Extraction of Sentences Describing Originality from Conclusion in Academic Papers

Bolin Hua, YoungKug Shin

Department of Information Management, Peking University, China

Abstract. Citation analysis-based strategies such as SCI, impact factor, and h-index reveals the influence of scientific papers, but it is difficult to demonstrate their originality. With the advancement of text mining technology and deep learning algorithms, it is feasible to extract the segment that illustrate originality (hereafter "originality points") from the paper and compare it with the originality points in previous literatures so as to detect the originality of a certain focal paper. The extraction of originality points in the paper is the first step in judging the originality of the paper. On the basis of summarizing the writing rules of the conclusion part of the literature, this paper summarizes the expression of sentences describing originality(SDO) of the papers in the conclusion and forms a vocabulary of guiding words for SDO of the papers, and then uses the rules to identify and extract SDO of the papers. In the experiment, we download the full text of papers on artificial intelligence from arXiv for the experiment, and the recognition accuracy and recall rate are 83.3% and 72.2%, respectively.

Keywords: academic literature, originality point recognition, originality feature words, knowledge extraction, originality evaluation

1 Introduction

For decades, scientometricians have proposed many sophisticated measurements to characterize the impact of scientific publications, such as the number of citations of a specific publication(Bornmann 2008) and the impact factor of the journal in which the paper is published(Garfield 1955). Yet, it is oftentimes difficult to reflect the originality and innovation of publications. Despite the fact that later science of science researchers employed citing relations to estimate these (e.g., Uzzi et al., 2013; Wang et al., 2020), current practice mainly relies on peer review.

Text mining techniques can be employed for evaluating the originality of a paper, which requires much less time and human effort compared to peer reviewing. The judgment of the originality of a paper includes subjective and objective reviews. Subjective reviews may come from the authors themselves (i.e., self-evaluation) or other scholars: The former is embodied in the description of originality and research conclusion of a paper, while the latter is mainly distributed in the citing content of citations. According to whether the cited literature appears in the reference or the main body of the citing literature, Ding and colleagues (2013) defined the "count one" and "count x" indices. He (2010) presented a prototype system CiteSeerX which aims to build a context-aware

citation recommendation system to recommend a set of citations for a paper with high quality.

Although measurements such as the number of citations, impact factor, and h-index have been introduced to reflect the influence/popularity of research papers, it is difficult to reflect the originality. To detect the originality of a research work and a paper, the current practice mainly relies on peer review. Peer reviews are subjective, and it is difficult to handle the evaluation task for a considerable number of scientific papers. While citation content analyses have been proposed to address this issue, most existing practices have purely focused on the motivation and sentiment of citations instead of the detection of the originality of a paper.

In the current paper, we address this gap and aim at developing methods for the automatic identification of SDO of a paper ("originality points") in scientific publications. Extraction of originality in a paper is the first step in judging the originality of the paper. This paper uses the full-text data of arXiv for the experiment, and studies the recognition and extraction of SDO of the papers in the conclusion part of scientific publications,

2 Related work

The expression of originality in academic literatures is diverse, and originality may appear in various parts of the research in different forms. Therefore, it is necessary to identify and extract SDO in academic literatures. The current methods of extracting information about originality in academic literatures can be divided into rule-based methods and machine learning-based methods.

2.1 Rule-based methods

The core idea of rule-based methods is to analyze the language features of the originality point, to select the feature items of the sentence for extraction, or to specify some rules for extraction. Kirschner (2015) presents the results of an annotation study that focused on the fine-grained analysis of argumentation structures in scientific publications by specifying four types of binary argumentative relations between sentences. Zhang et al. (2011) proposed a method of extracting sentence-level originality in the field of scientific and technological literature based on the relationship between domain-wise vocabulary and the ontology. Wen (2019) proposed a method of semantic recognition and classification. Specifically, he divided the scientific and technological abstracts into 6 categories according to syntax and semantic functions. Then he performed statistical analysis of the distributions of categories and sentence positions, sentence types, and sentence semantic positions. Li (2005) proposed an approach of originality detection based on the identification of sentence-level patterns. Zhang (2011) addressed the problem of multilingual sentence categorization and originality mining.

2.2 Machine learning-based methods

With the substantial increase in computing power and the rapid expansion of data scale, it has become possible to use deep learning methods in the big data environment to expand the semantics of text features and calculate the similarity of content. The computational efficiency has also been significantly improved, and scholars adopted deep learning methods for originality detection research. Markou (2003) reviewed various neural network methods (such as MLP, ART, RBF) that can be used for novel information detection based on the theoretical level. Kim et al (2018) presented a network-based method to detect the originality of a research paper. An autoencoder neural network is used as the originality detection model. Among the constructed networks, keyword-level graph features exhibit the best performance using regression analysis as the metric. Safdera (2020) proposed a set of methods to automatically identify and extract algorithmic pseudo-codes and the sentences that convey related algorithmic metadata using a set of machine-learning techniques.

These studies promote the innovative extraction and evaluation of papers, but there are still some shortcomings, such as:

1. Machine learning-based methods often need some labeled training data, but there is no corresponding dataset about originality of the papers;
2. Rule-based methods aim at a small amount of data, and how to design a method to process a large amount of data is quite challenging;
3. Most existing studies focused on the abstract of the paper, with only a few exploring the full text of scientific publications.

To tackle these problems, we design a method to extract SDO of the papers from the conclusion section, which combines rules and statistics. This method finds some features through statistical analysis of the description of originality points and then transforms these features into regular expressions to reduce the trouble of large-scale annotation data required by machine learning.

3 Methodology

3.1 Research framework

Our technical framework for the extraction of SDO of the papers in academic literatures includes three modules, namely data preparation, text processing, and extraction of SDO of the papers, as shown in Figure 1.

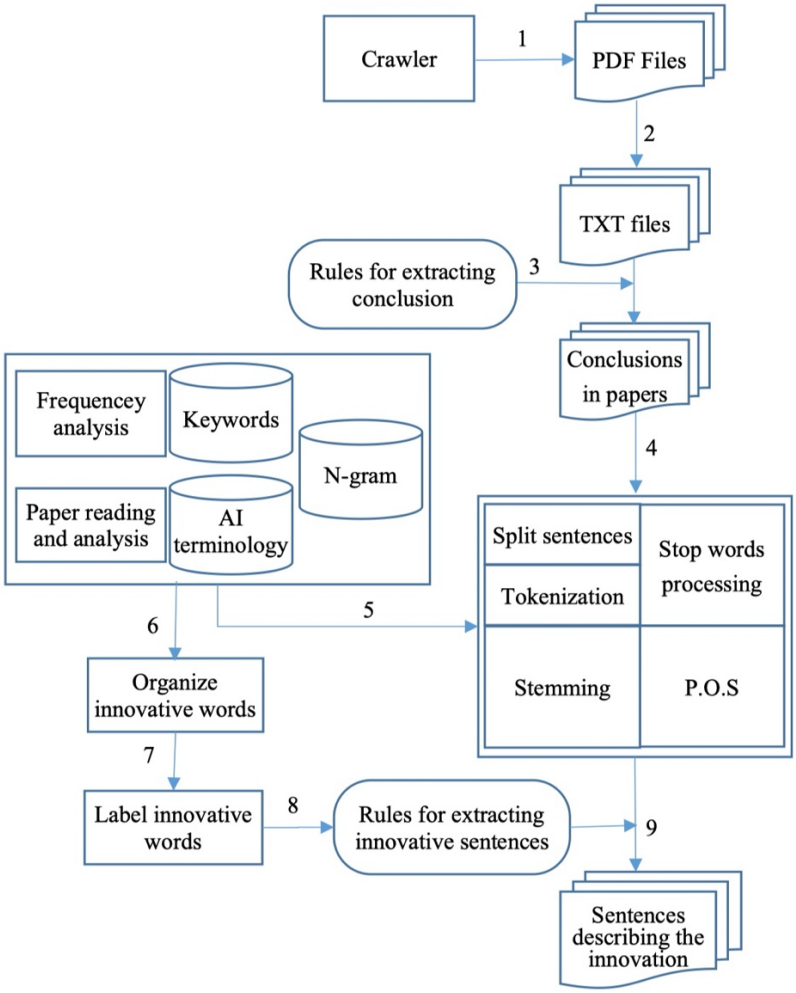


Fig. 1. Flow of SDO of the paper recognition.

The main processing part includes the following steps:

1. Using a web crawler, we obtained the full-text information of the papers from arXiv.
2. Then, we converted the format of the papers from PDF to TXT.
3. We summarized the characteristics of the conclusions in the literatures to extract them.
4. The conclusion section was split into sentences according to the characteristics of the text or the "full stop" character.
5. We processed the sentence set, such as word segmentation, stemming, stop words, part-of-speech tagging, and synonym merging.

6. In the module to recognize SDO of the paper, we collected and organized the words that comprise the SDO of the paper through literature research, word frequency analysis, domain dictionary, literature keyword collection.
7. We labeled the originality-related words and serialized the sentence in the conclusion section according to the originality guide vocabulary.
8. According to the result of sentence serialization, extracting rules of SDO of the paper were constructed and realized by regular expressions.
9. To extract the sentence describing originality from the conclusion of the paper by using rules.

Among them, the first and second steps belong to data acquisition module, step 3, 4 and 5 constitute data preprocessing module, and step 6, 7, 8 and 9 belong to extraction module.

3.2 Data collection and processing

3.2.1 Document format conversion and preprocessing

In order to extract SDO, it is necessary to convert the documents formatted as PDF into the TXT format. In practice, we adopt the pdfminer3k library in Python, an open-source package that converts PDF files into manageable TXT or Microsoft Word documents. When a PDF is parsed into a corresponding TXT document, people oftentimes encounter some issues, such as the lack of paragraph marks, the disappearance of the first-line indentation, and the forced disconnection of words. Therefore, the comprehensive application of line breaks, punctuation marks, hyphenation symbols, and sentence length was used to identify the paragraphs of TXT.

After extracting the conclusion section from the academic literatures, this paper used spaCy natural language processing software package for word segmentation, part-of-speech tagging, stemming, and stop words processing. To improve the accuracy of word segmentation, this paper introduces a keyword list, a domain glossary before word segmentation and uses Bi-gram and Tri-gram methods to identify phrases in the literature.

3.2.2 Extracting conclusions from academic literature

This paper recognized the conclusion or summary based on the chapter title, then split texts into sentences according to the length of the sentence and punctuation. After this, we divided sentences into words using professional dictionaries, keyword vocabulary, N-gram, and other methods for word segmentation, and finally generated a dataset in sentence units. The structure and function of most academic texts can be identified by chapter titles. For example, "Introduction" and "Introduction and Motivation" can be directly judged as the introduction; "Related Work" and "Context of this Research" can be directly judged as "related research". Due to the different expressions of the conclusion section in the literature, this paper manually screened the chapter titles of the experimental data and finally derive the characteristic vocabulary of the conclusion chapter titles in Table 1.

Table 1. Characteristic vocabulary of the title of the conclusion chapter

Chapter	Chapter Title Featured Vocabulary	Chapter End Featured Vocabulary
Conclusion	Conclusion, conclusions, discussion, summary, future, perspective, limitations, outlook, work, directions, results, concluding, remarks, suggestions, recommendations, comments, discussions	Acknowledgement[s], acknowledge, reference[s], \n\n

According to the above-mentioned starting feature vocabulary and ending feature vocabulary, the conclusion chapter extraction rules were constructed. We extract experimental data with these rules and finally obtained 18,563 conclusions. Then 17,653 conclusions were finally screened out through manual inspection.

3.3 SDO of the paper extraction

3.3.1 Constructing a dictionary of originative guiding words

Originality of academic literatures is mainly reflected in two aspects: characteristic words (guiding words) and common expressions. Aiming at the linguistic characteristics and style of scientific literature, the use of rule-based extraction methods could accurately identify the "knowledge claims" in the papers. Approximate 95% of originalities in papers are guided by characteristic words (Wen, 2014). Therefore, this paper combines the previous research results, domain keywords, domain terminology database, and word frequency statistical analysis to obtain the vocabulary list through manual screening and preliminary screening of originalities point feature guiding words. We then use WordNet to expand synonyms and finally select originative feature word sets.

The main basis for selecting the guiding words of originality in this paper comes from the usefulness, originality, enlightenment, scientificity and other elements described in the definition of scientific originality by some scholars. This article referred to the research results of Dahl (2009), Trine (2008), Parkinson (2011): We selected originative linguistic feature guiding words and divided them into the following types: referring to the author, referring to the article, iconic verb, iconic noun, and iconic adjectives. Since the subject terms of the field reveal the research focus of the field, the content of originality was closely related to the research subject. Given these, when constructing the originality guide vocabulary in this article, the field glossary and the keywords of the literature were introduced as the subject terms of the literature collection. In addition, the word frequency of the text in the conclusion part showed that most of the originative guiding words were distributed in the high-frequency range. Therefore, this article will compute word frequency on nouns and verbs and filter out originative feature words to construct an originative guiding vocabulary table.

After initially identifying the originative feature words, we use WordNet to expand synonymous word as the final selection of originative feature guiding words in this

article. According to their linguistic features, the originative feature guiding words are divided into 6 categories. The finally constructed originality point feature guiding words are shown in Table 2.

Table 2. guiding words of originality features

Type	Marking symbol	Word examples
Refers to authors	RF	I, We, Our
Refers to papers	TP	[In this|this|our|the] [paper|article|study|work|]
Verb	VB	Use, show, propose, provide, present, improve, observe, describe, investigate, prove, define, obtain, represent, design, aim, address, find, analyze, illustrate, conduct, appear, try, drive, and so on
Nouns	NN	problem, method, approach, work, result, performance, experiment, finding, insight, notion, and so on
Adjective	AD	new, novel, unused, caused, resulting, considered, known, observed, predicted, and so on
Keywords/subject terms	TW	algorithm, data, information, framework, knowledge, Acoustic, Bayesian network, beam search, CNN, RNN, LSTM, ontology, optimization, cluster, bi-lstm, classifier, crf, dnn, deep q-learning, embedding, robotic, transfer learning, recommender system, and so on

3.3.2 Identification of SDO of the papers

Recognition rules are constructed based on the relationship between domain thesaurus and ontology, and the method of the redundancy based on the overlapping degree of subject words is used to filter the candidate set of originality points (Zhang and Le, 2014). The vocabulary in the sentence is labeled according to the labeling symbols in Table 2, and then the labeling symbols in the sentence are separated come out and form a sequence of labeling symbols separated by spaces (according to the example in Figure 2, the labeling sequence in the sentence is: TP VB TW TW NN TW).

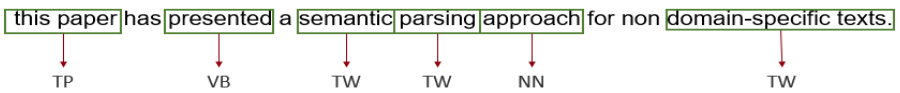


Fig. 2. An example of labeling originative feature guiding words

We comprehensively consider the labeling sequence and structure of SDO of the papers and consider the positions of different types of clue words and the combination of different clue words when constructing rules. We also set limited matching for some rules. Finally, the rules for writing regular expressions are as follows:

((RF)|(NN)|(TW)|(AD)|(TP)){0,3}(RF)((TP)|(AD)|(TW)|(NN)){0,3}(VB)((AD){0,1}(TW){0,6}(NN){0,2}){0,3}

4 Evaluation and Results

4.1 Experiment Data

This article used a web crawler to obtain publications in the field of "artificial intelligence" under CS (computer science) on arXiv which were labelled “CS. AI”. These experimental data were used for extraction of SDO from the papers. We collected basic information about the papers including title, author, publication time, URL (document PDF location). After that, the requests library was used to parse the URL to finally obtain 22,213 academic papers in PDF format. Figure 2 shows the annual distribution of literatures in the field of artificial intelligence.

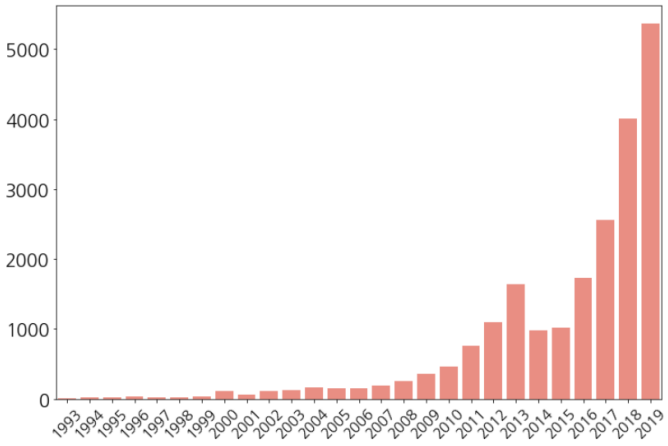


Fig. 3. Number of documents issued per year from 1993 to 2019

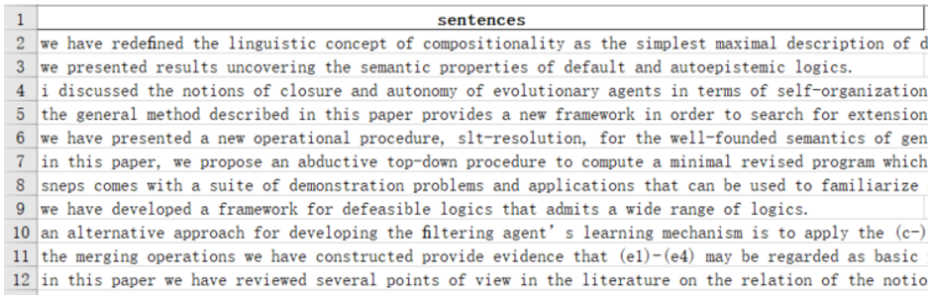
4.2 Experimental results

This paper selects sentences from the conclusion chapters of randomly chosen 200 papers from the collected documents for manual annotation and obtains 346 SDO of the papers out of 1,227 sentences. In order to test the performance of the SDO of the paper recognition rules constructed in this paper, the accuracy and recall rates in information retrieval are used to verify the recognition results. The results are shown in Table 3.

Table 3. Results of rules on experimental data

	Accuracy	Recall	F1-score
Rules	0.833	0.722	0.698

According to Table 3, the originality point identification rules constructed in this paper have an accuracy rate of 83.3% in the conclusion section. SDO of the paper were recognized from experimental dataset with the recognition rules. A total of 14,234 sentences that match the rules are used as input for originality objects and topic mining. Part of the extracted results is shown in Figure 4.



	sentences
2	we have redefined the linguistic concept of compositionality as the simplest maximal description of d
3	we presented results uncovering the semantic properties of default and autoepistemic logics.
4	i discussed the notions of closure and autonomy of evolutionary agents in terms of self-organization
5	the general method described in this paper provides a new framework in order to search for extension
6	we have presented a new operational procedure, slt-resolution, for the well-founded semantics of gen
7	in this paper, we propose an abductive top-down procedure to compute a minimal revised program which
8	sneps comes with a suite of demonstration problems and applications that can be used to familiarize
9	we have developed a framework for defeasible logics that admits a wide range of logics.
10	an alternative approach for developing the filtering agent's learning mechanism is to apply the (c-)
11	the merging operations we have constructed provide evidence that (e1)-(e4) may be regarded as basic
12	in this paper we have reviewed several points of view in the literature on the relation of the notio

Fig. 4. Results of innovative sentence extraction in conclusion chapter

A qualitative analysis of the content of SDO of the papers is carried out by observing approximately 100 papers selected randomly, and the commonly used SDO of the papers in the conclusion chapter are summarized. A part of the results is shown in Table 4.

Table 4. Examples of SDO of the papers in the conclusion section.

Type	SDO of the paper patterns
New method class	[This paper|We] [propose|introduce|present|develop] a [new|first|novel] [model|solution|algorithm|method……] that ……
Methodology	We [presented|introduced] a methodology [for|to] …… We demonstrated …… methodology for ……
concept/ Viewpoint	In this paper we have [redefined|defined|proposed] the [notion|concept] of …… The [concept|notion] of …… is defined ……
Proof class	[This paper|We] [demonstrate|prove] that
Problem class	We considered …… problem ……
Application class	In this paper, we [shown|studied] the application of ……

From Table 4, we can see there are mainly seven kinds of descriptions about the originality of the paper, which are new method class, methodology class, concept class, viewpoint class, proof class, problem class, and application class. Among them, the first two describe are method originality, the latter two refer to application originality, and

the middle three belong to theory originality. We will make a detailed analysis of the theme, object and the pattern of sentence describing originality through the following papers.

4.3 Analysis of experimental results

This article extracted SDO in the conclusion chapter according to the innovative sentence recognition rules. High-frequency innovation objects and subject terms will be analyzed.

4.3.1 Analysis of core nodes in SDO

According to the results of the dependency syntax analysis, the core nodes (ROOT) in the innovative sentences are counted, and the proportion of the core nodes is shown in Table 5. In the SDO from the conclusion chapter, the words present, propose, and introduce respectively represent 23%, 22%, 11%, which amounts to more than 50% of the entire core node, while the remaining core words account for a relatively small proportion. This shows that in the conclusion chapter, researchers mainly use these words to summarize or introduce the main points and originality of the article.

Table 5. Proportion of core nodes in SDO

Core node in SDO	Proportion	Core node in SDO	Proportion
present	23.937844	develop	2.871755
propose	22.324941	provide	2.517703
introduce	11.801731	demonstrate	1.848938
show	3.599528	investigate	1.809599
study	3.127459	consider	1.298190
describe	3.029111	……	……

4.3.2 Analysis of Innovation Objects

This paper takes the direct object of the core node as the innovation object of SDO, and counts the frequency of the innovation object. The proportion of the innovation object is shown in Table 6.

In the results of the proportion of innovation objects, approach, method, way and other words about method have a relatively high proportion. It can be seen that the innovation of methods in the field of artificial intelligence is the key research direction. However, the innovation of methodology only accounts for 0.6% of the total, which shows that there is relatively little research on methodological innovation.

Table 6. Proportion of innovation objects in SDO

Innovation Objects in SDO	Proportion	Innovation Objects in SDO	Proportion
approach	8.477577	methodology	0.609756
framework	7.317073	concept	0.590087
method	6.805665	solution	0.550747
model	4.956727	notion	0.531078
algorithm	4.661684	scheme	0.531078
problem	4.346971	dataset	0.49173
system	2.163651	mechanism	0.472069
architecture	1.593234	application	0.432730
technique	1.258851	strategy	0.432730
network	1.121164	information	0.393391
performance	0.983478	complexity	0.373721
way	0.668765	……	……

In addition to methodological innovation, according to the proportion of word frequency, framework, model, algorithm, system, problem, architecture, network, concept and dataset are key innovation objects in the field of artificial intelligence.

In short, the above experimental results show that in the field of artificial intelligence, the main focus is on method innovation (approach, method, etc.), as well as specific application innovation (model, algorithm, application, etc.), while there is less innovation in the theory itself (methodology, idea, theory, etc.).

5 Discussion and Conclusion

By employing arXiv scientific publications, this paper constructs recognition rule about SDO of the paper based on originative guiding words aiming to recognize the sentence-level originality point of academic literature. Implementing SDO of the paper recognition experiments on the literature on artificial intelligence topics on arXiv, we find that the proposed method is quite effective to extract SDO of the paper from papers. After obtaining SDO of the papers, people can evaluate papers by comparing the SDO of the papers in the different papers.

The results of this paper show that the method constructed is feasible and effective for sentence-level originality point identification and mining methods. Yet, as a research-in-progress paper, there are still several limitations, and we are going to implement the following related studies in the future:

1. Although the SDO of the paper recognition rules constructed in this article are effective in recognition of SDO of the papers, the formulation and maintenance of the rules cannot cover all papers. Therefore, in order to improve the accuracy of the recognition of SDO of the papers, the sequence will be marked in the follow-up as training data, machine learning methods are used to convert the extracted questions into classification questions.
2. The current paper only identifies sentences that reflect the originative content of the thesis. Next, the SDO of the papers will be analyzed and excavated, and the originative objects, topics, and specific methods will be extracted.
3. In this paper, we only extract the innovative description in the papers' conclusion section. We will extract information describing the originality in the research objectives, related works, and methodology from papers in the following research.

6 Acknowledgements

This work was supported in part by The National Social Science Foundation of China (Number: 17BTQ066).

References

1. Amplayo R.K., Hong S.L., Song M.(2018). Network-based approach to detect novelty of scholarly literature. Information Sciences, 422, 542-557.
2. Dahl T. 2009. The Linguistic Representation of Rhetorical Function: A Study of How Economists Present Their Knowledge Claims. Written Communication, 29(4), 370-391.
3. Ding Y, Liu X.Z., Guo C., Cronin B.(2013). The distribution of references across texts:Some implications for citation anlysis. Journal of information, 7(3), 583-592.
4. Garfield E(1955). Citation indexes for science. A new dimension in documentation through association of ideas. Science, 3159(122), 108-111.
5. He Q., Pei J., Kifer D., et al.(2010). Context-aware citation recommendation. In Proceedings of the 19th international conference on World wide web (WWW '10). Association for Computing Machinery, New York, NY, USA, 421–430.
6. Kirschner C., Eckle-Kohler J., Gurevych I.(2015). Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications. 2nd Workshop on Argumentation Mining (ARG-MINING 2015) Denver, Colorado, USA, June 4.
7. Markou M., Singh S. (2003). Novelty detection: A review-part 2: Neural network based approaches. Signal Processing, 83, (12), 2499–2521.
8. Parkinson J. (2011). The Discussion section as argument: The language used to prove knowledge claims. English for Specific Purposes, 30(3), 164-175.
9. Safdera I., Hassana S.U., Visvizi A. Noraset T.， Tuarob S. (2020). Deep Learning-based Extraction of Algorithmic Metadata in Full-Text Scholarly Documents. Information Processing and Management, 57, 102269

10. Shibayama, S., Wang, J. (2020). Measuring originality in science. Scientometrics, 122, 409-427.
11. Trine D (2008). Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics. Journal of Pragmatics, 40(7),1184-1201
12. Uzzi B., Mukherjee S., Stringer M., Jones B. (2013). A typical Combinations and Scientific Impact, Science, 342, 468-472.
13. Wen H.(2019). Semantic Recognition and Classification Method of originality Points in Scientific and Technological. Journal of The China Society for Scientific and Technical Information, 38(3), 249-256.
14. Wen Y.K., Wu G.Y(2014). Dynamic Mining of Fragmented Scientific Research originality Points, Digital Library Forum,7,25-32.
15. Zhang F., & Le X.Q. (2014). Research on originality Points Extraction from Scientific Research Paper Based on Field Thesaurus. New Technology of Library and Information Service, (9), 15-21.
16. Zhang Y., Tsai F.S., & Kwee A.T.(2011). Multilingual sentence categorization and novelty mining. Information Processing and Management, 47, 667–675.