Postprint: A Study on the Innovative Characteristics of Chinese and International Information Science Papers

Authors: Cao Shujin, Yan Xinyang, Zhang Qian, Zhuo Yiling

Date: 2023-04-01T00:00:00+00:00

Abstract

[目的/意义] This study employs a combined qualitative and quantitative methodology to analyze and compare the innovativeness of recent Chinese and international information science papers, revealing the innovative characteristics of research in the information science domain, uncovering knowledge relationships within innovation sentences in academic papers, enabling finer-grained analysis of paper innovativeness, providing a foundation for the deep utilization of domain innovation points, while enriching approaches for monitoring innovation in scientific papers and promoting scientific research innovation.

[方法/过程] Beginning with sentence-level innovation identification, we selected two Chinese and two international information science journals as samples, utilized information extraction and machine learning methods to extend innovation sentence extraction from abstracts to full texts, leveraged sentence structure and syntactic features to identify domain innovation content, examined characteristics of recent Chinese and international information science papers regarding innovation objects, themes, and categories, conducted comparative analysis, and finally performed qualitative content analysis on automatically classified paper collections to summarize the expression paradigms of innovation in Chinese and international information science papers.

[结果/结论] Regarding innovation expression, the distribution of innovation sentences in Chinese and international information science papers is largely consistent, with English journal papers demonstrating richer innovation expression. In terms of innovative characteristics, English information science journal papers exhibit more concentrated innovation themes, whereas Chinese themes are diverse and dispersed; innovation in specific methods represents a hotspot in the information science domain in recent years, while methodological innovation remains insufficient; the innovative features of both Chinese and English

information science journal papers reflect a trend of abundant achievements in applied and empirical research, but slow advancement in theoretical innovation.

Full Text

Research on Characteristics of Innovation in Chinese and International Information Science Papers

Cao Shujin, Yan Xinyang, Zhang Qian, Zhuo Yiling

School of Information Management, Sun Yat-sen University, Guangzhou 510006

Abstract: [Purpose/significance] This study employs a comprehensive approach combining qualitative and quantitative methods to analyze and compare the innovative features of Chinese and international information science papers in recent years, revealing the innovative characteristics of research in this field. It examines the knowledge relationships within innovative sentences in academic papers to enable more fine-grained analysis of paper innovation, provides conditions for the deep utilization of innovation points in research fields, enriches approaches for monitoring innovation in scientific papers, and promotes scientific research innovation. [Method/process] Starting from sentence-level innovation identification, we selected two Chinese and two English information science journals as samples. Using information extraction and machine learning methods, we extended the extraction of innovative sentences from abstracts to full texts, fully utilizing sentence structure and syntactic features to identify innovative content in the field. We explored the characteristics of Chinese and international information science papers in terms of innovation objects, themes, and categories, conducted comparative analyses, and finally summarized the innovation expression paradigms of Chinese and international information science papers through qualitative content analysis of automatically classified paper collections. [Result/conclusion] In terms of innovation expression, the distribution of innovative sentences in Chinese and international information science papers is basically consistent, though English journal papers express innovation more richly. Regarding innovative characteristics, English information science journal papers have more concentrated innovation themes, while Chinese themes are more diverse and dispersed. Innovation in specific methods represents a hotspot in recent information science research, while innovation in research methodology is insufficient. The innovative features of both Chinese and English information science journal papers reflect a trend of abundant applied and empirical research achievements alongside slow theoretical innovation.

Keywords: characteristics of innovation; academic papers; dependency parsing; sentence classification
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.01.011

Innovation is the primary driver of development, the core of scientific research,

and the essential requirement of academic papers. As R. Tan once noted regarding scientific research innovation: "The value of a person' s research does not lie in how much effort they put in; rather, the value of research lies in the innovativeness of the results" [1]. Innovativeness is a crucial criterion for evaluating the academic quality of scientific papers, a core trait that determines their academic level, and the primary basis for publication decisions [2]. Extracting and analyzing innovation points from domain academic papers can effectively reveal the progress of innovation in the field and the types and impacts of innovation points [10].

This study aims to conduct a comparative investigation of the innovative characteristics of information science papers, starting from sentence-level innovation identification and fully utilizing sentence structure and syntactic features. Using information extraction and machine learning methods, we analyze the aspects, objects, and themes of innovation to measure the innovative characteristics and progress of information science research in recent years from a broad yet precise perspective. The novelty of this study lies in extending the extraction of innovative sentences from existing abstracts to full texts, rather than remaining at the abstract level alone, and fully incorporating sentence structure information in the application of information extraction and machine learning methods. The purpose of this research is not only to reveal the innovation situation in a particular field but, more importantly, from a macro disciplinary perspective, to discover the knowledge relationships within innovative sentences in domain academic papers. This provides pathways for the deep utilization of innovation points in research fields, lays the foundation for fine-grained knowledge organization and retrieval, facilitates knowledge reasoning and discovery in big data environments, enriches approaches for monitoring innovation in scientific papers, and ultimately promotes scientific research innovation.

2 Literature Review

2.1 The Meaning and Manifestation of Academic Paper Innovation

It is difficult to provide a precise definition of "innovativeness."Merriam-Webster defines "novelty"as "something new or different from anything familiar"[11]. Innovative information refers to sentences containing new content, typically defined as the opposite of redundancy [12]. In J. Allan et al.' s research, innovativeness is described as new information based on the presence of new words in sentences [13]. B. Uzzi et al. argue that scientific innovation arises through original combinations that stimulate new insights [14]. From a knowledge combination perspective, innovativeness can be defined as reorganizing existing knowledge in unprecedented ways, a view accepted by scholars across disciplines [15]. These definitions share essentially the same core connotation: combining existing or new knowledge in a "new" way.

Regarding the innovativeness of academic papers, Zhou Luyang [16] identifies three basic meanings: (1) papers that are "different" from existing academic

literature, which can be either "partial improvements" or completely "new" ; (2) this difference must involve knowledge or information in the academic field; and (3) this different knowledge or information must be "valuable." This study' s identification of academic paper innovativeness involves mining and analyzing the innovative aspects expressed by authors in their papers.

2.2 Evaluation of Academic Paper Innovation

Traditional evaluation methods for academic innovation in humanities and social sciences journals mainly include two approaches: qualitative evaluation through peer review and quantitative evaluation based on bibliometrics. Bibliometric methods for evaluating paper innovation include single-feature indicator evaluation, measuring innovation through impact, indicator system evaluation, and content-based evaluation [17]. Although widely used, peer review suffers from drawbacks such as conflicts of interest, strong subjectivity, large operational space, time consumption, and low efficiency [18].

Some scholars use single-feature indicators such as authors' h-index, citation counts, reference influence, or journal impact factors to evaluate academic innovation [19]. Y. Lee et al. [20] calculated the similarity of each journal pair in references, ranked all journal pair similarities from smallest to largest, took the top 10th percentile, and used the negative value—larger values indicating greater novelty. L. Wu et al. [21] proposed a new innovation measurement indicator, Disruption, in *Nature*, quantifying paper novelty by dividing citation structures. Other scholars have integrated social network analysis and statistical methods, using correlation analysis, regression analysis, and structural equation modeling to verify relationships between various indicators and literature innovation [22]. Ye Jiyuan [23] analyzed the drawbacks and causes of quality and innovation evaluation systems for Chinese academic journal papers from social, economic, cultural, and political perspectives, proposing a new concept combining formal evaluation, content evaluation, and utility evaluation, which provides strong theoretical guidance for evaluating journal paper quality and measuring innovation in Chinese social sciences.

In summary, many scholars have used statistical methods to empirically analyze the relationships between author prestige, journal influence, citation counts, download counts, and paper innovation, reaching different conclusions. However, they generally agree that measuring paper innovation using only these indicators is unscientific. High author or journal influence only indicates better academic value, content quality, or impact, not necessarily high innovation, as citations and downloads are affected by temporal, author, and institutional factors. Equating paper influence with innovation is clearly one-sided; paper innovation may correlate with influence but is not equivalent to it. Furthermore, novelty does not equal innovation. Novelty is a necessary but not sufficient condition for innovation. Some papers may have new research objects and originality but low practical or applied value, or their argumentation process may be logically flawed, rendering their conclusions invalid. In such cases, even with

strong originality, their innovative value is insufficient. As Li Rusen et al. [24] propose, innovation points in scientific papers should be categorized into topic innovation, technical background innovation, methodology innovation, conclusion innovation, and overall innovation, with innovative achievements possessing three characteristics: "originality, novelty, and practicality." This reflects that paper innovation requires not only theoretical and methodological innovation but also that innovative results must be effective and useful.

Bibliometric methods for evaluating paper innovation are largely limited to quantitatively assessing papers from perspectives of authors, institutions, journals, and references, without truly focusing on paper content. Although qualitative peer review partially compensates for this deficiency, its inherent drawbacks affect innovation evaluation. If natural language processing and machine learning techniques can assist paper innovation evaluation—through semantic analysis to construct grammatical rules, extract innovation indicators, identify paper innovation points, build innovation knowledge bases, and use automatic classification or clustering to identify innovation themes and patterns—this would provide solid technical support and knowledge foundations for paper innovation evaluation. Natural language processing technology already has mature applications in calculating literature similarity, automatic topic identification, keyword extraction, and topic classification and clustering. Yang Jianlin and Qian Lingfei [25] constructed a set of formulas for measuring document topic novelty based on word frequency principles, inverse document frequency principles, and co-word analysis, empirically demonstrating their rationality and practicality. Liang Shuai and Gao Jiping [26] used F5000 paper review comments as text analysis objects, conducting text mining, keyword extraction, and content analysis on excellent paper reviews. Through feature word frequency and co-occurrence analysis, they identified four main characteristics of excellent papers: "innovation features," "value features," "research content," and "writing style."Therefore, judging paper innovation value should not be limited to novelty or originality; the usefulness and effectiveness of innovative achievements and the scientific nature of methodological innovation should all be evaluation criteria. Qian Lingfei et al. [27] used ontology theory and technology to construct academic innovation concept ontologies and academic innovation knowledge resource ontologies, instantiating these ontologies and introducing high-frequency keywords from CNKI journal bibliographic data to enrich ontology knowledge, define class relationships, and build an academic innovation ontology—providing a solid knowledge foundation for subsequent automatic measurement of academic innovation. He Wanying [17] constructed machine learning models for innovation evaluation, using paper data from core library and information science journals to empirically analyze multiple models, evaluating different machine learning models' performance to identify suitable models for innovation evaluation. These examples demonstrate that text mining technology and machine learning methods have considerable application potential and value in paper innovation evaluation.

2.3 Identification and Extraction of Academic Paper Innovation Points

An academic paper may contain multiple innovations or only a few, but regardless of quantity, any innovation—even a single point or sentence—should be identified and recognized [28]. However, the expression of innovation in academic papers is diverse, with innovation points appearing in various forms throughout different sections, making identification essential. As described above, most existing research focuses on innovation evaluation, with identification methods primarily approaching from an innovation degree evaluation perspective.

Cao Xiaochun [29], from an editorial perspective, argues that academic paper innovation broadly includes new topics, new arguments, new evidence, and new reasoning, with approaches such as filling gaps, supplementary development, direct confrontation, revolutionary change, and progressive introduction—all of which can be identified as academic paper innovation points if they possess certain characteristics. Beyond these perspectives, innovation can also manifest in research methods, ideas, and designs [28].

From a content perspective, the presence of innovative knowledge units indicates innovativeness, and these units constitute the paper' s innovation points [30]. T. Heinze et al. [31] suggest that innovation points can be identified and extracted as new theories, new phenomena, new methods, new instruments, or new integrations of existing theories from new angles. Zhou Luyang [16] progressively refines content innovation from new arguments and evidence to new theories, methods, objects, disciplines, data, and facts, proposing a set of methods for identifying academic literature innovation points. Some scholars believe innovation points can be located and identified through reference positions [32].

Scholars' proposed methods for identifying academic paper innovation points are mostly at the macro theoretical level, with gaps remaining in practical implementation. For extracting paper innovation points, one method involves comparing paper titles with existing papers in databases through similarity ranking to extract innovation points [33]. Another method extracts keywords from papers and calculates keyword frequency changes over time in retrieval systems to identify innovative keywords [25, 34]. A more comprehensive approach uses the KeyGraph algorithm [35] to extract research themes from papers, then calculates similarity between extracted themes and current disciplinary frontiers to identify innovative research themes [36]. Some researchers consider comparing paper contexts, mining new versus old information from text to identify innovations such as technologies or inventions [37].

Most of these methods are essentially consistent: comparing with historical subsets to obtain innovation points. However, this approach has drawbacks—creating comprehensive historical subsets is difficult and costly, potentially resulting in insufficient training data for identified innovations and degraded identification effectiveness [40]. Therefore, this study takes a different approach, selecting two Chinese and two English information science journals as samples, using in-

formation extraction and machine learning methods that incorporate sentence features themselves to identify domain innovative content, explore academic paper innovation characteristics, summarize recent information science paper innovation situations, and explore new methods for innovation point identification and extraction to assist domain scientific paper innovation monitoring and promote scientific research innovation.

3 Research Design

3.1 Data Source and Text Preprocessing

Considering paper writing standards and accessibility, Chinese journal papers from 2013-2018 in *Information Science* and *Data Analysis and Knowledge Discovery* (formerly *Modern Library and Information Technology*) were selected as data sources. English journal papers from 2013-2018 in *Information Processing & Management* and *Journal of Informetrics* were selected as data sources. Meeting minutes, moderator introductions, speeches, call for papers, topic selections, and other non-research papers were removed, resulting in final datasets of 2,487 and 1,050 papers respectively.

In the preprocessing stage, all papers were converted to plain text format. Research by Yu Husheng et al. [41] and T. Dahl [42] summarized the distribution characteristics of paper innovation points, concluding that abstracts, introductions, and conclusions are sections where innovation points are concentrated. Based on this and statistical results, we extracted the abstract, introduction, methodology, results, and conclusion sections where innovative sentences might appear, then segmented these sections into sentences.

3.2 Selection of Innovation Feature Guide Words

The linguistic features of paper innovation points are mainly reflected in guide words (feature words) and expression patterns [42-43]. For scientific literature' s linguistic and genre features, rule-based extraction methods can accurately identify "knowledge claims" in papers [44]. For the segmented sentences, we used the Stanford CoreNLP tool for tokenization, word frequency statistics, and part-of-speech tagging. Combining manual annotation results from randomly selected datasets, we selected and determined words closely related to innovation. The main basis for selecting innovation feature guide words came from the "Ten-point Scoring Standard for CSSCI Paper Evaluation," specifically the four elements of "degree of innovation, degree of completeness, degree of difficulty, and value of achievement," which we refined into "usefulness, novelty, effectiveness, and scientific nature." Novelty applies to problems, methods, and results —the three core elements—and essentially means the paper' s "difference" from existing literature, which can be either "partial improvement" or completely "new" [26]. Effectiveness and scientific nature primarily concern research methods, while usefulness and effectiveness concern research results.

After initial determination of innovation feature guide words, we introduced them into HowNet for synonym expansion as our final set of innovation feature guide words. Guide words include but are not limited to 标志性 nouns, adjectives, a few verbs, and phrases. Based on the selected innovation feature guide words and drawing on Zhang Fan and Le Xiaoqiu' s [44] approach, we used Stanford Parser syntactic tree parsing to build rules incorporating innovation guide words, extracting from sentence collections to form innovative sentence sets. This yielded 19,088 English sentences and 12,451 Chinese sentences. Innovation feature guide word examples and corresponding sentences are shown in Table 1 .

3.3 Innovative Sentence Syntactic Analysis

Using Stanford Parser for dependency syntactic analysis of innovative sentences, we built rules to extract innovation objects and themes. Dependency syntax, first proposed by French linguist Lucien Tesnière, posits that sentence components have governing and dependent relationships [51]. Governing words are called heads or governors, while dependent words are called modifiers or dependents [51]. Dependency syntactic analysis represents relationships between words in sentence pairs, and by locating predicate nodes with semantic annotation types as innovation feature verbs, we can further identify the theme words they govern—core theme words revealing innovations [8].

Based on this characteristic of dependency relationships, we established extraction rules for innovation objects and themes: (1) Identify core words revealing innovation points (ROOT words) from dependency relation pairs; (2) Filter direct object relationships (marked with "dobj" identifier, indicating a governing relationship) from dependency pairs containing core words, extracting the governed component as words revealing innovation objects; (3) Extend to find noun compound modification relation pairs closely associated (nearest distance) with ROOT words and innovation objects, using these as the sentence' s innovation theme. For Chinese sentences, if "topic" relation pairs exist, they are directly used as innovation themes. For sentences lacking the above-specified relation pairs, we formulated supplementary rules based on syntactic trees: (4) For English sentences, extract relation pairs labeled "JJ" (adjective) with subordinate modification labels "NN" (common noun), filter them according to the innovation guide word set, and use the modified component as the innovation object; for Chinese sentences, with richer labels, extend the rule to extract "ADJP" (adjective phrase) with subordinate modification labels "NP" (noun phrase), similarly filter, and use the modified component as the innovation object; (5) Extract the innermost "NP" label of "IP" (simple clause) as the innovation theme.

3.4 Innovative Sentence Classification

Li Ying and Zhou Li [52] refined and specified innovation points in papers, categorizing them into ten aspects: {new discovery, new method, new technology,

new viewpoint, new theory, new idea, new process, new application, new contribution, new concept}, summarizing their expression content and common feature words. Based on dependency syntactic analysis extracting innovative sentence features, we added dependency syntactic labels and vectorized them to improve identification accuracy. Building on Li Ying and Zhou Li' s [52] classification and according to the innovative aspects described, we categorized sentences into 4 major categories and 8 subcategories, with classification and content explanations shown in Table 2 .

Following this classification, we used the SVM algorithm to train and test on manually annotated training sets, then classified the remaining sentence corpora. Evaluation metrics are shown in Table 3 and Table 4 .

4 Results Analysis and Discussion

4.1 Innovative Sentence Statistics

Statistical analysis of the innovative sentence sets revealed an average of approximately 5 innovative sentences per Chinese journal paper and 18 per English journal paper. The distribution proportions across sections are shown in Figure 1 [Figure 1: see original paper] and Figure 2 [Figure 2: see original paper].

Comparing Figures 1 and 2, we find that the distribution of innovative sentences in Chinese and international information science papers is relatively consistent, with English papers showing more even distribution. Over 30% of innovative sentences come from the introduction section, partly because introductions are standard and longer than abstracts. While abstracts express innovation points concisely, introductions expand and enrich innovation expression, including not only appropriate introductions of innovative content but also the origins, purposes, and means of innovation, primarily through narrative expression [52]. Since authors "can customize the writing format of the main paper sections" [53], the "results" section is not mandatory, leading to diverse presentation forms and a relatively low proportion of innovative sentences. However, the proportion of innovative sentences in the results section of selected English information science journal papers remains higher than in abstracts, indicating more standardized writing in English information science journals. The conclusion section contains fewer innovation-revealing expressions, mainly stating innovation value, function, and significance, with innovation itself typically expressed implicitly and indirectly. Overall, the distribution gap of innovative sentences between Chinese and English information science journal papers is small, with English journal papers having higher average numbers per paper and richer innovation expression.

4.2 Analysis of Innovation Object Characteristics

The word frequency and ranking of innovation objects are shown in Figure 3 [Figure 3: see original paper] and Figure 4 [Figure 4: see original paper]. The

frequency ranking of innovation objects in both Chinese and English information science papers basically conforms to Zipf' s law: high-ranking innovation objects occupy the vast majority, while low-ranking ones are very rare. In contrast, the curve for English information science journal papers is steeper, reflecting more concentrated innovation objects. Due to the larger number of innovation objects in English papers, the long-tail effect is also more pronounced.

Further investigation of innovation objects reveals that "method" innovation accounts for a very large proportion in both Chinese and English information science journal papers, indicating that methodological innovation is a key research direction in recent information science. In relative terms, "method" innovation accounts for a higher proportion in English papers, with the combined frequency of "method" and "approach" exceeding half. Framework, algorithm, problem, data, and measurement are key objects of innovation in English information science journal papers. The overlapping parts are still substantial (e.g., data, problem, model, algorithm), representing the main innovation directions in recent information science. By comparison, English information science journal papers focus more on detailed innovations, with less theoretical innovation but greater attention to "framework" innovation (see Figures 5 [Figure 5: see original paper], 6 [Figure 6: see original paper], and Table 5). This aligns with conclusions by Liu Qijin et al. [9], who analyzed over 210,000 English documents in "computer" disciplines from the ACM Full-Text Database (1951-2012), finding that most innovations concentrated on method innovation (e.g., approach, method, way) and specific application innovations (e.g., algorithm, model, application), with relatively few theoretical innovations (e.g., idea) [9].

Method innovation in Chinese papers, while having the highest proportion, includes only about 1% research methodology innovation, with the proportion of high-level methodological innovation still low. Wei Ruibin [54] analyzed domestic co-word analysis research and similarly found few papers with overall methodological innovation. He argued that research method innovation requires breakthroughs at the principle level or improvements to certain processes, requiring researchers to have deep understanding of research methods and the ability to propose their own solutions, making such innovation difficult. Comparing the two Chinese information science journals (Table 6), we find significant differences in innovation objects, mainly related to journal column settings and positioning. *Information Science* includes theoretical research and business research columns, resulting in relatively high proportions of theoretical, technical, and methodological innovation. *Data Analysis and Knowledge Discovery* focuses on big data-based research and applications across industries that rely on complex mining analysis for knowledge discovery and prediction, supporting decision analysis and policy formulation. It is committed to providing theoretical guidance, technical support, and best practices, integrating computer science, scientometrics, social informetrics, webometrics, data science, management science, predictive analytics, and evidence-based policy analysis to help people discover knowledge from data, distill wisdom (insights) from knowledge, and design solutions from knowledge and wisdom [55]. This orientation makes

algorithm and experiment innovations prominent in its papers.

The two English information science journals show relatively consistent high-frequency innovation objects (Table 7), with differences again mainly related to journal positioning. Beyond methods and models, *Information Processing & Management* papers have broader, more comprehensive innovation objects, including algorithms, frameworks, research problems, experiments, features, and techniques, positioning the journal at the intersection of computer and information science. *Journal of Informetrics* focuses on quantitative research in information science, making indicators, data, and utilization particularly prominent innovation objects.

4.3 Analysis of Innovation Theme Characteristics

Co-occurrence analysis of innovation themes yields the networks shown in Figure 7 [Figure 7: see original paper] and Figure 8 [Figure 8: see original paper]. Edge thickness represents weight, and node color indicates modular classification. With similar numbers of nodes, the two networks differ greatly in edge count: the Chinese theme co-occurrence network has a density of 0.029, while the English network has a density of 0.649. English information science papers show frequent co-occurrence across themes, more intersections, and dense connections, with "method," "model," and "data" forming a very dense network core. In contrast, Chinese information science research themes are relatively dispersed and independent.

Recent English information science innovation themes can be roughly categorized into: (1) Bibliometrics (citation, indicator, journal, research evaluation, collaboration···); (2) Text mining (classification, text, topic, detection···); (3) Information systems (system, search, framework, recommendation···); and (4) Machine learning combined with natural language processing (algorithm, cluster, language, word, training, sentiment···). The bibliometrics category aligns with Liu Zhifeng et al.' s [56] findings. Analyzing *Journal of Informetrics* papers (2007-2017), they found main research themes including measurement indicators, scientific evaluation and ranking, research collaboration, and citation analysis [56], reflecting these as core, stable innovation themes in the field.

Recent Chinese information science innovation themes can be roughly divided into: (1) Library-related (digital libraries, university libraries, library services, resources···); (2) Algorithm categories (genetic algorithms, optimization algorithms, neural networks···); (3) Social media (social networking sites, topics···); (4) Community research (scientific/academic/health communities, members, influence···); (5) Enterprise knowledge management (knowledge transfer, individual capability, innovation···); (6) Ontology; and (7) Bibliometrics. Due to overly tight connections, English themes are harder to subdivide like Chinese themes, but it is clear that English information science journal papers focus on natural language processing and bibliometrics with concentrated themes, while Chinese information science journal papers have more dispersed and diverse

themes. Although this relates partly to journal selection, it also reflects the comprehensiveness of Chinese information science journals and the specialization of English journals, as well as different research topic characteristics.

4.4 Distribution of Innovation Categories

The category distribution of paper innovation sentences is shown in Figures 9 [Figure 9: see original paper] and 10 [Figure 10: see original paper], with core expression paradigms for each category summarized in Tables 8 and 9 . Chinese papers show uneven distribution across innovation categories: Categories 6 (viewpoint/concept innovation) and 7 (research method innovation) are relatively scarce, while Categories 4 (proposing new methods/techniques/ideas) and 8 (research problem/object innovation) account for higher proportions. In contrast, English papers show even more uneven distribution, with nearly 80% of innovative sentences concentrated in Category 1 (discovering new patterns/connections), Category 4, and Category 8, while Categories 2, 3, 5, and 6 are less common.

Both Chinese and English papers concentrate innovation categories in Category 4 and Category 8—proposing new methods/techniques/ideas and research problem/object innovation—echoing the "Analysis of Innovation Object Characteristics" section. The main differences between Chinese and English papers lie in the proportions of Category 2 (building/improving new models), Category 5 (proposing new countermeasures/suggestions/applications), and Category 7 (applying new methods, introducing new data). Chinese papers have relatively higher proportions of Category 2 and Category 5, reflecting higher proportions of review and qualitative evaluation papers. Chinese review papers propose countermeasures and applications for product, industry, discipline, or domain development by 梳理 domestic and international theoretical achievements and practical progress. Evaluation papers often comprehensively apply expert interviews, questionnaires, and AHP to construct evaluation indicator systems, with some combining mathematical statistics or natural language processing to empirically test model/method reliability and validity. This also reflects insufficient innovation in research methods in Chinese information science journal papers, with applying other disciplines' methods and introducing new data not being mainstream. English papers have a relatively higher proportion of Category 7, indicating that English information science journals publish more studies not only proposing new specific methods but also applying new research methods and introducing new data, giving them an advantage in quantitative empirical research and focusing on applied and practical research outcomes. This also reflects differences in thinking between Chinese and Western scholars in paper writing: Ji Rongqin [56] argues that "the developed mathematical logic in Western culture and Westerners' logical thinking patterns lead English articles to 倾向于引用 more facts, especially survey data and factual data, when stating arguments, while Chinese articles use data and experimental results less frequently to demonstrate viewpoints."

Categories 3 (obtaining new theories) and 6 (viewpoint/concept innovation) account for low proportions in both Chinese and English papers, further reflecting the slow progress of theoretical innovation in recent information science. Wei Ruibin and Liu Yu [57], through text analysis of titles from 434 information science doctoral dissertations (1996-2012), also concluded that domestic information science papers contain more applied and empirical research achievements but fewer pure theoretical research results, as theoretical innovation is the most valuable yet most difficult form of paper innovation.

4.5 Innovation Expression Paradigms

Through qualitative analysis of content from machine learning-trained classification sets, this study summarizes the innovation expression paradigms for eight categories of Chinese and English papers, as shown in Tables 8 and 9 . Regardless of category, all innovations are built upon literature review and evaluation, understanding theoretical foundations, tracing research problem development, identifying what previous research has and has not solved, and recognizing deficiencies to provide new directions and ideas for one' s own research innovation. Second, innovative methods must be genuine and effective: arguments must withstand verification, data must be authentic and valid, and proposed methods must pass experimental testing. Finally, innovative results must be valuable, as indicated by feature words such as improve, optimize, enhance, effective, theoretical value, and practical value.

5 Research Conclusions

Identifying innovative features and new research progress from papers is significant for scientific research within the field. This study starts from sentence-level innovation identification, using sentence structure and syntactic features with information extraction and machine learning to analyze innovation categories, objects, and themes, empirically revealing recent innovation characteristics and progress in information science. The study concludes: First, regarding innovation expression, the distribution of innovative sentences in Chinese and international information science papers is basically consistent, with the introduction being the section where innovation points are concentrated. However, English journal papers are more standardized in writing with richer innovation expression; both Chinese and English innovation expressions are relatively normative and follow certain patterns. Second, regarding innovation characteristics, analysis of innovation object features and category distribution reveals that innovation object frequency distribution conforms to Zipf' s law, with high-frequency innovation objects occupying the vast majority. Innovation in specific methods is the main research direction in recent information science, with data, problems, models, and algorithms also being focal points, while innovation in research methodology is insufficient. Comparatively, English information science journal papers focus more on detailed innovations, with less theoretical innovation but more attention to "framework" innovation. The innovative characteristics of

both Chinese and English information science journal papers reflect abundant applied and empirical research results alongside slow theoretical innovation, as theoretical innovation is the most valuable yet most difficult form of paper innovation. Analysis of paper innovation theme characteristics reveals that English information science journal papers focus on natural language processing and bibliometrics with concentrated themes, while Chinese information science journal papers have diverse and dispersed themes, reflecting the comprehensiveness of Chinese journals and the specialization of English journals, as well as differences in research topics.

Based on these conclusions, information science researchers should focus not only on specific-level method, problem, and technical innovations but also attempt more difficult, higher-level research method and theoretical innovations, emphasizing interdisciplinary approaches to find new breakthroughs and promote information science theoretical development from quantitative to qualitative change, laying theoretical foundations for more mature development. These conclusions can also provide references for future research innovation directions and innovation point expression in paper writing. This preliminary attempt also demonstrates that analyzing paper innovation points through information extraction and syntactic analysis is feasible and valuable. This approach enables more fine-grained paper innovation analysis, grasping research innovation progress across the entire field, achieving monitoring purposes, promoting scientific research innovation, and providing methodological references for content-based innovation evaluation as a supplement to existing innovation evaluation systems.

However, this study has limitations: it uses sentence structure and syntactic features for sentence-level innovation identification without deeply considering semantic connections between sentence contexts at the logical level, which may affect results. Additionally, the study only extracted papers from two Chinese and two English journals, with insufficient sample coverage. Sentence extraction results are affected by syntactic analysis, with inadequate consideration of negation words, and related extraction rules require improvement. Addressing these limitations, future research will further optimize innovation classification structures, improve paper innovative sentence extraction algorithms, enhance extraction comprehensiveness, consider relationships between sentences and incorporate publication timelines, select more domain or multidisciplinary journal papers for empirical research, and add temporal comparisons to more comprehensively explore paper innovation characteristics, facilitating better fine-grained knowledge organization and retrieval.

References

[1] TAN R. On the declaration of novelty in scientific journal articles [EB/OL]. [2019-10-30]. https://www.philstar.com/business/science-and-environment/2014/04/24/1315251/declaration-novelty-scientific-journal-articles. [2] XU Shurong. The role of scientific journal editors in enhancing

paper innovation [J]. Chinese Journal of Scientific and Technical Periodicals, 2014, 25(6): 761-764. [3] GABRILOVICH E, DUMAIS S T, HORVITZ E, et al. Newsjunkie: providing personalized news feeds via analysis of information novelty [C]//Proceedings of the 13th international conference on World Wide Web. New York: ACM, 2004: 482-490. [4] OBEID N, RAOB K N. On integrating event definition and event detection [J]. Knowledge and information systems, 2010, 22(2): 129-158. [5] TSAI F S, CHAN K L. Redundancy and novelty mining in the business blogosphere [J]. The learning organization, 2010, 17(6): 490-499. [6] BREJA M. A novel approach for novelty detection of web documents [J]. International journal of computer science and information technologies, 2015, 6(5): 4257-4262. [7] LI X, CROFT W B. Novelty detection based on sentence level patterns [C]//Proceedings of the 14th ACM international conference on Information and knowledge management. New York: ACM, 2005: 744-751. [8] ZHANG Fan, LE Xiaoqiu. Research on theme attribute instance recognition in domain scientific literature innovation sentences [J]. Data Analysis and Knowledge Discovery, 2015, 31(5): 15-23. [9] WEN Hao. Research on semantic identification and classification methods for innovation points in scientific abstracts [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(03): 27-34. [10] LIU Qijin, CHENG Qikai, LU Wei. Analysis of objects described by academic literature innovation modifiers [C]//Nanjing: 9th National Doctoral Forum on Information Science. 2019. [11] Merriam-Webster, Incorporated. Merriam-Webster Online Dictionary [EB/OL]. [2019-10-30]. http://www.merriam-webster.com. [12] NG K W, TSAI F S, CHEN L C L, et al. Novelty detection for text documents using named entity recognition [C]//International Conference on Information. Piscataway: IEEE, 2008. [13] ALLAN J, WADE C, BOLIVAR A. Retrieval and novelty detection at the sentence level [C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2003: 314-321. [14] UZZI B, MUKHERJEE S, STRINGER M J, et al. Atypical combinations and scientific impact [J]. Science, 2013, 342(6157): 468-472. [15] ARTHUR W B. The nature of technology: what it is and how it evolves [M]. New York: Simon and schuster, 2009. [16] ZHOU Luyang. On the indicator system for evaluating academic paper innovation factors [J]. Acta Editologica, 2006(01): 68-70. [17] HE Wanying. Research on academic paper innovation evaluation based on machine learning [D]. Nanjing: Nanjing University, 2019. [18] LIU Liping, LIU Chunli. Analysis and suggestions on the pros and cons of open peer review [J]. Chinese Journal of Scientific and Technical Periodicals, 2017, 28(05): 389-395. [19] WANG Xiaohui, WANG Kang. Research on scholar academic influence evaluation and indicator relationships based on journal papers—taking competitive intelligence research as an example [J]. Information Science, 2018, 36(02): 63-66, 87. [20] LEE Y, WALSH J, WANG J, et al. Creativity in scientific teams: unpacking novelty and impact [J]. Research policy, 2015, 44(3): 684-697. [21] WU L, WANG D, EVANS J A, et al. Large teams develop and small teams disrupt science and technology [J]. Nature, 2019, 566(7744): 378-382. [22] SONG Ge. Design and empirical research of the S-index for scientific achievement innovation [J].

Library and Information Service, 2016, 60(05): 77-86, 124. [23] YE Jiyuan. Quality and innovation evaluation of academic journals [J]. Journal of Zhejiang University (Humanities and Social Sciences), 2013, 43(02): 108-117. [24] LI Rusen, PENG Caihong, ZHAO Furong. Methods for judging the innovation of scientific papers [J]. Journal of Anshan Iron and Steel Institute, 2001(03): 234-236. [25] YANG Jianlin, QIAN Lingfei. A topic novelty measurement method based on keyword pair inverse document frequency [J]. Information Studies: Theory & Application, 2013, 36(03): 99-102. [26] LIANG Shuai, GAO Jiping. Identification of excellent paper characteristics based on F5000 paper review comments [J]. Studies in Science of Science, 2017, 35(03): 331-337. [27] QIAN Lingfei, ZHANG Jiyu, WANG Rong, et al. Research on constructing an academic innovation measurement ontology based on domain knowledge [J]. Modern Information, 2019, 39(05): 30-37. [28] CUI Jingyan, DENG Yuan, LI Chunmei, et al. Identification and evaluation of innovation and scientific nature in traditional Chinese medicine scientific papers [J]. China Medical Herald, 2018, 15(05): 172-176. [29] CAO Xiaochun. The responsibilities of academic journal reviewers [J]. Editing Friend, 2006(06): 62-63. [30] SUO Chuanjun. Research on paper aging and innovation from a knowledge transfer perspective [J]. Library and Information Service, 2014, 58(05): 5-12. [31] HEINZE T, SHAPIRA P, SENKER J, et al. Identifying creative research accomplishments: Methodology and results for nanotechnology and human genetics [J]. Scientometrics, 2007, 70(1): 125-152. [32] ZHU Daming. The main functions of references and the evaluation of academic paper innovation [J]. Acta Editologica, 2004(02): 91-92. [33] CANNON D C, YANG J J, MATHIAS S L, et al. TIN-X: target importance and novelty explorer [J]. Bioinformatics, 2017, 33(16): 2601-2603. [34] SHEN Yang. A novelty evaluation method based on keywords [J]. Information Studies: Theory & Application, 2011, 30(3): 164-175. [35] OHSAWA Y, BENSON N E, YACHIDA M. KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor [C]//Proceedings of IEEE international forum on research and technology advances in digital libraries. Washington: IEEE, 1998: 12-18. [36] YANG Jing, WANG Fang, BAI Rujiang. A single academic paper innovation evaluation method based on research theme comparison [J]. Library and Information Service, 2018, 62(17): 75-83. [37] THORLEUCHTER D. Finding new technological ideas and inventions with text mining and technique philosophy [M]//Data analysis, machine learning and applications. Springer, Berlin, Heidelberg, 2008: 413-420. [38] TANG W, TSAI F S, CHEN L. Blended metrics for novelty detection [J]. Expert systems with applications, 2010, 37(7): 5172-5177. [39] FU X, CH' NG E, AICKELIN U. An improved system for sentence-level novelty detection in textual streams [C]//3rd International Conference on Smart Sustainable City and Big Data (ICSSC). London: IET, 2015. [40] AMORIM M, BORTOLOTTI F D, CIARELLI P M, et al. Novelty detection in social media by fusing text and image into a single structure [J]. IEEE access, 2019, 7: 132786-132802. [41] YU Husheng, ZHANG Ruiqing, YAN Weimin. Reviewing the innovation of scientific papers [J]. Acta Editologica, 2006(05): 333-334. [42] DAHL T. The linguistic representation of rhetorical

function: a study of how economists present their knowledge claims [J]. Written communication, 2009, 26(4): 370-391. [43] PARKINSON J. The discussion section as argument: the language used to prove knowledge claims [J]. English for specific purposes, 2011, 30(3): 164-175. [44] ZHANG Fan, LE Xiaoqiu. Research on sentence-level innovation point extraction for domain scientific literature [J]. New Technology of Library and Information Service, 2014(09): 15-21. [45] LI Lu. Empirical analysis of the integration between information resource industry and cultural industry—based on Chinese listed companies' data from 1997-2012 [J]. Information Science, 2016, 34(03): 122-126. [46] LI Aiming. Research on query expansion method based on ontology and user query intention [J]. Information Science, 2015, 33(05): 68-71. [47] WEI Dezhi, CHEN Fuji, LIN Lina. A hot topic discovery model and algorithm based on time series [J]. Information Science, 2017, 35(10): 142-146. [48] CAO X, CHEN Y, LIU K J, et al. A data analytic approach to quantifying scientific impact [J]. Journal of informetrics, 2016, 10(2): 471-484. [49] ROUSSEAU R, ZHAO S X. A general conceptual framework for characterizing the ego in a network [J]. Journal of informetrics, 2015, 9(1): 145-149. [50] KLAVANS R, BOYACK K W. Mapping altruism [J]. Journal of informetrics, 2014, 8(2): 431-447. [51] ZHENG Jie. NLP Chinese natural language processing [M]. Beijing: Publishing House of Electronics Industry, 2017: 283-287. [52] LI Ying, ZHOU Li. The value and ideal model of innovation point presentation in scientific journal papers [J]. Chinese Journal of Scientific and Technical Periodicals, 2018, 29(10): 993-999. [53] GB 7713-1987. Presentation of scientific and technical reports, dissertations and academic papers [S]. Beijing: Standards Press of China, 1987. [54] WEI Ruibin. Research on methodological innovation in domestic library and information science based on content analysis—taking co-word analysis as an example [J]. Library and Information Service, 2016, 60(24): 107-114. [55] Editorial Department of Data Analysis and Knowledge Discovery. Introduction to Data Analysis and Knowledge Discovery [EB/OL]. [2019-11-02]. http://manu44.magtech.com.cn/Jwk_{{{infotech}}_{{wk3}}}/CN/column/column291.shtml. [56] LIU Zhifeng, LI Xin, CHENG Qikai, et al. Construction and analysis of an academic text keyword semantic function dataset—taking Journal of Informetrics as an example [J]. Library Tribune, 2019, 39(07): 64-74. [57] JI Rongqin. The influence of differences between Chinese and Western thinking patterns on Chinese and English academic paper writing [J]. Journal of East China Jiaotong University, 2006(06): 134-137. [58] WEI Ruibin, LIU Yu. Research on the innovation of doctoral dissertation topics based on title text analysis—taking domestic information science doctoral dissertations as an example [J]. Journal of Intelligence, 2017, 36(07): 122-127.

Author Contributions:

Cao Shujin: Proposed research questions, framework, and revised the paper;
Yan Xinyang: Designed experimental protocols, conducted experimental analysis and data processing;
Zhang Qian: Conducted experimental analysis and data processing, wrote the paper;

Zhuo Yiling: Conducted experimental analysis and data processing, wrote the paper.

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