A Comprehensive Evaluation Method for Paper Plagiarism Detection Postprint

Authors: Xie Zhaoxian, Shuzhen Ye, Huang Shenquan

Date: 2018-05-24T00:00:00+00:00

Abstract

Plagiarism detection constitutes a necessary component in the quality control process of university graduation theses. This study proposes a novel comprehensive methodology capable of identifying anomalies when internet-based plagiarism detection systems operate in an overly rigid manner, thereby triggering subsequent manual review. The primary objective is to reduce false positive rates and provide an appeals mechanism for papers incorrectly flagged as plagiarized despite containing no actual plagiarism. When substantial discrepancies arise in the similarity rates detected for the same thesis across different plagiarism detection platforms, this novel comprehensive approach incorporates human judgment weighting into the final determination of similarity results. This reduces the degree to which similarity rates are controlled by website-specific factors, ensuring that detection outcomes are not entirely dominated by either automated systems or human assessment. This dual hybrid detection paradigm, combining website-based analysis with human review, compensates for the impact of database limitations inherent to individual platforms on detection results, thereby enhancing both the accuracy and credibility of plagiarism detection outcomes.

Full Text

Preamble

A Comprehensive Plagiarism Detection Evaluation Method for Academic Papers

Xie Zhaoxian, Ye Shuzhen†, Huang Shenquan

(School of Mechanical & Electrical Engineering, Wenzhou University, Wenzhou, Zhejiang 325035, China)

Abstract: Plagiarism detection is essential for quality control of university graduation theses. This paper proposes a novel comprehensive approach to address the extreme inconsistencies inherent in internet-based plagiarism detection systems. This method identifies anomalies through automated detection before requiring manual re-evaluation, aiming to reduce miscarriages of justice by providing opportunities for appeal when non-plagiarized work is incorrectly flagged. When the same paper yields substantially different similarity rates across different detection platforms, this comprehensive method incorporates human judgment weights into the final plagiarism determination, reducing the influence of website-controlled factors. This hybrid approach ensures that detection results are not entirely controlled by either automated systems or human evaluators, compensating for database limitations and improving the accuracy and credibility of plagiarism detection outcomes.

Keywords: graduation thesis; duplicate checking website; manual detection; plagiarism

Classification: TP391

0 Introduction

Currently, many Chinese universities require students to submit graduation theses as a means of developing comprehensive knowledge understanding, verification skills, and academic writing abilities. However, despite instructors' diligent guidance in topic selection and supervision, they cannot thoroughly examine every aspect of each student' s research work, leaving gaps in detecting plagiarized content. Fortunately, with the development of the internet, many institutions now submit theses to online plagiarism detection platforms for preliminary screening to obtain objective similarity assessments.

While these platforms assist in education by identifying potentially plagiarized papers based on textual similarity, the reported percentages are determined by each platform' s proprietary database content. Consequently, different detection mechanisms produce varying similarity ratios for the same paper, and high similarity percentages do not necessarily indicate actual plagiarism. Traditional detection systems primarily function as text-matching tools, making them prone to misjudgment. This paper emphasizes reducing false positives by analyzing existing detection methods and proposing a novel plagiarism detection framework.

1 Problem and Design

The proliferation of plagiarism detection services has created both business opportunities and technical challenges. Numerous platforms have emerged, offering both free and paid services for authors and institutions to review academic papers. These services give rise to four possible detection outcomes: (1) actual plagiarism correctly identified; (2) no plagiarism incorrectly flagged; (3)

actual plagiarism missed; and (4) no plagiarism correctly identified. The first and fourth scenarios represent desired outcomes, while the third reflects limitations in database coverage or algorithmic scope. However, this paper focuses on the second scenario—false positives where original work is misidentified as plagiarized.

Furthermore, since each platform employs different algorithms and databases, detection results vary significantly across websites. This raises critical questions: Which result should be trusted? Is a higher similarity rate necessarily more accurate? We argue that stricter algorithms artificially inflate similarity scores. This paper addresses two key issues: assisting cases of false positives and determining reliable similarity thresholds. Our proposed solutions include: (a) analyzing methods to modify references and text to reduce misjudgment, such as altering mathematical notation and textual descriptions; and (b) employing statistical and inductive methods to analyze credible similarity ratios through controlled experiments.

2.2 System Architecture

As shown in [Figure 1: see original paper], the general process for online plagiarism detection involves three stages: input (submitting the paper), processing (algorithmic analysis), and output (generating similarity reports with percentages and source information). This can be simplified as an input-processing-output pipeline.

[Figure 2: see original paper] illustrates our experimental methodology for obtaining detection results from free platforms. The same paper is submitted to three systems (A, B, and C), yielding results a, b, and c respectively. By comparing these outcomes, we identify detection patterns and problems to derive solutions.

Based on this architecture, we developed an enhanced framework called New Check Repeat (NCP), depicted in [Figure 3: see original paper]. The NCP architecture comprises five components:

a) **Acquisition Layer:** Contains an input module where users submit manuscripts for detection.

b) **Data Layer:** Includes databases and full-text retrieval systems. The database serves as a data warehouse for organizing, storing, and managing content, while the full-text retrieval system indexes every word in documents to enable efficient searching.

c) **Processing Layer:** Contains a query calculation module that computes similarity rates using proprietary algorithms and identifies duplicated content and its sources.

d) **Core Business Layer:** Includes testing and manual judgment modules. The testing module compares similarity rates from different platforms,

while the manual judgment module enables human review of flagged content.

e) **Users:** The system serves academic publishers, research administrators, journal editors, universities, and the public.

2.1 Classification and Description of Domestic Plagiarism Detection Websites

Through practical testing of popular Chinese plagiarism detection platforms, we summarize their characteristics:

a) **CNKI:** High accuracy and fast processing, but expensive.
b) **VIP:** Low cost, fast detection, and relatively high accuracy, but cannot recognize tables and has limited foreign language paper coverage.
c) **Wanfang:** Low cost but limited database resources.
d) **PaperPass:** Enables real-time online editing with efficient duplication comparison and moderate accuracy, but does not support English detection.
e) **CopyCheck:** Provides extensive web data, similarity detection, and English support, but has limited resources and moderate accuracy.
f) **ChinaSou Article Mirror:** Free, fast detection, but unstable with low accuracy.
g) **Daya:** Free with unlimited daily queries and fast processing, but limited database makes results for reference only.
h) **Gezida:** Free with precise results and online editing, but no English support, incomplete database, and shows high citation rates with low plagiarism rates.
i) **PaperFree:** Accurately identifies potential plagiarism and improper citations with pay-per-sentence modification features, but less strict detection.
j) **Thesis Dog:** Free with unlimited word count and real-time database updates, but limited resources and lower accuracy.
k) **PaperTest:** Fast detection with modification suggestions, but limited literature coverage.

2.3 Method

Based on experimental analysis, we developed an improved detection workflow ([Figure 4: see original paper]) to enhance accuracy and prevent misjudgment. The specific process involves: submitting a paper to Website A to obtain similarity rate A-R, then to Website B to obtain B-R. The testing module evaluates these results:

If both A-R and B-R are below 20% (the threshold cited by Chinese academic guidelines where "plagiarism exceeding 20% incurs serious consequences"), the system concludes the paper has low similarity and calculates composite rate R2:

$$R2 = \alpha \times \max\{A-R, B-R\} + (1-\alpha) \times \min\{A-R, B-R\}$$

where represents confidence level. Generally, higher confidence yields more reliable intervals but reduces precision. The value is determined by thesis reviewers based on actual circumstances, typically set at 95% confidence according to statistical standards. When calculating R2, greater weight is assigned to the higher similarity rate because it suggests the platform has more relevant resources and thus higher credibility, while the lower rate receives weight (1-).

Otherwise, the system computes the difference C between A-R and B-R. If C is less than /2, the results are within acceptable error margins, and R2 is used. If C exceeds /2, manual judgment H is introduced, and the final rate R1 is calculated as:

$$R1 = 0.8 \times H + 0.2 \times [\alpha \times \max\{A-R, B-R\} + (1-\alpha) \times \min\{A-R, B-R\}]$$

In calculating R1, 80% weight is assigned to manual judgment H because similarity rates cannot exceed 20%. This ensures that even if automated systems flag complete plagiarism, expert review can override the decision based on experience.

Workflow Example: A paper yields A-R = 15% from Website A and B-R = 80% from Website B. Since B-R > 20%, we compute C = |15% - 80%| = 65%. With = 95%, /2 = 47.5%, and since C = 65% > 47.5%, manual judgment is triggered. If manual assessment yields H = 18%, the final similarity rate is:

$$R1 = 0.8 \times 18\% + 0.2 \times [0.95 \times 80\% + (1-0.95) \times 15\%] = 29.75\% > 20\%$$

3.1 Experimental Content

We selected a paper titled "Several Attempts to Help High School Students Overcome Mathematics Learning Difficulties" and applied four modification methods:

a) **Formula Modification:** Altering numbers, operators, algorithms, functions, and logic in mathematical expressions.
b) **Text Modification:** Changing words or sentences either meaningfully (altering original meaning) or meaninglessly (mechanical changes rendering text nonsensical).
c) **Formatting Modification:** Adjusting positions and sizes of text, images, and graphics in the layout. While this may reduce similarity rates, it produces poorly readable content with similar sentence structures, potentially causing misjudgment—a concern this paper addresses. Note that developing new detection algorithms is beyond this paper' s scope.

3.2 Analysis Results

We tested the paper using three free platforms: Daya, ChinaSou Article Mirror, and Gezida. [Figure 5: see original paper] shows normalized similarity rates across the four modification methods. Normalization calculates the ratio between modified and original similarity rates for each platform.

Key findings: Formatting modifications caused ChinaSou Article Mirror' s normalized rate to reach 1.6 (60% higher than the original), indicating high sensitivity to layout changes and thus lower credibility for such modifications. Formula modifications caused Daya' s normalized rate to drop to 0.57 (43% lower than original), showing Daya' s high sensitivity to formula changes and lower credibility for mathematical content modifications.

[Figure 6: see original paper] presents Gezida' s three metrics: self-written rate, rewrite rate, and citation rate. Self-written rate = (M - N)/M, where M is total valid segments and N is similar segments (including citations). Rewrite rate = (N - C)/M, where C is cited segment count. Citation rate = C/M. Notably, meaningless text modifications achieved 100% self-written rate and 0% rewrite rate, potentially causing actual plagiarism to be missed—a vulnerability for future research.

4 Conclusion

This paper proposes a quantitative plagiarism detection method to prevent misjudgment and compensate for limitations in internet-based detection. Experiments with deliberately modified papers reveal predictable patterns across platforms. Due to varying databases and algorithms, the same paper can produce substantially different results across websites, making purely automated judgments unreliable. Manual judgment becomes necessary when inter-platform results diverge significantly, reducing false positive factors and improving overall detection accuracy.

References

[1] Isoc D. Preventing plagiarism in engineering education and research [C]// Proc of International Symposium on FundAmentals of Electrical Engineering. 2014: 1-7. [2] https://baike.baidu.com/item/%E4%BA%92%E8%81%94%E7%BD%91/199186?fr=aladdin. (Internet [DB/OL]. [3] 郭平, 王可, 罗阿理, 等. 大数据分析中的计算智能研究现状与展望 [J]. 软件学报, 2015, 26 (11): 3010-3025. (Guo Ping, Wang Ke, Luo Ali, et al. Present situation and prospect of computational intelligence in big data analysis [J]. Journal of Software, 2015, 26 (11): 3010-3025.) [4] Irwanto M R, Zamara S B, Herdianto R, et al. SIPOC business model process to prevent plagiarism in an electronic journal [C]// Proc of the 3rd International Conference on Science in Information Technology. 2017: 492-497. [5] 大雅论文查重网站 [EB/OL]. http://dsa.dayainfo.com/. [6] 中国搜文章照妖镜论文查重网站 [EB/OL]. http://www.zhongguosou.com/zonghe/fanchaoxi.html. [7]

格子达论文查重网站 [EB/OL]. http://www.gezida.com/. [8] 知网 [EB/OL]. http://lwjc.cmscoo.cn/cncnki/?200. [9] 维普网 [EB/OL]. http://gocheck.cn. [10] 万方检测网站 [EB/OL]. http://www.wanfangdata.com.cn/. (Wanfangdata [DB/OL]. [11] PaperPass 论文查重网站 [EB/OL]. http://www.paperpass.com/. [12] 知识产权卫士-拷克网 [EB/OL]. http://www.copycheck.com.cn/index.html. [13] PaperFree 论文查重网站 [EB/OL]. http://www.paperfree.cn/. [14] 论文狗 [EB/OL]. http://www.lunwengo.net/. [15] PaperTest 论文查重网站 [EB/OL]. http://www.papertest.com/. [16] Abdi A, Shamsuddin S M, Idris N, et al. A linguistic treatment for automatic external plagiarism detection [J]. Knowledge-Based Systems, 2017, 135 (11): 135-146. [17] Gaspar P V, Velásquez Juan D. Docode 5: building a real-world plagiarism detection system [J]. Engineering Applications of Artificial Intelligence, 2017, 64 (9): 261-271. [18] Markus Eckerstorfer, Eirik Malnes. Manual detection of snow avalanche debris using high-resolution Radarsat-2 SAR images [J]. Cold Regions Science and Technology, 2015, 120 (12): 205-218. [19] 发表期刊查重率多少算合格——中国鸣网 [DB/OL]. http://mingmw.com/baike/bk/31523.html. (The publication of the journal weight rate is qualified-mingmw [EB/OL]. http://mingmw.com/baike/bk/31523.html.) [20] 丁正生. 概率论与数理统计简明教程 [M]. 北京: 高等教育出版社, 2005: 125-127. (Ding Zhengsheng. Brief tutorial on probability theory and mathematical statistics [M]. Beijing: advanced education press, 2005: 125-127.) [21] 徐洪文. 关于置信度选取问题的讨论 [EB/OL]. http://www.docin.com/p-729685972.html. (Xu Hongwen. Discussion on the selection of confidence degree. [EB/OL]. http://www.docin.com/p-729685972.html.) [22] 孔欣欣, 苏本昌, 王宏志, 等. 基于标签权重评分的推荐模型及算法研究 [J]. 计算机学报, 2017, (6): 1440-1452. (Kong Xinxin, Su Benchang, Wang Hongzhi, et al. Recommendation model and algorithm based on label weight score [J]. Acta computer science, 2017, (6): 1440-1452.) [23] 帮助高中生渡过数学学习困难期的几点尝试——论文网 [EB/OL]. http://www.lunwendata.com/thesis/2017/105980.html. (A few attempts to help high school students get through their math difficulties [EB/OL].)

Note: Figure translations are in progress. See original paper for figures.