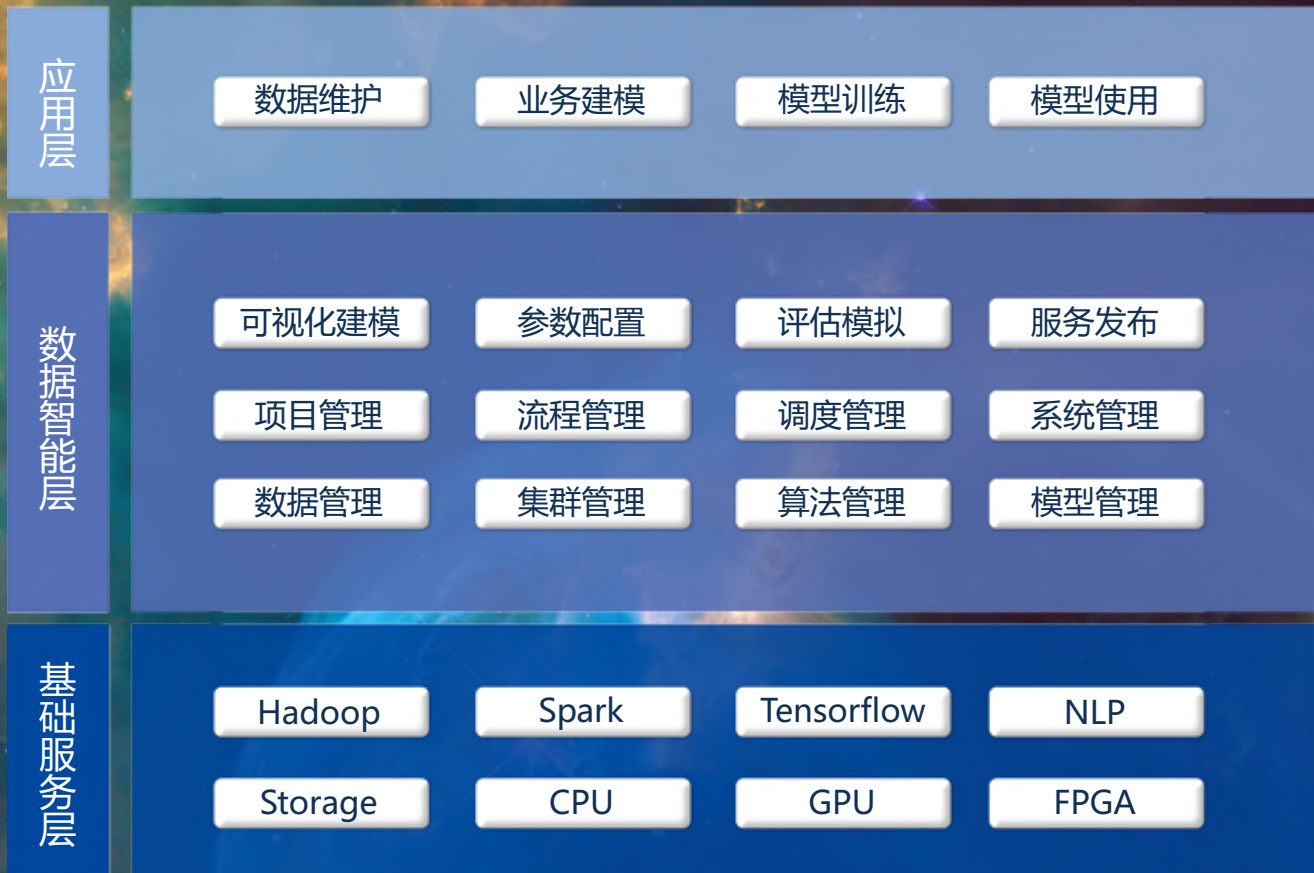
细粒度的权限体系和安全审核控制；



机器学习全过程细粒度监控；



多类型的机器学习服务发布组件；



宇宙的本质是计算。

感谢观看，敬请指导!

