Journal of Informetrics 18 (2024) 101587

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

An effective framework for measuring the novelty of scientific articles through integrated topic modeling and cloud model ☆

Zhongyi Wang [a], Haoxuan Zhang [b,c], Jiangping Chen [d], Haihua Chen [b,c,*]

[a] School of Information Management, Central China Normal University, Wuhan, 430079, China
[b] Department of Information Science, University of North Texas, Denton, 76203, TX, USA
[c] Intelligent Data Engineering and Analytics Lab, University of North Texas, Denton, 76203, TX, USA
[d] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, 61820, IL, USA

ARTICLE INFO

Keywords:
Novelty measurement
Cloud model
Topic modeling
BERTopic

ABSTRACT

Novelty is a critical characteristic of innovative scientific articles, and accurately identifying novelty can facilitate the early detection of scientific breakthroughs. However, existing methods for measuring novelty have two main limitations: (1) Metadata-based approaches, such as citation analysis, are retrospective and do not alleviate the pressures of the peer review process or enable timely tracking of scientific progress; (2) Content-based methods have not adequately addressed the inherent uncertainty between the qualitative concept of novelty and the textual representation of papers. To address these issues, we propose a practical and effective framework for **m**easuring the **n**ovelty of **s**cientific **a**rticles through **i**ntegrated **t**opic **m**odeling and **c**loud **m**odel, referred to as **MNSA-ITMCM**. In this framework, papers are represented as topic combinations, and novelty is reflected in the organic reorganization of these topics. We use the BERTopic model to generate semantically informed topics, and then apply a topic selection algorithm based on maximum marginal relevance to obtain a topic combination that balances similarity and diversity. Furthermore, we leverage the cloud model from fuzzy mathematics to quantify novelty, overcoming the uncertainty inherent in natural language expression and topic modeling to improve the accuracy of novelty measurement. To validate the effectiveness of our framework, we conducted empirical evaluations on papers from the Cell 2021 journal (biomedical domain) and the ICLR 2023 conference (computer science domain). Through correlation analysis and prediction error analysis, our framework demonstrated the ability to identify different types of novel papers and accurately predict their novelty levels. The proposed framework is applicable across diverse scientific disciplines and publication venues, benefiting researchers, librarians, science evaluation agencies, policymakers, and funding organizations by improving the efficiency and comprehensiveness of identifying novelty research.

1. Introduction

Novelty is a crucial criterion for evaluating scientific outputs. Novel scientific articles may signal the emergence of future technological advances and serve as precursors to social change (Dwivedi et al., 2023). The assessment of novelty in scientific articles is

☆ This study is supported by National Social Science Foundation of China (Grant Number 22BTQ102).

* Corresponding author at: Department of Information Science, University of North Texas, Denton, 76203, TX, USA.
E-mail addresses: wzywzy13579@163.com (Z. Wang), haoxuanzhang@my.unt.edu (H. Zhang), jpchen@illinois.edu (J. Chen), haihua.chen@unt.edu (H. Chen).

https://doi.org/10.1016/j.joi.2024.101587
Received 12 October 2023; Received in revised form 12 July 2024; Accepted 2 September 2024

Available online 12 September 2024

often distinguished by ex-post and ex-ante approaches. Ex-post measurements primarily focus on the impact of scientific articles, such as citation-based bibliometric methods. However, evaluating post-publication outcomes does not provide an independent measure of their novelty; rather, it intertwines the assessment with determining their essential contribution to the field (Foster et al., 2015). Furthermore, factors like citation intent, citation sentiment, citation significance, the Matthew effect, and even citation manipulation can introduce biases (Kunnath et al., 2021; Ghosal et al., 2021; Horbach et al., 2022). Ex-ante measurements involve qualitative assessments by domain experts in peer review. The exponential growth in article numbers has imposed a substantial workload on scholars, delaying publication. Additionally, peer review is subject to subjectivity and low reproducibility (Lin et al., 2023). Consequently, this study will concentrate on ex-ante novelty measurements, aiming to provide a reference index for peer review and promote scientific progress.

Ex-ante novelty measurements assess the difference between a new scientific article and the population of pre-existing scientific articles (Foster et al., 2021; Wang et al., 2024b). Two major approaches are used for ex-ante novelty measurements: qualitative and quantitative. *Qualitative approaches* measure novelty through peer review. However, although the number of papers has snowballed, few are novel studies (Fortunato et al., 2018). Using the peer review process to measure novelty has imposed an undue burden on researchers. Furthermore, finding qualified reviewers for novel studies is often time-consuming, leading to prolonged publishing timelines (Huisman & Smits, 2017). Therefore, conducting novel research is challenging and risky: incumbent forces will resist the study's results and be subjected to a longer handling time before publication (Liang et al., 2022). For groundbreaking research, subjective factors, such as scientific paradigms and conservatism, may bias reviewers and make novel information unacceptable to the established scientific community (Wang et al., 2017).

Quantitative approaches aim to address the above problems, and several novelty measurements have been developed. The goal of novelty measurements using quantitative approaches is to represent the uniqueness of a specific knowledge element in a scientific article. If an article contains or is associated with new knowledge elements, it indicates that the article delivers novel information. Researchers define *novelty* as the capturing of a recombination of knowledge elements in which a new or rare combination of knowledge elements is considered a sign of novelty (Uzzi et al., 2013). This definition emphasizes the recombination of pre-existing knowledge components in an unprecedented fashion, which is popular among the research community in different disciplines (Wang et al., 2017). For example, the novelty measurements of combining MeSH keywords (Boudreau et al., 2016), IPC codes (Verhoeven et al., 2016), and cited references (Matsumoto et al., 2021) have been proposed. However, these alternative approaches do not reflect the true novelty of the content, and it is questionable whether metadata can be used as a representation of the content (Leydesdorff et al., 2017).

Recently, word embedding-based methods have become the new direction of novelty measurements that incorporate the content of scientific articles. For instance, Jeon et al. (2023) extended the training of the fastText word vector model and employed the local outlier factor (LOF) anomaly detection technique to discover and measure the novelty of scientific articles. Luo et al. (2022) proposed combining the scientific research questions and methods in a paper to measure their novelty. Wang et al. (2022b) measured the novelty of each method knowledge element in scientific articles to help researchers find essential studies. Hou et al. (2022) represented scientific articles as distinct combinations of research questions, methods, and results. However, using the novelty of a specific knowledge element to represent the novelty of the whole article might cause bias. The reason is that, in addition to research questions and methods, scientific articles include other knowledge elements, such as research theory, dataset, and others, which also contribute to the novelty of a paper. On the other hand, the knowledge content discussed in scientific articles is subject to uncertainty factors (Li et al., 2021). For instance, authors might employ different expressions to convey the same meaning within their texts. Additionally, they may make inferential statements using ambiguous vocabulary (Yao et al., 2023). When measuring the novelty of scientific articles, the role of uncertainty factors should be considered. In summary, novelty measurements are highly heterogeneous and complicated.

This study aims to develop a practical and effective framework for pre-evaluating the novelty of scientific articles. To achieve this purpose, we develop and answer the following research questions:

- **How to appropriately represent the content of scientific articles?** The representation of the content of scientific articles should not only capture the semantic information, but also consider the impact of uncertainty factors. On the one hand, we adopt a topic modeling approach based on pre-trained language models (BERTopic) to extract the core knowledge and semantic relationships within the scientific articles. On the other hand, considering the uncertainties in natural language expression and the topic modeling process, we introduce the cloud model from fuzzy mathematics to enable the transformation between the qualitative novelty level and the quantitative topic representation of the scientific articles.
- **How to effectively measure the novelty of scientific articles?** The novelty of scientific articles is reflected in the atypical recombination of knowledge, which we view as topic-level combinations. We use the maximum marginal relevance algorithm to obtain topic combinations. This ensures the topic combinations are both relevant to the paper and diverse from each other. We then use the cloud model to process the topic combinations. Finally, we calculate the similarity between the article's cloud and the novelty standard cloud using the Hellinger distance algorithm, thereby determining the level of novelty.
- **How to objectively validate the novelty measurement method?** For the evaluation dataset, we selected papers from the ICLR 2023 conference in computer science and papers from the Cell 2021 journal in biomedical science. The evaluation criteria for novelty are from real-world researchers, making the dataset robust and reliable. For the comparison methods, we selected citation-based methods (Originality index (Trajtenberg et al., 1997) and Wang's novelty (Wang et al., 2017)) and the latest content-based methods (fastText+LOF and fastText+IF (Jeon et al., 2023)) that have been widely validated by scholars. Combining correlation

analysis and prediction error analysis, we have demonstrated that the proposed method can effectively identify papers of various types of novelty and can accurately predict the level of novelty of the papers.

The rest of the paper is organized as follows: Section 2 reviews related work on measuring novelty in scientific articles. Section 3 proposes a framework for measuring novelty in scientific articles. Section 4 empirically analyzes the proposed framework. Section 5 discusses the implications and limitations of the study. Section 6 summarizes the paper and discusses future work. Our dataset, the code used to reproduce the method, and other experimental results are available on GitHub.[1]

2. Related work

This research mainly focuses on measuring the novelty of scientific articles. Therefore, we review the existing studies from two aspects: (1) novelty and other relevant concepts, and (2) novelty measurements of scientific articles.

2.1. Novelty and other relevant concepts

To define novelty formally, we clarify the differences between novelty and other relevant terms, such as creativity, innovation, disruption and originality, before considering how to measure the novelty of scientific articles.

Novelty has been used in multiple studies and defined from different perspectives. For example, Foster et al. (2021) defined novelty as the difference between new scientific articles and the population of pre-existing scientific articles. Arts et al. (2021) defined novelty as the uniqueness of a specific knowledge element. If a scientific article contains new knowledge elements, it indicates that the article delivers novel information. Other researchers defined novelty as a new combination of knowledge elements (Boudreau et al., 2016). Although scholars hold different interpretations regarding the novelty of scientific articles, they all agree that novelty refers to the quality of presenting new information within these articles.

Creativity is the production of a novel and appropriate response, product, or solution to an open-ended task (Amabile, 1983). Following the definition, Bornmann et al. (2019) regarded creativity as the medium in which novel directions in research emerge. Other researchers interpreted creativity as a process of imagining novel combinations of elements from structured knowledge spaces (Lee et al., 2015). Therefore, novelty can be seen as the result of creativity which manifests itself in various creative acts.

Innovation is the application of new ideas to the products, processes, or any other aspect of a firm's activities (Rogers & Rogers, 1998). It is the process of introducing new ideas to the firm, thereby increasing its performance. Innovation is concerned with the process of commercializing or extracting value from ideas, especially the first commercialization of an idea. Novelty is the first occurrence of an idea for new things. Even though novelty and innovation are two very similar concepts, novelty differs from innovation in that the former need not be directly associated with commercialization (Rogers & Rogers, 1998). Some researchers believe that novelty is a prerequisite for innovation (Runco & Jaeger, 2012). In many cases, however, there is a considerable time lag between the two. Such lags reflect the different requirements for each of them. Novelty may be carried out anywhere, for instance, in universities; while innovations occur mainly in firms in the commercial sphere (Fagerberg, 2004).

Disruption denotes the degree to which patents and scientific publications induce changes in established fields (Wang et al., 2024a). Both disruptive and incremental innovations are forms of innovation, with incremental innovation consolidating existing domains while disruptive innovation replaces and reshapes existing knowledge. Funk and Owen-Smith (2017) proposed the CD_t index based on patents to measure the disruptive of new technologies, considering the shift in inventors' attention from existing foundational knowledge and the reshaping of technological networks. Wu et al. (2019) applied the CD_t index to the field of scientometrics, using citation networks of focal papers (DI_1 index) to assess whether scientific articles are disruptive or incremental. To identify the specific content of disruptive scientific research, Wang et al. (2022a) further optimized the DI_1 index by examining key scientific concepts at a granular knowledge level (ED index). In comparison to novelty, disruptiveness not only focuses on the origin of scientific research but also considers its utility and impact (Leibel & Bornmann, 2024).

Originality arises from deviations in existing knowledge, generating new ideas, methods, conclusions, and valuable outputs, or fostering further innovation (Hou et al., 2022). According to Shibayama and Wang (2020), *originality* refers to a scientific discovery that imparts unique knowledge to subsequent research, which was not available to previous generations. Some researchers argue that originality encompasses anything that introduces novelty to human knowledge (Dirk, 1999; Ziman, 2003), including innovative problem-solving approaches, research methods, and theoretical perspectives. The distinction between originality and novelty lies in the uniqueness and distinctiveness of scientific research in the former, while the latter focuses on contributing to and advancing scientific research within existing knowledge. However, differentiating between originality and novelty is often challenging in practical measurements, as originality is often implicit in research papers (Guetzkow et al., 2004). Consequently, originality and novelty are frequently used interchangeably in most cases (Shibayama & Wang, 2020; Yan et al., 2020; Wang, 2024).

In summary, novelty, as a result of creativity, serves as a prerequisite for innovation and disruption, and is synonymous with originality. While scholars may have different interpretations of novelty, there is a consensus that in scientific articles, it refers to the inherent quality of presenting new knowledge. Drawing from the definition of novelty in the Oxford English Dictionary, we define it as an attribute of knowledge that includes "something" new. Specifically, scientific articles with novelty involve new research questions, research methods, theories, concepts, or the recombination of existing knowledge elements. Therefore, in this paper, we

[1] https://github.com/haihua0913/MNSA-ITMCM.

will explore the novelty of scientific articles from the perspective of topic combinations. Topics can reflect various types of knowledge elements through a series of topic terms, and the recombination of topics embodies the concept of atypical combinations of knowledge elements.

2.2. Novelty measurements of scientific articles

The essence of scientific article evaluation is the assessment of academic value, which is concretely expressed in the measurement of novelty. How to evaluate the novelty of scientific articles is currently studied in the academic community under two categories: metadata-based measurements and content-based measurements.

2.2.1. Metadata-based measurements

Metadata-based novelty measurement of scientific articles can be categorized into two strategies: knowledge origins and atypical knowledge combinations. The first strategy focuses on assessing the novelty of keywords provided in scientific articles, which provides a simple and generic approach to capturing macro-level novelty (Chen & Fang, 2019). In the knowledge origins perspective, novelty is inferred based on the presence or association of specific keywords in a scientific article. However, the selection of keywords is subjective, and researchers may use different keywords to describe similar concepts, thereby introducing potential bias when assessing novelty.

On the other hand, alternative studies conceptualize novelty as atypical combinations of existing knowledge elements and measure it by analyzing the local or global co-occurrence of these combinations (Uzzi et al., 2013). These knowledge elements can include controlled word lists such as MeSH terms (Hofstra et al., 2020), IPC codes (Verhoeven et al., 2016), references (Matsumoto et al., 2020), and referenced journals (Wang et al., 2017) and others. However, the validity of metadata-based novelty measures, which do not consider the content of scientific articles and rely on a coarse granularity of knowledge elements, has been a subject of controversy (Fontana et al., 2020). Additionally, variations in the motivation behind citing literature and the content of citations can impact the calculation of novelty across different scientific articles (Bornmann et al., 2020; Ghosal et al., 2021).

2.2.2. Content-based measurements

Content-based measurements focus on extracting the critical information of the paper from its content, compensating for the shortcomings of metadata-based methods (Wang et al., 2024b). Representing the content of scientific articles is a critical issue, and various techniques are currently employed, including word embeddings, pre-trained language models, community detection, and topic modeling. Word embedding models transform words or texts into vectors in a continuous vector space, effectively capturing semantic and syntactic relationships. For example, Jeon et al. (2023) trained the fastText word embedding model on scientific paper titles and used density-based anomaly detection techniques to evaluate the isolation level of papers within their respective neighborhoods, thus calculating novelty.

In contrast, pre-trained language models such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2019) excel in contextual modeling, as they can map text to high-dimensional semantic spaces, capturing subtle semantic and contextual knowledge. Luo et al. (2022) treated scientific articles as combinations of research questions and research methods, leveraging the BERT model to assess the semantic novelty of these combinations. Similarly, Hou et al. (2022) used the SBERT model to measure the novelty of knowledge links within the network. Wang (2024) proposed an information-theoretic approach, utilizing GPT model to calculate the probability distribution of vocabulary in scientific articles, thereby measuring the novelty of papers at the word level.

Community detection involves analyzing network structures by considering scientific articles as nodes and constructing citation networks based on their citation relationships. Min et al. (2021) discovered unique network structural characteristics in groundbreaking scientific articles based on citation networks. Xu et al. (2020) applied the Leiden algorithm to partition research communities in the stem cell field. They identified emerging research topics by examining network structure-related indicators. Wang et al. (2024b) further refined the knowledge community defined by an article into a network of knowledge entities within the article. They predicted the article's quality by evaluating the novelty of entity pairs, tracking changes in network structure, and considering various additional features.

Topic modeling is an automated method aimed at discovering latent topics in textual data, allowing for the identification of topics, relationships between topics, and the distribution of text with respect to these topics. Savov et al. (2020) combined LDA topic modeling with SVM classifiers to classify the innovativeness of scientific articles by analyzing the occurrence time and trends of topics within the articles.

In conclusion, the current methods employed for evaluating the novelty of scientific articles possess certain limitations. Metadata-based approaches, which rely on external data rather than the content of the articles, fail to capture specific nuances among scientific articles. Furthermore, these methods struggle to encompass the incremental contributions made by scientific articles to existing research, thus constraining their capacity to measure novelty. Consequently, content-based methods exhibit significant potential in assessing the novelty of scientific articles. However, content-based approaches encounter challenges stemming from factors of uncertainty, including syntactic variations, conflicting terminology, and speculative discourse (Yao et al., 2023; Wang et al., 2023). A comprehensive resolution of these issues is imperative for effectively quantifying the novelty of scientific articles. This study aims to measure the novelty of scientific articles by focusing on the combinations of topics. Semantic representation of article content can be achieved through topic modeling, while comparing the topic combinations across different articles enables the calculation of their novelty. Unique topic combinations are regarded as possessing higher levels of novelty, aligning with the concept of atypical knowl-

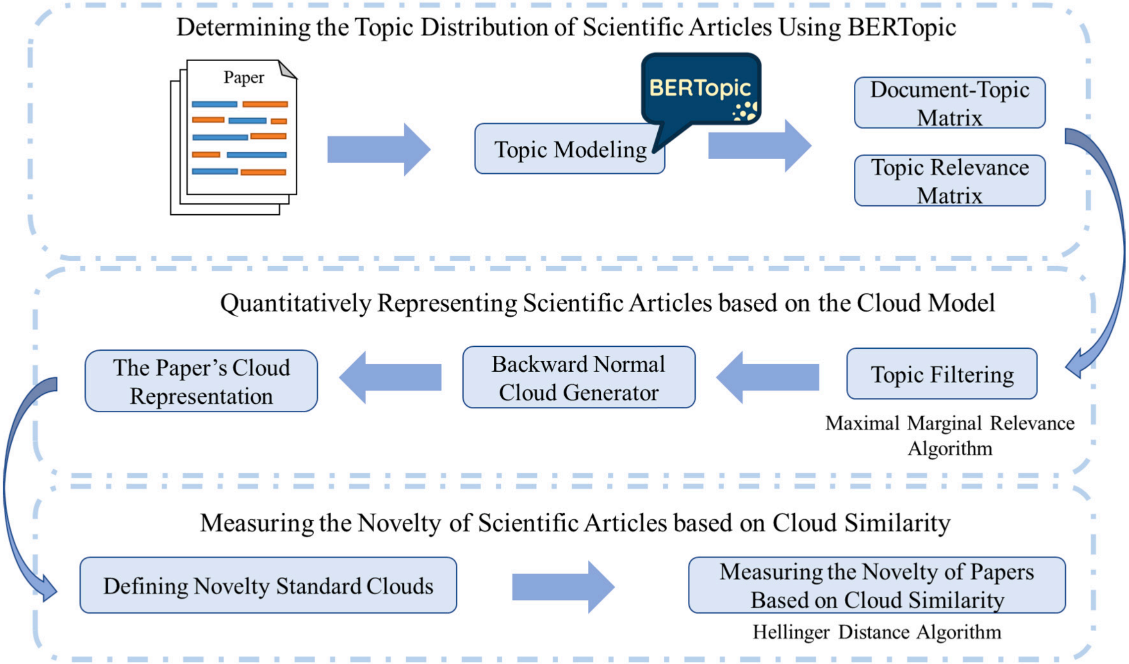

Determining the Topic Distribution of Scientific Articles Using BERTopic
Paper
Topic Modeling
BERTopic
Document-Topic Matrix
Topic Relevance Matrix
Quantitatively Representing Scientific Articles based on the Cloud Model
The Paper's Cloud Representation
Backward Normal Cloud Generator
Topic Filtering
Maximal Marginal Relevance Algorithm
Measuring the Novelty of Scientific Articles based on Cloud Similarity
Defining Novelty Standard Clouds
Measuring the Novelty of Papers Based on Cloud Similarity
Hellinger Distance Algorithm

Fig. 1. The workflow of MNSA-ITMCM.

edge recombination. Additionally, this research employs the cloud model to fuzzify the obtained topics, facilitating the quantitative representation of novelty levels for scientific articles.

3. Methodology

In this section, we describe the proposed framework for **m**easuring the **n**ovelty of **s**cientific **a**rticles through **i**ntegrated **t**opic **m**odeling and **c**loud **m**odel (**MNSA-ITMCM**). The overall workflow of MNSA-ITMCM is depicted in Fig. 1. The framework consists of three main steps: (1) determining the topic distribution of scientific articles using BERTopic, (2) quantitatively representing scientific articles based on the cloud model, and (3) measuring the novelty of scientific articles based on cloud similarity.

3.1. Determining the topic distribution of scientific articles using BERTopic

The first step is to transform the unstructured content into quantifiable data to measure the novelty of scientific articles. Savov et al. (2020) argued that an article's topic distribution could reflect its innovation level. Moreover, topics can provide insights into the novelty of knowledge elements (Wang et al., 2022b). Therefore, in this study, we will employ the topic modeling approach to facilitate the quantitative representation of scientific articles.

3.1.1. Quantitative representation of scientific articles using BERTopic

Topic modeling is a statistical-based text analysis method widely employed in academia. It allows for the automatic identification of latent topics within large-scale textual corpora, thus enabling the quantitative representation of scientific articles. Taking article titles as examples, the topic modeling approach assumes that titles are associated with multiple underlying topics, each represented by a set of words. By applying topic modeling to title texts, researchers can uncover the distribution of topics within the texts and explore the relationships between topics and articles. This analysis provides valuable insights into the topic structure of the texts and allows for a rigorous and quantitative representation of scientific articles.

Traditional topic models, such as latent Dirichlet allocation (LDA), require a prior specification of the number of clusters. However, their topic inference became less accurate when faced with challenges like insufficient co-occurrence information of document vocabulary or a lack of domain knowledge. Pre-trained language models (PLMs) provide the technical prerequisites to improve semantic consistency, classification accuracy, and topic interpretability of topics (Sia et al., 2020). BERTopic is an effective topic modeling technique that applies PLMs to enhance background knowledge and semantic understanding in the topic modeling process (Grootendorst, 2022).

3.1.2. Topic-based MMR algorithm for topic selection

We obtained the document-topic relevance matrix and topic-relevance matrix using BERTopic for topic modeling. Drawing upon the concept of combinatorial novelty, it is often the atypical combinations that stimulate innovation. Therefore, we incorporated the

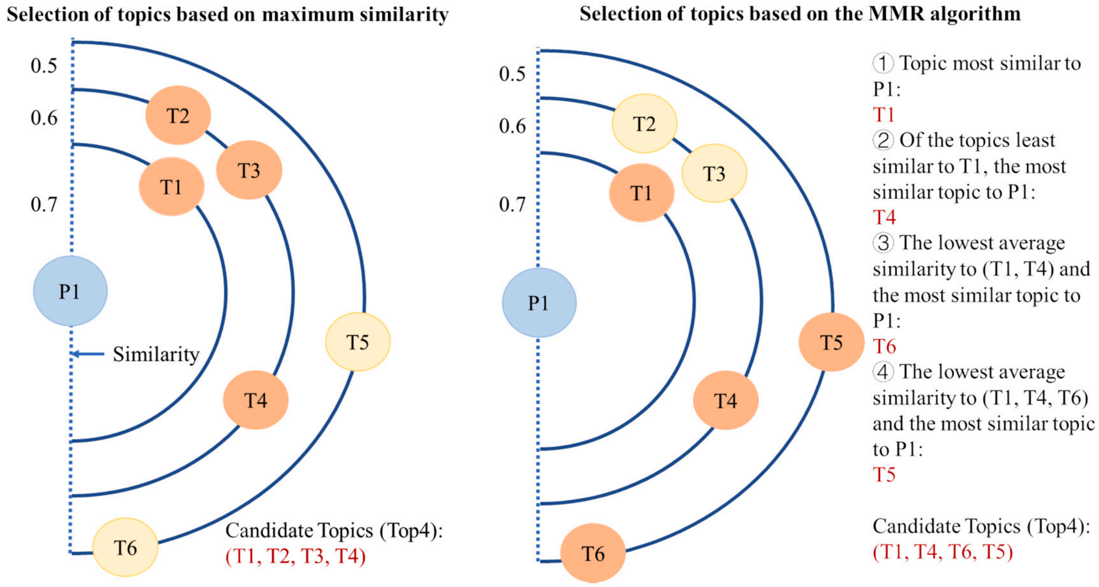

Selection of topics based on maximum similarity
0.5
0.6
0.7
P1
T1
T2
T3
T4
T5
T6
Similarity
Candidate Topics (Top4):
(T1, T2, T3, T4)
Selection of topics based on the MMR algorithm
0.5
0.6
0.7
P1
T1
T2
T3
T4
T5
T6
① Topic most similar to P1:
T1
② Of the topics least similar to T1, the most similar topic to P1:
T4
③ The lowest average similarity to (T1, T4) and the most similar topic to P1:
T6
④ The lowest average similarity to (T1, T4, T6) and the most similar topic to P1:
T5
Candidate Topics (Top4):
(T1, T4, T6, T5)

Fig. 2. Selection of candidate topics using the topic-based MMR algorithm. P represents scientific articles, while T represents the corresponding topics related to those articles.

perspective of combinatorial novelty theory into the composition of scientific article topics. Specifically, while ensuring the maximum similarity to scientific articles, selecting topics with greater diversity can help avoid being confined to specific research fields and uncover potential intersections across different domains. To achieve this goal, based on the document-topic matrix and topic relevance matrix, we enhanced the maximal marginal relevance (MMR) algorithm in information retrieval to select candidate topics with the highest diversity for scientific articles. Fig. 2 depicts the topic selection process, and its algorithmic flow is illustrated in Algorithm 1.

As illustrated in the left of Fig. 2, when there was a phenomenon of topic clustering, if we based the selection of candidate topics for scientific articles on maximum similarity, the obtained topics exhibited convergence in the semantic space. However, according to the Algorithm 1, the selected topics had a broader coverage in terms of relevance, aiding in the discovery of related and more diverse research domains. In other words, the topic-based MMR algorithm provided an interpretable and easily implementable method to identify non-conventional combinations of topics, thus offering a means to uncover atypical research areas.

3.2. Quantitatively representing scientific articles based on the cloud model

In previous studies exploring the novelty of scientific articles, researchers often overlooked the inherent fuzziness and randomness present in both natural language and the resulting topic distribution. Uncertainty within natural language encompasses fuzziness and randomness, including fuzzy concepts and quantifiers (Yao et al., 2023; Bornmann et al., 2020). For instance, words like “considerable,” “somewhat,” or “possible” exhibit fuzziness as their interpretations vary depending on context and individual understanding. Similarly, linguistic phenomena demonstrate randomness with different words or sentences conveying the same meaning. These characteristics extend to the topic distribution of scientific articles, where randomly generated distributions may differ for the same article across runs. Moreover, the polysemy of words, influenced by context, can lead to their potential allocation to various topics. In addition, topics possess a degree of fuzziness, comprised of multiple words that are not directly correlated. Randomness and fuzziness in natural language and topic models pose challenges in training and interpreting these models, necessitating further refinement of the obtained topic distribution. In this study, the cloud model from fuzzy mathematics is incorporated into the measurement of novelty to capture randomness and fuzziness. The cloud model, which is a probabilistic model, is used to handle uncertainty, facilitating the transformation between qualitative concepts and quantitative representations.

Definition 1 *(Cloud and Cloud Drops for Text).* Let **U** be a quantitative domain represented by numerical values, and let **C** be a qualitative text associated with **U**. Assuming that the quantitative topic $\mathbf{x} \in \mathbf{U}$ is a random realization of the qualitative text **C**, we define the determination of **x** on **C** as $\mu(x) \in [0, 1]$, where $\mu : \mathrm{U} \to [0, 1]$, $\mathrm{x} \in \mathrm{U}$, and $\mathrm{x} \to \mu(x)$. The distribution of **x** on the domain **U** is referred to as a cloud, denoted as **C(x)**, and each **x** is called a cloud drop W_{i}.

The fundamental concept of the cloud model is the “cloud,” which represents the randomness and fuzziness of information. The cloud model describes a cloud by defining three key parameters: Expectation, Entropy, and Hyper-entropy. Expectation represents the value on the domain corresponding to the centroid of the area covered by the cloud, and it is the most representative point for qualitative concepts. Entropy measures the uncertainty of the qualitative concept. On the one hand, it quantifies the randomness of the qualitative concept, reflecting the degree of dispersion of the cloud drops that represent the concept. On the other hand, it measures

Table 1
Descriptive information of the data collection.

Domain	Training Data		Testing Data	
	Source	Number of papers	Source	Number of papers
Computer Science	arXiv_cs (Jan 2020 - Sep 2022)	205,381	ICLR 2023	3,809
Biomedical Science	Cell (1974 - 2020)	15,204	Cell 2021	447

the fuzziness of the qualitative concept, reflecting the range of values the concept in the domain space can accept. Hyper-entropy is the uncertainty measure of entropy, i.e., the entropy of entropy. Researchers usually use hyper-entropy to describe the thickness of the cloud, determined by the combined effects of randomness and fuzziness. In this research, we consider the text of the ith scientific paper as a cloud (C_i). The cloud is quantitatively characterized by multiple candidate topics X_j and their respective similarity W_{ij} to the scientific paper. Thus, $C_i = (x_j, W_{ij})$. To determine the numerical properties of the cloud, including expectation, entropy, and hyper-entropy, we utilize the backward normal cloud generator (BNCG). The BNCG serves as an intermediate converter, transforming the cloud drops of each text into the numerical features (E_x, E_n, H_e) of a cloud.

Definition 2 *(Backward Normal Cloud Generator).* The BNCG employs equations (1), (2), (3), and (4) to calculate the numerical features (E_x, E_n, H_e) of the cloud.

$$E_{x_i} = \frac{1}{n}\sum_{j=1}^{n} \mu\left(W_{\mathrm{ij}}\right) \tag{1}$$

$$E_{n_i} = \sqrt{\frac{\pi}{2}}\frac{1}{n}\sum_{j=1}^{n}\left|W_{\mathrm{ij}} - E_{x_i}\right| \tag{2}$$

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(W_{\mathrm{ij}} - Ex_i\right)^2 \tag{3}$$

$$H_{e_i} = \sqrt{S_i^2 - En_i^2} \tag{4}$$

Here, n represents the number of cloud drops, W_{ij} represents the ith cloud drop of the jth text, and the sample variance S_i^2 represents the degree of association between the core topics and related topics in a text.

3.3. Measuring the novelty of scientific articles based on cloud similarity

After obtaining the cloud for each scientific article, we first defined three standard cloud models based on the degree of novelty - highly, moderate, and low. The parameters of these standard clouds are dependent on the specific datasets. We then employed the Hellinger distance algorithm to calculate the similarity between each article's cloud model and the standard cloud models. Hellinger distance is a measure of the similarity between two probability distributions, introduced by (Xu & Wang, 2023) for cloud similarity computation. It has strong differentiation ability and low computational complexity. This approach allows us to determine the novelty level of each article by comparing its cloud model to the predefined standard cloud models. See the Appendix A.2 for details on the Hellinger distance-based similarity calculation.

4. Experiments and results

4.1. Data collection and pre-processing

This study utilized the titles of publicly available academic papers as experimental data and selected papers from two different fields. The first dataset consisted of submission papers from the 2023 International Conference on Learning Representations (ICLR). ICLR is a top-tier conference in the field of artificial intelligence, covering research topics in data science, computer vision, speech recognition, and natural language understanding. The second dataset was collected from the Cell journal by Jeon et al. (2023), spanning the years 1974 to 2021. Cell is a prestigious biomedical journal, covering research topics in cellular biology, molecular biology, immunology, neuroscience, and other related areas.

Meanwhile, to train the topic model, we need to construct training corpora for each dataset. For the first dataset, ICLR 2023, we selected the titles of papers in the computer science category from the arXiv preprint database as the training data, spanning from 2020 to September 2022. We chose the arXiv preprint database because it houses a vast collection of conference papers in various computer science disciplines, including previous ICLR conference papers. For the second dataset, we only selected papers from the year 2021 as the test data, and papers from 1974 to 2020 as the training data. Finally, the collected datasets and their divisions are presented in Table 1.

4.2. Gold standard

We have established separate gold standards for the two datasets. The first dataset consists of ICLR conference papers, where the review information is publicly available on the OpenReview[2] website. Each paper is reviewed by two to five experts, who provide scores for the technical novelty and significance (TNS) as well as the empirical novelty and significance (ENS) of the paper. These scores reflect the novelty of the technical and empirical contributions, and range from 0 to 4, with higher scores indicating greater novelty. It is worth noting that for purely theoretical papers, the reviewers only provide the TNS rating and are unable to give an ENS rating. For each paper, we calculate the average of the TNS scores and the average of the ENS scores. This approach balances the scores from different reviewers, avoiding the influence of outliers. We then take the maximum of these two averages as the novelty score for that paper. Using the maximum value helps to capture the overall highest level of novelty, whether it is in the technical or empirical aspect. To obtain these data, we developed a Python script to extract the title text and novelty scores from the paper pages on OpenReview.

For the Cell dataset, numerous scholarly experts on the Faculty Opinions forum openly recommend papers and assign four positive novelty labels: *TECHNICAL_ADVANCE*, *NOVEL_DRUG_TARGET*, *NEW_FINDING*, and *HYPOTHESIS*. We have utilized the publicly available title text and gold standard data provided by Jeon et al. (2023).

4.3. Baselines

To evaluate the performance of our method and its consistency with other approaches, we conducted comparisons with several benchmark methods, including the Originality index (Trajtenberg et al., 1997), Wang's novelty (Wang et al., 2017), fastText+LOF, and fastText+IF (Jeon et al., 2023). The first two methods are based on citation data and have been widely used and validated in the field of scientometrics. However, their drawback is that they require time to accumulate citations. The fastText+LOF and fastText+IF methods are more recent approaches that leverage limited paper content, such as the title, to effectively identify the novelty of scientific articles, without relying on citation information.

- **Originality index**. The concept of originality, initially rooted in patent articles, measures novelty based on cited patents' classification codes. When applied to scientific articles, it uses cited journals instead. The formula for calculating a paper's originality (p) is as follows:

$$\text{Originality}(p) = 1 - \sum_{j=1}^{N_p} \left(\frac{NCITED_{pj}}{NCITED_p} \right)^2 \tag{5}$$

In the Formula (5), N_p signifies the number of unique journals cited by paper p, $NCITED_{pj}$ represents the frequency of paper p citing journal j, and $NCITED_p$ indicates the total number of references in paper p. The originality of a paper depends on the breadth of its references, with greater diversity in the sources leading to higher originality.

- **Wang's novelty**. Wang's novelty is based on the concept of combinatorial novelty, emphasizing that novel combinations of existing knowledge are essential for inspiring further research. This index quantifies knowledge recombination by examining co-citations of journal combinations. It calculates the ease of forming specific combinations, with higher difficulty indicating a lack of prior co-citations before a paper's publication but co-citations occurring afterward. The calculation formula is as follows:

$$\text{Wang's novelty}(p) = \sum_{\text{newpair}(J_i,J_j)} (1 - \cos_{J_i,J_j}) \tag{6}$$

In the Formula (6), cos_{J_i,J_j} represents the cosine similarity of co-citations between journals J_i and J_j, used to assess the ease of forming that journal combination. The novelty of scientific articles is the sum of the difficulty values for all journal combinations.

- **FastText+LOF/IF**. FastText+LOF/IF methods represent fastText+LOF and fastText+IF, which were proposed by Jeon et al. (2023). The fundamental idea of the two methods is that if a research paper is identified as an outlier in a vector space, it is considered novel. These methods utilize the fastText pre-trained word vector model to create a vector space for scientific paper titles. They employ density-based anomaly detection techniques to measure the isolation level of the detected object from its neighboring context. IF, on the other hand, is an outlier detection method based on isolation forests. It constructs a random binary tree to partition the data and quantifies the anomaly level of data points based on the length of their paths within the tree.

4.4. Experimental setup

We conducted experiments on Featurize,[3] an online machine learning platform, using TensorFlow 2.13.0 and Python 3.10. The experiments were run on an Intel Xeon Gold 5218R CPU and an RTX 3090 GPU. In the topic modeling, the document embedding model used, all-mpnet-base-v2 (Reimers & Gurevych, 2019), has been pre-trained on a large corpus of over 1 billion sentences, which

[2] https://openreview.net/group?id=ICLR.cc&referrer=%5BHomepage%5D(%2F).

[3] https://featurize.cn/.

Table 2
The experimental parameters for the BERTopic model.

Parameter	Value
Topic Model Version	BERTopic v0.16.0

Dimensionality Reduction (UMAP)	n_neighbors = 15
	n_components = 5
	min_dist = 0.0
	metric = ’cosine’

Clustering (HDBSCAN)	min_cluster_size = 150
	metric = ’euclidean’
	cluster_selection_method = ’eom’

Tokenizer (CountVectorizer)	stop_words = ”english”
	min_df = 2
	ngram_range = (1, 3)

Weighting Scheme	c-TF-IDF
Fine-tune Topic representation	KeyBERT

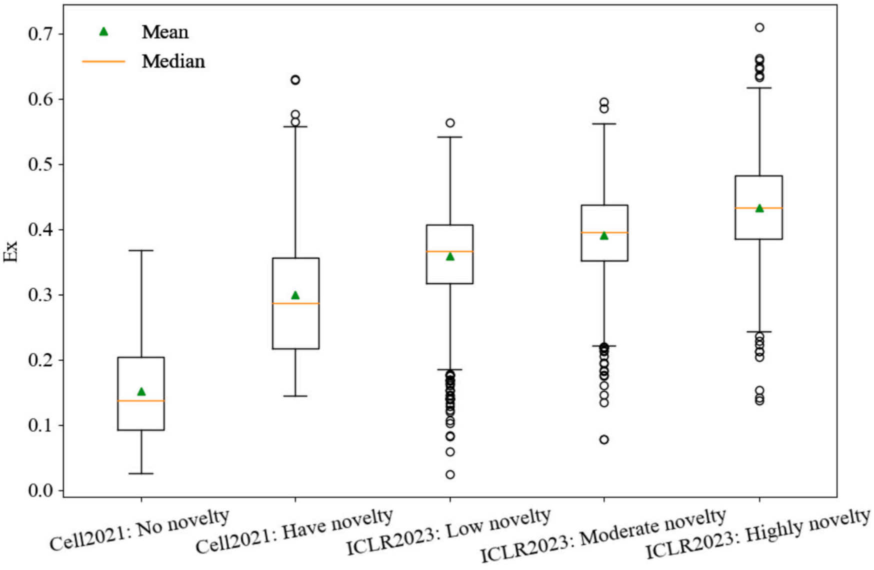


Fig. 3. The boxplots of the E_x values for the Cell 2023 and ICLR 2023 datasets.

includes the titles and abstracts of scientific articles. Given this broad training data, the model is well-suited to represent the textual content in the present experiment. The experimental parameters for the BERTopic model are presented in Table 2.

In the fastText model training, we followed Jeon et al. (2023)’s procedure, training word vector models individually for all years with parameter settings consistent with their experiments.

4.5. *Experimental results and analysis*

4.5.1. *Novelty measurement of scientific articles based on standard cloud*

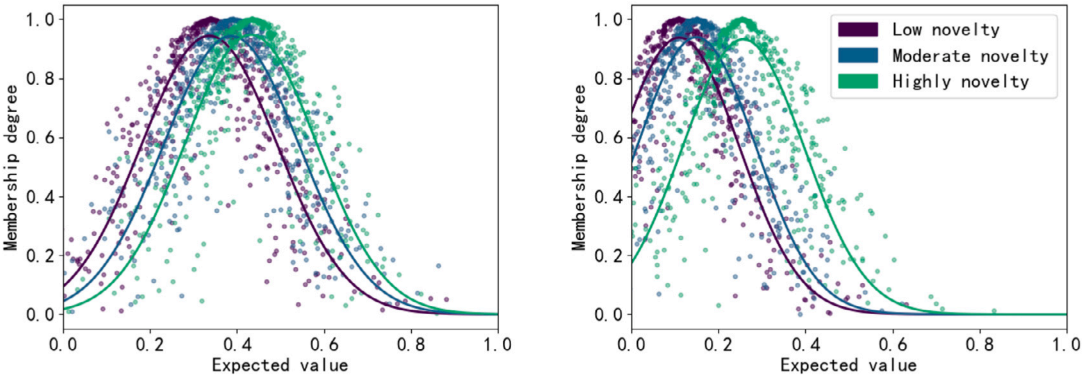
Based on the aforementioned experimental setup, we applied BERTopic for topic modeling on two distinct training datasets: arXiv_cs (2020.1-2022.9) and Cell (1974-2020). This resulted in the creation of two corresponding topic models. Subsequently, we constructed document-topic matrices and topic relevance matrices for the test datasets of the two corpora (ICLR 2023 and Cell 2021). The calculation of relevance was performed using the cosine similarity measure on the embedded vectors obtained through the BERTopic model’s interface. Then, based on these matrices, we utilized Algorithm 1 to select the top five candidate topics for each set of scientific articles, ensuring both relevance and maximal diversity.

Next, we established the novelty standard cloud criteria based on the test datasets. Prior to this, we divided the Cell 2021 dataset into papers with and without novelty labels. For the ICLR 2023 dataset, we categorized the papers into three groups based on their previously obtained novelty scores: highly novelty (top 25%), moderate novelty (middle 25-50%), and low novelty (bottom 50-100%).

The boxplots in Fig. 3 display the E_x values for the Cell 2023 and ICLR 2023 datasets, grouped by novelty level. As the level of novelty increases, the E_x values (including the quartiles, mean, and median) consistently rise in both datasets. This suggests that papers with higher novelty have greater E_x values. Therefore, we divided the novelty standard clouds for both datasets into three levels: highly, moderate, and low novelty. Specifically, we set the high novelty E_x as the upper quartile, the moderate novelty E_x as the median, and the low novelty E_x as the lower quartile, to mitigate the influence of outliers. Additionally, the E_n and H_e values

Table 3
Numerical features of the novelty standard cloud.

The Novelty Standard Cloud	ICLR 2023			Cell 2021		
	E_x	E_n	H_e	E_x	E_n	H_e
Highly Novelty	0.435	0.140	0.049	0.254	0.123	0.044
Moderate Novelty	0.389	0.140	0.049	0.150	0.123	0.044
Low Novelty	0.339	0.140	0.049	0.109	0.123	0.044



(a) ICLR 2023 (b) Cell 2021

Fig. 4. The novelty standard clouds in ICLR 2023 and Cell 2021.

Table 4
The descriptive statistical analysis of the novelty levels for ICLR 2023 and Cell 2021.

	Number of papers	Mean	Std	Min	Max
ICLR 2023	3,809	1.01	0.859	0	2
Cell 2021	447	1	0.873	0	2

were set to the dataset's mean, as these metrics exhibit individual variability. The numerical features of the novelty standard clouds for the two datasets are shown in Table 3, and the corresponding visualization of the clouds is presented in Fig. 4.

Lastly, we calculated the similarity between each scientific article set and five standard clouds using the Algorithm 2. The novelty level of each scientific article set was determined based on the standard cloud with the highest similarity. The descriptive statistical analysis of the novelty levels for ICLR 2023 and Cell 2021 is presented in Table 4.

4.5.2. Correlation analysis

To demonstrate the consistency of the MNSA-ITMCM with the baselines and its ability to measure different types of novelty in scientific articles, we conducted a correlation test. Due to the sparse distribution of novelty labels in the Cell 2021 dataset and the significant fluctuations in both the originality index and Wang's novelty index, the Spearman correlation analysis method was required. The advantage of using Spearman correlation analysis lies in its ability to handle non-normally distributed data, outliers, and non-linear relationships. The resulting Spearman correlation coefficient, ranging from -1 to 1, could be used to measure the monotonic relationship between variables.

Additionally, Bornmann et al. (2019) mentioned that the recommendations of scientific articles on the Faculty Opinions platform are influenced by authority, particularly reflected in citation counts. Therefore, to validate whether citation counts have an impact on the novelty labels in the Faculty Opinions platform, we utilized the Semantic Scholar API (Kinney et al., 2023) to retrieve the citation count for the Cell 2021 papers. The data was collected until September 11, 2023. The Spearman correlation between the different methods and the novelty labels in the Cell 2021 dataset is depicted in Fig. 5.

Here are the key conclusions that can be drawn from observing Fig. 5:

- MNSA-ITMCM shows significant positive correlations with all four types of novelty papers in the Cell2021 dataset. Specifically, *TECHNICAL_ADVANCE* ($r = 0.268$, $p < 0.001$), *NOVEL_DRUG_TARGET* ($r = 0.215$, $p < 0.001$), *NEW_FINDING* ($r = 0.465$, $p < 0.001$), and *HYPOTHESIS* ($r = 0.318$, $p < 0.001$). The Originality index is also significantly positively correlated with *NOVEL_DRUG_TARGET* ($r = 0.149$, $p < 0.001$) and *NEW_FINDING* ($r = 0.190$, $p < 0.001$), but not the other two novelty types. The Wang's novelty is significantly positively correlated with *NOVEL_DRUG_TARGET* ($r = 0.177$, $p < 0.001$), NEW_FINDING ($r = 0.305$, $p < 0.001$), and *HYPOTHESIS* ($r = 0.209$, $p < 0.001$), but not with *TECHNICAL_ADVANCE*. This suggests that the MNSA-ITMCM can effectively identify all types of novel papers in the Cell 2021 dataset.

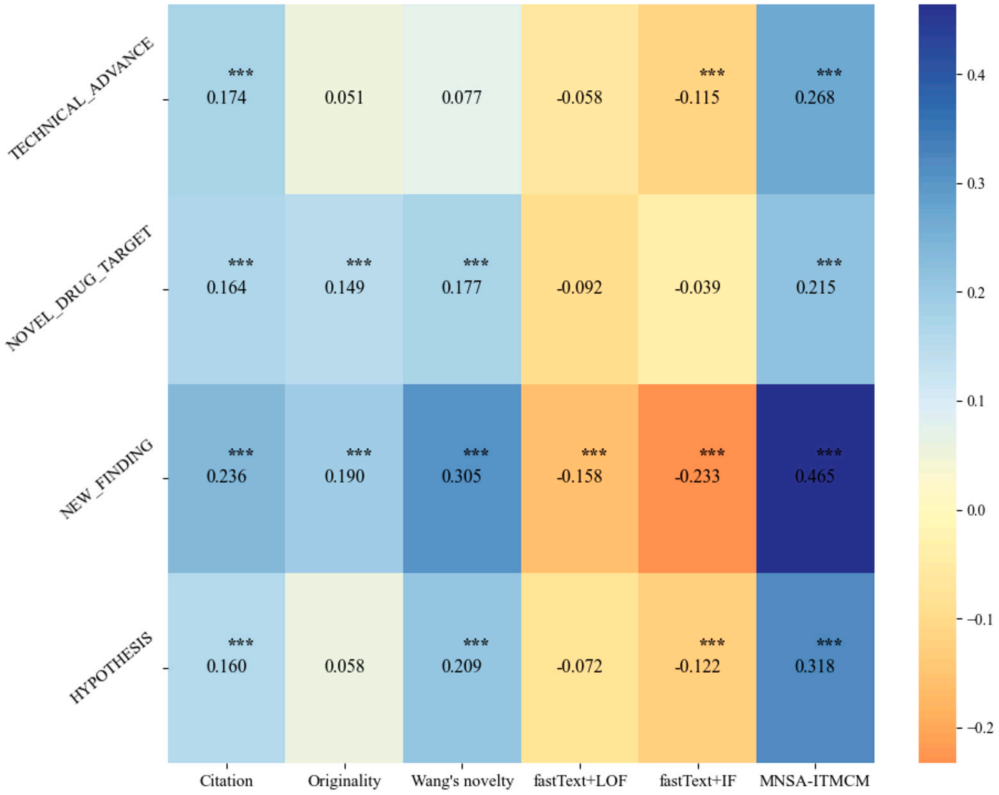


Fig. 5. The Spearman correlation between different methods and the novelty labels in the Cell 2021 dataset. "***" indicates a significant correlation at the 0.001 level (two-tailed).

- The *NEW_FINDING* type of novelty papers has the strongest correlations with multiple approaches. Compared to the other novelty types, the correlation coefficients for *NEW_FINDING* are the highest across multiple approaches. This indicates that when papers present original data, models, or methods, their novelty is more prominent and can be better captured by various approaches.
- The novelty of papers in the Cell 2021 dataset is significantly influenced by their authority. Novelty and authority are not equivalent, but the novelty labels in Cell 2021 come from recommendations by researchers at Faculty Opinions, which inevitably reflects the influence of authority, as evidenced by citation counts. Novel papers can attract widespread attention, which also brings them more citations. However, it should be noted that subsequent citations require time to accumulate, so citation-based measures cannot effectively evaluate the novelty of papers currently under review or recently published.
- The fastText+LOF/IF method produces results that are the opposite of other approaches. This method clusters similar scientific papers in vector space and identifies outliers as novel papers. However, the results exhibit significant negative correlations with some novelty labels. This suggests that during the fastText model training, certain research areas may have been prominent "hotspots", and words associated with these areas had higher frequencies. Consequently, when identifying outliers, paper titles containing these high-frequency words were classified as more similar. Nevertheless, the presence of these high-frequency words in the title does not necessarily indicate a lack of novelty. Rather, these papers may have made new contributions in important, active research areas, as evidenced by their significant positive correlation with citation counts. In contrast, the MNSA-ITMCM method, based on the MMR algorithm, constructs topic combinations that balance similarity and diversity. This allows it to capture multiple important, active research topics related to the paper. Highly novel papers often integrate knowledge from multiple "hotspots" in a novel way, making contributions that are not only novel, but also more impactful.

4.5.3. Prediction error analysis

On the ICLR 2023 dataset, we compared the accuracy of three methods, MNSA-ITMCM, fastText+LOF, and fastText+IF, in predicting the novelty levels of papers. Due to the lack of citation data for rejected papers in the ICLR 2023 dataset, we did not consider the Originality index or Wang's novelty. To facilitate comparison, we divided the continuous novelty scores obtained from fastText+LOF and fastText+IF into three ordinal novelty levels (0, 1, 2) based on the proportions of papers in the top 25%, 25-50%, and 50-100% novelty ranges. The distribution of prediction errors for the three methods on the ICLR 2023 dataset is shown in Fig. 6.

The results suggest that MNSA-ITMCM provides the most accurate predictions of paper novelty in the ICLR 2023 dataset. When the prediction error was at 0, MNSA-ITMCM correctly predicted 46.71% of the total articles, outperforming fastText+LOF (38.04%) and fastText+IF (39.17%). Furthermore, when the prediction error was within ±1, the accuracy rates improved to 83.83% for MNSA-ITMCM, 76.42% for fastText+LOF, and 75.66% for fastText+IF. However, these methods still face some challenges on the ICLR

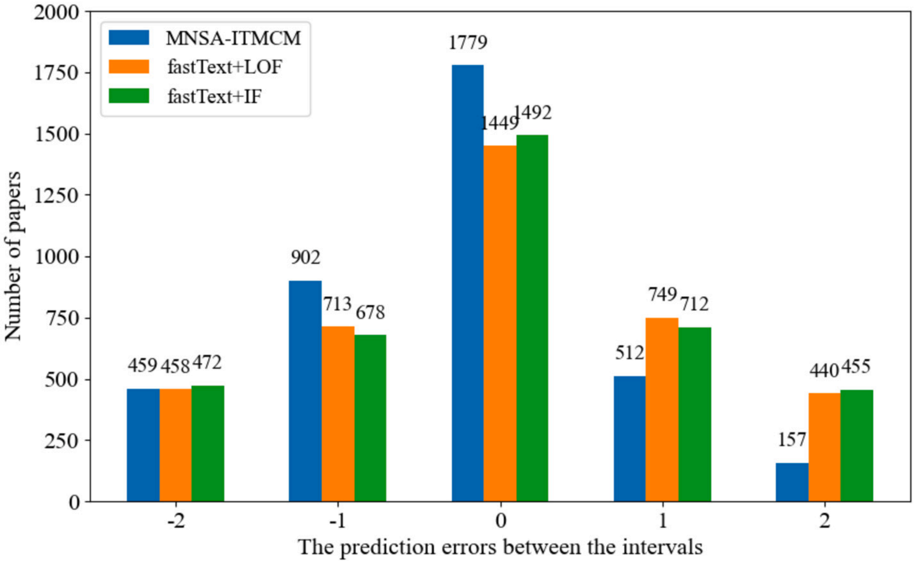


Fig. 6. The prediction errors between the intervals obtained from the three methods and the gold standard intervals.

Table 5
Novelty and topic combination of sample papers in the ICLR 2023 dataset.

Novelty level	Title	Cloud (E_x, E_n, H_e)	Topic combination
2	Learning Fair Graph Representations via Automated Data Augmentations	(0.459, 0.180, 0.088)	13(graph, networks, neural); 25(fairness, fair, bias); 7(3d, point, pose); 10(graph, neural, network); 39(ai, artificial, human)
2	Improving Out-of-distribution Generalization with Indirection Representations	(0.460, 0.096, 0.039)	11(domain, shot, learning); 16(logic, automata, semantics); 24(anomaly, detection, intrusion); 0(speech, language, translation); 22(explainable, explanations, interpretable)
1	Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning	(0.327, 0.196, 0.095)	1(reinforcement, control, learning); 4(covid, social, media); 14(games, bandits, bandit); 40(spiking, neuromorphic, neural); 2(method, equations, finite)
1	Self-Supervised Category-Level Articulated Object Pose Estimation with Part-Level SE(3) Equivariance	(0.379, 0.186, 0.083)	7(3d, point, pose); 25(fairness, fair, bias); 26(action, recognition, video); 35(re, person, identification); 38(eeg, brain, using)
0	Towards Identification of Microaggressions in real-life and Scripted conversations	(0.251, 0.132, 0.037)	0(speech, language, translation); 19(blockchain, smart, iot); 26(action, recognition, video); 4(covid, social, media); 24(anomaly, detection, intrusion)
0	Hazard Gradient Penalty for Survival Analysis	(0.229, 0.117, 0.058)	31(optimization, objective, algorithm); 4(covid, social, media); 2(method, equations, finite); 26(action recognition, video, temporal); 14(games, bandits, bandit)

2023 dataset, which may be because the research topics of conference papers are more focused, and the differences in novelty are more subtle compared to journal papers. Nevertheless, using only the paper titles, the MNSA-ITMCM, fastText+LOF, and fastText+IF methods were able to achieve relatively accurate novelty predictions, providing a promising direction for further exploration using more comprehensive textual content. This not only helps to improve the efficiency and quality of paper reviews but also offers valuable insights for the application of novelty prediction techniques in scientific article evaluation.

4.5.4. Case study

We randomly selected 2 papers from each novelty level of the ICLR 2023 and Cell 2021 datasets, resulting in a total of 12 case study papers, as shown in Tables 5 and 6. The analysis of the sample results indicates that the higher the novelty of the papers, the higher the E_x value (the average similarity between the paper and the topic combinations). Generally, the more novel the topic combinations, the more novel the papers. However, the average similarity we measured reflects the similarity between the papers and these topic combinations, rather than the novelty of the topic combinations themselves. After screening through the MMR algorithm, the topic combinations possess both similarity to the papers and internal diversity. When the papers have a high similarity to such topic combinations, it suggests that the papers have absorbed knowledge from a broader research field and the reorganized knowledge has depth and breadth, making contributions that are not only novel but also significant.

On the other hand, we found that multiple topics repeatedly appear in papers of different novelty levels, such as topics 24, 0, 26 in Table 5 and topics 9, 100, 74 in Table 6. This highlights that the novelty of a paper does not depend solely on specific topics, but rather on the interconnected reorganization of various subjects. For instance, the paper titled “Learning Fair Graph Representations via Automated Data Augmentations,” explores important research themes including graph representation learning, fairness, and the

Table 6
Novelty and topic combination of sample papers in the Cell 2021 dataset.

Novelty level	Title	Cloud (E_x, E_n, H_e)	Topic combination
2	NeuroPAL: A Multicolor Atlas for Whole-Brain Neuronal Identification in C. elegans	(0.379, 0.229, 0.104)	8(elegans, the elegans, lin); 9(memory, synaptic, term); 100(history, population, ancient); 45(neural, brain, motor); 74(sars, cov, covid 19)
2	Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses	(0.458, 0.184, 0.026)	74(sars, cov, covid 19); 100(history, population, ancient); 63(influenza, virus, hemagglutinin); 52(cardiac heart, cardiomyopathy, myocardial); 82(herpes, herpes simplex virus, simplex)
1	Insular cortex neurons encode and retrieve specific immune responses	(0.172, 0.022, 0.008)	4(cell, cell receptor, mhc); 9(memory, synaptic, term); 16(gut, microbiota, intestinal); 57(interferon, ifn, human interferon); 45(neural, brain, motor)
1	Kr-h1 maintains distinct caste-specific neurotranscriptomes in response to socially regulated hormones	(0.222, 0.072, 0.034)	9(memory, synaptic, term); 101(erythroid, differentiation, friend); 100(history, population, ancient); 51(hormone, receptor, estrogen); 16(gut, microbiota, intestinal)
0	Sensational channels	(0.083 0.083, 0.031)	23(channel, ca2, calcium sodium); 100(history, population, ancient); 74(sars, cov, covid 19); 9(memory, synaptic, term); 59(leukemia, murine, virus)
0	Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps	(0.101, 0.101, 0.068)	3(cancer, tumor, breast); 16(gut, microbiota, intestinal); 86(polyoma, polyoma virus, virus); 40(p53, tumor, mdm2); 52(cardiac heart, cardiomyopathy, myocardial)

societal aspects of artificial intelligence. Similarly, some topics that are actually irrelevant to the papers also appear in the results, such as topic 7. This may be because “graph” and “3d point cloud” are common data structure terms, and their co-occurrence probability in the training corpus is relatively high, leading to the incorrect identification of this topic.

5. Discussion and implications

With the proliferation of scientific publications at a rapid pace, it is increasingly difficult for researchers to find highly novel articles and technological directions with infinite potential in the era of big data (Dwivedi et al., 2023). One reason is that most of them are repeated work, few of which are of high quality and novelty. The other reason is that researchers do not have time to dive into every paper in detail to make their judgment. Our research can help researchers quickly identify critical research content and estimate an article’s degree of novelty. In addition, the proposed novelty measurement framework can potentially improve the efficiency and quality of the peer review, which in turn can accelerate the output of scientific research and create the conditions for technology to be translated into productivity. Nowadays, researchers spend much of their time reviewing articles for academic journals and conferences. However, the quality can hardly be assured since each reviewer has different criteria. As we know, for most top journals and conferences, the research novelty is essential in identifying high-quality articles (Dwivedi et al., 2022). Therefore, our method can generate a reference recommendation based on research content and novelty for reviewers.

As for librarians, especially academic librarians, the method presented in this paper can improve the quality and efficiency of their academic search service, an important step in the modernization of academic library technology. Junior researchers usually consult the subject librarian regarding the novelty or value when starting a new project. They may also ask the librarian to provide a list of novel studies related to their projects. With the help of our method, librarians can quickly locate these resources, and the result is more reliable and comprehensive. Moreover, most library retrieval systems and academic search engines are metadata-based and only search at the document level. The topic modeling developed in our research can accurately identify different types of research topics, which can serve as semantic information for improving academic retrieval. By embedding the algorithm in library retrieval systems and academic search engines, we can provide fine-grained knowledge retrieval.

Science evaluation institutions and funding agencies can also benefit from our research. The publication is one of the essential indicators that science evaluation institutions and funding agencies use to evaluate the research abilities of universities or scholars. Currently, metadata, such as the number of publications, citation count, and impact factor of a journal, are mainly used for the evaluation. However, all of the metadata-based evaluation methods have bias. For example, the cumulative nature of the number of citations over time, inactive but promising research directions, non-native English publications, authors from underdeveloped countries, and other factors all play a role in the impact of the research. Our proposed method is based on the paper’s content, which can eliminate bias in the metadata-based evaluations.

Policymakers must improve the efficiency and effectiveness of research policy implementation under limited resources (Abramo & D’Angelo, 2020; Xu et al., 2021). However, metrics based on metadata often have a time lag that may cause policymakers to miss potentially novel scientific topics. Therefore, we propose an ex-ante assessment method that helps policymakers anticipate emerging scientific fields and thus rationalize resource allocation.

6. Conclusion and future work

In this study, we combined topic modeling and cloud models to measure the novelty of scientific articles. The knowledge contained in the articles was treated as a combination of topics, which were quantitatively represented using the BERTopic topic modeling approach. We then generated corresponding clouds for each scientific paper and calculated their novelty levels by comparing the Hellinger distances between these clouds and a standard novelty cloud. Our empirical analysis compared our method with citation-based novelty metrics and the fastText+LOF/IF methods on two datasets: ICLR 2023 and Cell 2021. The results show that our method has a significant positive correlation with various types of novel papers, and can measure the novelty levels of papers reasonably accurately using only the title information. Furthermore, our method is applicable across different scientific disciplines and publication types, such as journal and conference papers. These findings not only contribute to improving the efficiency and quality of paper reviews, but also provide valuable insights for the application of novelty prediction techniques in scientific paper evaluation.

This study also has some limitations. First, we focused solely on the paper titles, which may not fully capture the knowledge content of the papers. Additionally, some authors may use metaphors in their titles, which could introduce biases in our results. Second, we require an appropriate number of historical papers to build the topic model, and the performance of topic modeling is closely related to the quality of the corpus. Finally, we have tested our method on small-scale datasets in specific domains. To ensure broader applicability, it is necessary to validate our approach using larger-scale datasets covering a wider range of academic disciplines.

In future research, further optimization will be pursued in the following areas. (1) Exploring the use of more comprehensive paper content, beyond just the titles, to better capture the knowledge and novelty of the papers. (2) Expanding the topic model corpus to include an appropriate number of historical papers, and studying the impact of corpus size and quality on the performance of the novelty measurement approach. (3) Validating the proposed method on larger and more diverse datasets covering a wider range of academic disciplines.

CRediT authorship contribution statement

Zhongyi Wang: Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Haoxuan Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jiangping Chen:** Writing – review & editing, Methodology, Conceptualization. **Haihua Chen:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Conceptualization.

Appendix A

A.1. Topic selection using MMR

Algorithm 1 Topic Selection using MMR.

```
Require: Candidate topics $T1, T2, T3, ..., Tn$ related to scientific articles $P1$;
    Similarity $(P1, T1)$, $(P1, T2)$, $(P1, T3)$, ..., $(P1, Tn)$ between $P1$ and each topic;
    Similarity $(T1, T2)$, $(T1, T3)$, ..., $(Ti, Tj)$ between each topic
Ensure: First $n$ selected candidate topics
Initialize an empty list $SelectedTopics$
Set $SelectedTopics[0]$ as the topic with the highest similarity to $P1$
Remove $SelectedTopics[0]$ from the candidate topic set
for $i$ from 1 to $n-1$ do
    $maxSim \leftarrow -\infty$
    $minAvgSim \leftarrow \infty$
    $selectedTopic \leftarrow$ None
    for each $topic$ in the candidate topic set do
        $similarityToP1 \leftarrow$ Calculate similarity between $P1$ and $topic$
        $minSimilarityToSelected \leftarrow \infty$
        for each $selected$ in $SelectedTopics$ do
            $similarityToSelected \leftarrow$ Calculate similarity between $topic$ and $selected$
            if $similarityToSelected < minSimilarityToSelected$ then
                $minSimilarityToSelected \leftarrow similarityToSelected$
            end if
        end for
        if $similarityToP1 > maxSim$ and $minSimilarityToSelected < minAvgSim$ then
            $maxSim \leftarrow similarityToP1$
            $minAvgSim \leftarrow minSimilarityToSelected$
            $selectedTopic \leftarrow topic$
        end if
    end for
    Add $selectedTopic$ to $SelectedTopics$
    Remove $selectedTopic$ from the candidate topic set
end for
return $SelectedTopics$
```

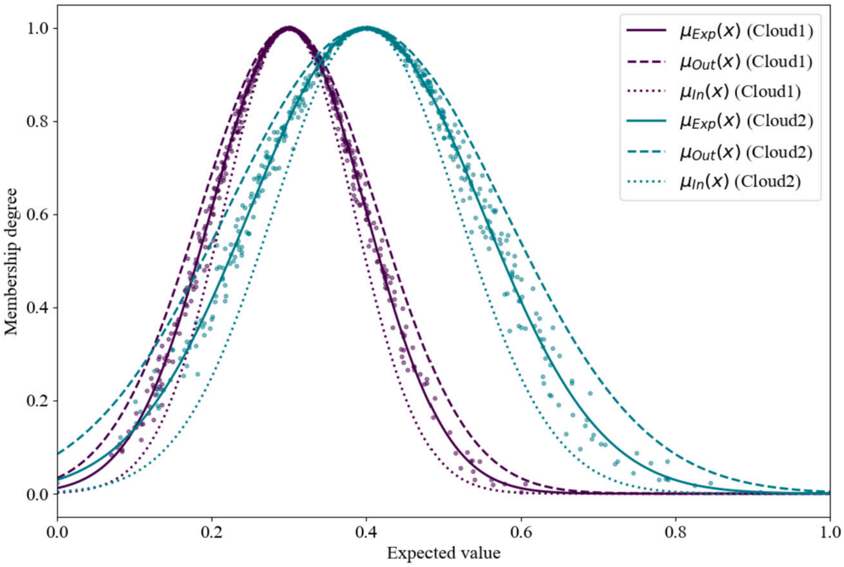


Fig. 7. Schematic of two normal clouds and their characteristic curves.

A.2. Calculation formulas for the novelty of scientific articles based on cloud similarity

The Hellinger distance between two discrete probability distributions P and Q is defined as:

$$H(P,Q)=\sqrt{1-\sum_{i=1}^{n}\sqrt{p_i q_i}} \tag{7}$$

Where p_i and q_i are the probabilities of the i-th outcome in the distributions P and Q, respectively. n is the number of possible outcomes.

The normal cloud model can be characterized by three characteristic curves that describe its overall distribution: the outer envelope curve $\mu_{\mathrm{Out}}(x)$, the inner envelope curve $\mu_{\mathrm{In}}(x)$, and the expected value curve $\mu_{\mathrm{Exp}}(x)$. The formulas for these three curves are as follows:

The outer envelope curve:

$$\mu_{\mathrm{Out}}(x)=\exp\left(-\frac{(x-E_x)^2}{2(E_n+3H_e)^2}\right) \tag{8}$$

The expected value curve:

$$\mu_{\mathrm{Exp}}(x)=\exp\left(-\frac{(x-E_x)^2}{2E_n^2}\right) \tag{9}$$

The inner envelope curve:

$$\mu_{\mathrm{In}}(x)=\exp\left(-\frac{(x-E_x)^2}{2(E_n-3H_e)^2}\right) \tag{10}$$

In Fig. 7, we present the normal cloud 1 (0.3, 0.1, 0.005) and normal cloud 2 (0.4, 0.15, 0.01), along with their corresponding three types of characteristic curves. The scattered points in the figure are set to 500 for each cloud, indicating that the default number of cloud drops for each cloud is 500.

Based on the Hellinger distance formula, we can calculate the distances between the outer envelope curves, inner envelope curves, and expected value curves of the two clouds, denoted as D_{exp}, D_{in}, and D_{out}, respectively. The formulas are as follows:

The distance between the expected value curves:

$$D_{\mathrm{exp}}=\sqrt{1-\sqrt{\frac{2E_{n1}E_{n2}}{E_{n1}^2+E_{n2}^2}}\exp\left(-\frac{(E_{x1}-E_{x2})^2}{4(E_{n1}^2+E_{n2}^2)}\right)} \tag{11}$$

The distance between the inner envelope curves:

$$D_{\mathrm{in}}=\sqrt{1-\sqrt{\frac{2\sigma_{\mathrm{in}1}\sigma_{\mathrm{in}2}}{\sigma_{\mathrm{in}1}^2+\sigma_{\mathrm{in}2}^2}}\exp\left(-\frac{(E_{x1}-E_{x2})^2}{4(\sigma_{\mathrm{in}1}^2+\sigma_{\mathrm{in}2}^2)}\right)} \tag{12}$$

Table 7
Spearman correlation analysis of topic extraction experiment results on the ICLR 2023 dataset.

	Result 1	Result 2	Result 3	Result 4	Result 5
Result 1	1***				
Result 2	0.882***	1***			
Result 3	0.907***	0.928***	1***		
Result 4	0.900***	0.879***	0.887***	1***	
Result 5	0.887***	0.906***	0.925***	0.901***	1***

Table 8
Spearman correlation analysis of topic extraction experiment results on the Cell 2021 dataset.

	Result 1	Result 2	Result 3	Result 4	Result 5
Result 1	1***				
Result 2	0.951***	1***			
Result 3	0.944***	0.969***	1***		
Result 4	0.954***	0.955***	0.957***	1***	
Result 5	0.941***	0.968***	0.961***	0.955***	1***

where $\sigma_{\text{in}1} = E_{n1} - 3H_{e1}$ and $\sigma_{\text{in}2} = E_{n2} - 3H_{e2}$.

The distance between the outer envelope curves:

$$D_{\text{out}} = \sqrt{1 - \sqrt{\frac{2\sigma_{\text{out}1}\sigma_{\text{out}2}}{\sigma_{\text{out}1}^2 + \sigma_{\text{out}2}^2}} \exp\left(-\frac{(E_{x1} - E_{x2})^2}{4(\sigma_{\text{out}1}^2 + \sigma_{\text{out}2}^2)}\right)} \tag{13}$$

where $\sigma_{\text{out}1} = E_{n1} + 3H_{e1}$ and $\sigma_{\text{out}2} = E_{n2} + 3H_{e2}$.

Therefore, the cloud similarity algorithm based on Hellinger distance is as follows.

Algorithm 2 Cloud Similarity Algorithm Based on Hellinger Distance.

```
Require: Cloud C1 $(E_{x1}, E_{n1}, H_{e1})$, Cloud C2 $(E_{x2}, E_{n2}, H_{e2})$
Calculate the distance $D_{\text{exp}}$ between the expectation curves of $C1$ and $C2$ using Equation (11).
Calculate the distance $D_{\text{in}}$ between the inner envelope curves of $C1$ and $C2$ using Equation (12).
Calculate the distance $D_{\text{out}}$ between the outer envelope curves of $C1$ and $C2$ using Equation (13).
Calculate the cloud similarity: $\text{Sim}(C1, C2) = 1 - \frac{1}{3}(D_{\text{exp}} + D_{\text{in}} + D_{\text{out}})$
return $\text{Sim}(C1, C2)$
```

A.3. Robustness testing of topic combinations

To assess the robustness of the topic combination extraction methodology, we conducted five repeated experiments on the ICLR 2023 and Cell 2021 datasets. Specifically, we performed topic modeling and topic combination selection without altering the underlying word embedding model. For each paper, we then computed the average topic probability across the identified topic combinations.

By examining the correlations of these average topic probabilities across the five experimental iterations, we evaluated whether the results exhibited significant differences, which would suggest the selected topic combinations were arbitrary (Table 7 and Table 8). The results revealed strong positive correlations, with coefficients exceeding 0.8, for both datasets. This indicates a high degree of consistency in the topic combination outcomes, with minimal variability despite differences in the topic modeling process. These findings demonstrate the robustness of the topic combination approach, producing stable and reproducible results across multiple experimental runs. The consistency in the topic combination analysis lends confidence to the reliability and validity of the insights gleaned from these datasets, underscoring the reproducibility and trustworthiness of the methodology.

References

Abramo, G., & D'Angelo, C. A. (2020). A novel methodology to assess the scientific standing of nations at field level. *Journal of Informetrics*, *14*, Article 100986. https://doi.org/10.1016/j.joi.2019.100986.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, *45*, 357. https://doi.org/10.1037/0022-3514.45.2.357.

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*, Article 104144. https://doi.org/10.1016/j.respol.2020.104144.

Bornmann, L., Tekles, A., Zhang, H. H., & Fred, Y. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, *13*, Article 100979. https://doi.org/10.1016/j.joi.2019.100979.

Bornmann, L., Wray, K. B., & Haunschild, R. (2020). Citation concept analysis (CCA): A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics*, *122*, 1051–1074. https://doi.org/10.1007/s11192-019-03326-2.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*, 2765–2783. https://doi.org/10.1287/mnsc.2015.2285.

Chen, L., & Fang, H. (2019). An automatic method for extracting innovative ideas based on the Scopus® database. *Knowledge Organization*, *46*. https://doi.org/10.5771/0943-7444-2019-3-171.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from arXiv:1810.04805.

Dirk, L. (1999). A measure of originality: The elements of science. *Social Studies of Science*, *29*, 765–776. https://doi.org/10.1177/030631299029005004.

Dwivedi, Y. K., Hughes, L., Cheung, C. M., Conboy, K., Duan, Y., Dubey, R., Janssen, M., Jones, P., Sigala, M., & Viglia, G. (2022). How to develop a quality research article and avoid a journal desk rejection. Retrieved from https://doi.org/10.1016/j.ijinfomgt.2021.102426.

Dwivedi, Y. K., Sharma, A., Rana, N. P., Giannakis, M., Goel, P., & Dutot, V. (2023). Evolution of artificial intelligence research in technological forecasting and social change: Research topics, trends, and future directions. *Technological Forecasting & Social Change*. https://doi.org/10.1016/j.techfore.2023.122579.

Fagerberg, J. (2004). Innovation: A guide to the literature. Retrieved from https://doi.org/10.1093/oxfordhb/9780199286805.003.0001.

Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, *49*, Article 104063. https://doi.org/10.1016/j.respol.2020.104063.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, *359*, Article eaao0185. https://doi.org/10.1126/science.aao2998.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, *80*, 875–908. https://doi.org/10.1177/0003122415601618.

Foster, J. G., Shi, F., & Evans, J. (2021). Surprise! Measuring novelty as expectation violation. Retrieved from https://doi.org/10.31235/osf.io/2t46f.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*, 791–817. https://doi.org/10.1287/mnsc.2015.2366.

Ghosal, T., Tiwary, P., Patton, R., & Stahl, C. (2021). Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies*, *2*, 1511–1528. https://doi.org/10.1162/qss_a_00170.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Retrieved from arXiv:2203.05794.

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the humanities and the social sciences? *American Sociological Review*, *69*, 190–212. https://doi.org/10.1177/000312240406900203.

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*, 9284–9291. https://doi.org/10.1073/pnas.1915378117.

Horbach, S. P., Oude Maatman, F. J., Halffman, W., & Hepkema, W. M. (2022). Automated citation recommendation tools encourage questionable citations. *Research Evaluation*, *31*, 321–325. https://doi.org/10.1093/reseval/rvac016.

Hou, J., Wang, D., & Li, J. (2022). A new method for measuring the originality of academic articles based on knowledge units in semantic networks. *Journal of Informetrics*, *16*, Article 101306. https://doi.org/10.1016/j.joi.2022.101306.

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics*, *113*, 633–650. https://doi.org/10.1007/s11192-017-2310-5.

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). Measuring the novelty of scientific publications: A fasttext and local outlier factor approach. *Journal of Informetrics*, *17*, Article 101450. https://doi.org/10.1016/j.joi.2023.101450.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., et al. (2023). The semantic scholar open data platform. Retrieved from arXiv:2301.10140. https://doi.org/10.48550/arXiv.2301.10140.

Kunnath, S. N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, *2*, 1170–1215. https://doi.org/10.1162/qss_a_00159.

Lee, Y. N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*, 684–697. https://doi.org/10.1016/J.RESPOL.2014.10.007.

Leibel, C., & Bornmann, L. (2024). What do we know about the disruption index in scientometrics? An overview of the literature. *Scientometrics*, *129*, 601–639. https://doi.org/10.1007/s11192-023-04873-5.

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2017). The relative influences of government funding and international collaboration on citation impact. *The Journal of the Association for Information Science and Technology*, *70*, 198–201. https://doi.org/10.1002/asi.24109.

Li, X., Peng, S., & Du, J. (2021). Towards medical knowmetrics: Representing and computing medical knowledge using semantic predications as the knowledge unit and the uncertainty as the knowledge context. *Scientometrics*, *126*, 6225–6251. https://doi.org/10.1007/s11192-021-03880-8.

Liang, Z., Mao, J., & Li, G. (2022). Bias against scientific novelty: A prepublication perspective. *The Journal of the Association for Information Science and Technology*, *74*, 114–199. https://doi.org/10.1002/asi.24725.

Lin, J., Song, J., Zhou, Z., Chen, Y., & Shi, X. (2023). Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion*, *98*, Article 101830. https://doi.org/10.1016/j.inffus.2023.101830.

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, *16*, Article 101282. https://doi.org/10.1016/j.joi.2022.101282.

Matsumoto, K., Shibayama, S., Kang, B., & Igami, M. (2020). *A validation study of knowledge combinatorial novelty*. Working Paper.

Matsumoto, K., Shibayama, S., Kang, B., & Igami, M. (2021). Introducing a novelty indicator for scientific research: Validating the knowledge-based combinatorial approach. *Scientometrics*, *126*, 6891–6915. https://doi.org/10.1007/s11192-021-04049-z.

Min, C., Bu, Y., & Sun, J. (2021). Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting & Social Change*, *164*, Article 120502. https://doi.org/10.1016/j.techfore.2020.120502.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 9.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing, association for computational linguistics*. Retrieved from http://arxiv.org/abs/1908.10084.

Rogers, M., & Rogers, M. (1998). *The definition and measurement of innovation, Vol. 98*. Parkville, VIC: Melbourne Institute of Applied Economic and Social Research.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*, 92–96. https://doi.org/10.1080/10400419.2012.650092.

Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management*, *57*, Article 102168. https://doi.org/10.1016/j.ipm.2019.102168.

Shibayama, S., & Wang, J. (2020). Measuring originality in science. *Scientometrics*, *122*, 409–427. https://doi.org/10.1007/s11192-019-03263-0.

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1728–1736). Association for Computational Linguistics (Online).

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology*, *5*, 19–50. https://doi.org/10.1080/10438599700000006.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*, 468–472. https://doi.org/10.1126/science.1240474.

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*, 707–723. https://doi.org/10.2139/ssrn.2382485.

Wang, H. (2024). A content-based novelty measure for scholarly publications: A proof of concept. Retrieved from arXiv:2401.03642. https://doi.org/10.48550/arXiv.2401.03642.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*, 1416–1436. https://doi.org/10.2139/ssrn.2710572.

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2022a). Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *The Journal of the Association for Information Science and Technology*, *74*, 150–167. https://doi.org/10.1002/asi.24719.

Wang, Z., Peng, S., Chen, J., Kapasule, A. G., & Chen, H. (2023). Detecting interdisciplinary semantic drift for knowledge organization based on normal cloud model. *Journal of King Saud University: Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2023.101569.

Wang, Z., Qiao, X., Chen, J., Li, L., Zhang, H., Ding, J., & Chen, H. (2024a). Exploring and evaluating the index for interdisciplinary breakthrough innovation detection. *Electronic Library*. https://doi.org/10.1108/EL-06-2023-0141.

Wang, Z., Wang, K., Liu, J., Huang, J., & Chen, H. (2022b). Measuring the innovation of method knowledge elements in scientific literature. *Scientometrics*, *127*, 2803–2827. https://doi.org/10.1007/s11192-022-04350-5.

Wang, Z., Zhang, H., Chen, H., Feng, Y., & Ding, J. (2024b). Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network. *Journal of King Saud University: Computer and Information Sciences*, *36*, Article 102119. https://doi.org/10.1016/j.jksuci.2024.102119.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*, 378–382. https://doi.org/10.1038/s41586-019-0941-9.

Xu, C., & Wang, K. (2023). *The similarity measurement of normal cloud concept based on Hellinger distance and expectation curve with entropy* (pp. 554–563).

Xu, H., Winnink, J. J., Yue, Z., Zhang, H., & Pang, H. (2020). Multidimensional scientometric indicators for the detection of emerging research topics. *Technological Forecasting & Social Change*, *120490*. https://doi.org/10.1016/j.techfore.2020.120490.

Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021). A topic models based framework for detecting and forecasting emerging technologies. *Technological Forecasting & Social Change*, *162*, Article 120366. https://doi.org/10.1016/j.techfore.2020.120366.

Yan, Y., Tian, S., & Zhang, J. (2020). The impact of a paper's new combinations and new components on its citation. *Scientometrics*, *122*, 895–913. https://doi.org/10.1007/s11192-019-03314-6.

Yao, M., Wei, Y., & Wang, H. (2023). Promoting research by reducing uncertainty in academic writing: A large-scale diachronic case study on hedging in science research articles across 25 years. *Scientometrics*, 1–18. https://doi.org/10.1007/s11192-023-04759-6.

Ziman, J. (2003). Emerging out of nature into history: The plurality of the sciences. *Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences*, *361*, 1617–1633. https://doi.org/10.1098/rsta.2003.1233.