A Study on Decision Models for Plagiarism Detection Results in Initial Review of Scientific Papers

Authors: Yang Jitao, Guo Baishou, Guo Boshou

Date: 2023-07-03T00:00:00+00:00

Abstract

【Objective】 This study aims to investigate the evaluation model for plagiarism detection results in the initial review of scientific papers, in order to propose a scientifically dialectical acceptance/rejection strategy that makes the initial review process more fair and impartial.

【Method】 First, the survey statistics method is employed to analyze the current status of initial review plagiarism detection. Next, mathematical analysis is utilized to deconstruct the total similarity ratio of papers. Finally, papers are classified according to the distribution of similar fragments, and different evaluation models are proposed for different types.

【Result】Survey statistics reveal that in the current initial review processes of various journals, whether a paper's total similarity ratio exceeds the specified threshold already plays a decisive veto role, which is biased. Based on the importance of different sections of a paper and the distribution of similar fragments across these sections, this study classifies initial review papers into continuous distribution type, single-end distribution type, double-end distribution type, and middle distribution type. The continuous distribution type is further divided into continuous uniform type and continuous non-uniform type; the single-end distribution type is subdivided into head distribution type and tail distribution type. On this basis, targeted evaluation strategies are proposed for different types, and both a practical workflow for dialectical acceptance/rejection in paper plagiarism detection and a computer program flowchart for dialectical plagiarism detection acceptance/rejection are provided.

【Conclusion】 Paper review is one of the important components of journal publishing work, and continuously exploring and optimizing every procedure of the review process is particularly essential. With the continuous development of computer and information technology, AI plagiarism detection systems for

papers based on dialectical analysis thinking will certainly be developed, and long-standing issues troubling journals, such as academic quality control and the definition of academic misconduct, will be solved more reasonably and optimally, making the review process of journal papers more reasonable, objective, and impartial.

Full Text

A Study on the Judgment Model for Plagiarism Detection Results in the Initial Review of Scientific Papers

YANG Jitao[1], **GUO Baishou**[2]*)*

[1] *College of Economics & Management, Northwest A&F University, Editorial Office of* Shaanxi Agricultural Sciences*, No. 3 Taicheng Road, Yangling, Shaanxi 712100, China*

[2] *College of Agronomy, Northwest A&F University, Editorial Office of* Acta Agriculturae Boreali-occidentalis Sinica*, No. 3 Taicheng Road, Yangling, Shaanxi 712100, China

Abstract

[**Objective**] This study aims to investigate judgment methods for plagiarism detection results in the initial review of scientific manuscripts, proposing scientifically dialectical acceptance/rejection strategies to enhance fairness and justice in manuscript preliminary evaluation. [**Methods**] First, survey statistics were employed to analyze the current state of initial-review plagiarism detection. Second, mathematical analysis was used to deconstruct the total duplication ratio of manuscripts. Finally, manuscripts were classified according to the distribution of similar fragments, with different judgment models proposed for different types. [**Results**] Survey statistics reveal that in current journal preliminary reviews, whether a manuscript's total duplication ratio exceeds the prescribed threshold has been granted veto power, which is biased. Based on the importance of each manuscript section and the distribution of similar fragments, preliminary-review manuscripts can be categorized into continuous distribution type, one-head distribution type, two-head distribution type, and intermediate distribution type. The continuous distribution type is further divided into continuous uniform and continuous non-uniform subtypes, while the one-head distribution type is subdivided into head-distribution and tail-distribution variants. Targeted judgment strategies are proposed for each type, along with a practical workflow for dialectical selection in plagiarism detection and a computer program flowchart for dialectical plagiarism checking. [**Conclusion**] Manuscript review constitutes a crucial component of journal publishing, making continuous exploration and optimization of every review process essential. With ongoing development in computer and information technology, AI-based plagiarism detection systems grounded in dialectical analysis will inevitably be developed, offering more rational and optimized solutions to long-standing challenges in

academic quality control and academic misconduct identification, thereby rendering the review process more reasonable, objective, and fair.

Keywords: Scientific papers; Academic misconduct; Plagiarism detection results; Duplication ratio; Fairness and justice

Since the inception of academic publishing and scholarly communication, misconduct such as duplicate submission, multiple publication, plagiarism, and appropriation has persistently troubled publishers, journal societies, and editorial offices. Before the introduction of network and digital information technology to the publishing field, preventing academic misconduct relied primarily on reviewers and editors manually consulting relevant literature and leveraging their disciplinary expertise to conduct comparative analyses—a process that was not only inefficient but also failed to achieve desired outcomes.

In recent years, network and computer information technology has been widely applied in editing and publishing. Relevant institutions have developed internet-based academic misconduct detection systems that provide document checking services for users. Relying on these systems, journal publishers have adopted similarity detection as an essential component of manuscript review, with similarity quantitatively described by the duplication ratio (or replication ratio). Today, duplication ratio detection (or document checking) has become the first step for scholars publishing journal articles, concluding research projects, and university students defending their theses, making successful passage of duplication ratio checks critically important.

Currently, four major similarity detection systems dominate domestically, belonging to CNKI, Wanfang Data, Paperpass, and CQVIP. Each is rooted in its extensive paper database, offering users rapid detection capabilities for academic misconduct including plagiarism, appropriation, fabrication, and falsification, and gradually playing an increasingly important role in the preliminary review of conference papers, dissertations, and journal articles.

Nearly all academic journals have adopted independent online office and editorial management systems, with online submission serving as the primary—indeed, for most journals, the sole—submission channel. Although various editorial management systems exist, most are compatible with academic misconduct detection modules. Consequently, journals currently rely on detection tools embedded in editorial systems to identify plagiarism and appropriation in preliminary manuscripts.

Early academic misconduct literature detection was limited to dissertation evaluation in universities, later gradually being introduced into journal article preliminary review. Academic misconduct detection software either functions as part of literature databases or is embedded in journal editorial systems. For instance, Beijing Qinyun Technology Development Co.'s editorial system integrates Wanfang Data's detection software; Xi'an Sancai Company's system integrates CNKI's detection software; and Beijing Magtech's editorial system, the earliest deployed domestically, has undergone over a decade of improvement and

upgrading, now featuring mature technology, comprehensive functions, and numerous users. This system includes literature search interfaces, metadata service interfaces, reference checking and verification, intelligent reviewer recommendation interfaces, and primarily collaborates with the international CrossRef database to provide literature comparison detection.

Analysis of the above situation reveals that regardless of which approach journal societies (editorial offices) adopt for manuscript preliminary review, their criterion for determining whether a manuscript involves academic misconduct hinges on whether the total duplication ratio provided by the detection system exceeds the journal's established threshold. When a manuscript's total duplication ratio is low, it can generally be assumed that its research content is relatively novel, highly innovative, and unlikely to involve plagiarism, thus probably not constituting academic misconduct. This demonstrates that in current journal preliminary review, whether the total duplication ratio exceeds the threshold has effectively been granted veto power.

Using editorial system detection tools to obtain total duplication ratios provides journal editors with tremendous convenience in identifying academic misconduct during preliminary review. However, as this model has become universally applied in editorial practice, its drawbacks and deficiencies have gradually emerged. Judging academic misconduct solely based on the journal's established total duplication ratio threshold sometimes appears hasty and mechanical, frequently causing misjudgments and consequently missing some innovative manuscripts. Therefore, it is necessary to treat academic misconduct determination cautiously in preliminary review, dialectically analyzing the total duplication ratio through deconstruction to reflect objectivity in the preliminary stage.

2.1 Conventional Preliminary Review Judgment Models

Currently, journal societies or editorial offices generally employ two approaches: (1) First, check new submissions for plagiarism using integrated detection software; if the total duplication ratio (denoted as r) exceeds the journal's prescribed limit (hereafter termed the duplication ratio threshold), the manuscript is rejected; if not, it proceeds to content review. (2) First conduct content review (including examination of publication scope, format, and academic value), rejecting non-compliant manuscripts; for those passing content review, preliminary plagiarism detection is then performed, with rejection if the total duplication ratio exceeds the threshold and passage if it does not.

Regarding duplication ratio threshold settings, the authors conducted preliminary research surveying 103 agricultural comprehensive journals included in the *Chinese Academic Journal Impact Factor Annual Report (Natural Science and Engineering Technology) 2021 Edition*. Results indicate that among surveyed journals, duplication ratio thresholds (r_0) ranged from 5% to 35%. Specifically, 3 journals set $r_0 = 5\%$, 8 journals set $5\% < r_0 \leq 10\%$, 14 journals set $10\% <$

$r_0 \leq 15\%$, 34 journals set $15\% < r_0 \leq 20\%$, 25 journals set $20\% < r_0 \leq 25\%$, 18 journals set $25\% < r_0 \leq 30\%$, and 1 journal set $30\% < r_0 \leq 35\%$. Analysis of threshold settings among selected journals is presented in Figure 1 [Figure 1: see original paper].

2.2 Deconstruction of Total Duplication Ratio

Regardless of genre or format, papers can be divided into several relatively independent sections, which can be classified into two categories: core sections and non-core sections. Core sections involve key, innovative, and distinctive content that differentiates the paper from others, while non-core sections involve literature review of research background and comparative analysis with other studies. Consequently, when core and non-core sections contain identical amounts of duplicated content, their negative impact on the evaluated manuscript may differ. Even when the core section contains less duplicated content than the non-core section, the damage to the paper could be fatal. Therefore, it is necessary to deconstruct the total duplication ratio, clarifying the relationship between each section's duplication ratio and the total, and conduct dialectical examination and analysis based on each section's proportion of the total duplication ratio and its contribution to the paper's importance.

2.3 Determination of Weighting Coefficients

For illustrative purposes, experimental research papers are used as an example for total duplication ratio deconstruction. Such papers typically consist of five sections: introduction, materials and methods, results and analysis, discussion and conclusion, and references (abstract content, primarily extracted from "materials and methods" and "results and analysis," can be omitted as it is replaceable by these sections), denoted as sections 1–5 respectively. Clearly, sections 2 and 4 constitute the core of experimental research papers, while sections 1, 4, and 5 can be considered non-core. Let r_1 , r_2 , r_3 , r_4 , and r_5 represent the duplication ratios obtained from independent detection of sections 1–5. Obviously, the total duplication ratio r cannot be obtained by directly summing r_1 through r_5 . Weighting corrections must be applied to r_1 –r_5 before summation to calculate r. Let N represent the manuscript's total character count, and n_1, n_2, n_3, n_4, and n_5 represent the character counts of sections 1–5 respectively. The weighting coefficients for these five sections are: $_1 = n_1/N$, $_2 = n_2/N$, $_3 = n_3/N$, $_4 = n_4/N$, $_5 = n_5/N$. Let the weighted duplication ratios for sections 1–5 be r_1, r_2, r_3, r_4, and r_5 respectively, yielding: $r_1 = {}_1r_1 = n_1r_1/N$, $r_2 = {}_2r_2 = n_2r_2/N$, $r_3 = {}_3r_3 = n_3r_3/N$, $r_4 = {}_4r_4 = n_4r_4/N$, $r_5 = {}_5r_5 = n_5r_5/N$. Therefore: $r = r_1 + r_2 + r_3 + r_4 + r_5 = n_1r_1/N + n_2r_2/N + n_3r_3/N + n_4r_4/N + n_5r_5/N = (n_1r_1 + n_2r_2 + n_3r_3 + n_4r_4 + n_5r_5)/N$. If a_1, a_2, a_3, a_4, and a_5 represent the proportions of sections 1–5 duplication ratios relative to the total duplication ratio, then $a_1 = r_1/r$, $a_2 = r_2/r$, $a_3 = r_3/r$, $a_4 = r_4/r$, $a_5 = r_5/r$. Subsequently, based on the numerical values of a_1, a_2, a_3, a_4, and a_5 and each section's importance to the paper, manuscripts

can be classified through dialectical analysis without blanket policies. Different duplication ratio thresholds can be set for different manuscript types, enabling more fair and just preliminary review decisions.

2.4 Dialectical Analysis and Manuscript Classification

Although the weighting coefficient method described above is a quantitative deconstruction process, it involves substantial workload, requiring separate detection of all five sections and statistical analysis of total and section character counts. Based on deconstruction and weighted analysis of the total duplication ratio, a more convenient classification method can be derived: classifying manuscripts according to the distribution of similar fragments to determine academic misconduct in preliminary review.

A random experimental research paper was selected from the editorial system, and its similarity distribution obtained from the academic misconduct detection system is shown in Figure 2 [Figure 2: see original paper]. The similarity fragment distribution in Figure 2 reveals that the total duplication ratio is derived by accumulating duplicated character counts from the manuscript's head, mid-front, middle, mid-rear, and tail sections, divided by the total character count. These sections roughly correspond to sections 1–5 of experimental research papers. Examination of Figure 2 can visually reveal the amount of duplicated content in each section and the severity of suspected academic misconduct (yellow indicates "mild plagiarism," orange indicates "moderate plagiarism," red indicates "severe plagiarism," and green indicates "no plagiarism," meaning no overlap with other literature). Manuscripts can then be classified accordingly, and academic misconduct determined in the preliminary review stage based on classification type.

3. Problems in Current Preliminary Review Plagiarism Detection

Careful examination of prevailing practices in journal preliminary review reveals that editors typically base their decisions on the total duplication ratio provided by editorial systems combined with individual journal thresholds, without conducting in-depth dialectical analysis of plagiarism detection results. Specific problems with this approach include: failure to investigate how the total duplication ratio is formed, neglect of relationships between section duplication ratios (introduction, materials and methods, results and analysis, conclusion and discussion, etc.) and the total ratio, disregard for distribution patterns of similar fragments across sections, and lack of awareness regarding which sections' high duplication ratios truly constitute academic misconduct. Evidently, determining academic misconduct solely based on whether the total duplication ratio exceeds a single journal threshold is not only simplistic and hasty but also biased.

For a scientifically novel manuscript, the presence of a certain proportion of

duplicated content does not necessarily mean the paper has completely lost scientific value; the source of duplicated content must be traced. Moreover, text overlap in different sections of a manuscript typically indicates distinctly different natures of academic misconduct. As previously discussed, a scientific paper's total duplication ratio is formed through weighted accumulation of its sections' duplication ratios. Therefore, to achieve fairness and justice, the total duplication ratio must be subdivided, with dialectical analysis of each component section's duplication ratio and its negative impact on the manuscript, combined with comprehensive examination of similarity fragment distribution patterns, to render preliminary review decisions as prudent and objective as possible.

4.1 Classification and Selection Principles for Manuscripts Under Review

Based on similarity fragment distribution patterns in similarity distribution diagrams, manuscripts under review can be classified into four types: continuous distribution type, one-head distribution type, two-head distribution type, and intermediate distribution type. The continuous distribution type is further divided into continuous uniform and continuous non-uniform subtypes, while the one-head distribution type is subdivided into head-distribution and tail-distribution variants. Once the manuscript type is identified, targeted solutions can be developed through case-specific analysis, establishing different selection principles for each type.

4.1.1 Continuous Distribution Type and Selection Principles

Both Figure 3 [Figure 3: see original paper] and Figure 4 [Figure 4: see original paper] represent continuous distribution types. In this type, similar fragments are distributed across nearly all parts of the manuscript, indicating the entire text contains content similar to other literature, lacking significant academic value and innovation. Rejection at the preliminary review stage is recommended.

4.1.2 One-Head Distribution Type and Selection Principles

As shown in Figure 5 [Figure 5: see original paper], the "head-distribution" variant of the one-head type concentrates duplicated content primarily in the introduction section. Introductions generally synthesize and review published literature, clarifying advantages and limitations of previous research, identifying unresolved issues, and presenting the author's research objectives and expectations. When cited literature is relatively old, it may have been repeatedly used by other papers in the same field, potentially being flagged as duplicated content during detection. Additionally, for highly-cited literature within an industry, authors often quote directly, naturally resulting in higher duplication ratios.

Figure 6 [Figure 6: see original paper] illustrates the "tail-distribution" variant of the one-head type, where similar fragments are mainly distributed in the discussion and conclusion or references sections. The discussion section functions

to compare and analyze the author's results with previous research, explaining similarities and differences and exploring reasons for divergent results. Therefore, like the introduction, the discussion inevitably requires extensive citation of reported literature as evidence, making considerable content repetition understandable. The conclusion section distills universal scientific laws based on results and analysis. In tail-distribution manuscripts, the results and analysis sections contain virtually no duplicated content, indicating high innovation; conclusions should not exhibit high duplication. Conversely, if they do, it suggests the author failed to identify truly meaningful patterns from experimental results and requires revision. High duplication in the references section likely stems from outdated citations, which is not fatal to the manuscript, as replacement with recent relevant literature resolves the issue.

Thus, even when the total duplication ratio of one-head distribution manuscripts exceeds the threshold, it does not negate the objectivity, rationality, and innovation of the manuscript's material, methodology, results, and conclusions. During preliminary review after content evaluation, the duplication ratio threshold for this type can be set higher, for example, $r_0 = 25\%$. When total duplication ratio $r > 25\%$, preliminary rejection may be appropriate; when $r \leq 25\%$, preliminary revision should be granted, with authors requested to consult recent literature and rewrite the introduction or discussion and conclusion sections to reduce the total duplication ratio.

4.1.3 Two-Head Distribution Type and Selection Principles

The two-head distribution type (Figure 7 [Figure 7: see original paper]) is relatively common among manuscripts under preliminary review. Similar fragments in this type concentrate primarily in the introduction, discussion and conclusion, and references sections, while the manuscript's core—materials and methods, results and analysis—contains no duplicated content from other literature. If content review is passed, revision value remains. Given that the introduction, discussion and conclusion, and references may account for approximately half the manuscript's total length, the duplication ratio threshold for two-head distribution types can be set higher than for one-head types, for example, $r_0 = 35\%$. When total duplication ratio $r > 35\%$, preliminary rejection is recommended; when $r \leq 35\%$, preliminary revision should be implemented, with authors requested to carefully revise and reduce the total duplication ratio.

4.1.4 Intermediate Distribution Type and Selection Principles

The intermediate distribution type of similar fragments is shown in Figure 8 [Figure 8: see original paper], which can be further subdivided into mid-front, middle, and mid-rear variants. The location of similar fragments in intermediate distribution manuscripts corresponds to the materials and methods and results and analysis sections of experimental research papers—the core and soul of the manuscript. Consequently, the duplication ratio threshold for this type should be set lower, for example, $r_0 = 5\%$. If total duplication ratio $r > 5\%$, revision

significance is minimal, and preliminary rejection may be considered; if $r \leq 5\%$, preliminary revision remains viable, though rechecking after revision is recommended.

4.2 Practical Workflow for Dialectical Selection in Preliminary Review

Based on the above dialectical analysis, a standardized preliminary review workflow has been compiled (Figure 9 [Figure 9: see original paper]), with duplication ratio thresholds for one-head, two-head, and intermediate distribution types adopting the recommended values from Sections 4.1.2, 4.1.3, and 4.1.4 respectively. This workflow is easy to master and highly practical. Naturally, journal societies or editorial offices can also determine different duplication ratio thresholds for different manuscript types based on their own characteristics, replacing the current single threshold with multiple thresholds to enable case-specific analysis, prevent valuable manuscripts from being overlooked, and effectively safeguard authors' publication rights.

A computer program flowchart based on dialectical plagiarism detection (Figure 10 [Figure 10: see original paper]) has been developed from Figure 9. Programming and embedding this into academic misconduct detection systems according to Figure 10 can render plagiarism detection more intelligent.

Preliminary review plagiarism detection is no trivial matter, directly determining whether a manuscript can proceed to peer review, re-review, and final decision stages. Therefore, carelessness in preliminary review, though seemingly inconsequential, profoundly impacts authors' research work and careers. Journal editors must treat every manuscript's preliminary review with utmost caution.

The dialectical selection workflow proposed in this study, based on deconstruction analysis of total duplication ratios, provides a reference for journal manuscript preliminary review, enabling more fair and just preliminary evaluation and striving to protect authors' rights while continuously improving journal service levels.

As agricultural journal editors, the duplication ratio threshold survey was based on 103 comprehensive agricultural science and technology journals. Due to limitations, not all domestic scientific journals were included, making the recommended thresholds for different manuscript types somewhat limited in scope, serving merely as an initial contribution. Different scientific journals can determine scientifically reasonable type-specific thresholds based on their own disciplinary classification, professional content, and industry characteristics.

Manuscript review constitutes a crucial component of journal publishing, making continuous exploration and optimization of every review process essential. With ongoing development in computer and information technology, AI-based plagiarism detection systems grounded in dialectical analysis will inevitably be developed, offering more rational and optimized solutions to long-standing

challenges in academic quality control and academic misconduct identification, thereby rendering the review process more reasonable, objective, and fair.

References

[1] Cong Lixian. The academic, copyright, and social significance of similarity detection[J]. China Publishing Journal, 2019(15): 40-44.
[2] Yan Jing. Research on legal regulation of paper duplication ratio detection[J]. Journal of Hengyang Normal University (Social Science Edition), 2021, 42(1): 85-92.
[3] Sun Juan, He Li, Song Yonggang, et al. The role and implementation path of academic journals in research integrity construction[J]. Chinese Journal of Scientific and Technical Periodicals, 2021, 32(2).
[4] Tian Xin, Ma Hanqing, Zheng Junwei, et al. Comparative study of five major online peer review system platforms at home and abroad[J]. Chinese Journal of Scientific and Technical Periodicals, 2014, 25(11): 1363-1368.
[5] Tang Hong, Zhu Yinzhou. Analysis of acceptance/rejection of manuscripts exceeding text duplication limits in academic misconduct detection[J]. Chinese Journal of Scientific and Technical Periodicals, 2020, 31(3): 281-287.
[6] Cao Mengyuan, Wang Yihan, Li Nan. Analysis and measures for preventing academic misconduct in scientific journals[J]. Standard Practice, 2022(1)(II): 92-94.
[7] Li Dan. Rational use of academic misconduct detection systems[J]. Academics, 2012(12): 129-133, 286.

Author Contributions

1) **YANG Jitao:** Identified the problem, reviewed and analyzed relevant literature, drafted and completed the initial manuscript, and revised the paper.

2) **GUO Baishou:** Proposed the research concept and provided final review and revision of the paper.

Author Biographies

YANG Jitao (ORCID: 0000-0003-2896-8614), M.S., Associate Editor, Deputy Editor-in-Chief of *Shaanxi Agricultural Sciences*, E-mail: snkx001@163.com
GUO Baishou (ORCID: 0000-0003-0119-2510), Associate Editor, Deputy Editor-in-Chief of *Acta Agriculturae Boreali-occidentalis Sinica*, E-mail: xbnx04@nwsuaf.edu.cn

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