Analyzing Mathematical Content to Detect Academic Plagiarism

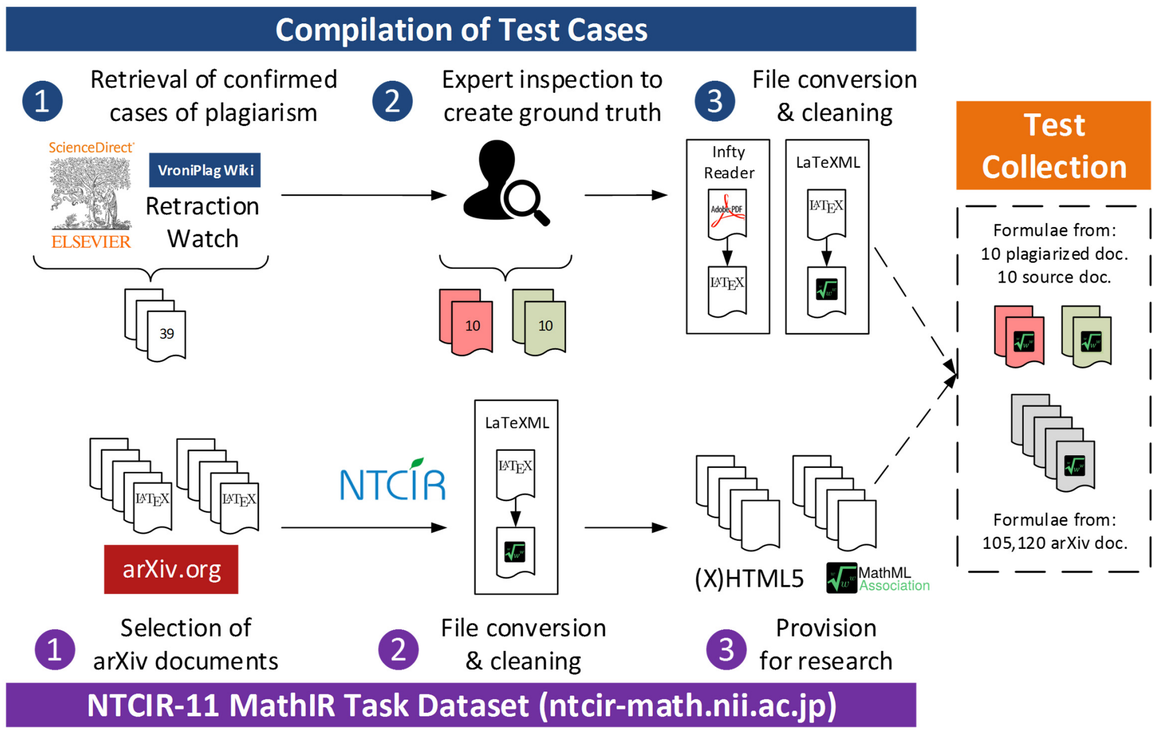
Norman Meuschke[1], **Moritz Schubotz**[1], **Felix Hamborg**[1], **Tomas Skopal**[2], **Bela Gipp**[1]

[1]University of Konstanz, Germany [2] Charles University Prague, Czech Republic

Problem

- **Academic plagiarism**: use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected
- **Productive Plagiarism Detection Systems** use text retrieval.
  - Find copy & paste plagiarism typical for students
  - Miss disguised plagiarism frequent among researchers
- **Text retrieval methods perform poorly for documents in math-heavy disciplines**, because these disciplines interweave natural language and mathematical expressions.
- **Citation-based Plagiarism Detection (our prior work) performs poorly math-heavy disciplines**, because documents cite fewer sources.

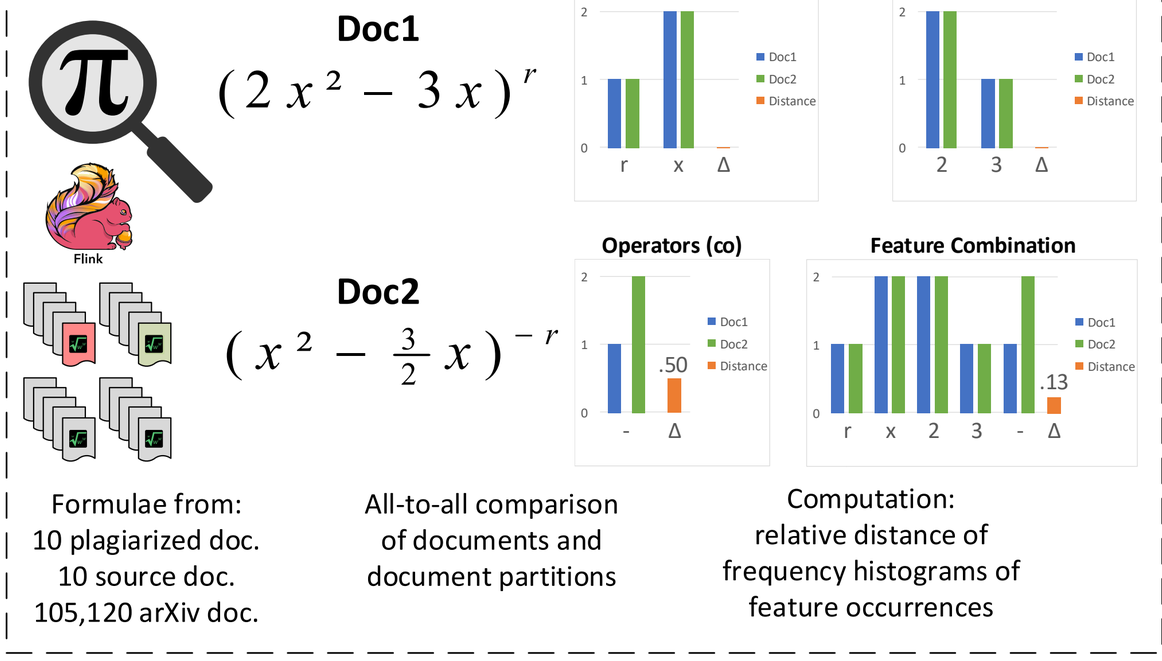
Test Collection



Method

- **Analysis of similarity in mathematical content**
  - *Features:* identifiers (ci), numbers (cn), operators (co), feature combination
  - *Descriptors:* frequency histograms of feature occurrences
  - *Granularity*: *i)* full document, *ii)* document partitions
  - Similarity measure: relative distance of feature occurrence frequencies for individual features $d_{\mathrm{ci,cn,co}}$ and combination of all features D

Mathosphere Framework

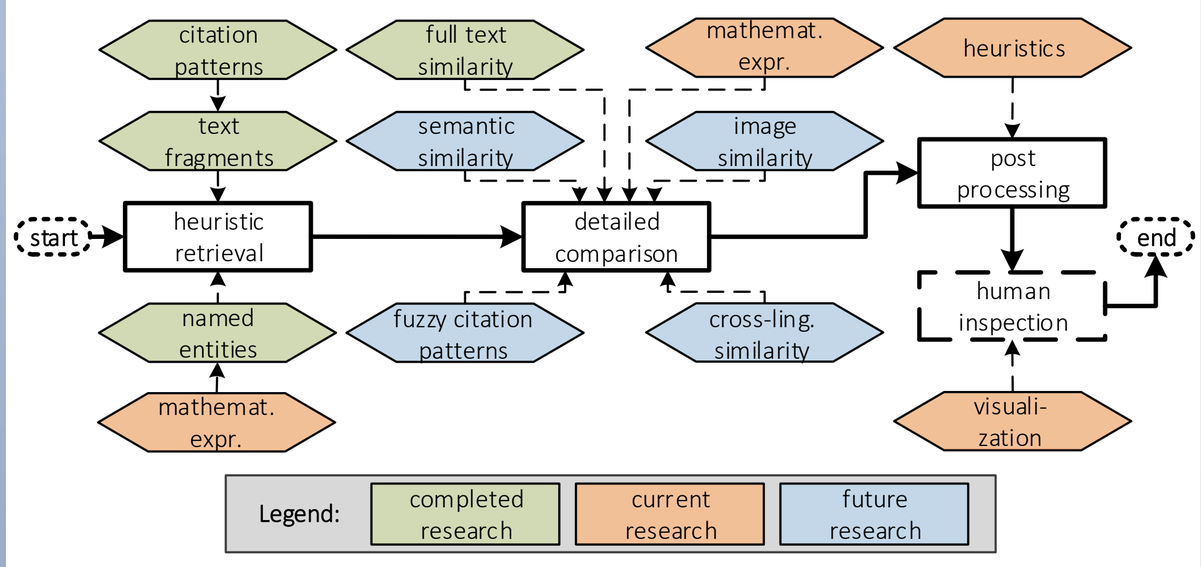


- **Test Collection:**
  - 10 confirmed plagiarism cases and their source documents embedded in NTCIR-11 MathIR Task dataset (105,120 arXiv documents, 60 million formulae)

Results

	full document source retrieved at rank				partitions source retrieved at rank			
Case	**D**	**d_{ci}**	**d_{cn}**	**d_{co}**	**D**	**d_{ci}**	**d_{cn}**	**d_{co}**
C1	3,606	1	27,857	30,784	1	1	85,418	99,201
C2	1	1	88,891	90,962	1	1	12,266	10,277
C3	11,628	2	28,415	3,144	1	16	34,966	5,757
C4	2,581	1	1,950	86	189	6	54,560	18,374
C5	1	1	5,790	22,408	1	6	92,951	16,180
C6	25,498	12	19,862	38,145	7,976	3	24,405	72,687
C7	1	1	4,690	1,627	19,900	1	67,614	14,758
C8	1	1	39,215	11,576	1	1	21,152	9,475
C9	1	1	13,591	35,393	1	1	11,519	32,687
C10	1	1	76,678	30,673	1	1,223	89,703	3,280
	MRR							
	0.60	**0.86**	< 0.01	< 0.01	**0.70**	0.57	< 0.01	< 0.01

Future Work



Acknowledgements

DAAD German Academic Exchange Service **Travel Support**

DFG German Research Foundation **Grant: GI 1259/1**

GAČR GRANTOVÁ AGENTURA ČESKÉ REPUBLIKY **Project: 17-22224S1**

More Information

Paper, Code, Data: **purl.org/mathpd**

Contact: **isg.uni.kn**