Universität Konstanz

Analyzing Non-textual Content Elements to Detect Academic Plagiarism

Norman Meuschke

Doctoral Defense, March 5, 2021

Outline

Introduction

- Motivation
- Research Objective & Research Tasks

Results for Research Tasks

- State of the Art & Research Gap
- Detection Approaches
  - Citations
  - Images
  - Mathematical Content
- Evaluation
- System Prototype

Conclusion & Outlook

Introduction

Defining Academic Plagiarism

The use of ideas, words, or other work without appropriately acknowledging the source to benefit in a setting where originality is expected.*

* Definition adapted from: Fishman, T., "'We Know It When We See It' Is Not Good Enough: Toward a Standard Definition of Plagiarism That Transcends Theft, Fraud, and Copyright," in Proceedings of the 4th Asia Pacific Conference on Educational Integrity (4APCEI) , 2009, p. 5.

Forms of Academic Plagiarism

low

Obfuscation

high

- **Lexis-preserving plagiarism**
  - Verbatim copying without citation
- **Syntax-preserving plagiarism**
  - Synonym substitution
  - Technical disguise
- **Semantics-preserving plagiarism**
  - Paraphrase
  - Translation
- **Idea-preserving plagiarism**
  - Appropriation of ideas or concepts
  - Reusing the sequence of arguments or ideas
  - Reusing materials other than text
- **Ghostwriting**
  - Contract Cheating

Prevalence of Academic Plagiarism

Students

Many studies since the 1950s — hard to consolidate due to diverse definitions, objectives, and methods

Rough trends of average prevalence reported [1]:

~20-30%	North America & Western Europe
~30-60%	Australia, Eastern Europe & Russia
~60-85%	Middle East & Asia

(Insufficient data for South America)

Researchers

Few systematic studies but much empirical evidence

~15%	average estimate of 372 journal editors for the % of plagiarized submissions [2]
2,670	Journal articles retracted for plagiarism in RW database (11% of total) [3]
210 **786**	Reports on doctoral theses with strong evidence of plagiarism by the VroniPlag [4] and Dissernet [5] projects

Consensus in the literature:

Academic plagiarism is a pressing problem among students and researchers.

[1] Studies reviewed in: Ison, D. C., "An Empirical Analysis of Differences in Plagiarism Among World Cultures," Journal of Higher Education Policy and Management, vol. 40, no. 4, pp. 291–304, Jul. 2018.
[2] Smart, P. & Gaston, T., "How Prevalent Are Plagiarized Submissions? Global Survey of Editors," Learned Publishing, vol. 32, no. 1, pp. 47–56, Jan. 2019.
[3] http://retractiondatabase.org [4] https://vroniplag.wikia.org [5] https://www.dissernet.org/

Problem of Detecting Academic Plagiarism

- The systems can find
  “[...] **a good bit of text overlap**.”

- Their performance is
  “[...] **only partially satisfactory** [...]”
  for synonym replacements

- “[...] **quite unsatisfactory for paraphrased and translated texts.**”

Plagiarism forms more characteristic of researchers

Foltýnek *et al. International Journal of Educational Technology in Higher Education* (2020) 17:46
https://doi.org/10.1186/s41239-020-00192-4

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Testing of support tools for plagiarism detection

Check for updates

Tomáš Foltýnek[1,2*], Dita Dlabolová[1], Alla Anohina-Naumeca[3], Salim Razı[4], Július Kravjar[5], Laima Kamzola[3], Jean Guerrero-Dib[6], Özgür Çelik[7] and Debora Weber-Wulff[8]

* Correspondence: tomas.foltynek@mendelu.cz
[1]Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czechia
[2]University of Wuppertal, Wuppertal, Germany
Full list of author information is available at the end of the article

Abstract

There is a general belief that software must be able to easily do things that humans find difficult. Since finding sources for plagiarism in a text is not an easy task, there is a wide-spread expectation that it must be simple for software to determine if a text is plagiarized or not. Software cannot determine plagiarism, but it can work as a support tool for identifying some text similarity that may constitute plagiarism. But how well do the various systems work? This paper reports on a collaborative test of 15 web-based text-matching systems that can be used when plagiarism is suspected. It was conducted by researchers from seven countries using test material in eight different languages, evaluating the effectiveness of the systems on single-source and multi-source documents. A usability examination was also performed. The sobering results show that although some systems can indeed help identify some plagiarized content, they clearly do not find all plagiarism and at times also identify non-plagiarized material as problematic.

Keywords: Text-matching software, Software testing, Plagiarism, Plagiarism detection tools, Usability testing

Research Objective & Research Tasks

Devise, implement, and evaluate automated approaches capable of identifying previously non-machine-detectable forms of disguised academic plagiarism.

RT1 Identify the strengths and weaknesses of state-of-the-art methods and systems to detect academic plagiarism.

RT2 Devise detection approaches that address the identified weaknesses.

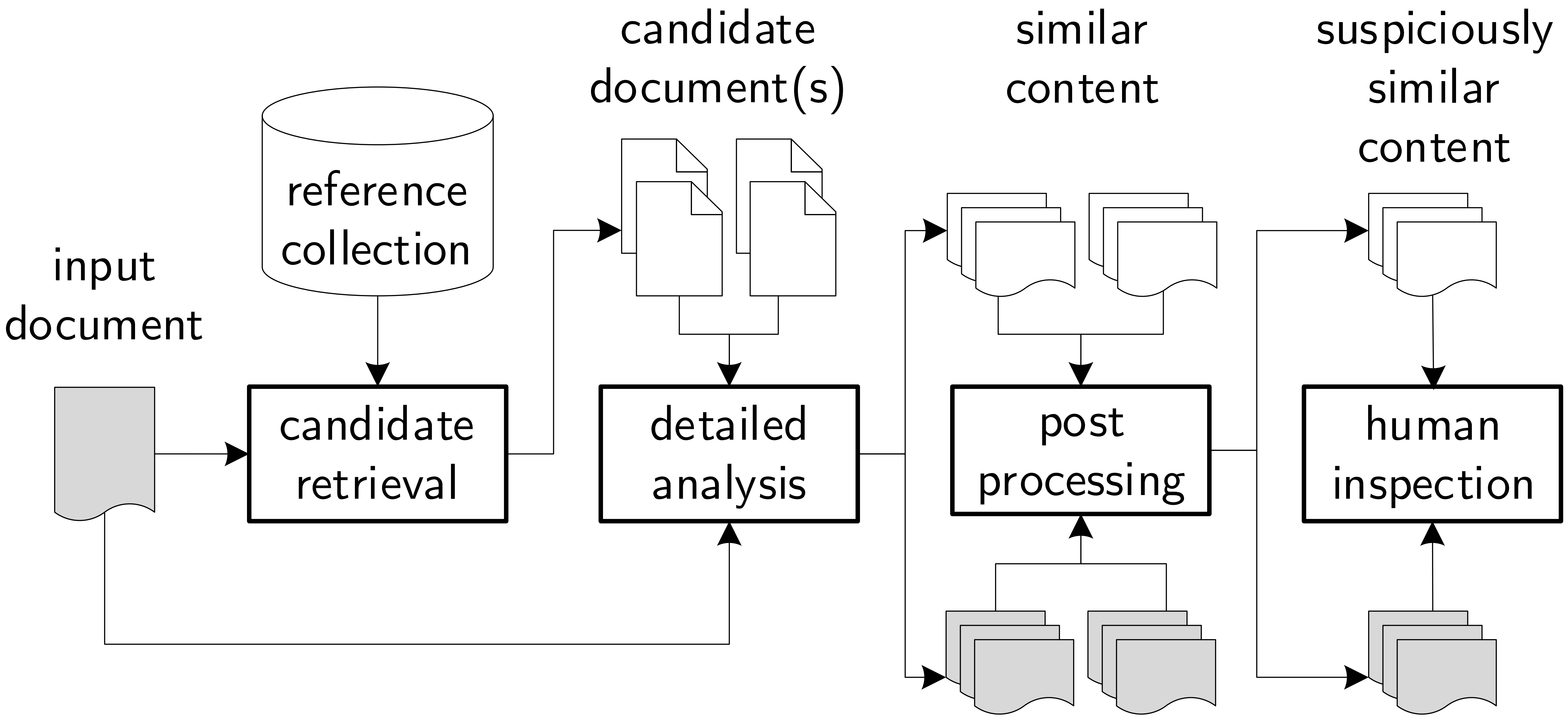
RT3 Evaluate the effectiveness of the proposed detection approaches.

RT4 Implement the proposed detection approaches in a plagiarism detection system capable of supporting realistic detection use cases.

Results for Research Tasks

State of the Art in Plagiarism Detection Research

External Plagiarism Detection Process



State of the Art in Plagiarism Detection Research

Obfuscation: low → high

Lexis-preserving plagiarism
- Verbatim copying without citation

Syntax-preserving plagiarism
- Synonym substitution
- Technical disguise

Mature Technologies
- n-gram fingerprinting, VSM, PoS analysis …
- Candidate retrieval $R \approx 60\%$
- $F_1 \approx 88\%–96\%$ (1:1 comparisons)

Semantics-preserving plagiarism
- Paraphrase
- Translation

Idea-preserving plagiarism
- Appropriation of ideas or concepts
- Reusing text structure
- Reusing sequence of ideas or arguments

Intense Research
- Text-based semantic analysis (LSA, ESA, token embeddings, KGA, …), Machine Learning, Deep Learning, Machine Translation
- Candidate retrieval $R \approx 60\%$
- $F_1 \approx 50\%–60\%$ (1:1 comparisons)

Ghostwriting
- Contract Cheating

Open Research Problem

RT1

Identified Research Gap

- **Candidate Retrieval** and **Detailed Analysis** methods capable of improving the identification of:
  - **Strong paraphrases**
  - **Sense-for-sense translations**
  - **Structural and idea plagiarism**

Research Approach

In addition to **text**, analyze:

- **Citations**
- **Images**
- **Mathematical content**

RT2 Devise detection approaches that address the identified weaknesses.

RT3 Evaluate the effectiveness of the proposed detection approaches.

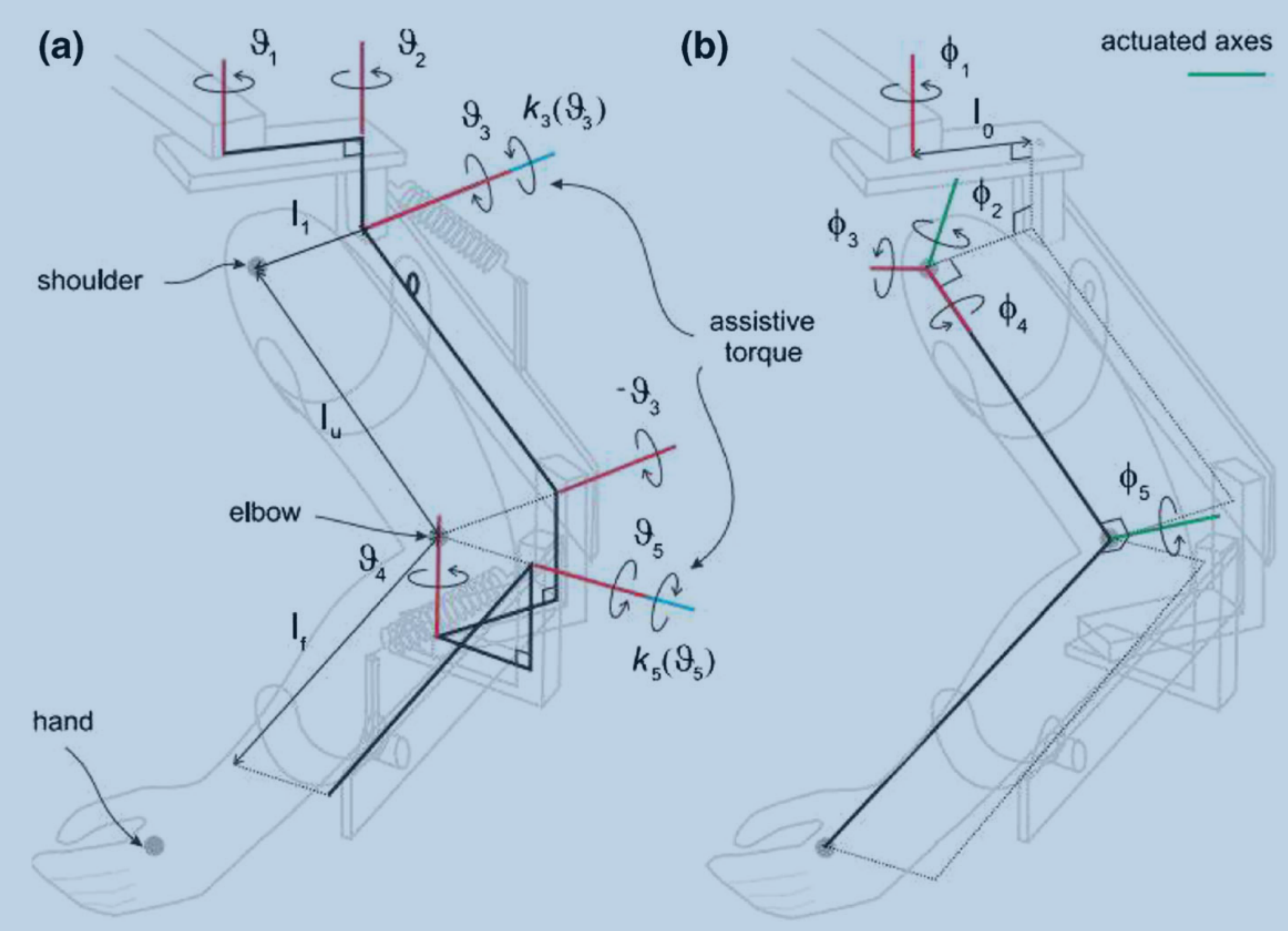


Fig. 1 Kinematic system relations: **a** rehabilitation support and **b** human arm [8]

where $\Theta = [\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4, \vartheta_5]^T$ are the joint variables, $\boldsymbol{B}_a(\cdot)$ and $\boldsymbol{C}_a(\cdot)$ denote 5-by-5 inertial and Coriolis matrices. $\boldsymbol{F}_a(\cdot)$ and $\boldsymbol{G}_a(\cdot)$ are the frictional and gravitational vectors. The vector $\boldsymbol{K}_a(\cdot)$ denotes the moments arising from gravity compensation provided by the two springs, which is the function of θ_3 and θ_5, respectively, thus making $\boldsymbol{K}_a(\cdot)$ take the form of $[0, 0, k_3(\vartheta_3), 0, k_5(\vartheta_5)]^T$.

2.2 Human arm

Spasticity in stroke patients typically produces a resistance to arm extension associated with the overactivity of muscles, like the biceps, wrist and finger flexors, and with loss of activity of muscles such as the triceps, anterior deltoid, wrist and finger extensors [27]. In order to provide effective treatment, it is the latter group of muscles that must be activated during the functional reaching tasks; therefore, the triceps and anterior deltoid are selected for FES stimulation according to clinical need [8]. It is first assumed that FES stimulation to the triceps produces a moment about an axis orthogonal to both the forearm and upper arm, and stimulation to the anterior deltoid generates a moment about an axis that is fixed corresponding to the shoulder. The actuated joints variables corresponding to the stimulated muscles are denoted as ϕ_5 and ϕ_2, respectively, as shown in Fig. 1b, and the remaining degrees of freedom are encompassed by ϕ_1, ϕ_3, ϕ_4.

The dynamics of the human arm with FES applied to the two muscles, similar to the model of the mechanical support, as shown in Fig. 1b, is represented by

$$\boldsymbol{B}_h(\Phi)\ddot{\Phi} + \boldsymbol{C}_h(\Phi, \dot{\Phi})\dot{\Phi} + \boldsymbol{F}_h(\Phi, \dot{\Phi}) + \boldsymbol{G}_h(\Phi) = \boldsymbol{\tau}(\boldsymbol{u}, \Phi, \dot{\Phi}) \quad (7)$$

where $\Phi = [\phi_1, \phi_2, \phi_3, \phi_4, \phi_5]^T$ denote the anthropomorphic joints, comprising of those stimulation-actuated dynamics and those unactuated, and $\boldsymbol{\tau}(\cdot)$ are the input torques produced from stimulated muscles, thus taking the form

$$\boldsymbol{\tau}(\boldsymbol{u}, \Phi, \dot{\Phi}) = [0, \tau_2(u_2, \phi_2, \dot{\phi}_2), 0, 0, \tau_5(u_5, \phi_5, \dot{\phi}_5)]^T. \quad (8)$$

2.3 Muscle model

The muscle models utilized for performance evaluation and the development of model-based controllers about both upper and lower limb vary a lot structurally. However, the most widely assumed structure, by far,

Citation-based Plagiarism Detection – Summary

- **First non-textual PD approach**
  - Analyzed confirmed plagiarism cases
- **Devised set-based and sequence-based methods** to identify observed patterns and can handle:
  - Transpositions
  - Insertion or substitutions
  - Repetitions
- **Applied the methods to a large-real-world collection** of biomedical articles
  - Citation-based methods outperformed text-based methods for disguised forms of plagiarism

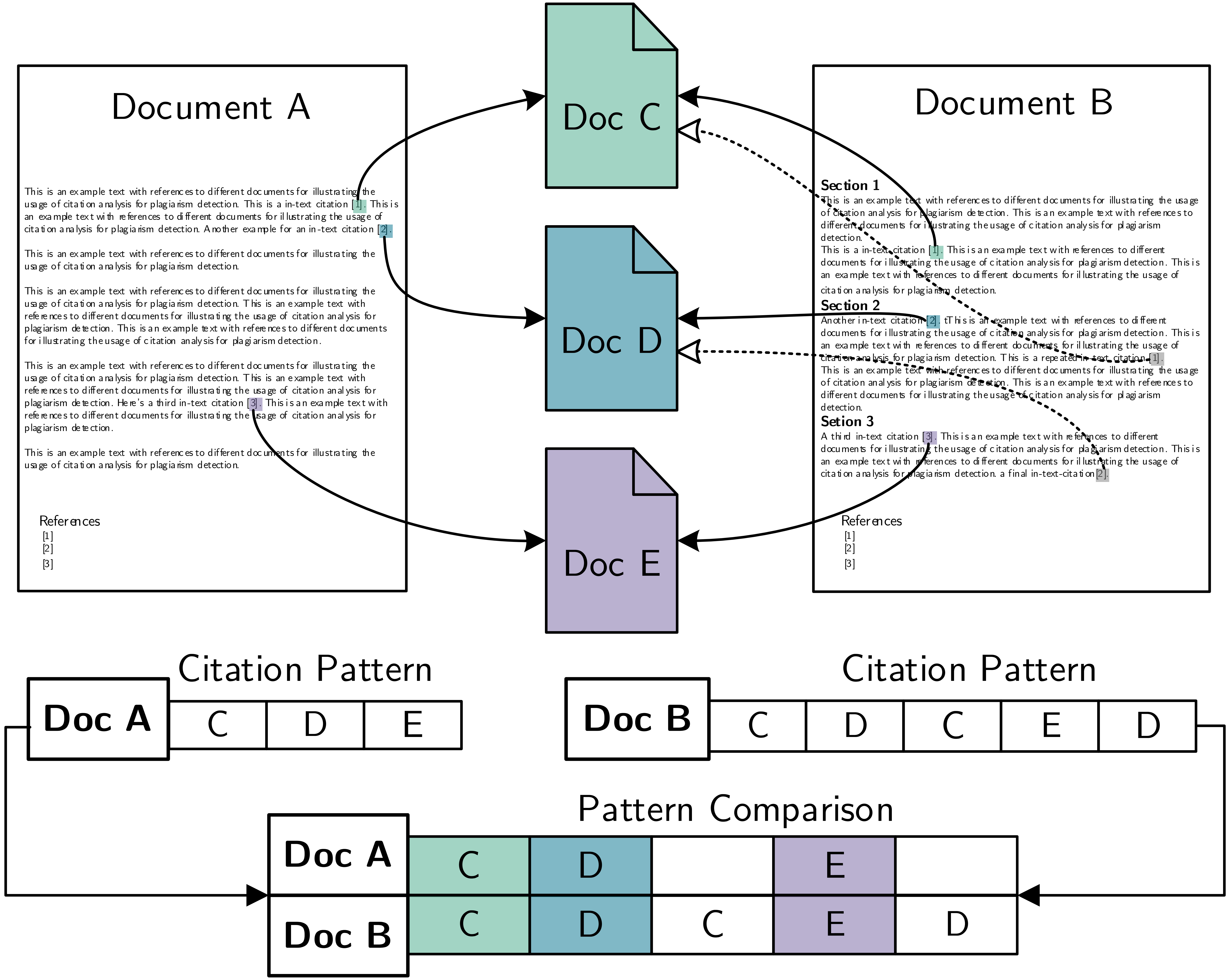
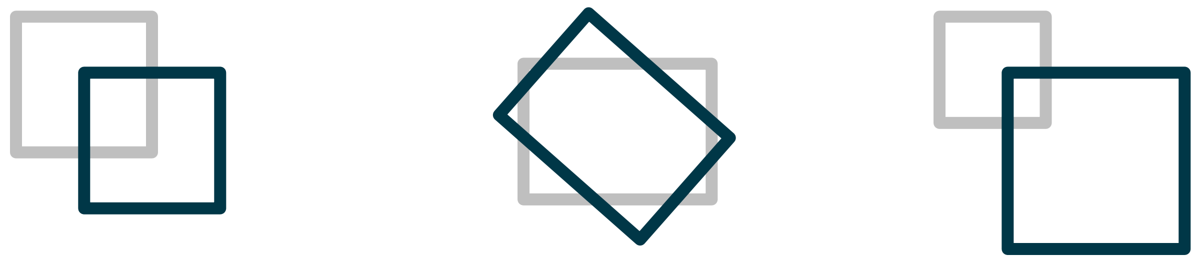


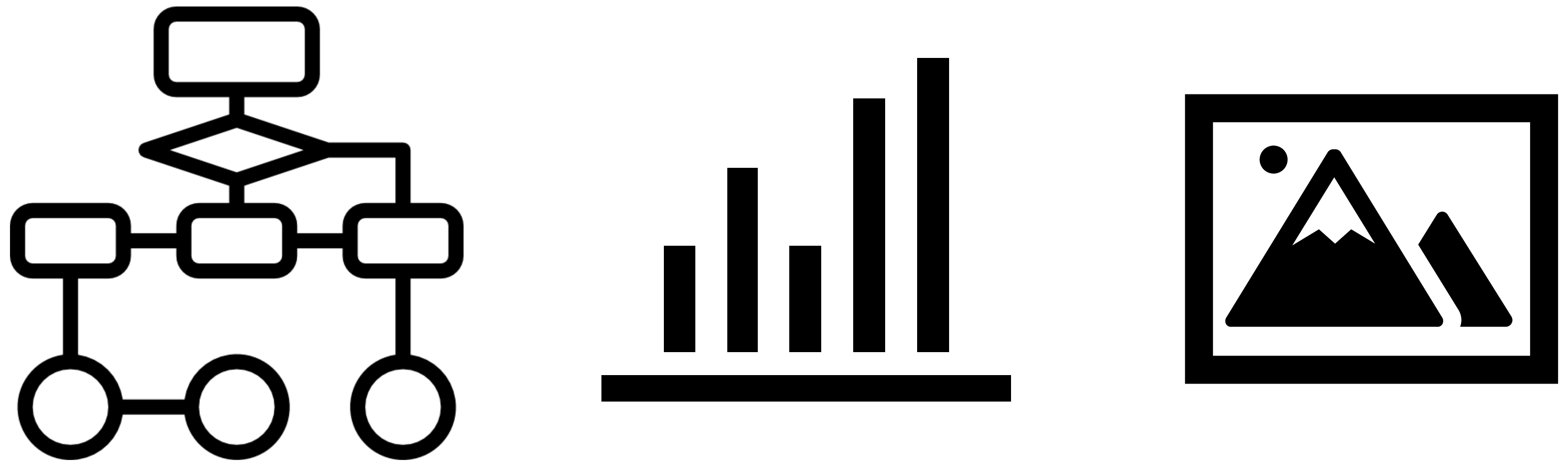
Image-based Plagiarism Detection — Summary

Related Work

- Retrieval approaches for
  - Copied
  - Cropped
  - Affinely transformed images



- Focus on specific image types

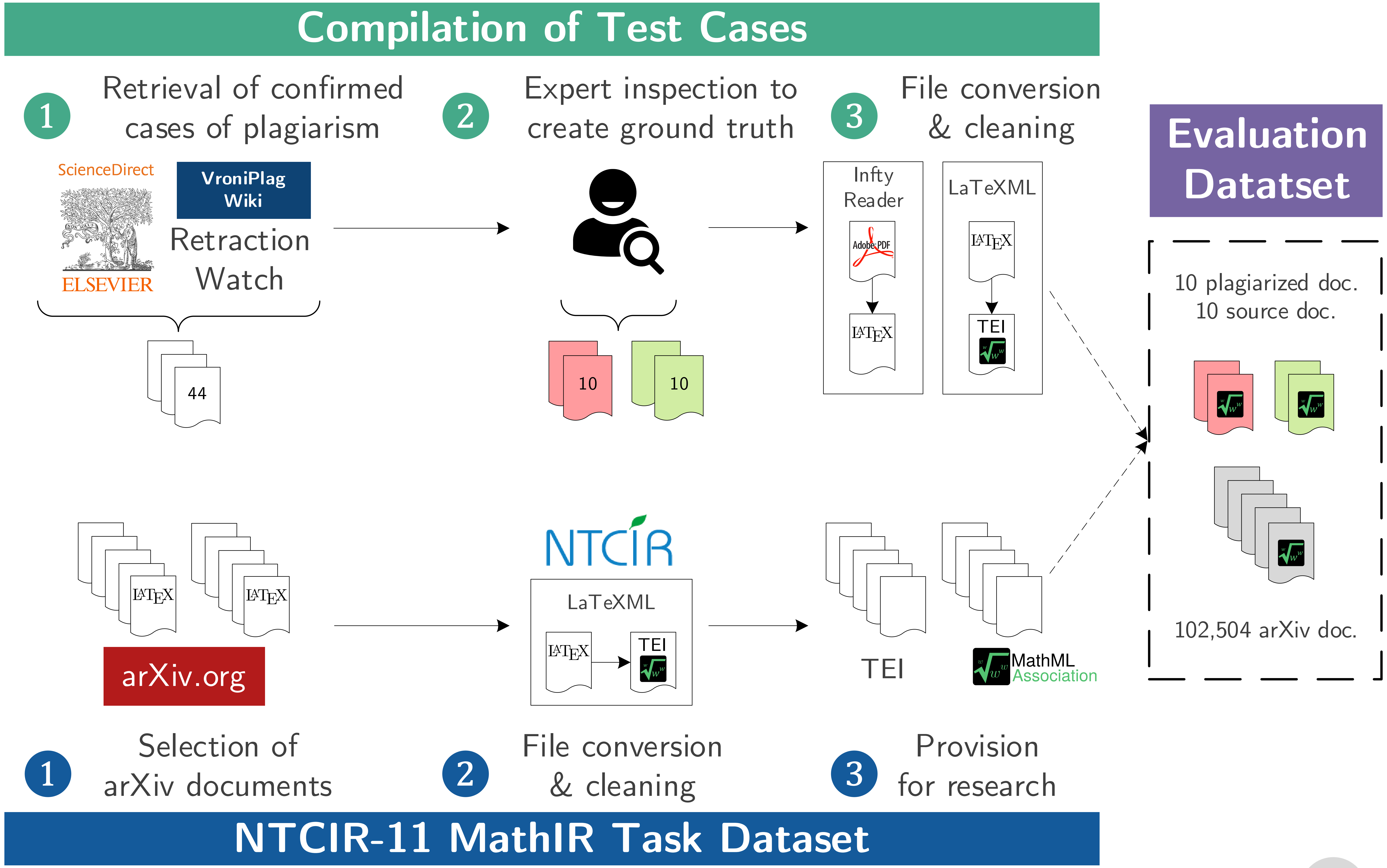


Contributions

- Use-case-specific detection methods for typical image types in academic documents, e.g., bar-charts and flow-charts
- Image-based detection process that:
  - Combines analysis methods for image types typical for academic documents
  - PD-specific relevance scoring
  - Is efficient and extensible

Math-based Plagiarism Detection

- First study on the topic – Starting point: confirmed plagiarism cases



Math-based PD — Observations for Plagiarism Cases

- **Identical expressions**
- **Equivalent expressions**, e.g., commutativity, distributivity, and associativity
- **Order changes** for near-identical formulae
- **Splits or merges** of expressions
- **Different presentation** of structurally and semantically identical expressions
- **Different concepts**, e.g., summation over vector components vs. matrix multiplication

Math-based PD — Features for Preliminary Experiments

- Retrieval experiments using essential presentational elements of mathematical notation:

- **Identifiers**

$$\eta_2^T B_U(\tilde{\eta}_1)\dot{\eta}_2 + \eta_2^T \frac{\dot{B}_U(\tilde{\eta}_1)}{2}\eta_2 + \eta_2^T F_s(\tilde{\eta}_1) + \eta_2^T \left(\bar{C}_{UC}(\tilde{\eta}_1)\hat{\Phi}_C^{(1)} + B_{UC}(\tilde{\eta}_1)\,\hat{\Phi}_C^{(2)}\right)$$
$$= \eta_2^T \left(\frac{\dot{B}_U(\tilde{\eta}_1)}{2} - \bar{C}_{UC}(\tilde{\eta}_1, \eta_2)\right)\eta_2 - \eta_2^T F_v(\eta_2) \leq \eta_2^T \left(\frac{1}{2}\dot{B}_U(\tilde{\eta}_1) - \bar{C}_U(\tilde{\eta}_1, \eta_2) - \bar{F}_v\right)\eta_2$$

- **Numbers**

$$\eta_2^T B_U(\tilde{\eta}_1)\dot{\eta}_2 + \eta_2^T \frac{\dot{B}_U(\tilde{\eta}_1)}{2}\eta_2 + \eta_2^T F_s(\tilde{\eta}_1) + \eta_2^T \left(\bar{C}_{UC}(\tilde{\eta}_1)\hat{\Phi}_C^{(1)} + B_{UC}(\tilde{\eta}_1)\,\hat{\Phi}_C^{(2)}\right)$$
$$= \eta_2^T \left(\frac{\dot{B}_U(\tilde{\eta}_1)}{2} - \bar{C}_{UC}(\tilde{\eta}_1, \eta_2)\right)\eta_2 - \eta_2^T F_v(\eta_2) \leq \eta_2^T \left(\frac{1}{2}\dot{B}_U(\tilde{\eta}_1) - \bar{C}_U(\tilde{\eta}_1, \eta_2) - \bar{F}_v\right)\eta_2$$

- **Operators**

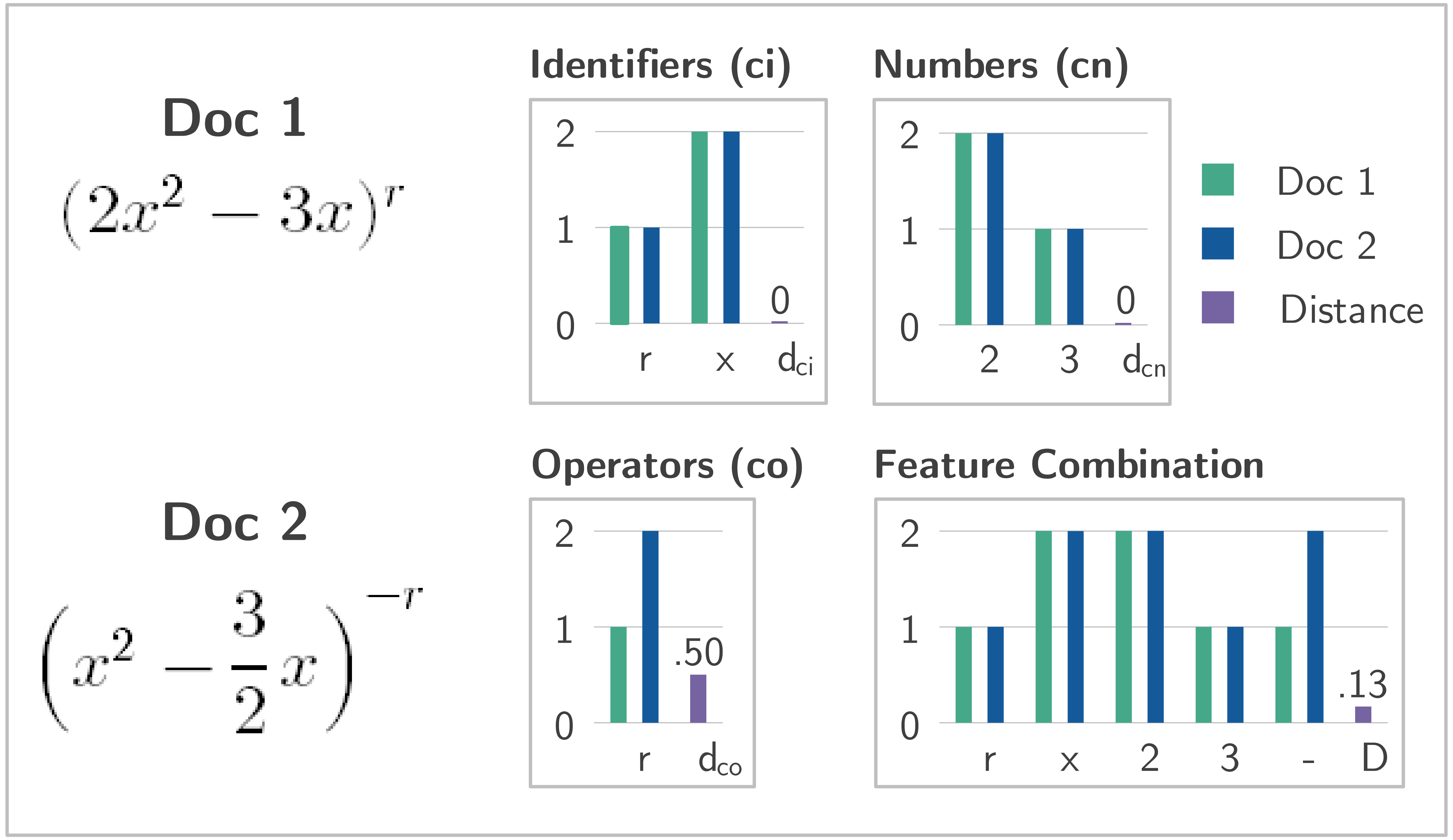
$$\eta_2^T B_U(\tilde{\eta}_1)\dot{\eta}_2 + \eta_2^T \frac{\dot{B}_U(\tilde{\eta}_1)}{2}\eta_2 + \eta_2^T F_s(\tilde{\eta}_1) + \eta_2^T \left(\bar{C}_{UC}(\tilde{\eta}_1)\hat{\Phi}_C^{(1)} + B_{UC}(\tilde{\eta}_1)\,\hat{\Phi}_C^{(2)}\right)$$
$$= \eta_2^T \left(\frac{\dot{B}_U(\tilde{\eta}_1)}{2} - \bar{C}_{UC}(\tilde{\eta}_1, \eta_2)\right)\eta_2 - \eta_2^T F_v(\eta_2) \leq \eta_2^T \left(\frac{1}{2}\dot{B}_U(\tilde{\eta}_1) - \bar{C}_U(\tilde{\eta}_1, \eta_2) - \bar{F}_v\right)\eta_2$$

- **Combination**

$$\eta_2^T B_U(\tilde{\eta}_1)\dot{\eta}_2 + \eta_2^T \frac{\dot{B}_U(\tilde{\eta}_1)}{2}\eta_2 + \eta_2^T F_s(\tilde{\eta}_1) + \eta_2^T \left(\bar{C}_{UC}(\tilde{\eta}_1)\hat{\Phi}_C^{(1)} + B_{UC}(\tilde{\eta}_1)\,\hat{\Phi}_C^{(2)}\right)$$
$$= \eta_2^T \left(\frac{\dot{B}_U(\tilde{\eta}_1)}{2} - \bar{C}_{UC}(\tilde{\eta}_1, \eta_2)\right)\eta_2 - \eta_2^T F_v(\eta_2) \leq \eta_2^T \left(\frac{1}{2}\dot{B}_U(\tilde{\eta}_1) - \bar{C}_U(\tilde{\eta}_1, \eta_2) - \bar{F}_v\right)\eta_2$$

Math-based PD — Analysis for Preliminary Experiments

- No candidate retrieval
- Basic order-agnostic "bag of features" comparisons of **presentational features**
- Entire **documents** and **partitions**



Math-based PD — Results of Preliminary Experiments

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

Focused on identifiers for devising detailed analysis methods

RT2

Math-based PD — Detailed Analysis Methods

Identifier Histograms (Histo)

- Order-agnostic "bag of identifiers"
- Similarity = relative difference in occurrence frequency

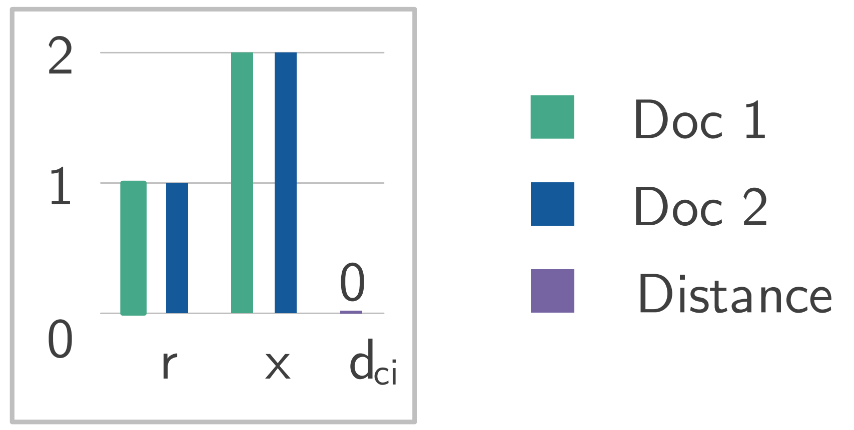
Doc 1

$$(2x^2 - 3x)^r$$

Doc 2

$$\left(x^2 - \frac{3}{2}x\right)^{-r}$$

Identifiers (ci)



Greedy Identifier Tiles (GIT)

- Individually longest blocks of 5 or more matching identifiers in same order normalized by number of identifiers in document

$$C_{U,i,j} = \sum_{k=1}^{n} c_{\mathcal{J}_U(i),\mathcal{J}_U(j),k}\,\dot{\phi}_k$$

$$C_U(i,j) = \sum_{k=1}^{n} c_{N_U(i),N_U(j)}\,\dot{\phi}_k$$

Longest Common Identifier Sequence (LCIS)

- Identifiers in same order but not necessarily contiguous normalized by number of identifiers in document

$$C_U(i,j) = \sum_{k=1}^{n} c_{N_U(i),N_U(j)}\,\dot{\phi}_k$$

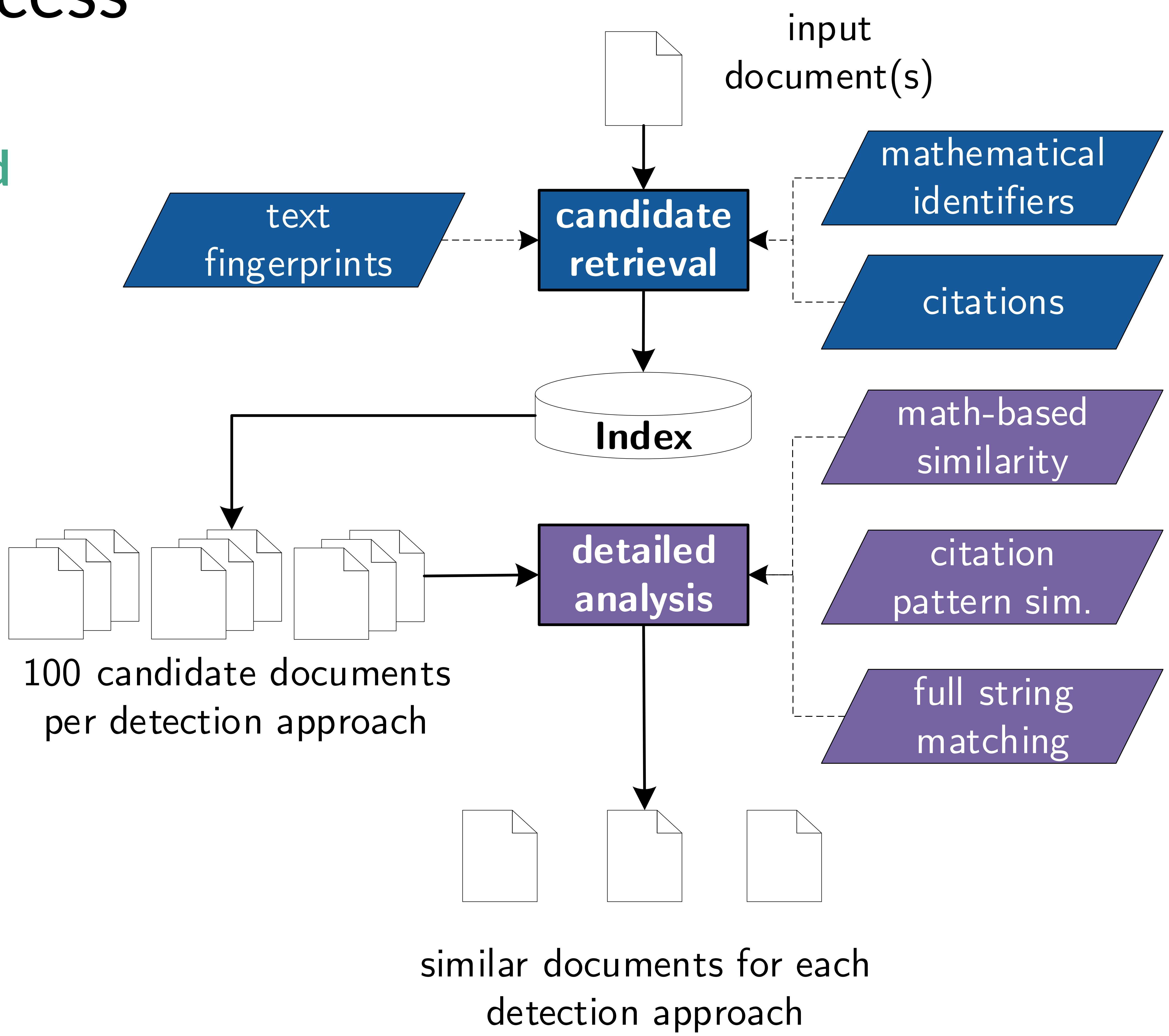
$$C_{U,i,j} = \sum_{k=1}^{n} c_{\mathcal{J}_U(i),\mathcal{J}_U(j),k}\,\dot{\phi}_k$$

Math-based PD — Evaluation Process

Comparison of **math-based, citation-based and text-based** detection methods

Lucene Scoring for candidate retrieval

- Combined tf-idf & Boolean retrieval model
- Features:
  - Identifiers (boost: number of occurrences)
  - Citations
  - Text-fingerprints (selected character 3-grams)



Math-based PD — Results Candidate Retrieval

- **Effectiveness of math-based candidate retrieval must be improved**
- **Detection methods complement each other**
  - No single method retrieves all cases.
  - Any combination of two methods achieves 100% recall.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	***R***
Mathematics	+	+	+	−	−	−	+	+	+	+	0.7
Citations	+	+	−	+	+	+	+	+	+	+	0.9
Text	+	+	+	+	+	+	−	+	+	+	0.9

Legend: C1...C10 IDs of test cases, ***R*** Recall

Math-based PD — Results Detailed Analysis

	Mathematics						Citations									Text	
	Histo		LCIS		GIT		BC			LCCS			GCT			Enco	
Case	r	s	r	s	r	s	r	s	s^*	r	s	s^*	r	s	s^*	r	s
C1	1	.68	1	.40	1	.21	1	.06	.15	1	.06	.10	-	-	.04	1	.13
C2	1	.60	1	.39	1	.12	10’	.05	.28	1	.33	.42	-	-	-	1	.16
C3	3	.29	1	.88	1	.78	-	-	-	-	-	-	-	-	-	1	.36
C4	(1)	(.36)	(99)	(.37)	(3)	(.03)	-	-	.35	-	-	.44	-	-	.25	1	.15
C5	(1)	(.57)	(86)	(.30)	(1)	(.23)	5	.02	.18	7’	.02	.23	-	-	.05	1	.45
C6	(19)	(.14)	(98)	(.40)	(1)	(.15)	2	.04	.32	1	.11	.44	-	-	.22	1	.27
C7	2	.52	98	.25	1	.09	-	-	.04	-	-	.05	-	-	-	(4)	(.02)
C8	1	.76	1	.65	1	.37	1	.11	.37	-	-	.25	-	-	-	1	.32
C9	1	.69	1	.51	1	.27	1	.03	.26	1	.08	.39	-	-	-	1	.68
C10	1	.85	1	.81	1	.63	1	.03	.03	1	.04	.04	-	-	-	1	.51
MRR	.58 (.79)		.60 (.60)		.79 (.93)		.48 (.48)			.60 (.60)			.00 (.00)			.90 (.93)	

Legend:
r rank at which the source document was retrieved, s similarity score, s^* citation-based similarity score without extraction errors,
(...) candidate retrieval step did not retrieve the source document, it was added manually to evaluate the detailed analysis step,
– no similarity score computed due to method-specific exclusion criteria, 10’ mean rank considered since ranks were tied,
<u>###</u> similarity score above the method-specific significance threshold, MRR Mean Reciprocal Rank

5. For all but one case (C7), at least one detection method yielded clearly suspicious scores.

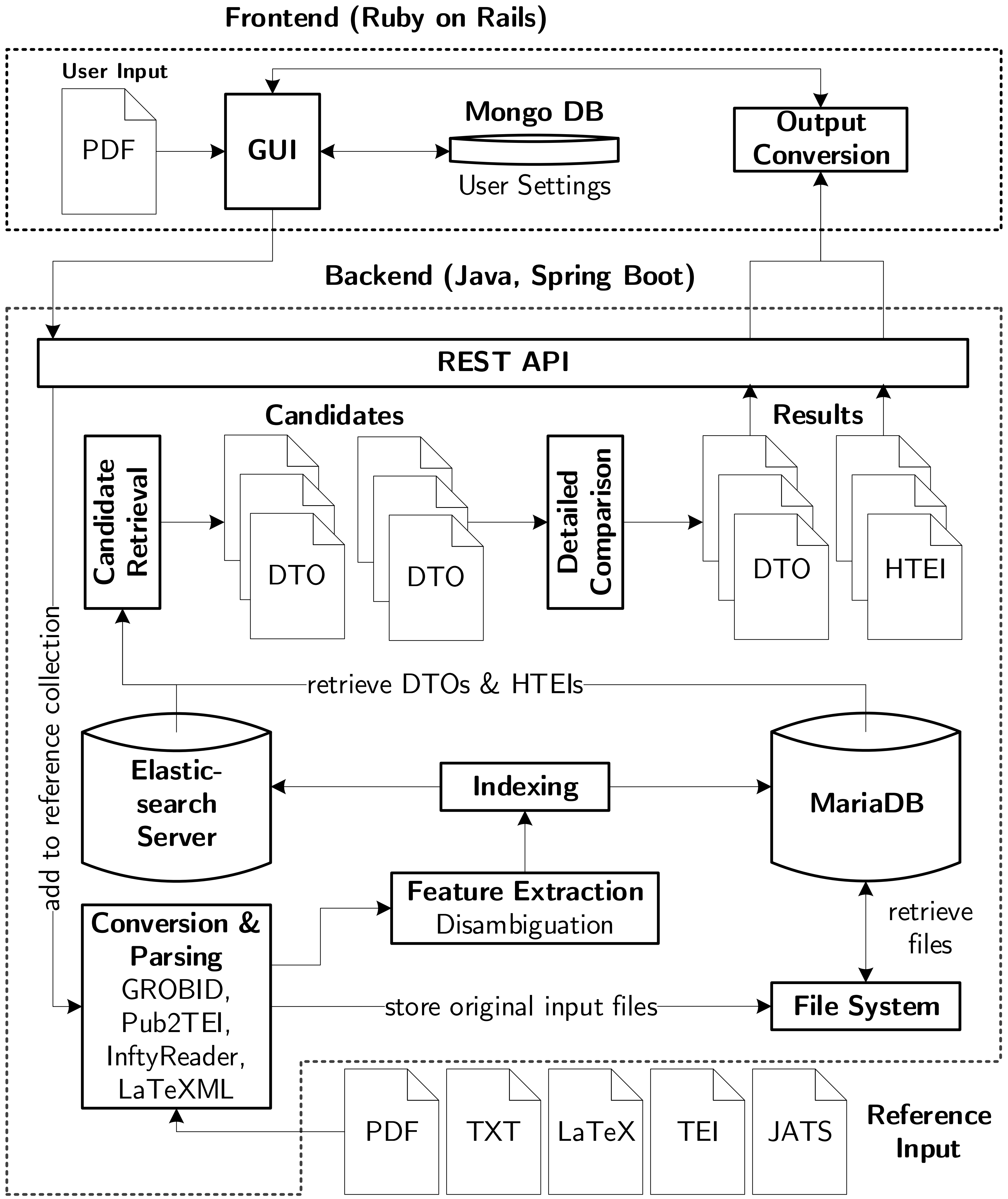
Math-based PD — Exploratory Search

- **Retrieve consolidated candidate set** (100 documents) using best-performing math-based, citation-based, and text-based methods for all 102,524 documents
- **Detailed analysis** of all candidate documents
- **Manual Investigation** of top-10 results

Math-based PD — Results Exploratory Search

Rank	Case ID	Rating
1	C3	Confirmed plagiarism case
2	C11	Author-confirmed case
3	C12	Notable legitimate content reuse
4	C13	False-positive detection
5	C10	Confirmed plagiarism case
6	C14	False-positive detection
7	C15	Notable legitimate content reuse
8	C16	Notable legitimate content reuse
9	C17	Notable legitimate content reuse
10	C18	Notable legitimate content reuse

Plagiarism Detection System Prototype — HyPlag



HyPlag Frontend Demo

Video

Conclusion & Outlook

Key Contributions — 1

RT1 **Identify the strengths and weaknesses of state-of-the-art methods and systems to detect academic plagiarism.**

- Most comprehensive literature review on plagiarism detection technology to date (376 papers, 25 year-period)

RT2 **Devise detection approaches that address the identified weaknesses.**

- Initiated the research on analyzing non-textual content in addition to text for PD use case
- Introduced two novel detection approaches: citation-based PD and mathematics-based PD
- Extended prior work on image-based PD

Key Contributions — 2

RT3 Evaluate the effectiveness of the proposed detection approaches.

- 5 Evaluations using confirmed cases of plagiarism and exploratory searches in large-scale collections
- Non-textual detection methods complement text-based methods and often outperform them for disguised plagiarism forms
- Identified 10 previously undiscovered cases of plagiarism

RT4 Implement the proposed detection approaches in a plagiarism detection system capable of supporting realistic detection use cases.

- HyPlag integrates the analysis of citations, images, mathematical content, and text
- Backend enables hybrid plagiarism detection for large-scale collections

Future Work (Selection)

1. Extend and improve detection methods

- Extending Math-based PD & related information extraction and retrieval technologies
- Improving the hybrid approach, e.g., neural language models, sequential pattern analysis

GI 1259/5 (GI 1259/1)

2. Create productive hybrid plagiarism detection system

- Improve frontend
- Extend reference collection

BERGISCHE UNIVERSITÄT WUPPERTAL

Bundesministerium für Bildung und Forschung

LIS

3. Research confidential, decentralized PD

- Devise confidential similarity analysis and visualization
- Develop distributed, blockchain-backed detection process

GI 1259/6

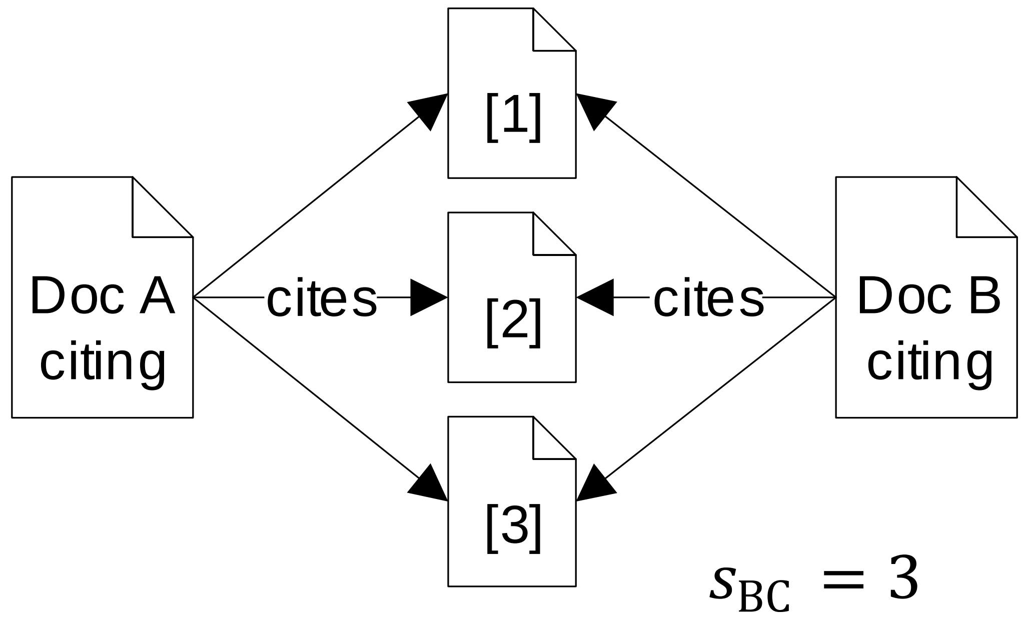
SFB “Structural Transformation of Trust”

Thank You for Your Attention!

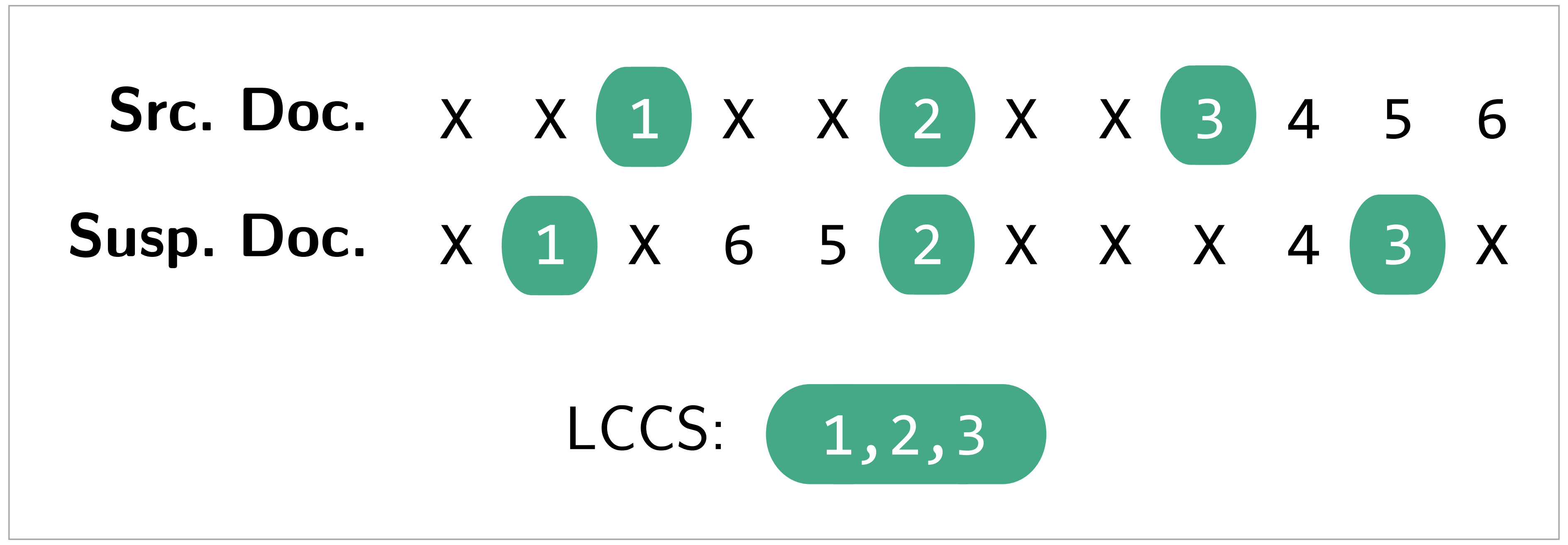
Rushed or Unmentioned Topics
Citation-based PD
Detection Methods
Preliminary Experiments
Large-scale Evaluation Methodology
Results Retrieval Effectiveness
User Utility
Computational Efficiency
Image-based PD
Detection Methods
Detection Process
Relevance Scoring
Evaluation Results
Math-based PD
Categorization of Detection Methods
Determination of Significance Thresholds
Newly Discovered Case
HyPlag
Full System Demo

I'm happy to answer Your Questions!

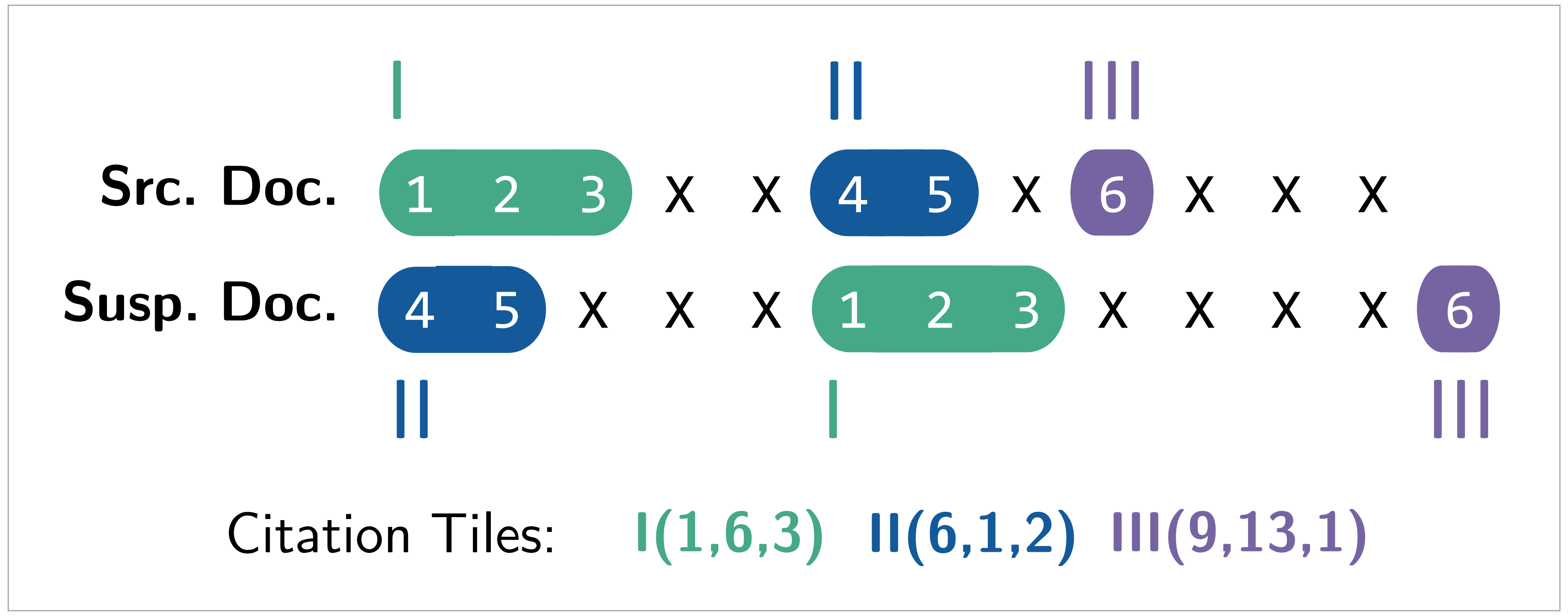
Citation-based Detection Methods



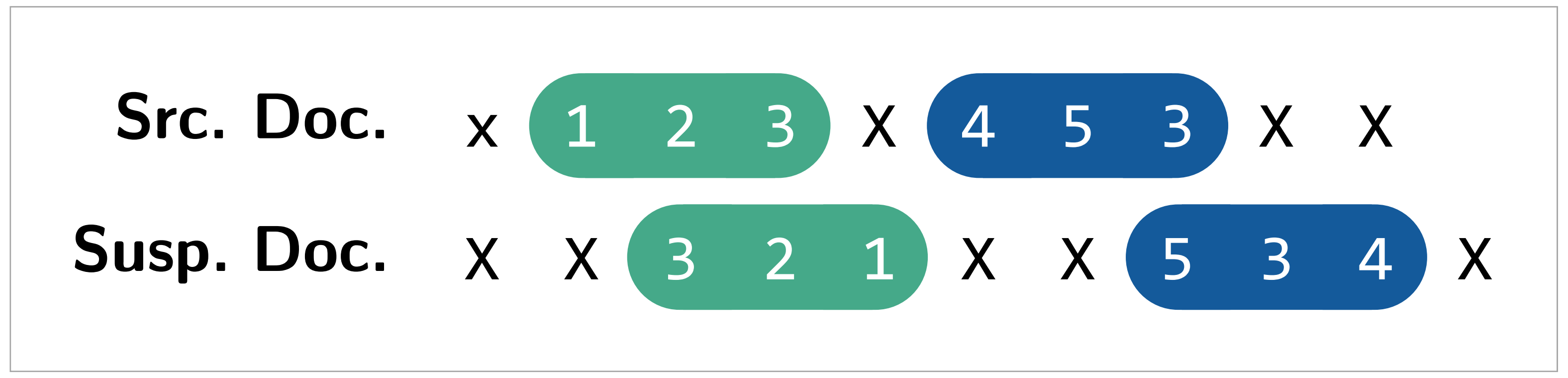
Bibliographic Coupling



Longest Common Citation Sequence

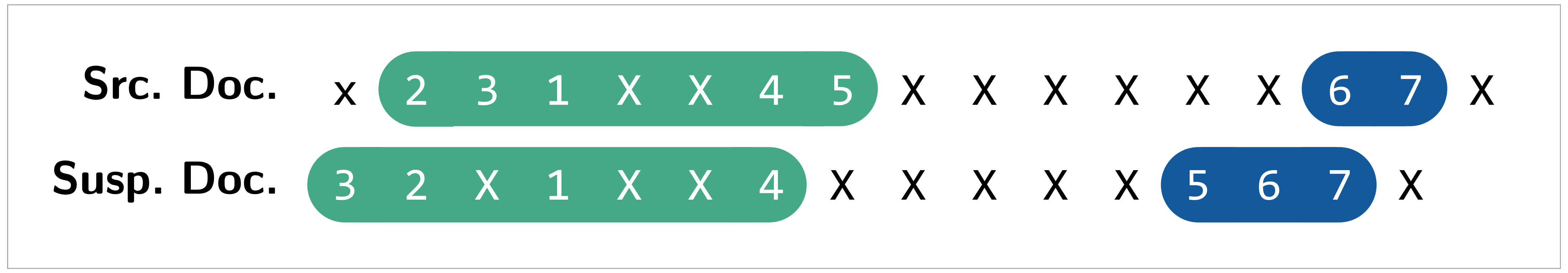


Greedy Citation Tiling



Citation Chunking

(consecutive citations only)



Citation Chunking

(depending on previous citations)

Citation-based Plagiarism Detection – Preliminary Experiments

Analysis of translated plagiarism in doctoral thesis of K.T. zu Guttenberg

Page	Documents	Citation Patterns
30	Bouton01	
	Guttenberg06	
39	CRS92_Pream.	
	Guttenberg06	
44	Tushnet99	no shared citations
223	Vile91	
	Guttenberg06	
224	CRS92_Art.V	
	Guttenberg06	
225	Vile91	
	Guttenberg06	
226 f.	CenturyFnd99	no shared citations
229 - 231	CRS92_Art.V	
	Guttenberg06	
	Vile91	
232 - 233	CRS92_Art.V	
	Guttenberg06	
	Vile91	
234	Vile91	
	Guttenberg06	
235 -	CRS92_Art.V	
239	Guttenberg06	
240 -	CRS92_Art.V	
242	Guttenberg06	
242 -	CRS92_Art.V	
244	Guttenberg06	
246 -	Vile91	
247	Guttenberg06	
267 -	Murphy00	
268	Guttenberg06	
300	Buck96	no shared citations

Example of a cleaned citation pattern:

242 -	CRS92_Art.V	
244	Guttenberg06	

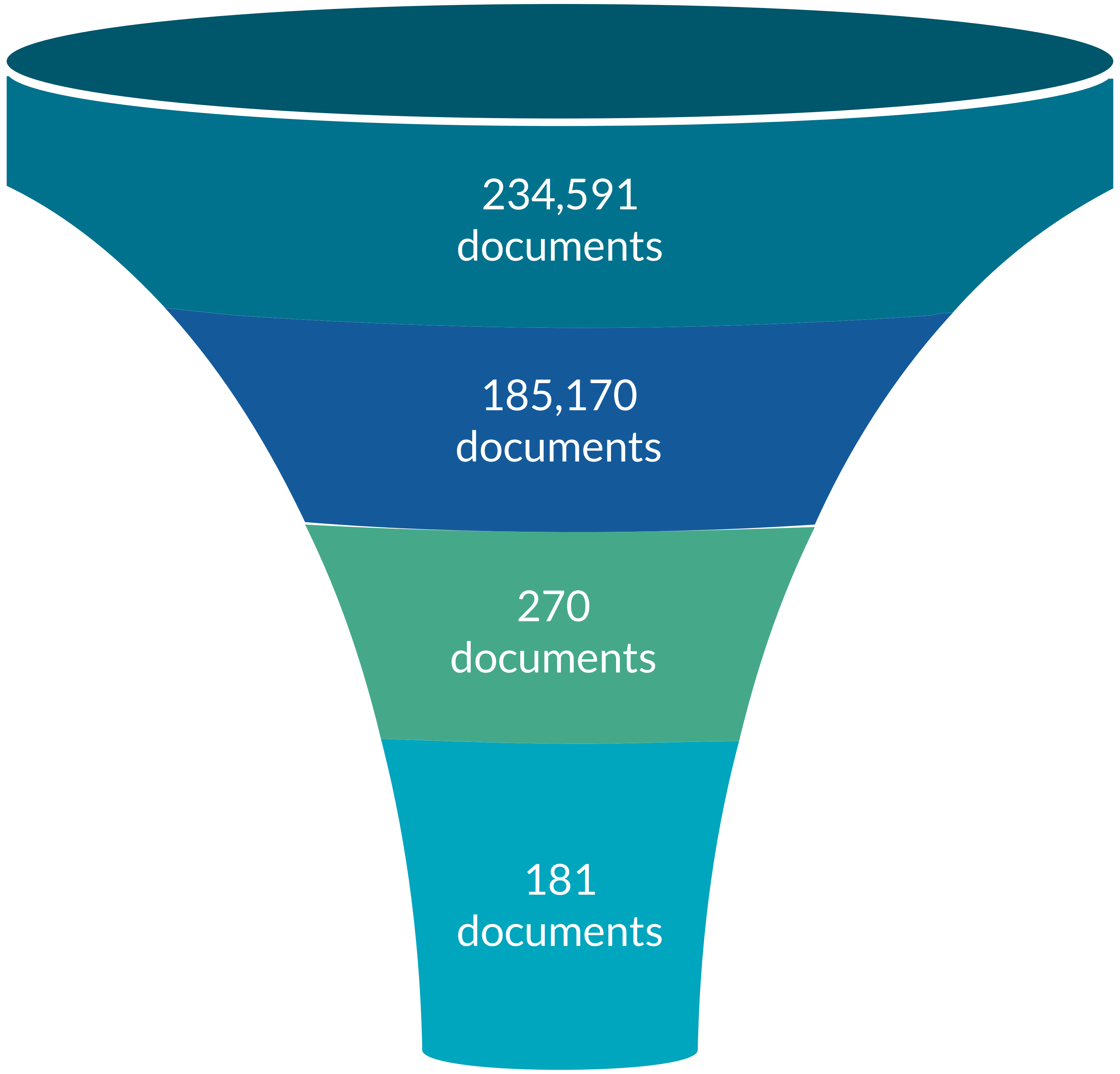
242 -	CRS92_Art.V	
244	Guttenberg06	

Citation-based Plagiarism Detection – Evaluation Methodology

PMC

PubMed Central Open Access Subset
- Full-text articles from medicine and life sciences openly available in an XML format

Results Pooling
- Pooling the top-30 results for 7 citation-based and 2 text-based detection methods

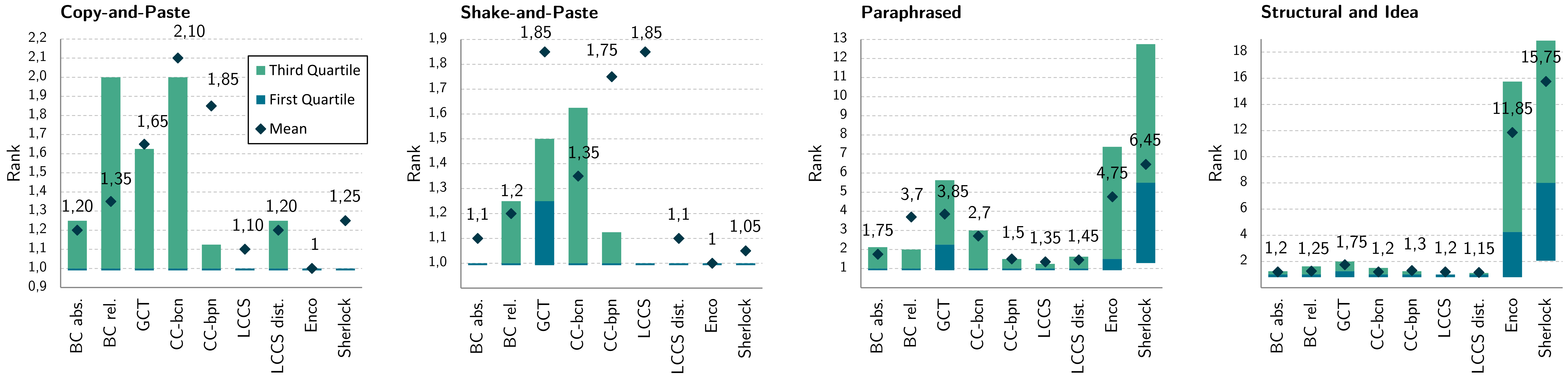


Preprocessing
- 49,421 documents excluded: no text available (scans), duplicates, no references or citations, etc.

Relevance Judgment
- 5 medical experts, 10 medical and life science graduate students, 11 undergraduate students (various majors)
- Numerical scoring (0 = false positive 5 = very strong suspicion)
- Expertise-weighted average

Citation-based PD — Results Retrieval Effectiveness

Distribution of ranks for the 10 document pairs with the highest suspiciousness scores per category



Follow-up for identified suspicious documents:

- 4 retracted articles
- 5 author-confirmed cases of plagiarism

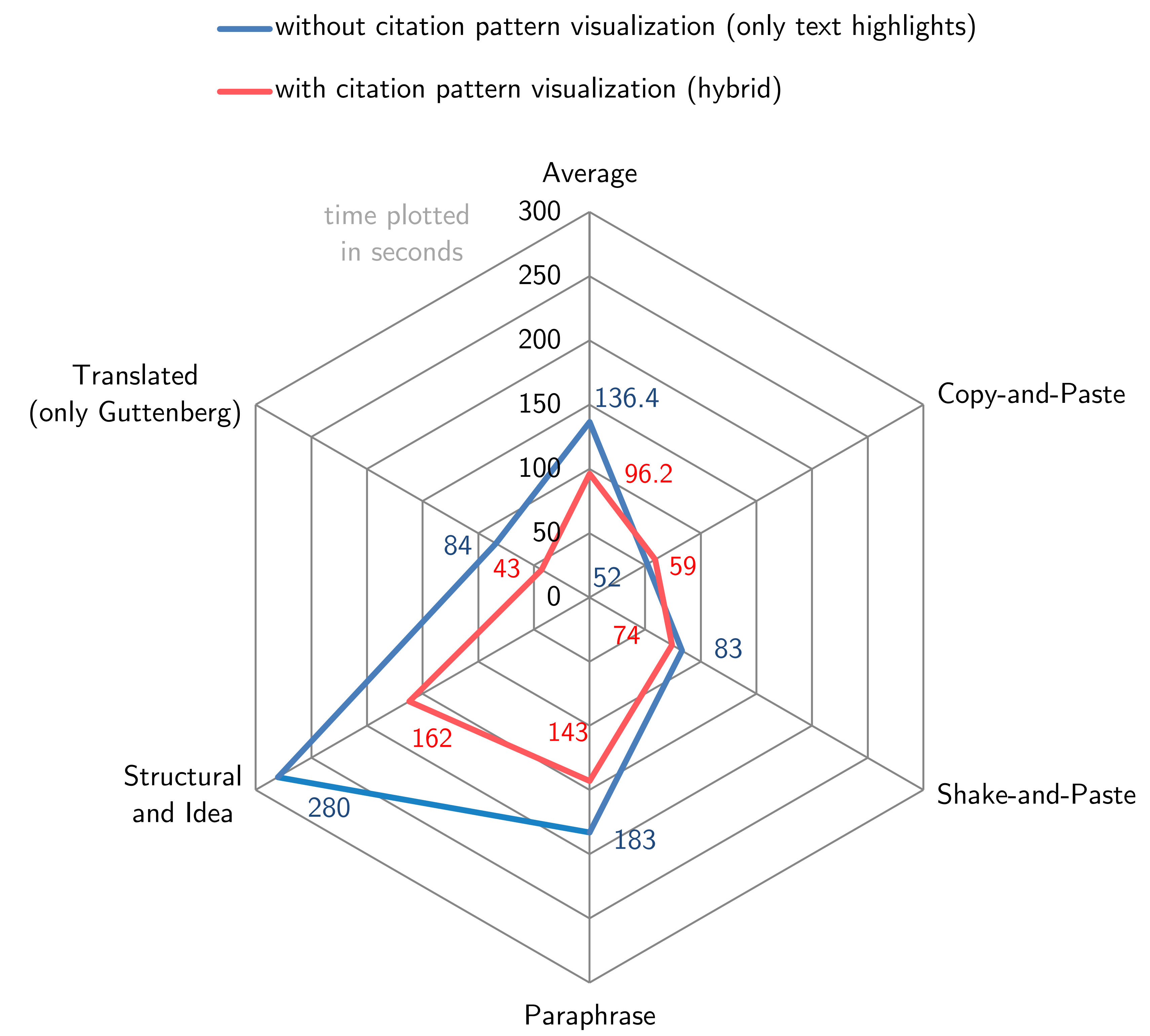
Citation-based Plagiarism Detection – User Utility

Visualization users (N=26; 13 for transl.) perceived as most beneficial for analyzing plagiarism forms (D=461 document pairs)

	Copy-and-Paste	Shake-and-Paste	Para-phrased	Structural and Idea	Trans-lated*
Text-based	51%	27%	6%	1%	-
Citation-based	1%	5%	32%	86%	54%
Hybrid	47%	68%	62%	13%	46%

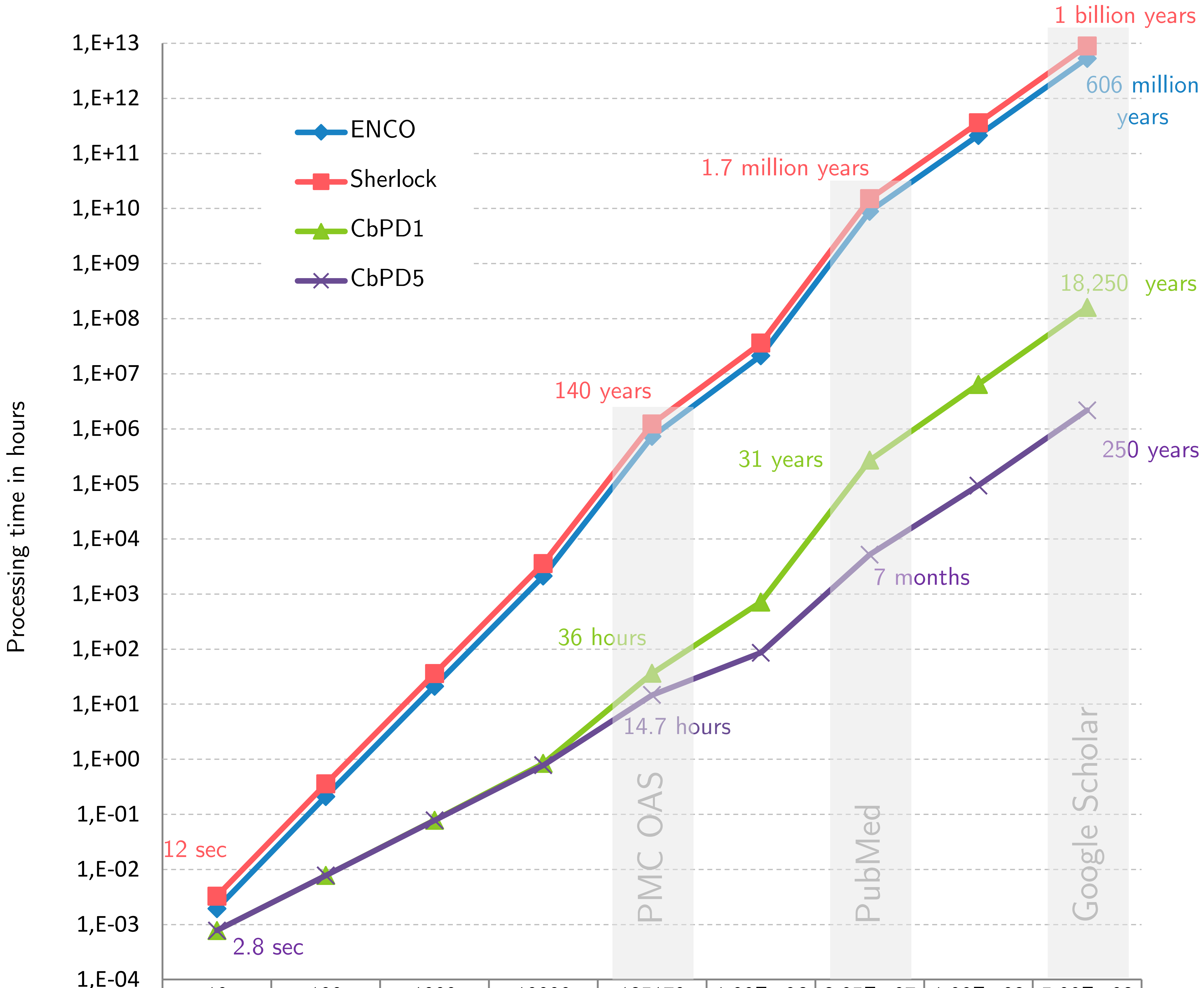
* examination of Guttenberg thesis only

Avg. time required for verifying first two plag. Instances (N=8, D=8x25)



Citation-based Plagiarism Detection – Computational Efficiency

Average case processing times of detection methods by collection size.

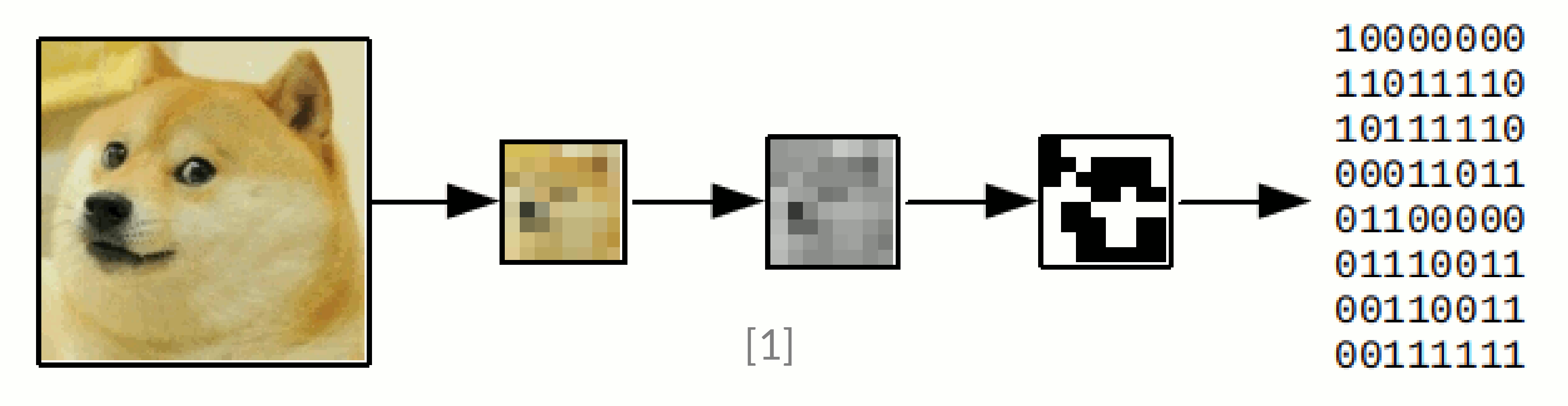


	10	100	1000	10000	185170	1,00E+06	2,05E+07	1,00E+08	5,00E+08
ENCO	1,9E-03	2,1E-01	2,1E+01	2,1E+03	7,3E+05	2,1E+07	8,9E+09	2,1E+11	5,3E+12
Sherlock	3,3E-03	3,6E-01	3,6E+01	3,6E+03	1,2E+06	3,6E+07	1,5E+10	3,6E+11	9,0E+12
CbPD1	7,8E-04	7,8E-03	7,8E-02	8,4E-01	3,6E+01	7,2E+02	2,7E+05	6,4E+06	1,6E+08
CbPD5	7,8E-04	7,8E-03	7,8E-02	7,8E-01	1,5E+01	8,6E+01	5,2E+03	9,3E+04	2,2E+06

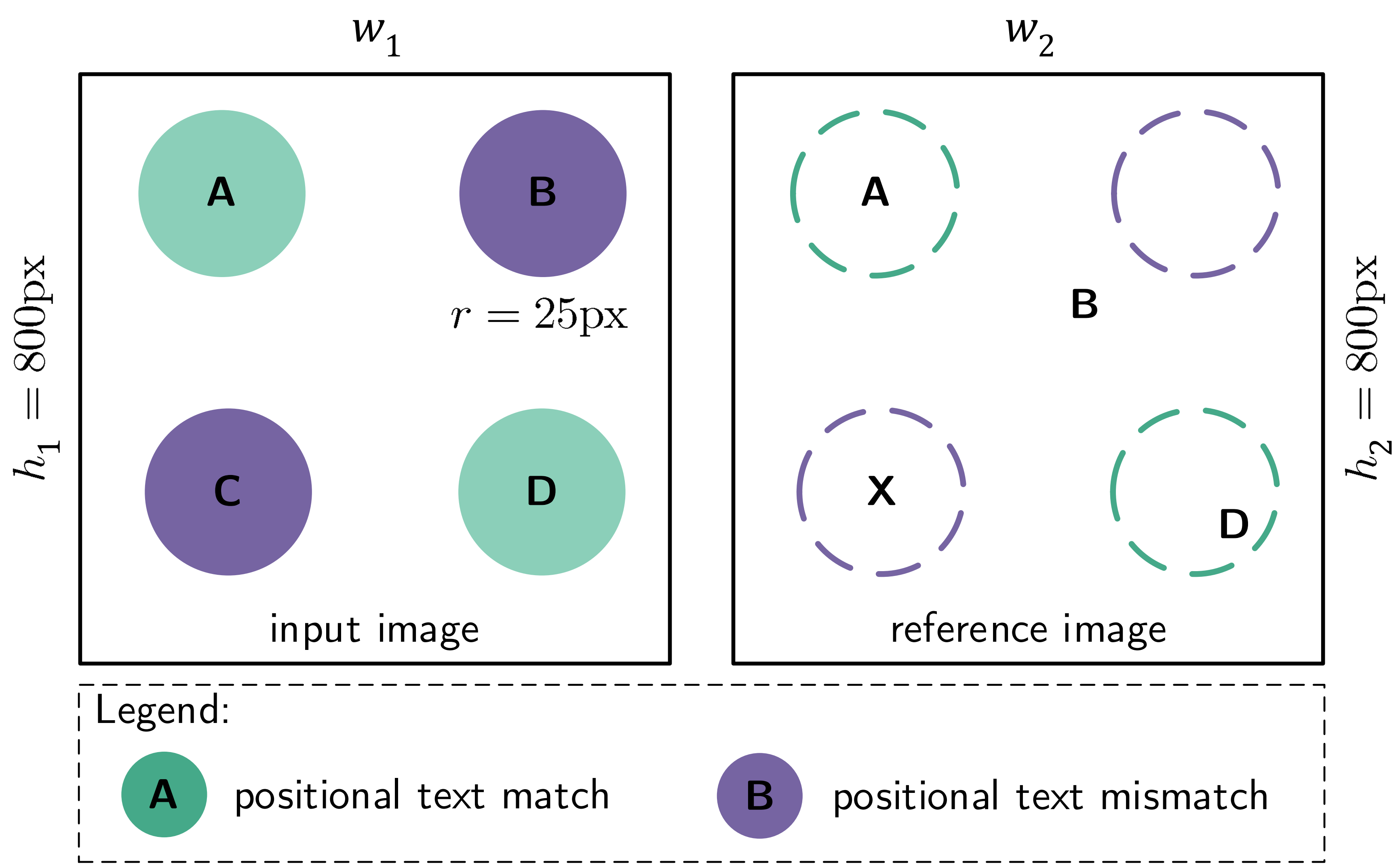
The first row lists the collection size. The rows underneath show processing times in hours (partially extrapolated).

Image-based Plagiarism Detection Methods

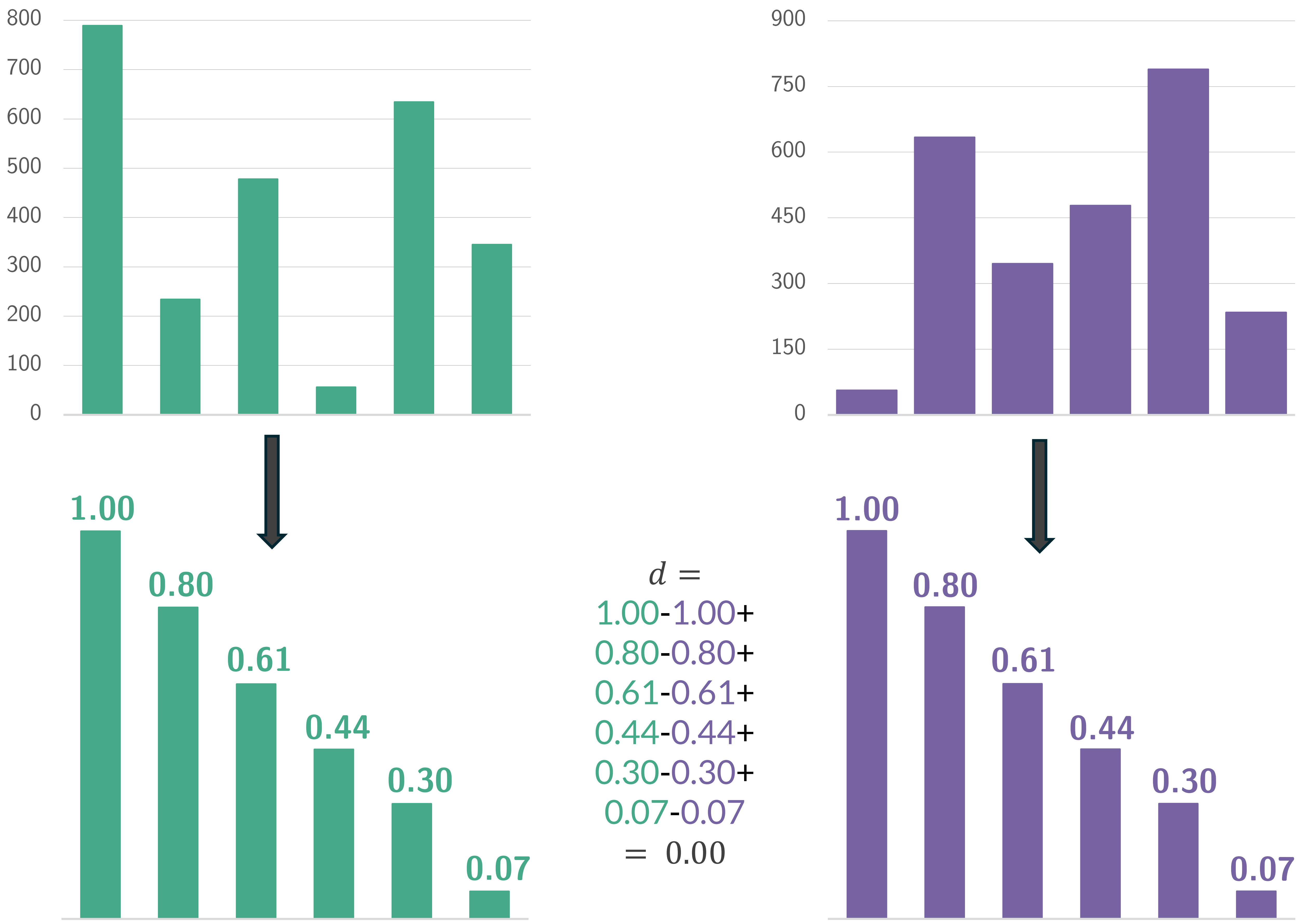
Perceptual Hashing



(Positional) Text Matching



Ratio Hashing



[1] Image Source: https://medium.com/taringa-on-publishing/why-we-built-imageid-and-saved-47-of-the-moderation-effort-b7afb69d068e

Image-based Plagiarism Detection Process

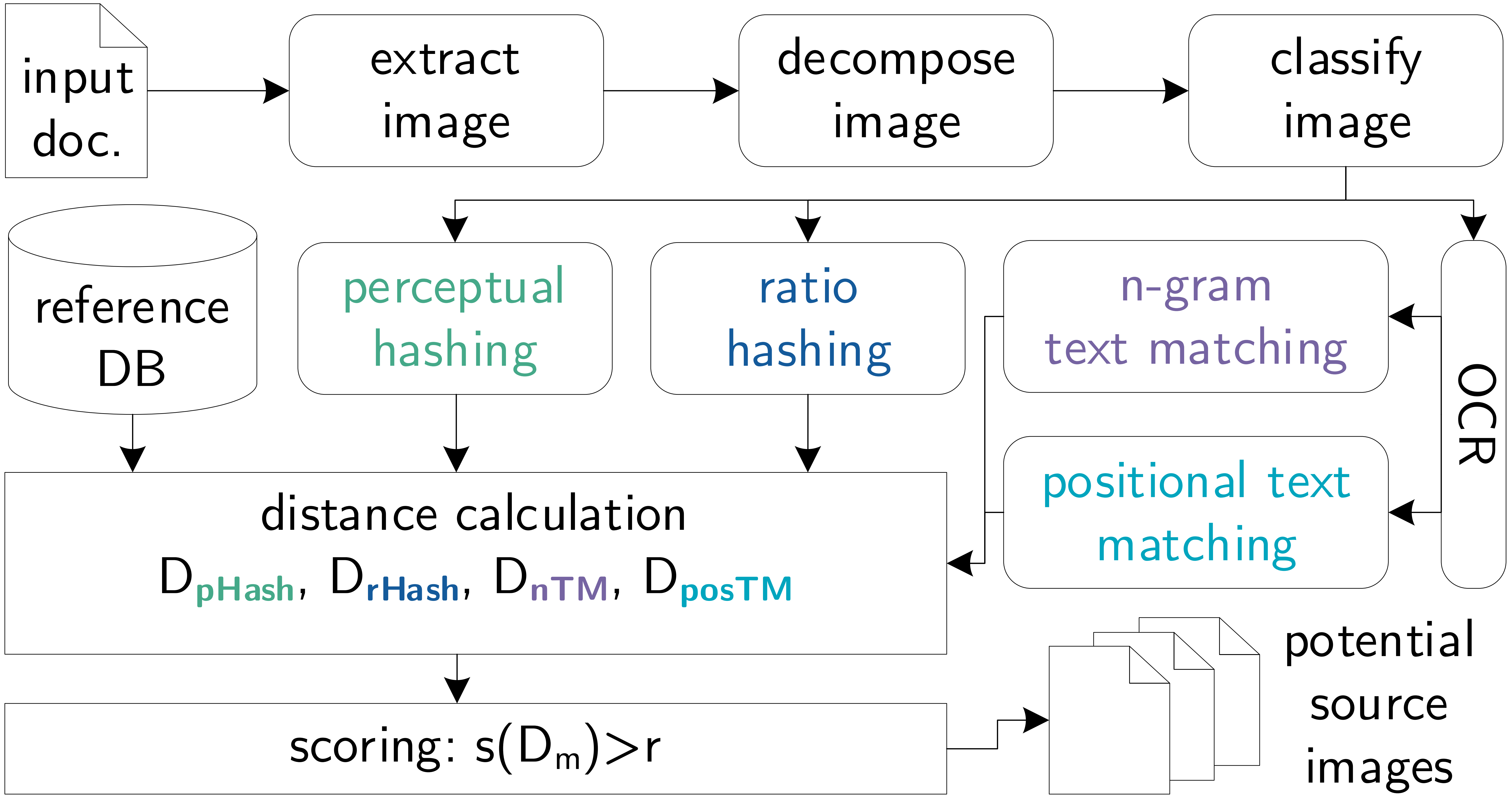


Image-based PD — Relevance (Suspiciousness) Scoring

Requirements on suspicious images:

1. Highly similar images are clear outliers.
2. The outlier group is small.

Final similarity score

$$s = \frac{\bar{d}}{1+\bar{d}} \qquad \bar{d} = \frac{\max(d_i' \in D_{m,1}')}{t}.$$

Margin of least similar outlier image to remainder of collection:

s=0.5 … 2x distance to input image

s=0.75 … 3x distance to input image

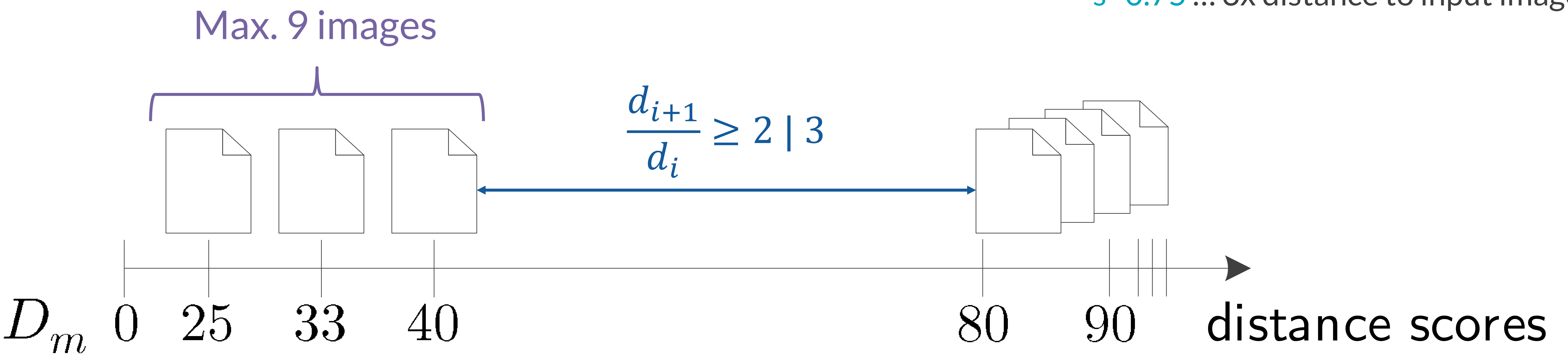


Image-based Plagiarism Detection — Evaluation Results

Similarity scores for input images.

#	Image Type	Alteration	pHash	nTM	posTM	rHash
1	Illustration	near copy	**0.87**	< 0.5	< 0.5	-
2	Illustration	near copy	**1.00**	0.79	0.77	-
3	Illustration	near copy	**0.86**	< 0.5	< 0.5	-
4	Illustration	weak	**0.78**	< 0.5	< 0.5	-
5	Illustration	weak	**0.57**	< 0.5	< 0.5	-
6	Illustration	moderate	< 0.5	**0.87**	< 0.5	-
7	Illustration	strong	< 0.5	< 0.5	< 0.5	-
8	Bar Chart	near copy	0.62	0.64	0.77	**0.92**
9	Table	near copy	< 0.5	< 0.5	< 0.5	-
10	Table	near copy	0.62	**0.71**	0.55	-
11	Table	near copy	< 0.5	**0.92**	< 0.5	-
12	Table	weak	< 0.5	**0.79**	< 0.5	-
13	SEM Image	near copy	< 0.5	< 0.5	< 0.5	-
14	Line Chart	weak	< 0.5	< 0.5	< 0.5	-
15	Line Chart	strong	< 0.5	**0.70**	< 0.5	-

$P = 1$ $\quad R = \frac{11}{15} = 0.73$ $\quad F_1 = 0.84$

Ranks at which the detection process retrieved source images.

#	Image Type	Alteration	pHash	nTM	posTM	rHash
1	Illustration	near copy	**1**	> 10	> 10	-
2	Illustration	near copy	**1**	**1**	**1**	-
3	Illustration	near copy	**1**	> 10	> 10	-
4	Illustration	weak	**1**	> 10	> 10	-
5	Illustration	weak	**1**	> 10	> 10	-
6	Illustration	moderate	**1**	**1**	> 10	-
7	Illustration	strong	**1**	> 10	> 10	-
8	Bar Chart	near copy	**1**	**1**	**1**	**1**
9	Table	near copy	> 10	> 10	> 10	-
10	Table	near copy	**1**	**1**	**1**	-
11	Table	near copy	**1**	**1**	> 10	-
12	Table	weak	> 10	**1**	> 10	-
13	SEM Image	near copy	**1**	> 10	> 10	-
14	Line Chart	weak	> 10	> 10	> 10	-
15	Line Chart	strong	> 10	**1**	> 10	-

Math-based PD — Detailed Analysis Methods

	Global Similarity Assessment	**Local Similarity Assessment**
Set-based (Order-agnostic)	Identifier Histograms	Identifier Histograms (outperformed)
Sequence-based (Order-observing)	Longest Common Identifier Sequence	Greedy Identifier Tiling

Math-based PD — Determining Significance Thresholds

- **Goal**: Derive approximation for maximum similarity by chance
- **Analysis**: of score distribution for 1M (hopefully) unrelated document pairs (no common authors, do not cite each other)
- Threshold = score of highest ranked document pair without noticeable topical relatedness

	Histo	LCIS	GIT	BC	LCCS	GCT	Enco
s	$\geq .56$	$\geq .76$	$\geq .15$	$\geq .13$	$\geq .22$	$\geq .10$	$\geq .06$

Math-based PD — Newly Discovered Case

Source Documents (S1, S2)

also [23]). Some thermodynamic quantities associated with the cosmological horizon are

$$T=\frac{1}{4\pi r_c}\left(-(n-1)+(n+1)\frac{r_c^2}{l^2}+\frac{n\omega_n^2Q^2}{8r_c^{2n-2}}\right),$$
$$S=\frac{r_c^n\,\mathrm{Vol}(S^n)}{4G},\qquad \phi=-\frac{n}{4(n-1)}\frac{\omega_nQ}{r_c^{n-1}},\tag{3.2}$$

where ϕ is the chemical potential conjugate to the charge Q. In the BBM prescription, the gravitational mass, subtracted the anomalous Casimir energy, of the RNdS solution is

$$E=-M=-\frac{r_c^{n-1}}{\omega_n}\left(1-\frac{r_c^2}{l^2}+\frac{n\omega_n^2Q^2}{8(n-1)r_c^{2n-2}}\right).\tag{3.3}$$

The Casimir energy E_c, defined as $E_c=(n+1)E-nTS-n\phi Q$ in this case, is found to be

$$E_c=-\frac{2nkr_c^{n-1}\,\mathrm{Vol}(\sigma)}{16\pi G}.\tag{3.9}$$

When $k=0$, the Casimir energy vanishes, as the case of asymptotically AdS spaces. This is expected since

which has a same form as the case of SdS solution. Thus we can see that the entropy (3.2) of the cosmological horizon can be rewritten as[7]

$$S=\frac{2\pi l}{n}\sqrt{|E_c|\big(2(E-E_q)-E_c\big)},\tag{3.5}$$

where

$$E_q=\frac{1}{2}\phi Q=-\frac{n}{8(n-1)}\frac{\omega_nQ^2}{r_c^{n-1}}.\tag{3.6}$$

Suspicious Document

Some thermodynamic quantities associated with the cosmological horizon are

$$T=\frac{1}{4\pi r_c}\left(-(n-1)k+(n+1)\frac{r_c^2}{l^2}+\frac{n\omega_n^2Q^2}{8r_c^{2n-2}}\right),$$
$$S=\frac{r_c^n\,\mathrm{Vol}(\sigma)}{4G},\tag{5}$$
$$\phi=-\frac{n}{4(n-1)}\frac{\omega_nQ}{r_c^{n-1}},$$

where ϕ is the chemical potential conjugate to the charge Q.

The Casimir energy E_c, defined as $E_c=(n+1)E-nTS-n\phi Q$ in this case, is found to be

$$E_c=-\frac{2nkr_c^{n-1}\,\mathrm{Vol}(\sigma)}{16\pi G},\tag{6}$$

when $k=0$, the Casimir energy vanishes, as the case of asymptotically AdS space. When $k=\pm1$, we see from Eq. (6) that the sign of energy is just contrast to the case of TRNAdS space.[30]

Thus we can see that the entropy Eq. (5) of the cosmological horizon can be rewritten as

$$S=\frac{2\pi l}{n}\sqrt{\left|\frac{E_c}{k}\right|\big(2(E-E_q)-E_c\big)},\tag{7}$$

where

$$E_q=\frac{1}{2}\phi Q=-\frac{n}{8(n-1)}\frac{\omega_nQ^2}{r_c^{n-1}}.\tag{8}$$

RT3

Issues Arising from the Limited Detection Capabilities

Likely, we only see the tip of the iceberg.

Prevalence of plagiarism is probably significantly larger.

Building a better sonar for underwater icebergs.

The lower part of the iceberg is typically more dangerous.

- Current detection tools focus on students who plagiarize due to a lack of time or skill.
- Researchers typically have more skills, time, and incentives to obfuscate plagiarism.
- Plagiarism in research publications has higher potential damage
    - Systematic reviews (!)
    - Wasted effort