Artificial Intelligence-Based Pattern Recognition of "Paper Mill"

Tianyi Hu[1]、Jianhua Liu[2]、Haihong E[1]、Jun Zhang[1]、Junpeng Ding[1]、Xiaodong Qiao[2]
1.Beijing University of Posts and Telecommunications
2.Beijing Wanfang Data Co., Ltd

Reported by: E Haihong
ehaihong@bupt.edu.cn

CONTENT

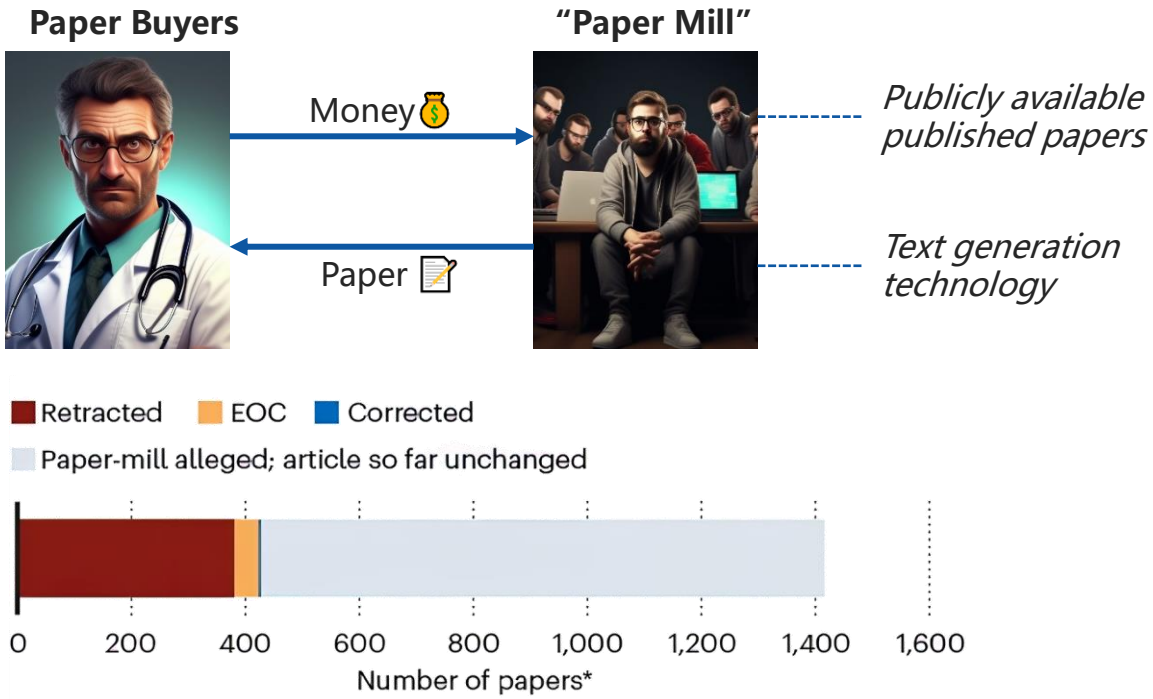
01

Background And Research Target

Part one

01.Background

☑ Research Background



Academic Misconduct Growing in "Paper Mill"
Numerous potential "Paper Mill" have yet to be

☑ Current Detection Tools

- Problematic Paper Screener: **Fingerprints-based search**
- Tongji University Fig_check: Academic Integrity Risk Assessment, Paper Mill Feature Recognition, and Image Reuse Risks
- 24hReview: Pre-review and paper risk assessment services are provided for each journal.

02.Research Target

- Lack of large-scale relevant datasets for research integrity tasks.
- Existing text detection models do not consider structural information.
- Existing detection methods are limited to a single dimension.

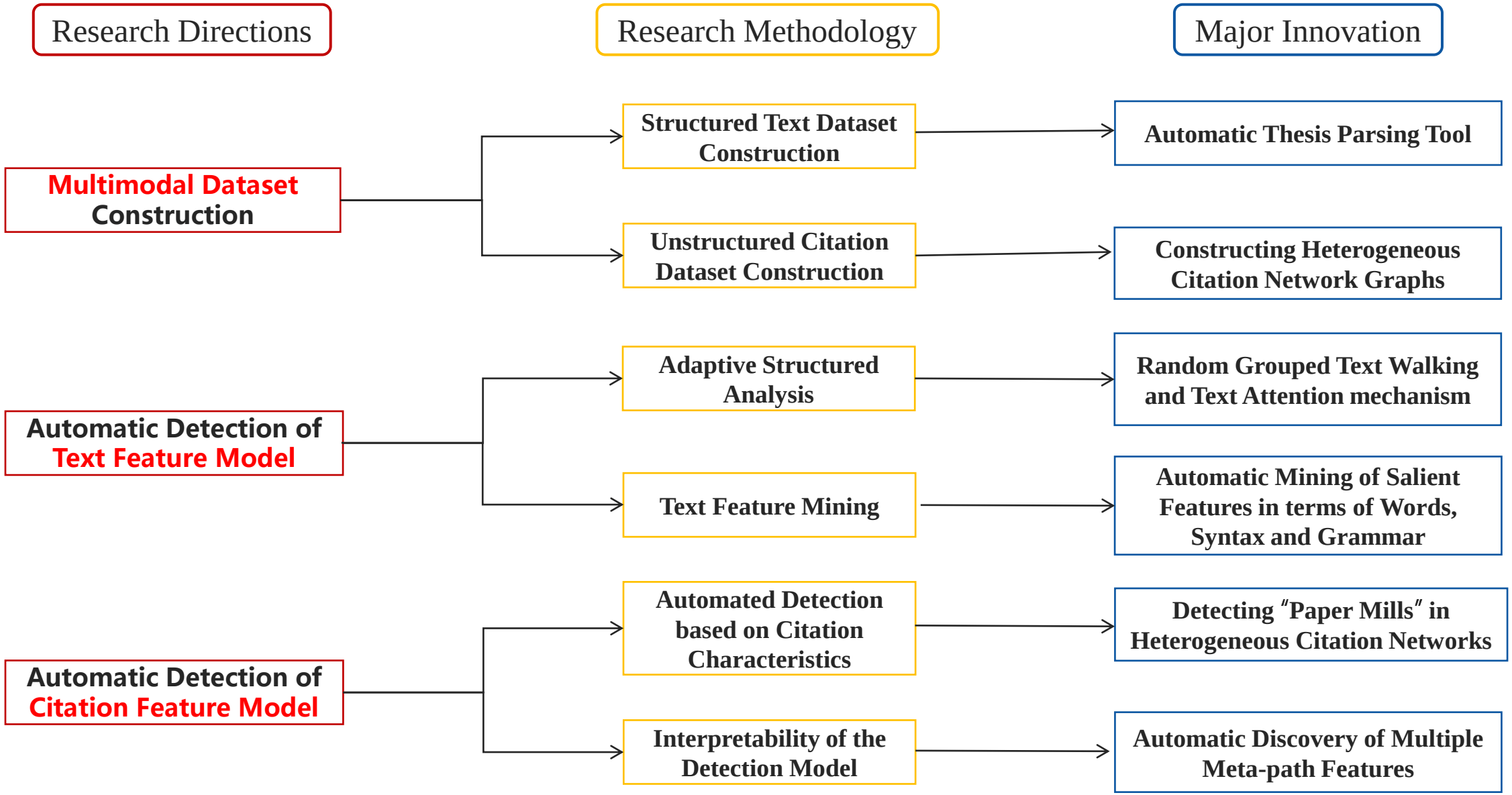
- ☑ **Constructing a Multimodal Dataset based on publicly available retracted papers from "Paper Mill".**
- ☑ **Developing structured parsing tools and Structured Detection Models.**
- ☑ **Research on "Paper Mill" Multimodal Detection Model, including structured content detection, citation structure detection.**

02

Research Method

Part two

01. Research Directions and Major Innovations



02. Relevant Dataset Construction (1/2)

Dataset Requirement 1: Automated Structured Parsing of Paper Files

- **High-quality Paper Datasets** are required for Text Detection Model Training.
- Existing **PDF tools** can not be Automated to get the **Paper's Fine-grained Structural Data**

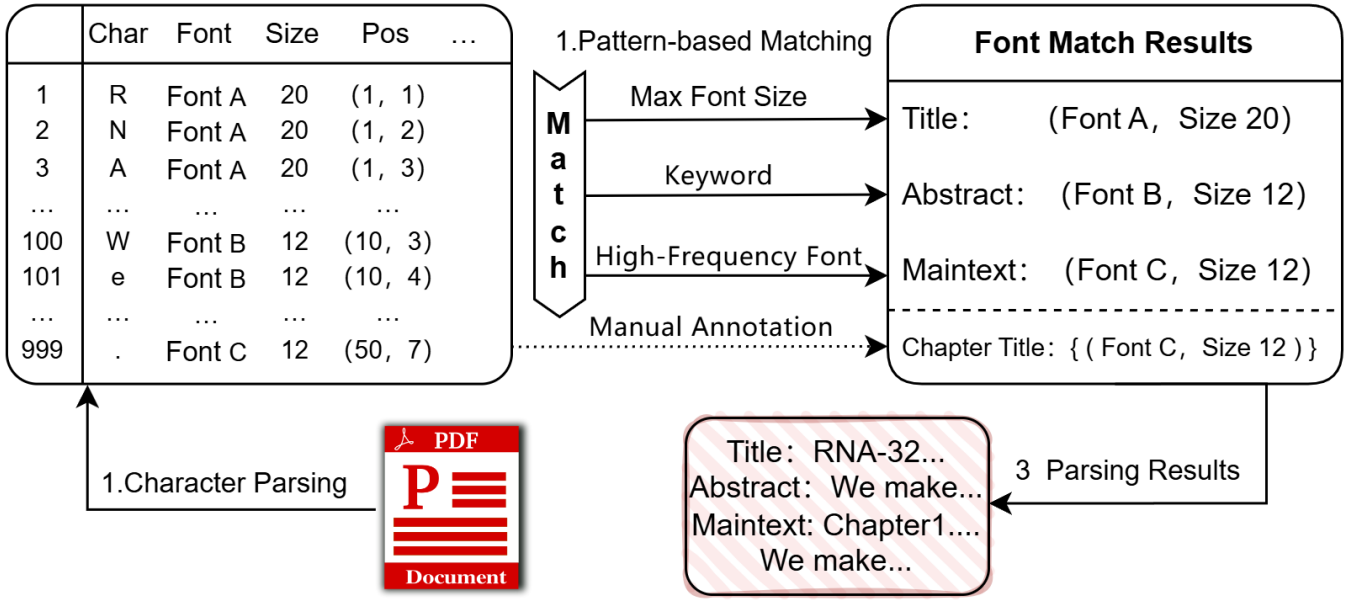
Resource

- **Papers** clearly marked as Paper Mills by **Retraction Watch**
- Collect **Files and Metadata** using multiple open access platforms

Dataset

- A **structured dataset** including title, abstract, main text from 1,535 papers
- The tool is **open-sourced** in https://github.com/TianYi2000/PaperTool

Automatic PDF Parsing



02. Relevant Dataset Construction (2/2)

Dataset Requirement 2: Citation Network Construction

- The existing dataset only contains individual citation information and has not **formed a Citation Network.**

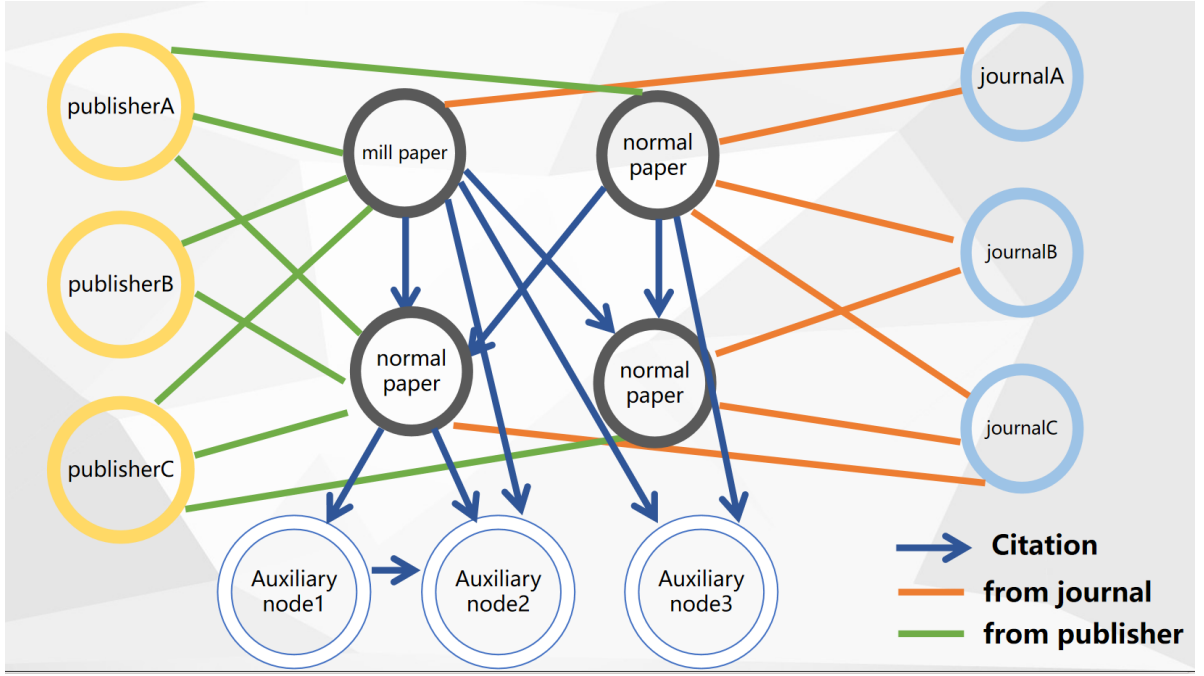
Construction Process

- Starting from "Paper Mill" nodes, **expand the citation network outward through citation relationships** via the *OpenCitations* website.
- It includes **Meta-Data** as well.

Dataset Size

- Mill Nodes：816
- Normal Nodes:25,084
- Edges：5,095,488
- Other Nodes(Journal, Publisher)：5 million

Heterogeneous Citation Network



03. Text Feature Detection Model (1/3)

- Existing research work conduct deep research with eyes, but not **Automatic**, including:
  - Characteristics of **Manuscript Submission and Review Behaviour**
  - Text Content Features: **Similarity of text, Raw Data, Background Content**, etc.
- **Existing Detection Methods:**

Scholar One	• Large number of papers submitted for testing **from the same IP address**
Papermill Alarm	• Building a database of **titles and abstracts** for "Paper Mill" • Comparison of the paper to be tested with the search **Database for Judgement**
Problematic-Paper-Screener	• **Manual collection of features** present in inappropriate papers as 'Fingerprints' to identify inappropriate behaviour.

Single feature

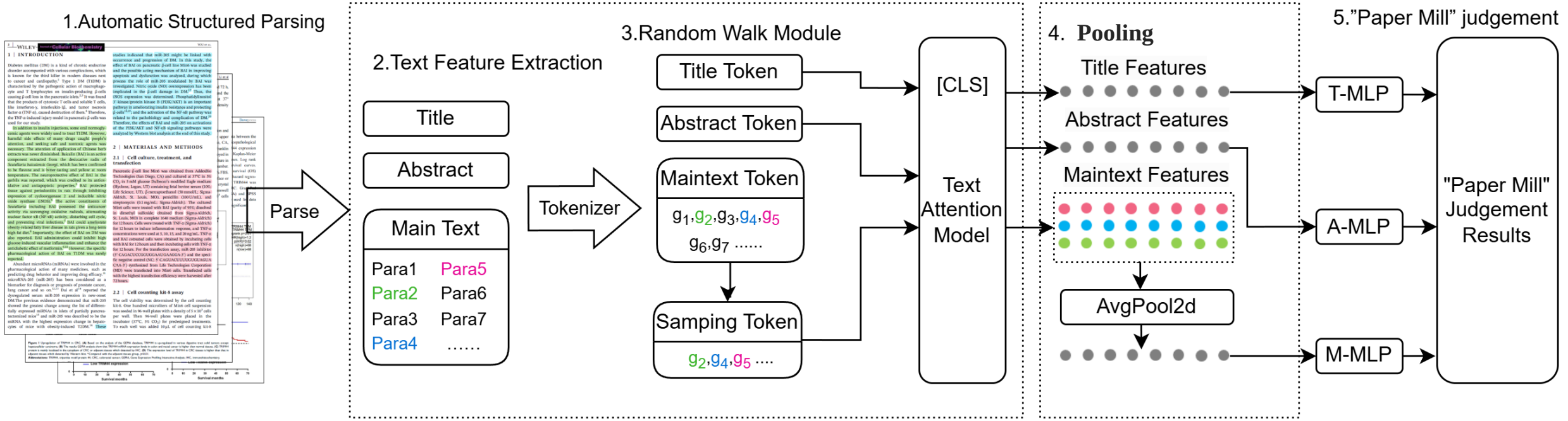
Simple model

Limited detection capability

Target	• A Multi-dimensional Automated Detection Model

03. Text Feature Detection Model (2/3)

- Proposed "Paper Mill" Detection Model based on **Random Wandering strategy and Text Attention** (RWTA)
- **End-to-end Processing** : Input original manuscript files, Output possibility of existence of Paper Mill traces.
- Including: **Automatic Structured Parsing of Papers**、**Text Feature Extraction**、**Random Walk Module**、**Pooling**、**" Paper Mill" judgement**



03. Text Feature Detection Model (3/3)

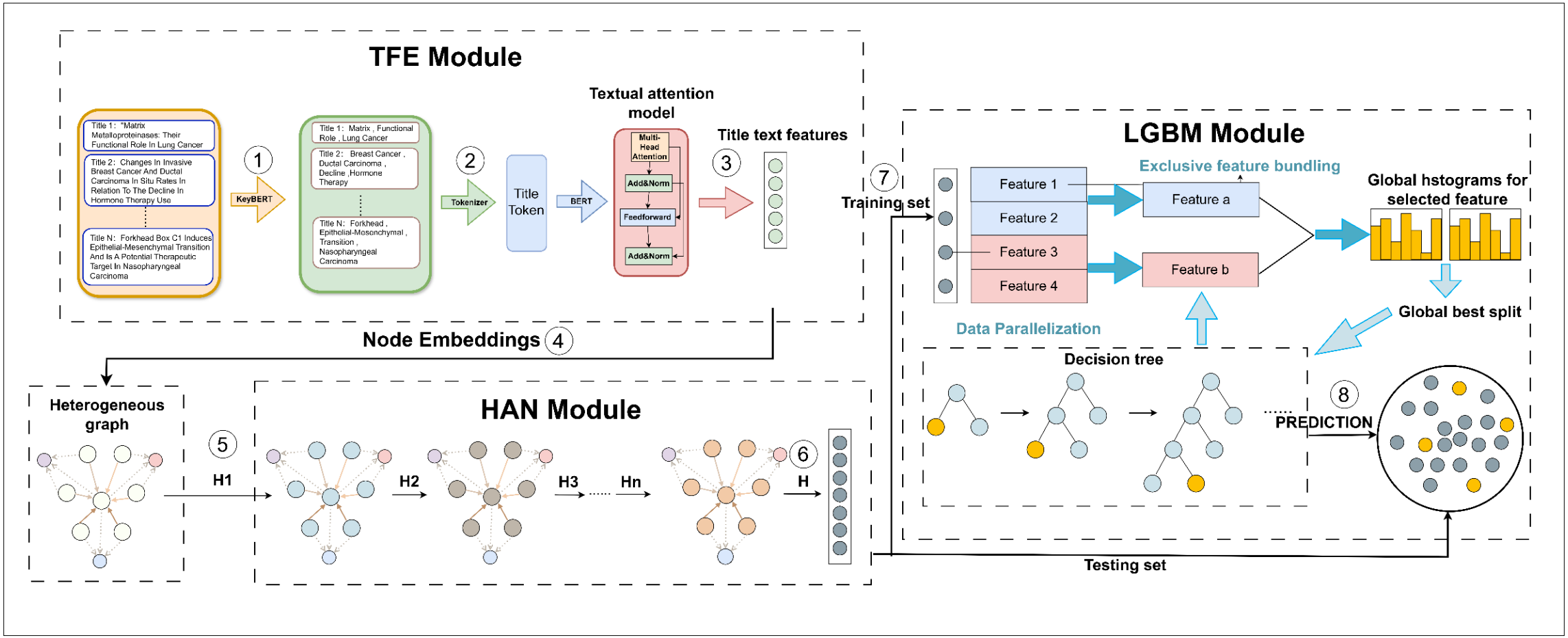
- Compared to some classic machine learning models, our proposed methods achieved the **best results across three dimensions**.

Method	Title			Abstract			Main Text		
	precision	recall	f1	precision	recall	f1	precision	recall	f1
LDA-LR	0.381	0.1039	0.1633	0.5489	0.4372	0.4867	0.4380	0.2597	0.3261
LDA-DT	0.4228	0.2251	0.2938	0.5714	0.4502	0.5036	0.5841	0.5411	0.5618
TF-LR	0.8010	0.6797	0.7354	0.8404	0.7749	0.8063	0.8402	0.7965	0.8178
Tfidf-LR	0.7978	0.6320	0.7053	0.8232	0.7056	0.7599	0.8507	0.7403	0.7917
TF -RF	0.8471	0.6234	0.7182	0.8737	0.7186	0.7886	0.9318	0.7100	0.8059
Tfidf-RF	0.8045	0.6234	0.7024	0.8627	0.7619	0.8092	0.9247	0.7446	0.8249
RWTA (Bert)	0.7510	0.7835	**0.7669**	0.7957	0.8095	0.8026	0.8476	0.7706	0.8073
RWTA (DistBert)	0.7093	0.7957	0.7500	0.7838	0.8788	0.8286	0.8390	0.8571	0.8480
RWTA (RoBERTa)	0.6911	0.7424	0.7158	0.8088	0.8788	**0.8423**	0.8578	0.8355	**0.8465**

Hu, T. Y.; Liu, J. H.; E, H. H.; Ding, J. P. Qiao, X. D. Research on Automatic Detection Feature Model for Sniffing 'Paper Mill'. Data Mining and Knowledge Discovery.

04. Citation Feature Detection Model (1/3)

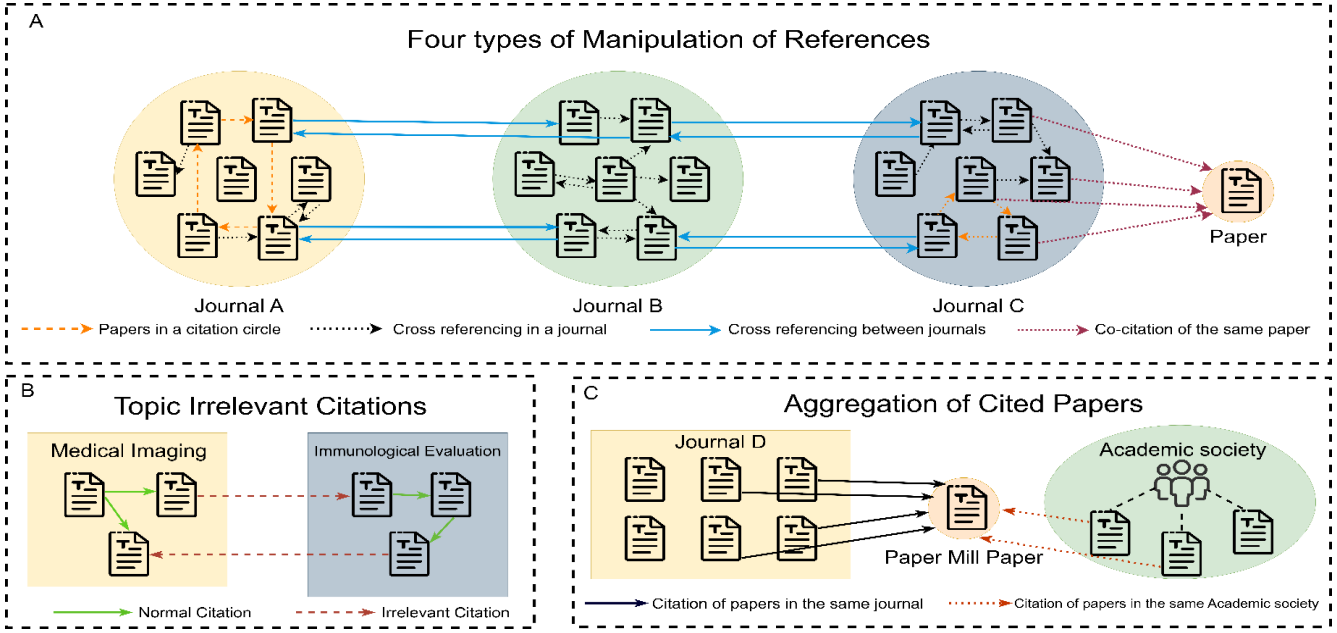
- **A Citation Network-based Model** for detecting Paper Mill: **Features of Nodes** + **Various Meta-paths.**
- **Interpretable**: **Provide scientific explanations** based on Meta-paths and Weights of neighbouring nodes.
- Three Modules: **Text Feature Extraction Module**、**HAN(Heterogeneous Graph Attention Network) Module**、**Classifier LGBM Module**



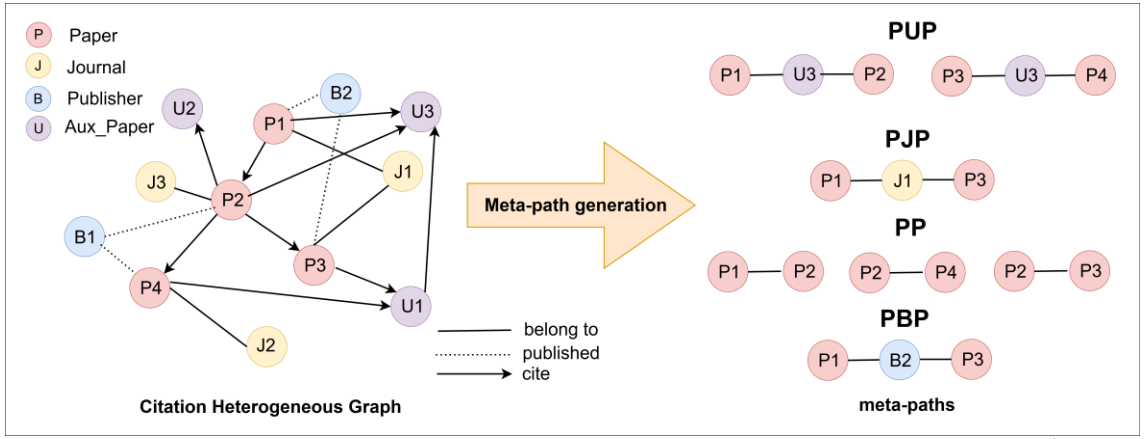
04. Citation Feature Detection Model (2/3)

- **We have designed various meta-paths to assist in the automatic detection of "Paper Mill" within heterogeneous citation networks. (shown in the lower right corner)**
- **Based on the above referencing meta-paths We have found "Paper Mill" citation patterns: (shown at bottom left)**

1. **Manipulation of References:** Papers cite each other or circularly
2. **Topic Irrelevant Citations:** Citations have little to do with the topic of the paper
3. **Aggregation of Cited Papers:** Increase in the number of citations in a short period of time



- **Meta-path**: A path linking two objects on a network schema.



04. Citation Feature Detection Model (3/3)

Comparison Experiment

- On the citation network dataset, the **PDCN model achieved the best performance.**

Model	Precision	Recall	F1-score	NMI	ARI
RGCN	0.047	0.794	0.089	0.013	-0.01
HGT	0.095	0.303	0.145	0.023	0.091
GIN	0.333	0.954	0.494	0.325	0.438
RGAT	0.029	0.845	0.058	0.001	0.007
PDCN	0.819	0.795	0.805	0.626	0.788

Ablation Experiment

- **All three modules** of the PDCN model are important.
- The citation structure of "Paper Mill" papers is **characterised by significant features.**

Model	Precision	Recall	F1-score	NMI	ARI
H	0.101	0.865	0.186	0.091	0.129
T+H	0.057	0.857	0.108	0.038	0.039
T+L	0.231	0.041	0.069	0.018	0.058
H+L	0.428	0.439	0.434	0.234	0.414

Zhang Jun; Liu Jianhua; E Haihong; Hu Tianyi; Ding Junpeng; Qiao Xiaodong. Citation Network-Based Paper Mill Detection Model.JDIS (in progress)

03

Future Study

Part three

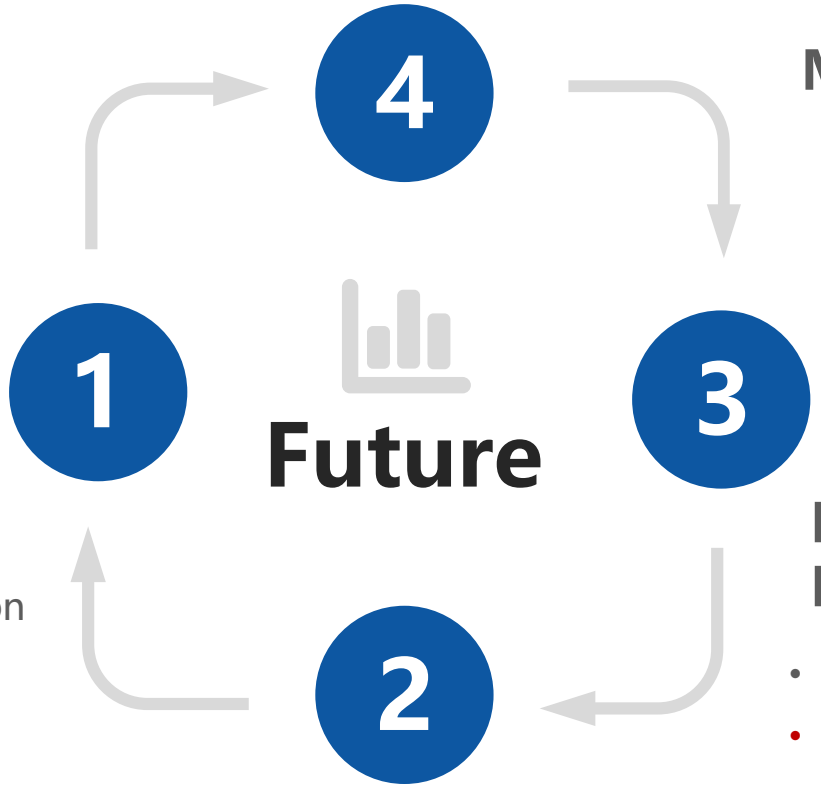
Future Direction of Works

Paper Content Knowledge Graph Extraction

- Textual Content-->**Structured Knowledge Graphs**
- Determine the soundness and **Consistency of the paper's structure**

Temporal Citation Network

- **Adding temporal information** to Citation Heterogeneous Networks
- Capturing citation patterns **between papers over time**
- Create **different Heterogeneous Citation Networks** for each year



More citation patterns

More illegal citation patterns exist, such as:

- Creating fake citations
- Using inappropriate citations
- Presence of citations unrelated to the topic of the paper

Multimodal Graphic Relevance Detection

- Use of image captions and related text data
- **Training a Multimodal Graphic Correlation Detection Model** for the Paper Mill Recognition Task

THANK YOU FOR LISTENING

Artificial Intelligence-Based Pattern Recognition of "Paper Mill"

Reported by: E Haihong
ehaihong@bupt.edu.cn

Tianyi Hu[1]、Jianhua Liu[2]、Haihong E[1]、Jun Zhang[1]、Junpeng Ding[1]、Xiaodong Qiao[2]
1.Beijing University of Posts and Telecommunications
2.Beijing Wanfang Data Co., Ltd