Dataset URL Finder

从学术论文中自动提取数据集URL

王思宇 薛佳音 王镜凯

2025.5.26

目录

1 项目概述

2 系统架构

3 Paper Preprocessor

4 URL Digger

5 URL Evaluator

6 技术挑战与解决方案

7 总结

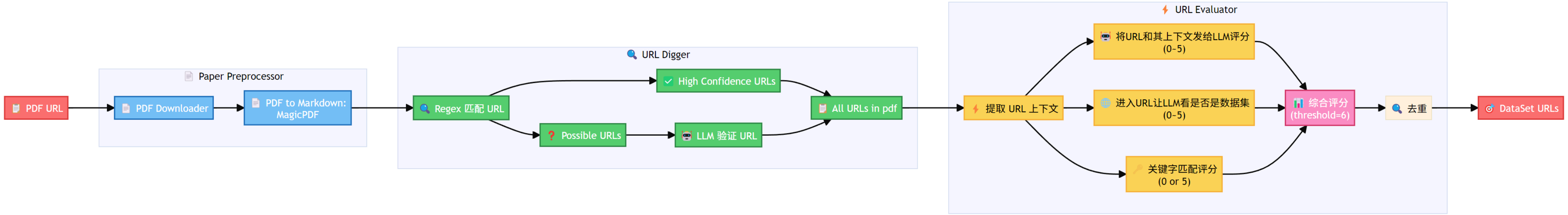
项目概述

项目概述

- **目标**：从学术论文PDF中自动识别和提取数据集相关的URL
- **核心技术**：
    - PDF处理与文本提取
    - 正则表达式 + 大语言模型提取URL
    - 多维度评分机制

系统架构

系统架构


PDF URL

Paper Preprocessor

PDF Downloader

PDF to Markdown: MagicPDF

URL Digger

Regex 匹配 URL

High Confidence URLs

Possible URLs

LLM 验证 URL

All URLs in pdf

URL Evaluator

提取 URL 上下文

将URL和其上下文发给LLM评分 (0~5)

进入URL让LLM看是否是数据集 (0~5)

关键字匹配评分 (0 or 5)

综合评分 (threshold=6)

去重

DataSet URLs

文档预处理 --> 链接智能提取 --> 评分验证

Paper Preprocessor

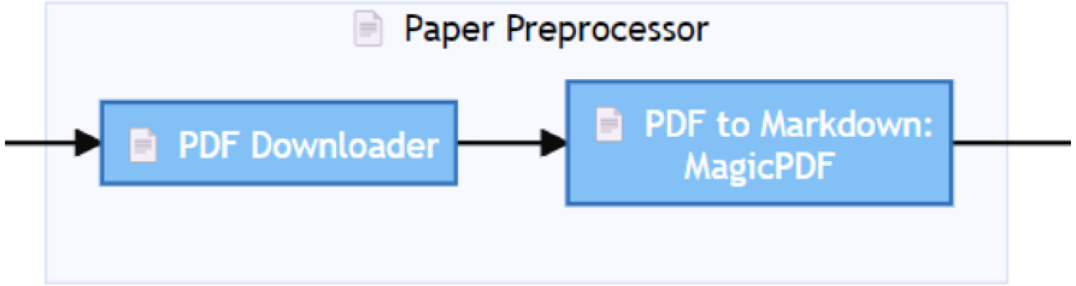
Paper Preprocessor

🎯 核心目标

完成PDF下载任务和文字提取任务

🔧 技术实现

- **PDF下载**：使用requests库
- **文字提取**：使用magic-pdf转换PDF→Markdown



⚠ 存在挑战

跨行URL处理问题

magic-pdf转换时，跨行的URL可能在换行处插入额外空格，影响URL完整性识别

💡 解决思路

- 后续通过智能匹配算法处理
- 大模型验证机制补偿

URL Digger

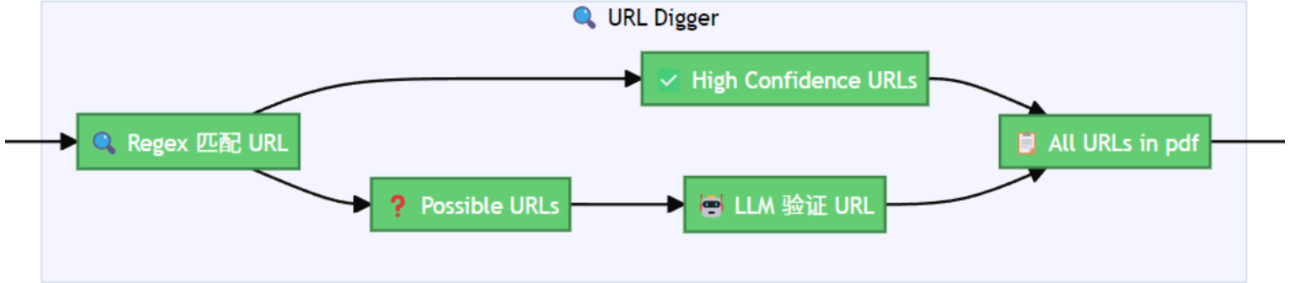
URL Digger

🎯 核心目标

从PDF文字中提取可能的数据集URL

处理流程

1. 正则表达式扫描全文
2. 分类：高置信度 vs 可能URL
3. LLM验证可能URL
4. 合并生成完整URL列表



🔍 双重匹配策略

高置信度URL匹配

- **完整格式**：`http://`，`https://`，`www.`
- **正则表达式**：标准URL格式识别
- **直接通过**：无需二次验证

可能URL匹配

- **简化格式**：`baidu.com`，`github.io`
- **正则表达式**：域名格式识别,可能有假阳性，eg. `Fig.4b`
- **LLM验证**：大模型判断合法性

URL Evaluator

URL Evaluator: 三重评分判定真正的数据集URL

🤖 上下文分析

LLM评分 (0-5分)

- 分析URL周围文本
- 判断与数据集的相关性
- 考虑学术语境

🌐 网站内容分析

BeautifulSoup + LLM (0-5分)

- 访问实际网站
- 分析页面内容
- 确认是否为数据集

🔑 关键词匹配

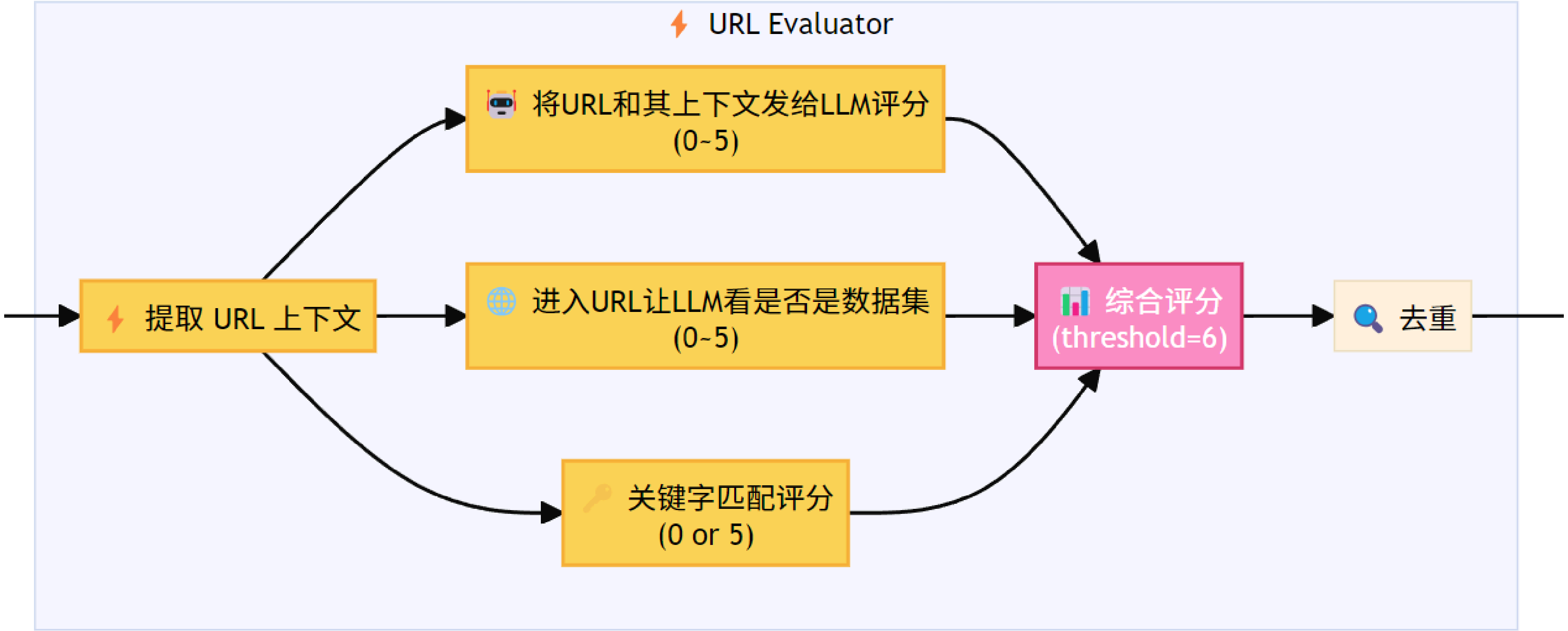
规则评分 (0分或5分)

- dataset, kaggle, data
- github.com/datasets
- 其他数据集平台标识

URL Evaluator

🎯 最终决策

阈值设定：总分 ≥ 6分 → 判定为数据集URL



去重

提取出的数据集URL可能有重复,如:

```
https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/
www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/
```

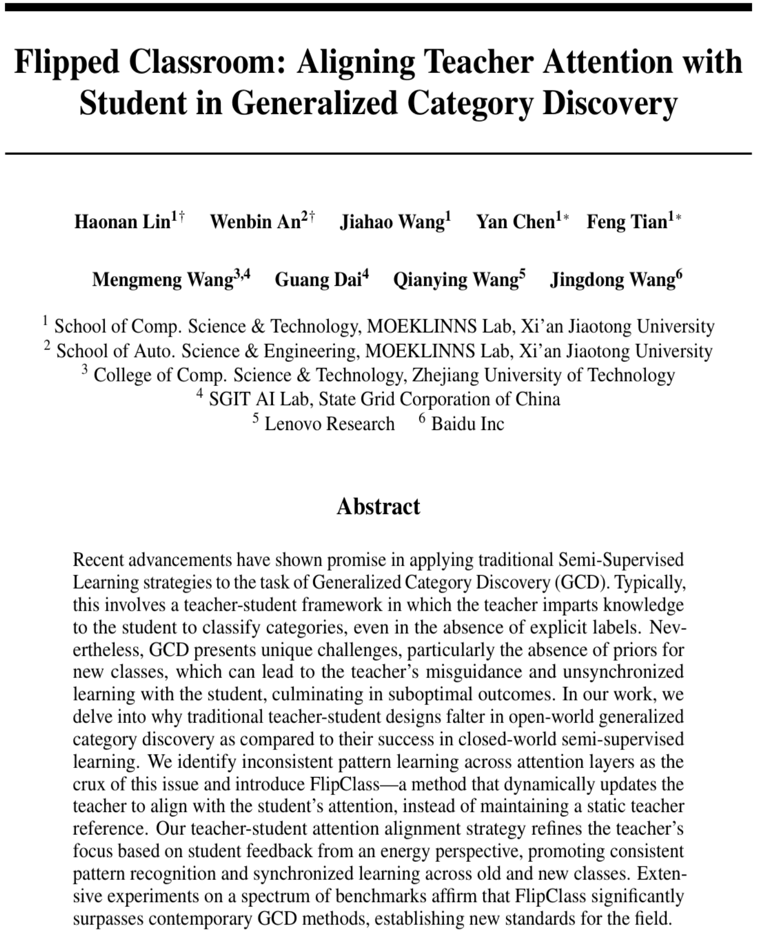
> **解决方案**
>
> 1.设定一个相似度阈值，对每两个url做检验，如果两个url最长相同子串长度占比超过这个阈值就记录下来
>
> 2.记录下来的url对让ai看看是不是真重复了
>
> 3.如果真重复了就删掉其中任意一个

成果展示

我们的架构从右边这篇论文提取出的结果:

```
https://www.cs.toronto.edu/~kriz/cifar.html
https://www.kaggle.com/c/imagenet-object-localization-challenge/overview/
https://www.image-net.org/download.php
https://www.vision.caltech.edu/datasets/cub_200_2011/
https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset
https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/
www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset
www.cs.toronto.edu/~kriz/cifar.html
```

经验证，基本符合人工寻找数据集的结果。



Flipped Classroom: Aligning Teacher Attention with Student in Generalized Category Discovery

Haonan Lin[1†] Wenbin An[2†] Jiahao Wang[1] Yan Chen[1*] Feng Tian[1*]

Mengmeng Wang[3,4] Guang Dai[4] Qianying Wang[5] Jingdong Wang[6]

[1] School of Comp. Science & Technology, MOEKLINNS Lab, Xi'an Jiaotong University
[2] School of Auto. Science & Engineering, MOEKLINNS Lab, Xi'an Jiaotong University
[3] College of Comp. Science & Technology, Zhejiang University of Technology
[4] SGIT AI Lab, State Grid Corporation of China
[5] Lenovo Research [6] Baidu Inc

Abstract

Recent advancements have shown promise in applying traditional Semi-Supervised Learning strategies to the task of Generalized Category Discovery (GCD). Typically, this involves a teacher-student framework in which the teacher imparts knowledge to the student to classify categories, even in the absence of explicit labels. Nevertheless, GCD presents unique challenges, particularly the absence of priors for new classes, which can lead to the teacher's misguidance and unsynchronized learning with the student, culminating in suboptimal outcomes. In our work, we delve into why traditional teacher-student designs falter in open-world generalized category discovery as compared to their success in closed-world semi-supervised learning. We identify inconsistent pattern learning across attention layers as the crux of this issue and introduce FlipClass—a method that dynamically updates the teacher to align with the student's attention, instead of maintaining a static teacher reference. Our teacher-student attention alignment strategy refines the teacher's focus based on student feedback from an energy perspective, promoting consistent pattern recognition and synchronized learning across old and new classes. Extensive experiments on a spectrum of benchmarks affirm that FlipClass significantly surpasses contemporary GCD methods, establishing new standards for the field.

总结

总结

🔍 三大部分

- **Paper Preprocessor:** 尽可能准确地将pdf转文本
- **URL Digger:** 混合匹配策略，正则+AI双重保障
- **URL Evaluator:** 三维评分机制+去重，得到真正的数据集URL

🚀 未来展望

- 使用更好的大模型进一步增加准确性
- 支持更多文档格式（Word、LaTeX等）
- 优化LLM调用策略，降低成本

感谢聆听！

- ✉ - ⌂ -