Construction and Implementation of an Automated Method for Research Design Fingerprint Identification in Scientific Papers: Postprint

Authors: Qian Li, Zhang Xiaolin, Wang Qian

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] Automatically identifying and extracting research design fingerprints from scientific papers can provide important methodological and operational support for researchers in project design, evaluation of research method effectiveness, diagnosis of research process issues, and identification and assessment of research results. [Method/Process] Based on the conceptual model of research design fingerprints in scientific papers, this study proposes a hybrid machine learning method based on multi-rule patterns, designs and implements a fingerprint recognition algorithm, and analyzes and verifies the feasibility and effectiveness of the recognition algorithm using journal literature data from the data mining domain as a case study. [Results/Conclusion] Except for research data and research trends, the acceptance rates for the accuracy of other research design fingerprint identifications were generally above 80%; the acceptance rates for coverage, except for research tools and research data, were generally above 80%.

Full Text

Preamble

Building and Implementing an Automatic Identification Method for Research Design Fingerprints in Scientific Papers

Qian Li[1], *Zhang Xiaolin*[1], *Wang Qian*[2]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005

Abstract

[**Purpose/Significance**] Automatically identifying and extracting research design fingerprints from scientific papers can provide researchers with crucial methodological and operational support for project design, evaluation of research method validity, diagnosis of research process issues, and assessment of research results. [**Method/Process**] Based on the conceptual model of research design fingerprints in scientific papers, this study proposes a hybrid multi-rule pattern machine learning approach, designs and implements a fingerprint identification algorithm, and validates the feasibility and effectiveness of the algorithm using journal literature data from the data mining domain as a case study. [**Result/Conclusion**] Except for research data and research trends, the recognition accuracy acceptance rate for other research design fingerprints basically reaches over 80%, and the coverage acceptance rate basically reaches over 80% except for research tools and research data.

Keywords: research design fingerprint; semantic annotation; knowledge extraction; machine learning
Classification Number: TP391; G25
DOI: 10.13266/j.issn.0252-3116.2018.02.018

Scientific papers, as important strategic resources for science and technology development, document research knowledge clues including scientific truth verification processes, experimental observation results, and research conclusions. The research designs involved in papers—including research questions, methods, processes, tools, and parameter settings for related methods and technologies—provide valuable methodological and operational foundations for subsequent researchers. These become critical bases for researchers' project design, research method validity assessment, research process problem diagnosis, and research result identification and evaluation. Researchers hope to have tools that can effectively answer questions such as "Who used what methods to solve this problem?" and "Which methods and their technical and parameter settings can better solve this problem?" However, in an environment where the number of scientific papers is rapidly increasing, current data mining theories and technologies that focus on keywords or abstract-level knowledge discovery are insufficient to meet these needs. Therefore, it is necessary and urgent to design and implement a theoretical and technical system for automatically identifying and extracting research design fingerprints from papers.

Building upon previously completed conceptual and identification models for research design fingerprints in scientific papers [1], this study further explores the methods and implementation for automatic identification. The research structure is as follows: Define the connotation and characteristics of research design fingerprints; Review related methods for automatic identification; Propose and design an automatic identification method to address the problem; Validate the feasibility and effectiveness through experimental data.

2. Connotation and Characteristics of Research Design Fingerprints

A research design fingerprint refers to the important knowledge units in a scientific paper that uniquely represent and describe various research stages and entities of a scientific research design, including nine fingerprint types: research hypothesis, research purpose, research background, research method, research data, research tool, research result, research conclusion, and research trend. It possesses four main characteristics: Knowledge uniqueness—these important knowledge units have unique research design fingerprint features under the premise of complying with research ethics, with core dimensions including author and article title; Research thinking—the fingerprint can concisely reveal "the overall design idea of a scientific research design"; Knowledge structure—it can structurally describe "scientific research methods, processes, and results," extracting, organizing, and associating important knowledge; Backbone network—a scientific paper can use research design fingerprints to visually depict "backbone knowledge in scientific research" similar to a network backbone diagram.

3. Related Research Methods

Research design fingerprints, as special semantic labels embedded in scientific paper content, involve methods similar to semantic annotation and knowledge extraction, using computer programs to automatically identify and extract knowledge components such as research methods.

3.1 Ontology-Based Knowledge Engineering Methods

In the construction of scientific papers represented by military documents, Guo Zhongwei, Zhou Xianzhong, and Huang Zhitong constructed Schema libraries for various military documents, using rhetorical predicates on the Schema to extract corresponding knowledge and ultimately construct document content [22]. For general scientific papers, methods such as Varga et al.'s MnM [4], Handschuh et al.'s S-Cream [5], and the AKT project's Melita [6] have achieved certain results in their respective fields. Y.F. Guo proposed a minimally supervised learning method using discourse rhetoric and lexical features for medical literature review, which achieved only 29% accuracy and 50% recall for "research method" identification [10].

3.2 Rule-Based Pattern Matching Methods

H. Houngbo used rules to extract methods described in scientific papers, achieving 85% accuracy, but did not study other fingerprint types [14]. C.D. Manning used information extraction patterns to identify technical methods and classification topic phrases, but accuracy was only 20% [15]. D. Kiela [16] and Y.F. Guo [17-18] used topic attributes in scientific papers (including position,

tense, verbs, grammar) to identify research methods. J.E. Kohler used indicator words like "method," "analysis," "algorithm," "approach," and "mode" to identify research methods in abstracts [19]. Liu Yining and Zheng Yanning designed an academic definition extraction system using mixed pattern rules, grammatical rules, and word frequency statistics [20]. Ding Junjun and Zheng Yanning analyzed quantitative relationships and sentiment information in attribute descriptions of academic journals [21].

3.3 Collaborative Editing Methods

The SemLibEU project (2012) developed Pundit [23], enabling users to build structured data while annotating web pages, supporting group sharing and collaborative knowledge building. The Open University developed SWEET [24], a lightweight Web API for semantic annotation.

3.4 Grammar-Based Methods

S. Gupta proposed using syntactic dependency trees to annotate and extract technical knowledge points from scientific papers [25]. S. Bethard used linguistics to identify events and semantic types in question-answering systems with high accuracy [26]. The ReVerb semantic annotation system [27] introduced grammatical and lexical constraints for binary relations expressed by verbs, significantly improving both recall and accuracy compared to TextRunner and WOE. The AKSW research group at Leipzig University proposed the FOX framework [28], integrating linked data cloud platforms and NLP algorithms to extract RDF triples from free text.

Overall, existing methods have achieved certain effectiveness for semantic knowledge identification in specific research environments but remain highly domain-dependent and insufficiently adaptable to unsupervised learning environments. Without domain knowledge organization systems (KOS) and manually defined rules, they cannot be applied to research design fingerprint identification, and rule definition requires high professional expertise. Therefore, this study starts from the essence of scientific papers, following objective patterns such as writing guidelines, journal checklists, experimental checklists, and expression habits of research design fingerprints, to design the identification algorithm.

4. Design of Automatic Research Design Fingerprint Identification Method

Given that full-text scientific papers are unstructured text making fingerprint features difficult to identify, this study proposes using natural language processing (NLP) technologies for automatic parsing, knowledge reorganization, and structured representation. The core algorithm is based on knowledge object extraction and word feature extraction.

4.1 Automatic Discovery of Research Design Fingerprint Clues from Document Paragraphs

4.1.1 Clue Discovery Method Based on Knowledge Object Extraction Knowledge extraction [29] refers to identifying, discovering, and extracting concepts, types, facts, relationships, constraint rules, and problem-solving steps from digital resources. Using Stanford CoreNLP [30], this study discovers and extracts fingerprint feature clues including clue words and patterns through part-of-speech tagging, named entity recognition, parsing, grammatical analysis, coreference analysis, and guided pattern learning. The implementation method discovers clues from three aspects: term extraction, grammatical analysis, and fact extraction.

(1) Clue word discovery based on scientific terms. First, domain scientific term standard libraries are used to extract clue words at the sentence level. Second, a parser identifies and extracts free terms at the sentence level, combining part-of-speech rules to select free term blocks as candidate clue words following the principle of maximum semantic chunks of continuous parts of speech. Finally, a term similarity algorithm based on cosine similarity [31] calculates similarity between free terms and standard scientific terms (threshold = 0.8) for further standardization.

The cosine similarity formula is:

$$sim(x, y) = \frac{\sum_{i=1}^{m} x[i] \cdot y[i]}{\|x\|\|y\|} = \frac{dot(x, y)}{\|x\|\|y\|}$$

(2) Clue rule discovery based on sentence grammar. NLP techniques parse sentences for tokenization, part-of-speech tagging, and maximum semantic chunk extraction, followed by structured storage. Analysis of this syntactically structured data forms series of clue rules to assist fingerprint type identification. For example, a rule for identifying research method fingerprints: (JJ|NN|NNS|NNP|NNPS)+(method|approach|measure|…), where the semantic block (JJ|NN|NNS|NNP|NNPS) can be identified as a research method fingerprint.

(3) Clue pattern discovery based on facts. J. Bessin emphasized that fact extraction is a core component of big data analytics, aiming to identify important semantic descriptions and their relationships [32]. This study uses "action verbs forming facts" in scientific papers as an entry point to discover relationships between research design fingerprints and factual behaviors. If fingerprint feature words from the corpus successfully match the subject, object, action, and behavior of a fact, a fingerprint clue rule pattern is formed.

4.1.2 Clue Discovery Method Based on Feature Indicator Words Scientific paper descriptions of research design fingerprints follow certain expression habits and depend on context. This method uses three types of indicator words:

(1) Indicator noun-based clues. A noun term can identify a fingerprint type. For example, in "The finite projective geometry method was first applied…", the indicator "method" identifies "finite projective geometry" as a "research method" fingerprint.

(2) Indicator action word-based clues. An action word can identify a fingerprint type. In the same example, "applied" indicates the sentence's fingerprint type as "research method." Syntactic analysis shows "applied" is passive voice, confirming the knowledge object as a research method fingerprint.

(3) Indicator co-occurrence word-based clues. Co-occurring words can identify fingerprint types. The co-occurrence of "method" and "applied" strongly indicates a "research method" fingerprint, as scientific papers often describe methods as "apply a XXX method to solve XXX problem."

4.2 Sentence-Level Automatic Research Design Fingerprint Identification

4.2.1 Algorithm Design and Implementation The sentence-level algorithm comprehensively judges the most likely fingerprint type using semantic indicator rules, semantic action word rules, semantic co-occurrence rules, sentence location, rhetorical type, corpus rules, and context fingerprint types. The sentence fingerprint feature vector (SSV) includes 10 dimensions: SentenceID, Text, CoreTerms, CorpusWords, CorpusWordsType, SectionType, Location, Action, ActionType, ActionTense.

The core algorithm has two phases: Construction based on semantic indicator words using grammatical, definitional, and Be-verb rules; Construction based on demonstrative pronoun features using pronouns to identify current sentence fingerprint types and suggesting nearest context sentence fingerprint types.

4.2.2 Comprehensive Evaluation Method for Sentence-Level Fingerprint Types Voting is the primary method, where each rule type represents a voter with different weights (see Table 1). Each voter scores weight*0 (opposition) or weight*1 (support). The final score is the sum of all voters, sorted from high to low, with the highest identified as the most likely fingerprint type:

$$Sentence\_FP\_Score = 2{*}IndicatingWordsValue{+}1{*}ActionWordsValue{+}2{*}Co{-}occurrenceValue{+}0.5{*}Lo$$

4.3 Term-Level Automatic Research Design Fingerprint Identification

4.3.1 Algorithm Design and Implementation The term-level algorithm comprehensively judges fingerprint types from five aspects: corpus features, sentence fingerprint features, action word features, SOX attribute features, and rhetorical structure features. The term fingerprint feature vector (TSV) includes 9 dimensions: Term, isCorpus, CorpusType, ParagraphType, Location, SentenceFP, Role, Action, ActionTense.

The core algorithm: Calculate CorpusWordsValue (+2 if term is in corpus); Calculate SentenceFPValue (+1 if matches sentence fingerprint type); Calculate ActionWordsValue (+1 if matches action word fingerprint type); Calculate SOXValue (+0.5 if core vocabulary); Calculate ORB?Value (+0.5 if matches rhetorical structure).

4.3.2 Comprehensive Evaluation Method for Term-Level Fingerprint Types Similar to sentence-level but uses sentence fingerprint type as a parameter. The voting score algorithm is:

$$Term\_FP\_Score = 2*CorpusWordsValue+1*SentenceFPValue+1*ActionWordsValue+0.5*SOXValue+$$

5. Experiments and Results

5.1 Corpus and Experimental Data Preparation

5.1.1 Corpus Data Materials

The corpus includes: Professional terms from the "Twelfth Five-Year" science and technology support plan project—Scientific and Technological Knowledge Organization System (STKOS) in engineering artificial intelligence; Domain KOS using IEEE thesaurus [34,35] for data mining; Research design fingerprint indicator words from WordNet [36], VerbNet [37], computer science research paper corpora analysis [38], and journal publication guidelines. The experiment created 235 indicator terms.

5.1.2 Scientific Paper Full-Text Data Materials

Using Elsevier's "rich media" HTML format, 100 full-text scientific papers were manually downloaded using "Data Mining" as the search term (small sample for validation, with plans to expand later).

5.2 Experimental Results Evaluation and Analysis

The implementation effects are shown in Figures 5 [Figure 5: see original paper], 6 [Figure 6: see original paper], and 7 [Figure 7: see original paper]. To validate the algorithm's effectiveness on 9 fingerprint types, 10 experts evaluated 50 papers using comparative analysis. The results are shown in Table 2 , with acceptance calculated as: Acceptance = (Votes × Minimum Accuracy) / Σ(Votes × Minimum Accuracy). Expert acceptance analysis is shown in Table 3 .

Overall results show the proposed method meets experimental expectations: accuracy and coverage for 7 fingerprint types (research methods, etc.) exceed 80%, while the remaining two types mainly fall in the 70%-80% range.

Lower accuracy/coverage for research data, hypothesis, and trend fingerprints may be due to: Less explicit descriptive features compared to stronger features

like research methods and conclusions; Small dataset affecting annotation corpus and rule pattern quantity.

6. Conclusion and Outlook

This study proposes a multi-rule hybrid machine learning algorithm for automatic research design fingerprint identification. Empirical analysis shows significant effectiveness for most fingerprint types, particularly research methods and conclusions. Future work will: Analyze factors affecting identification and improve evaluation methods; Conduct broader demonstrations across science, engineering, agriculture, and medicine; Use deep learning to more comprehensively mine features for research data, hypothesis, and trend fingerprints; Apply the method to large-scale literature to build rich associations and researcher/institution fingerprint knowledge bases, enhancing knowledge discovery.

References

[1] Qian Li, Zhang Xiaolin, Wang Qian. Research on the description framework of research design fingerprint based on scientific literature [J]. Journal of Academic Libraries, 2015(1): 14-20.
[2] Girju R, Beamer B, Rozovskaya A, et al. A knowledge-rich approach to identifying semantic relations between nominals [J]. Information Processing & Management, 2010, 46(5): 589-610.
[3] Wang D, Liu X, Luo H, et al. A novel framework for semantic entity identification and relationship integration in large-scale text data [J]. Future Generation Computer Systems, 2016, 64(C): 198-210.
[4] Vargas-Vera M, Motta E, Domingue J, et al. MnM: Ontology-driven semi-automatic and automatic support for semantic markup [C]//International Conference on Knowledge Engineering and Knowledge Management. London: Springer-Verlag, 2002: 379-391.
[5] Handschuh S, Staab S, Ciravegna F. S-Cream—Semi-automatic CREAtion of metadata [C]//Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. London: Springer-Verlag, 2002: 358-372.
[6] Advanced Knowledge Technologies [EB/OL]. [2017-09-26]. http://www.iam.ecs.soton.ac.uk/projects/akt/.
[7] Ciravegna F, Dingli A, Petrelli D, et al. User-system cooperation in document annotation based on information extraction [C]//International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web. London: Springer-Verlag, 2002: 122-137.
[8] Dill S, Eiron N, Gibson D, et al. A case for automated large-scale semantic annotation [EB/OL]. [2016-10-20]. http://www.websemanticsjournal.org/index.php/ps/article/viewFile/30/28.
[9] Ciravegna F, Chapman S, Dingli A, et al. Learning to harvest information for the semantic web [C]//Proceedings of the 1st European Semantic Web Symposium. Greece: Heraklion, 2004: 312-326.
[10] Guo YF, Silins I, Stenius U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical

literature review [J]. Bioinformatics, 2013, 29(11): 1440-1447.
[11] Su Mu, Xiao Renbin. Research on dynamic knowledge extraction method based on sentence clustering recognition [J]. Journal of Computer Science, 2001, 24(5): 487-495.
[12] Xu Yong, Song Rou. A method for classifying knowledge points in encyclopedia entries based on HMM [J]. Computer Engineering and Applications, 2005, 41(4): 35-38.
[13] Soldatova LN, Liakata M. An ontology methodology and CISP-the proposed core information about scientific papers [EB/OL]. [2016-09-24]. https://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/ReportCISPshort.pdf.
[14] Houngbo H, Hospic E, Mercer R E. Method mention extraction from scientific research papers [C]//24th International Conference on Computational Linguistics. New York: Curran associates, 2012.
[15] Gupta S, Manning CD. Analyzing the dynamics of research by extracting key aspects of scientific papers [C]//Proceedings of 5th International Joint Conference on Natural Language Processing. New York: Curran associates, 2011: 1-9.
[16] Kiela D, Guo Y, Stenius U, et al. Unsupervised discovery of information structure in biomedical documents [J]. Bioinformatics, 2015, 31(7): 1084-1092.
[17] Guo YF, Silins I, Stenius U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review [J]. Bioinformatics, 2013, 29(11): 1440-1447.
[18] Guo YF, Reichart R, Korhonen A. Improved information structure analysis of scientific documents through discourse and lexical constraints [C]//Proceedings of NAACL-HLT. New York: Curran associates, 2013: 928-937.
[19] Eckle-Kohler J, Nghiem TD, Gurevych I. Automatically assigning research methods to journal articles in the domain of social sciences [J]. Proceedings of the American Society for Information Science and Technology, 2013, 50(1): 1-8.
[20] Liu Yining, Zheng Yanning, Hua Bolin. Implementation and experimental analysis of academic definition extraction system [J]. Information Studies: Theory & Application, 2011, 34(12): 15-19.
[21] Ding Junjun, Zheng Yanning, Hua Bolin. Rule-based academic concept attribute extraction [J]. Information Studies: Theory & Application, 2011, 34(12): 10-14, 33.
[22] Guo Zhongwei, Zhou Xianzhong, Huang Zhitong. Design of content planning in combat document automatic generation system [J]. Fire Control & Command Control, 2002, 27(4): 51-54.
[23] Pundit-Semantic annotation tool [EB/OL]. [2017-03-20]. http://thepund.it/.
[24] SWEET [EB/OL]. [2017-03-20]. http://sweet.kmi.open.ac.uk/.
[25] Gupta S, Manning CD. Identifying focus, techniques and domain of scientific papers [EB/OL]. [2017-03-20]. https://www.researchgate.net/publication/267232558_{{{Identifying}}{{Focus
[26] Bethard S, Martin JH. Identification of event mentions and their semantic class [C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney: Emnlp, 2006: 146-154.

[27] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction [C]//Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 1535-1545.
[28] FOX-Federated Knowledge Extraction Framework (AKSW) [EB/OL]. [2017-03-20]. http://aksw.org/Projects/FOX.html.
[29] Zhang Zhixiong, Wu Zhenxin, Liu Jianhua, et al. Analysis of current main technical methods for knowledge extraction [J]. New Technology of Library and Information Service, 2008, 24(8): 2-11.
[30] Manning CD, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit [C]//Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Maryland: Curran associates, 2014: 55-60.
[31] Lee D, Park J, Shim J, et al. An efficient similarity join algorithm with cosine similarity predicate [C]//International Conference on Database and Expert Systems Applications. Heidelberg: Springer, 2010: 422-436.
[32] Bessin J, Das A. Big data analytics federal business analytics [EB/OL]. [2017-03-20]. https://www.xerox.com/downloads/services/white-paper/big-data-analytics.pdf.
[33] Sun Tan, Liu Zheng. Construction ideas of knowledge organization system for foreign scientific and technological papers [J]. Library and Information, 2013(1): 2-7.
[34] IEEE Baidu Baike [EB/OL]. [2017-04-10]. http://www.baike.com/wiki/IEEE.
[35] IEEE{{{thesaurus}}{{2013}}} [EB/OL]. [2017-04-10]. https://www.ieee.org/documents/ieee{{{thesaurus
[36] About WordNet [EB/OL]. [2017-03-20]. http://wordnet.princeton.edu/.
[37] Martha Palmer [EB/OL]. [2017-03-20]. http://verbs.colorado.edu/~mpalmer/projects.html.
[38] Posteguillo S. The schematic structure of computer science research articles [J]. English for Specific Purposes, 1999, 18(2): 139-160.

Author Contributions:
Qian Li: Responsible for writing and revising the paper;
Zhang Xiaolin: Responsible for content design and review;
Wang Qian: Responsible for paper revision.

Note: The promotional text about the "Library Tour" winter course at the end of the original document has been omitted as it is not part of the academic paper.

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