

# 基于AI对文件分析的流程挖掘系统

## 简介

本项目旨在探索如何利用人工智能技术，从非结构化或半结构化的企业文件（如发票、合同、邮件或报告）中进行流程挖掘。与传统流程挖掘依赖结构化事件日志和预设算法的方式不同，该方法通过AI模型对多样化的数据进行理解与分析，从而发现和可视化真实存在的业务流程。此方法特别适用于缺乏标准化流程或数字化基础设施的中小企业，致力于扩展流程挖掘的应用边界，让流程洞察更易获得、适用面更广。

---

## 系统的简单运作方式

### 在我们系统中的流程的定义

流程是一系列由处理节点构成的具备一定现实意义的业务逻辑的模拟。流程具有一定的输入和输出，可以互相组合，互相替代。在我们这个简单的MVP中，我们只考虑由处理节点构成的流程，不考虑由流程构成的处理节点，也就是不考虑流程的相互组合。

处理节点是指流程的一种组成部分。处理节点由执行角色，处理步骤，输入，输出和决策这五个关键属性组成。

执行角色是指具有执行某种处理步骤，或者能够做出某种决策的人员，职位，部门等。

处理步骤指代了处理节点具体对于某个输入进行的处理。在我们业务流程的拟合中，我们不关心处理步骤，因为我们已经得到了处理的结果——也就是整个系统的输入，一个文件。

因为我们不关心处理步骤，因此具体的输入和输出都只是逻辑上的输入和输出，只需展示其内容，而无需展示其具体的要求，也无需展示其它必要的内容。

### 文件的结构化认知（目前只处理PDF，DOC，EXCEL三个管线）

通过不同的技术从文件中获取文件的内容。将文件转化为AI可理解和分批次处理的相对结构化的数据。

### 文件的自动分类系统

使用大模型，建立一个文件分类和标签化系统。

分类的主要目标是将输入文件分类到某个流程里，然后去判断其是流程处理过程中的产物，还是需要保存的流程的最终输出。

主要的实现方法是首先审查用户现有的流程。判断该文件是不是属于某个已经存在的流程，是否有必要新增执行角色和处理节点，是否能完善现有的流程，是否属于之前上传的某种文件。我们在系统中创建和记录处理节点和流程。逐步的优化流程和节点的数据呈现模式。

# 基于结构化数据和流程的自然语言查询和可视化

在完成流程本身的分类和分析后，我们展示流程和流程所属的文件，并且结合AI，让用户能够随时通过自然语言获得关于流程和文件本身的洞见，并且呈现我们根据流程和处理节点的各个阶段的文件分类结果。

## 具体产物


一个AI辅助的流程挖掘原型系统，具备以下核心功能：

- 本项目将产出一个完整的流程挖掘系统原型，支持上传 PDF、Word、Excel 等常见格式的企业文件，通过 AI 对文件内容进行结构化分析、分类、流程归属判断，并生成由处理节点组成的业务流程图谱。用户可以在系统中查看、审阅和修改识别出的流程与节点，并借助自然语言交互获取关于流程结构和文件归属的实时洞察。
- 系统具备持续学习能力，可根据用户不断上传的文件，逐步完善处理节点与执行角色的识别精度，优化流程建模的准确性。此外，系统将搭载一个文件分类与标签推荐引擎，支持自动将文件标记为流程的输入、中间产物或输出结果，辅助用户理解其业务文档在整体流程中的位置。
- 在流程图构建完成后，用户可使用自然语言提出如“这个流程中有哪些节点由人工完成？”“哪些流程以PDF发票作为最终输出？”等问题，系统将结合流程结构与文档内容做出可视化回答，提升用户对流程运行机制的认知。

- 一份**技术研究论文**，内容包括：

此外，项目还将产出一份技术研究论文，系统总结本系统在非结构化数据环境下的适用性与技术路径。论文内容将包括：AI 在非结构化数据流程挖掘中的适配性与挑战，所采用的模型结构与方法设计，在真实中小企业数据场景中的实验测试结果与分析，以及对未来流程挖掘技术演进方向的思考与展望。

## 参考资料和库



<https://tika.apache.org/>  
Apache Tika – Apache Tika

VikParuchuri/  
marker



Convert PDF to markdown + JSON quickly with high accuracy

17 Contributors

217 Issues

24k Stars

2k Forks

<https://github.com/VikParuchuri/marker?tab=readme-ov-file>

GitHub - VikParuchuri/marker: Convert PDF to markdown + JSON quickly with high accuracy

Convert PDF to markdown + JSON quickly with high accuracy -...



 <https://arxiv.org/abs/2310.03376>

## Procedural Text Mining with Large Language Models

Recent advancements in the field of Natural Language Processing, particularly the development of large-scale language models that are pretrained on vast amou...