# 科技项目数据挖掘决策架构

2017年10月18日

## 一、概述

本系统采用数据挖掘、自然语言处理和文本处理技术对科技申报项目进行知识挖掘。科技申报项目主要指项目申报书（以pdf和word为主）和数据库专家信息等数据。旨在挖掘统计历年项目的**项目实体**、**科研单位实体**、**项目科研人实体**、**经费实体**、**区域实体**、**评审专家实体**和**词频关键词实体**七大实体。进而统计分析以供决策参考。

## 二、详细挖掘目标

【1】**项目实体特征**：项目名称|项目所属领域|项目申报单位|项目负责人|项目归口|项目起止年限|项目经费预算|项目立项经费|项目结题|申报年度|项目总参与人数

其中，项目立项经费和项目结题需要从数据库提取。针对项目实体特征提取后，主要分析年度项目申报总数量和同比、各领域项目数总数和同比、各单位申报项目数量和比例、各归口项目数量和比例、总体经费预算和立项经费对比、结题率统计、各年度项目参与人数与环比

【2】**科研单位实体特征**：科研单位名称|科研单位机构代码|科研单位所属区域|科研单位性质|科研单位归口|科研单位成立时间|科研单位注册资本（企业）|科研单位研发人员数量|科研单位总资产|科研单位负债|参与合作单位|年度

本实体主要分析各年度科研单位占比和同比、各区域科研单位数量和对比、科研单位性质占比、各归口科研单位量和对比、不同年度区间科研单位占比、不同资本区间科研企业占比、各年度科研人员参与量和同比、科研单位总资产占比、科研单位负债率、各年度总科研单位数和同比。

【3】**项目科研人实体特征**：申报人名称|申报人性别|申报人出生年月|申报人单位|申报人学历|申报人专业领域|申报人职称|参与人名称|参与人性别|参与人出生年月|参与人单位|参与人学历|参与人专业领域|参与人职称|申报年度

本实体对象主要是申报人和参与人相关信息统计，包括总参与人数和环比|性别分布|人员年度区间占比|人员学历占比|专业领域占比|职称占比等等

【4】**经费实体特征**：自筹费用|总经费|申请财政费用|（设备支出预算|材料支出预算|测试加工支出预算|燃料动力支出预算|差旅费支出预算|会议支出预算|国合交流支出预算|出版等支出预算|劳务支出预算|专用支出预算|其他预算）立项经费相关数据。

其中，立项经费需要从数据库提取。本实体主要分析各经费和总经费的占比、同比、环比情况。

【5】**区域实体特征**：区域项目数量|区域申报人数|区域申报经费|区域立项经费|区域申报单位数量|区域科研总人数|区域科研人数性别比例|区域科研人数学历比例|区域科研人数职称比例|区域结题情况|区域各项目类别占比

【6】**评审专家实体特征**：专家姓名|专家出生年月|专家单位|专家性别|专家区域|专家职称|专家领域|专家学历等等

其中所有专家信息均需要数据库采集，主要对专家情况进行占比、对比

【7】**实体和词频关键词实体特征**：研究目标文本|研究内容文本|研究国内外现状文本|研究应用文本|研究预期效益文本

对申报书主要内容进行处理，主要提取项目各特征的高频词和低频词，可以作为词库和热点研究对象。提取专业术语和关键词形成词库。提取各部分词频分布。提取词频词云。

## 三、技术路线和难点

### 3.1 技术路线：

1、本项目研发路线是对服务器申报书数据600G左右进行采集，然后将其转化为txt文本，将图片等转化的小文本剔除。

2、对所有大量文本进行格式化编码，以及分词和词性标注。

3、分析申报书提取特征，针对保留的词性进行保留（比如：申报人“张三”属于名词，保留下来。但是遇到“茶杯”等也是名词，存在干扰。），删除无关词性和停用词

4、指定提取特征词性表，依据词性表进行筛选保留文本信息。

5、对保留目标词性的干扰信息进行技术手段处理。

6、对特征文本，采用技术手段和工具，参照第二节【详细挖掘目标】进行处理

7、对数据库需要特征文本（比如：专家信息，立项经费等）信息提取合并

8、采用相关算法，提取挖掘目标数据单独存储

### 3.2难点

1 申报书格式准确转化txt问题，影响文本质量和结果

2 运行性能问题，大量数据，单机跑的性能跟不上

3 主要工作在文本数据预处理上，核心的干扰词性（名词、动词等）的排除

## 四、展望

展望一：科技行业实体和实体关系库构建

展望二：科技项目类别分类问题

展望三：科技热点分析和相关决策

展望四：科技项目查重系统构建