# 一、项目背景

高效、准确地获取和掌握信息是高质量战略策略制定的前提，近年来，军事单位对于更新频率高、多渠道来源的海量情报信息分析与知识挖掘诉求日益强烈。

随着自然语言处理、知识图谱等技术的不断成熟，军事场景海量信息的自动分析与挖掘应用也逐渐成熟。通过对文本进行深层理解与挖掘，深入语义层面抽取文本中包含的关键信息，为用户提供精准化的知识与情报服务。

信息抽取(Information Extraction，IE)相关的研究正是在这种背景下产生的，其主要目的就是将非结构化的自然语言文本转化成半结构化或结构化的数据，并以数据库形式存储，一方面可以用于对文本的快速阅读和理解，帮助人们更方便的获取知识，另一方面可以用于深入地挖掘分析，对本体构建、垂直搜索、自动问答等自然语言处理相关领域起着非常重要的作用。其中，领域实体抽取（Named Entity Recognition, NER）和领域关系抽取(Relation Extraction，RE)作为信息抽取的子任务和关键技术，是实现相关目标的必经途径。

知识图谱（Knowledge Graph），本质上是基于图的语义网络，表示实体和实体之间的关系。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系。在金融行业一般会把个人（法人），账户（案宗），公司等实体概念定义为知识图谱中的实体，把个人和公司，公司和公司，账户和账户之间的各种业务关系，如任职高管，投资，控股，增发，转账等关系来表征实体之间的关联关系。

在军事场景，行业数据整合与知识图谱构建同样势在必行。建议构建信息数据传导模型，需要行业专家和人工智能协作，以专家+大数据的方式构筑知识图谱生长模式，从而更好地利用行业海量各种结构化数据和非结构化信息。

# 二、项目目标

通过一定量文本语料数据的标注训练，输出信息与知识抽取的模型，主要包括实体抽取、关系抽取与事件抽取。实现自动对情报相关文本中重要人物、重要组织机构等主体以及主体之间关系的识别与关系网络构建；同时，实现对相关主体特定类型事件的发现与识别。

# 三、技术原理

本系统以知识图谱技术为核心，知识图谱技术主要分为构建和应用两部分，融合了机器学习、自然语言处理、信息抽取和检索，认知计算、知识表示和推理等方向的交叉应用形式。1994年图灵奖获得者，创建了知识工程这一人工智能领域重要分支的费根鲍姆定义知识工程为将知识集成到计算机系统从而完成只有特定领域专家才能完成的复杂任务。

在大数据时代， 知识工程是从大数据中自动或者半自动的抽取获取知识，在大数据时代，建立基于知识的系统，以提供智能认知计算。大数据对智能服务的需求，已经从单纯的搜集获取信息，转变为自动化的知识服务。我们需要利用知识工程为大数据添加语义/知识，使数据产生智慧（smart data），完成从数据到信息到知识，最终到智能应用的转变过程，从而实现对大数据的洞察、提供用户关心问题的答案、为决策提供支持、改进用户体验等目标。知识图谱在下面应用中已经凸显出越来越重要的应用价值：

-知识融合：当前领域内大数据具有分布异构的特点，通过知识图谱可以对这些数据资源进行语义标注和链接，建立以知识为中心的资源语义集成服务；

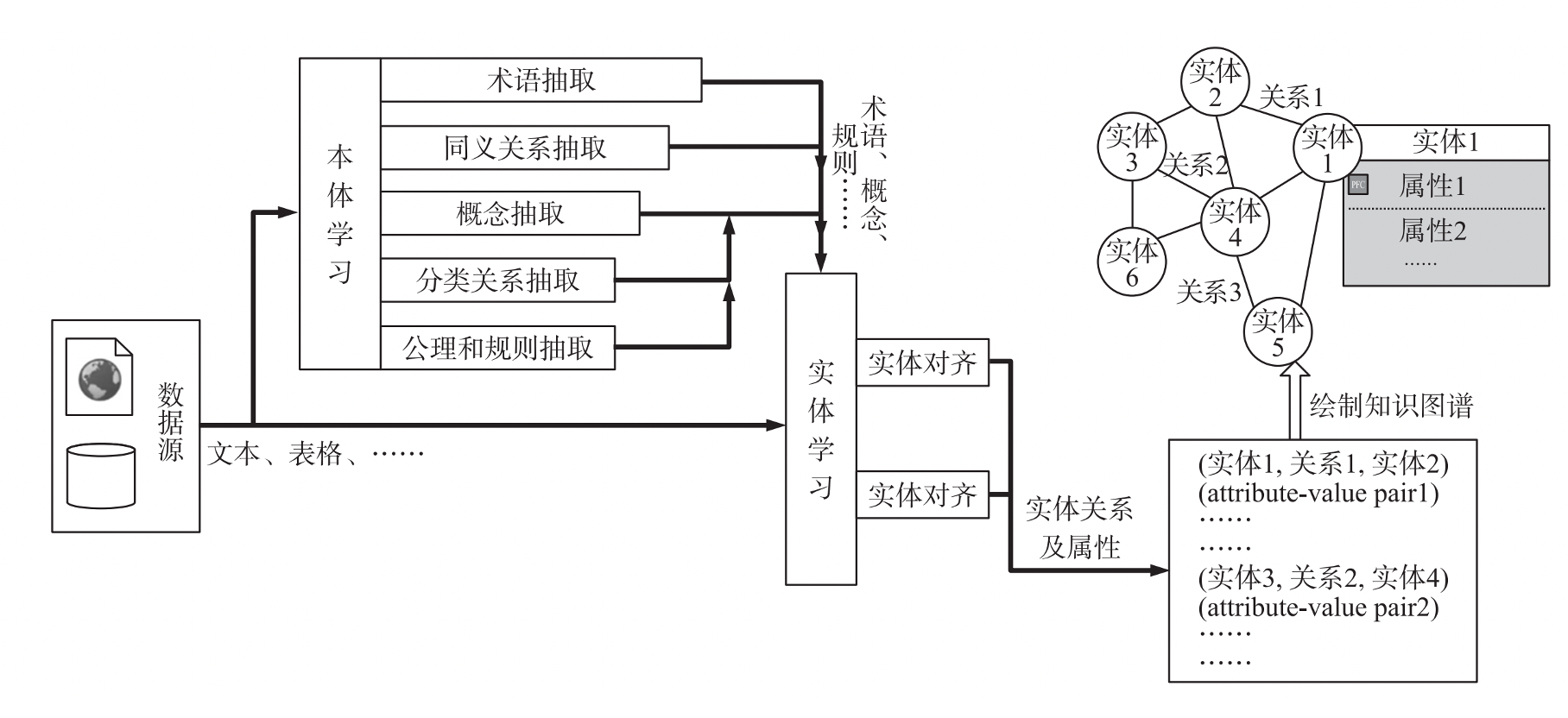
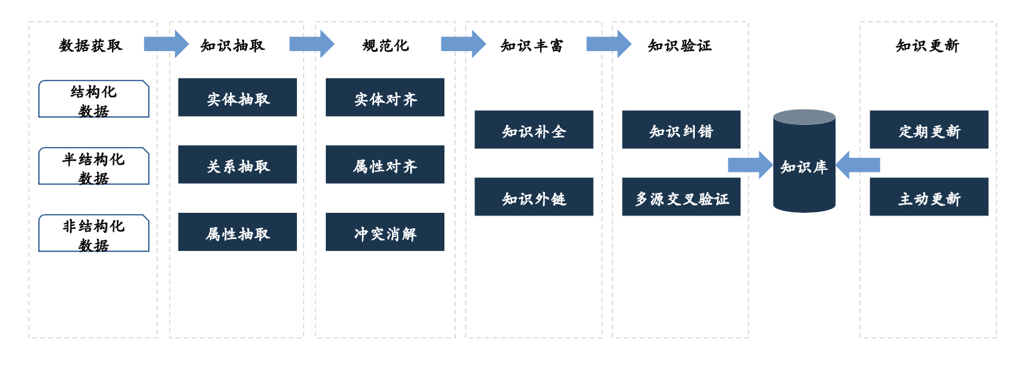
-语义搜索和推荐：知识图谱可以将用户搜索输入的关键词，映射为知识图谱中客观世界的概念和实体，如对公司控制人搜索结果直接显示出满足用户需求的结构化信息内容，而不是单纯的网页结果；

-问答和对话系统：基于知识的问答系统将知识图谱看成一个大规模知识库，通过理解将用户的问题转化为对知识图谱的查询，直接得到用户关心问题的答案；

-大数据分析与决策：知识图谱通过语义链接可以帮助理解大数据，获得对大数据的洞察，提供决策支持。

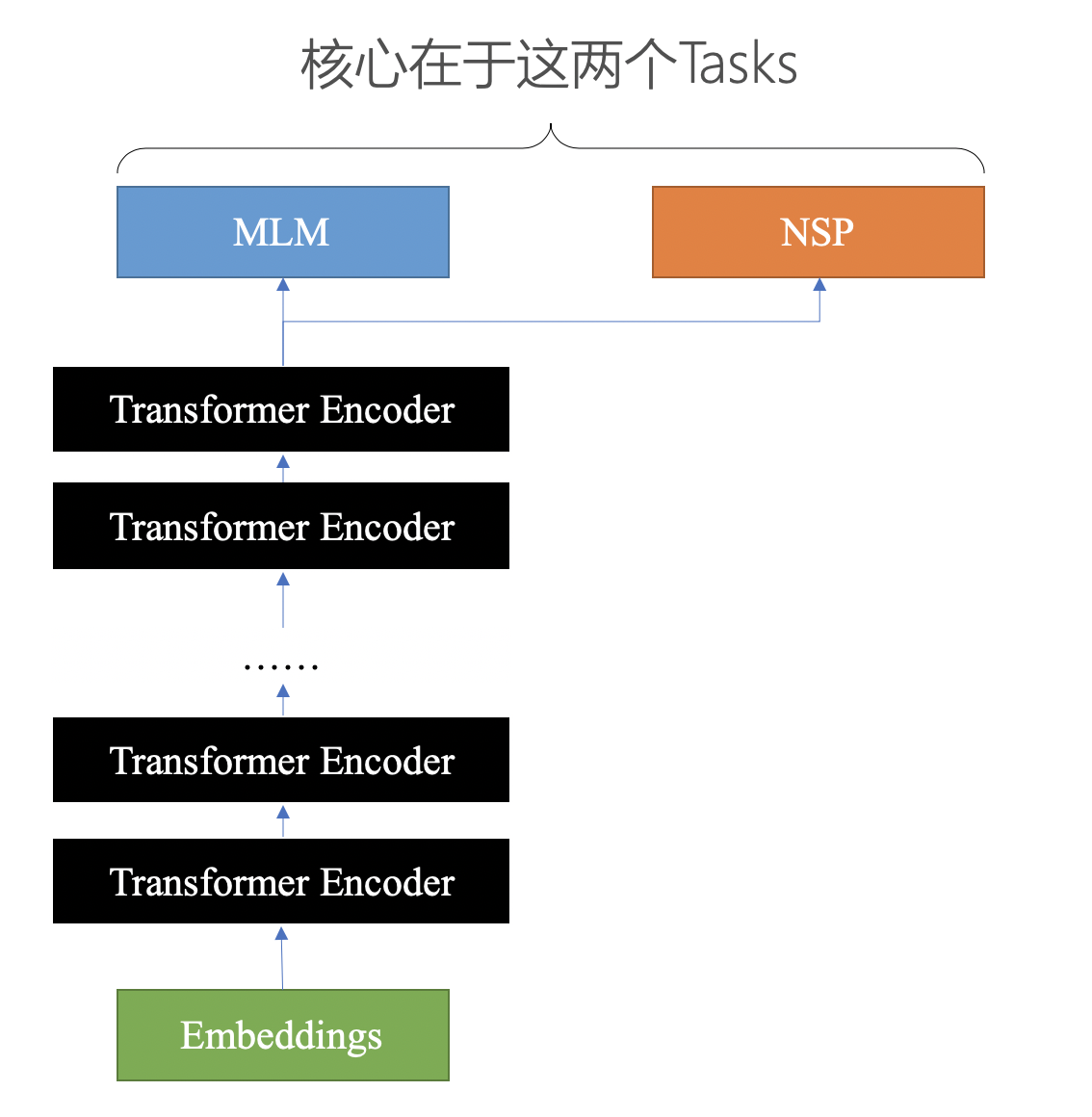
-关联实体特征建模： 例如金融知识图谱构建后可以进行公司维度实体的关联特征挖掘，以支持关联方受益方分析，最终控制人，交易筛查，产业链分析，同行业比较分析，单一舆情影响行业/范围分析，异常波动舆情关联和各项审核模型建模。利用知识图谱，传统的专家规则可以有更多的信息源输入，下游建模的模型也可以从规则和简单逻辑推理，升级改造成机器学习模型。

常规知识图谱构建流程如下图所示：



第四范式知识图谱构建所使用的核心组件包括：BERT、DropBlock、tri-gram、LAD等。

BERT主要用在我们的图谱构建平台中，主要包含MLM(Masked Language Model)和NSP(Next Sentence Prediction)两个任务。它是一个语言表示模型，它代表Transformer的双向编码器表示。与最近的其他语言表示模型不同，BERT旨在通过联合调节所有层中的上下文来预先训练深度双向表示。因此，预训练的BERT表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。



DropBlock是我们建模时一种防止过拟合的优化模块，dropout正则化方法广泛用于全连接层，且效果较好，很多由dropout引申的卷积网络正则化方法如DropConnect,maxout,StochasticDepth,DropPath等都是通过给卷积网络加入噪声来防止训练数据的过拟合，大部分成功的方法都需要噪声结构化。DropBlock与它们相比，更通用，可用在卷积网络的任意处。

Tri-gram（三元语言模型Trigram language model）主要用于分词时的优化，三元语言模型对于一元语言模型和二元语言模型都是一个巨大的进步。高阶 n-gram 对更多的上下文敏感，但是数据拥有更多的稀疏性。

LDA主要用于文本聚类，它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。

除了基于大量文本语料基础上的机器学习建模能力。为了解决知识图谱构建初始阶段的冷启动问题，第四范式知识图谱构建平台还支持基于规则引擎的专家加速中心。

在加速中心，一方面业务专家可以导入已有的数据白名单（如专业名词列表），通过批量导入的方式，可以快速高效地让机器识别已知的专业名词，将专家知识快速导入系统，当在语境中碰到这些词语时机器自动完成识别，可极大提高模型训练效率。另一方面，针对那些较规范明确的场景（如邮件收发人识别），我们提出了自然语言处理（NLP）+正则表达式的高级规则引擎技术。

通过高级规则引擎，可以支持纯文本输入，会对文本进行分词，命名实体识别，词性分析和句法分析，然后根据配置好的规则模版对文本进行正则匹配，从中抽取出需要的三元组。基于NLP规则引擎的非结构化数据抽取具有以下优势：

* 最大化的降低书写规则门槛，对于简单关系规则希望能够做到人人能理解，人人能写。
* 特定单个关系的规则增添成本可降低到最少
* 支持触发词选择，支持我们**独创**的词性过滤、实体类型过滤、句法依存过滤等筛选。
* 通过该引擎抽取的三元组，可以追溯到识别的具体规则和文本，具有很强可解释性，方便后续调试迭代
* 可以通过依存句法分析和词性，统计出一个领域数据集中存在最多的SPO和关系类型，方便决策该进行何种关系抽取

通过高级规则引擎，针对各种模式相对固定和规范的场景，快速完成知识抽取模板的定义，降低类似场景重复标注训练的工作量，将重点工作精力投入到语义模糊、语境不固定、多样性较高的样本处理过程中，极大提高整体模型训练效率。

第四范式知识图谱构建系统能力，主要包含数据接入模块、中间件模块、模型训练、模型预测、图谱生成以及上层服务模块，架构如下图所示：



# 四、系统能力

主要包含图谱配置、数据接入、数据处理、图谱构建、图谱查验、系统管理几大部分。

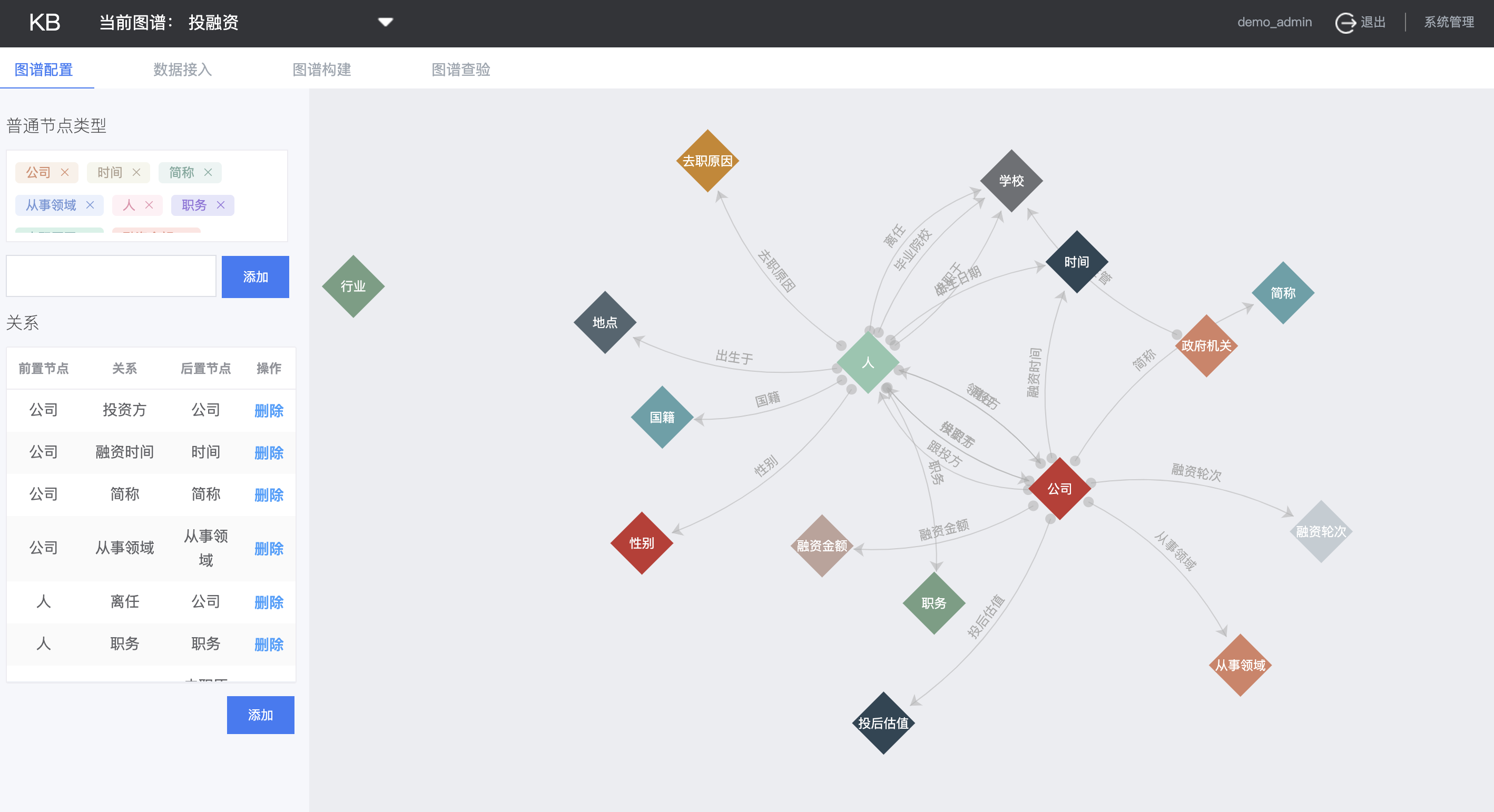
平台需要具备效果优秀且功能简洁、清晰的特点，用户通过系统可以将业务知识通过简单明确的操作直接转化成为机器识别模型，满足业务诉求。





##### 图谱配置

图谱结构即我们认知的结构，结合业务诉求针对性进行图谱结构的构建，主要包括节点类型配置、关系类型配置和事件配置。



#### 节点类型配置

我们关注的主体在图谱上体现为节点，节点有不同类型，通过配置可以完成对不同类型主体的定义。

#### 关系类型配置

在添加完节点类型后，需要指定两个节点直接的关系，每条关系由三部分组成，分别为头节点、关系、尾结点。不同类型节点之间的关系可能不同，可配置完成定义。

#### 事件配置

定义一个事件，定义该事件中各个要素之间的关系，如事件发生的时间、地点、人物、事件类型等。可灵活自定义事件结构。



##### 数据接入

支持多种数据源类型的接入和数据导入，包括但不限于本地上传、FTP、JDBC协议连接关系型数据库、流式数据接入等。兼容不同类型文件的接入，如csv、json等常见文件和数据格式。平台支持以任务的方式执行数据的引入，通过任务的管理来监控整个数据引入的过程，支持定时任务并能够按场景对数据集进行划分并能够完成高效的数据集检索。



具体能力说明如下：

* 支持通过浏览器对本地文件进行上传；
* 支持多种数据类型和数据文件格式：平台可以对接多种常见数据类型和常见数据格式，包括文本文件、CSV；

为了训练一个良好的模型，必须有质量高的标注数据集，高质量标注数据集需要具有包括足够多数据量、更丰富的类型、更精细的数据细节等。系统提供本地上传、HDFS上传或者FTP上传三种方式接入用户标注数据集，同时对已上传的标注数据，系统提供标注信息可视化展示，以及标注信息的修改功能，用户可快速针对上传标注数据进行标注查验与标注信息的调整，保障训练数据集的标注质量。对于上传的标注数据需满足以下需求训练效果可更佳：

######  更多细节

模型训练质量越好模型效果越佳，比如训练篇章识别模型，由于汉字的结构比数字和英文都要复杂，因此，为了体现更多的细节信息，需提供更加贴近真实使用场景的数据标注。

######  类型丰富

为保障模型更佳的泛化能力，能覆盖更多场景，用户在提供标注数据集时，更丰富的数据集可提供更佳的模型，如为了保证模型的实用性，用户需根据的使用场景，尽量多的提供业务场景更加常见的内容，最好针对业务场景进行相应的统计分析，统计各种数据类型的频率，对出现频率最高的相应场景提供更多的标注数据。

######  数据充分

提供更加丰富的数据，保障更佳的模型效果。如针对印刷体，甚至手写体，需要提供尽量全面的字体，如正规印刷的宋体、黑体、楷体等，不规范的手写体也需提供尽量全面的，基本上能够比较全面地覆盖业务识别数据。

######  数据标注规范

系统针对标注数据格式有格式要求，用户上传标注信息文件需统一采用JSON格式，平台支持的标注信息文件包括分类标注信息、文字标注信息等。

##### 3.数据处理

数据处理是将业务逻辑准确表达为数据定义的过程，决定后续的模型效果。数据处理繁琐耗时，为了帮助用户更轻松高效的完成任务，平台基于数据科学家多年的业界经验，将最常用、最有效的数据处理过程进行了产品封装，提供包括数据格式转换，数据表拼接、数据表的分割和抽样，以及数据统计分析等功能。

1）依靠系统功能，无需外部脚本，提供丰富的数据清洗工具,至少包括字符替换、去空格、自定义数据提取。

2）依靠系统功能，无需外部脚本，提供多种数据拆分方式，至少包括按比例分割、随机分割、按规则拆分。

3）依靠系统功能，支持数据表的直接拼接、时序拼接。

4）支持通过自定义脚本的方式进行数据预处理，至少包括SQL、Python。

5）数据格式转换

6）数据表拼接、数据表分割、数据表抽样、统计分析

7）通过配置或自定义脚本定义数据处理过程将一个或多个原始的数据表处理为包含样本属性和样本基础的表

8）数据表的抽样和分割

具体能力说明如下：

###### 数据清洗

1. 支持数值替换，即将具体内容或符合正则表达式的内容替换掉
2. 支持数值长度截取，即按照指定的开始截断位置和截断长度来提取数值中的部分内容
3. 支持数值中的空格去除，即去除数值中的两侧空格符
4. 支持数值中的内容提取，即按照指定的匹配内容或正则表达式来提取数值中的部分内容
5. 支持数据清洗过程中的信息统计

###### 数据表分割

1. 平台支持以多种方式对数据表进行按行的分割处理，从而可以将原数据表分割相同schema的两个数据表。支持的分割方式包括：
2. 按比例分割、随机抽样后的按比例分割、分层随机抽样后的按比例分割
3. 按SQL语法的用户自定义逻辑的分割
4. 按用户指定的列排序后的按比例分割。
5. 多种分割方式可帮助用户维持分割后数据集之间的分布同质性，以及避免建模过程中发生数据穿越（用较新的数据建模用于预测较历史的数据）。

##### 3.图谱构建

###### 标注管理

为方便用户快速查看已有标注数据集，系统提供标注数据集的列表展示功能，用户可清晰查看已上传的数据集名称、数据集类型、数据集状态、创建时间等信息，同时用户可通过数据列表对单个数据集进行数据管理，系统提供对数据集删除、查看管理功能。同时为了保障用户可快速对已有数据集状态进行查看，系统提供数据集状态展示功能，包括引入中、可用、不可用，用户可及时获知当前数据状态。



系统提供对单个标注数据集详情展示的功能，通过平台用户可对单个标注数据集内容进行一一展示，展示内容包括数据集基本信息、列表以及内部标注数据的查看。数据详情中包含数据集的具体信息，能够对数据执行删除操作。

系统提供标注数据集基本信息展示，主要包含：预料名称、当前状态、类型、数量、创建时间等。

应用管理功能实现模型服务及接口调用相关的发布和运维操作，需包含的主要功能有：

*  支持权限、用户、机构等相关信息管理
*  支持多批次多模型任务的并行处理及资源分配
*  支持日常运维操作，并提供可视化监控管理
*  支持模型及服务的自动化部署
* 提供标签的增、删、改、查功能

###### 模型构建

可视化的一站式建模体验，通过拖拉拽的方式完成模型的构建、训练以及验证的全过程。平台作为可视化、一站式、全流程的运营智能化基础平台，降低机器学习的准入门槛，为之后的智能化运维打下必要的基础，加速向智能化数据中心的转型.

###### 模型评估

平台支持多维度、多度量的评估系统，支持以可视化的方式输出模型评估的各项指标，针对分类场景输出包括AUC 分值、PR曲线、KS 值、F1值等，以及各判定阈值下的混淆矩阵以及对应的准确率、精确率、召回率，以及分段的各指标统计等，针对回归场景输出包括MSE、MAE、Y\_MEAN、PREDICTION\_MEAN等评估指标。

模型评估是指通过多种专业效果衡量指标，评价模型预测能力的过程。对于训练好的模型，使用一份有实际标注的样本数据进行模型预估，再通过模型评估功能，基于实际标注和模型预估的标注对模型效果进行整体衡量。

1. 分段的各指标统计：按预测分值score大小等区间划分时统计对应的准确率、召回率、本组AUC等等；按预测分值score从高到低排序等量划分时统计对应的准确率、召回率以及某个字段内容的对应统计值等；
2. 根据指定的某一字段，按其特征分布来统计每一特征对应的group AUC值，可以根据数值大小在后续针对某个特征做模型调整；

###### 模型训练

1. 供训练模型的维护功能；
2. 提供模型训练过程的记录功能；
3. 能够支持根据干预信息重新训练模型的功能
4. 能够支持小样本的模型训练；
5. 提供手动和自动模式下的模型训练触发方式，能够对模型的增量变动进行记录，增量含自学习和人工干预下产生的新内容；
6. 提供模型训练的自动更新和回滚的功能，能够在新模型的召回率
7. 模型上线、下线管理维护、模型服务状态管理

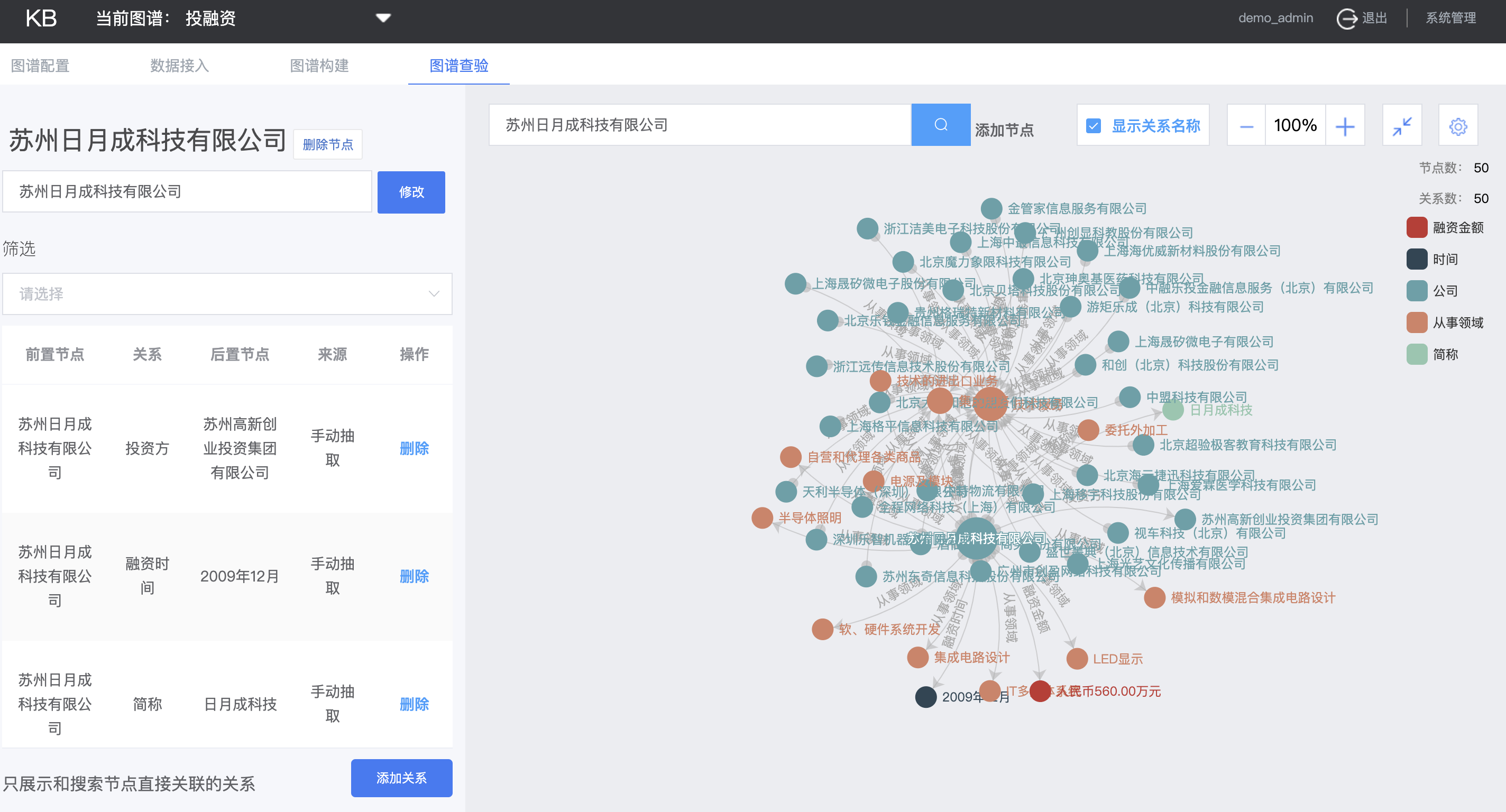
###### 模型效果要求

模型的效果通常由F1值来评估，F1值是准确率与召回率的综合指标。

主要分为实体抽取、关系抽取、事件抽取的要求，期望实体抽取F1值在80%以上，关系抽取与事件抽取的F1值在70%以上。

##### 4.图谱查验

在图谱查验中，展示图谱内的全部内容。可以搜索想要查询的节点，根据节点来检索需要的信息。默认展示的是以被检索节点为中心的二阶子图，图谱同时展示节点、关系与事件之间的关系，页面上的节点大小根据其连接数量得出，越大的节点在图谱中关联的边越多。通过图谱查验可以对谱图进行探索。



##### 5.系统管理

主要包含用户管理、任务分发管理、模型管理、图谱管理、数据管理等功能。

