ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00743

Advances and Trends in Semantic Enhancement of Scientific Papers: A Postprint

Authors: Song Ningyuan, Pei Lei, Wang Chunying

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

Abstract

[Purpose/Significance] As the scientific communication system shifts to electronic media, the traditional approaches to organizing and presenting scientific paper content have revealed numerous limitations. Semantic enhancement of scientific papers, which can innovate the organization and presentation of scientific paper content, serves as a key solution to these issues and has attracted attention from research institutions and academic publishers, yielding a series of theoretical and practical achievements. By organizing and summarizing these achievements and identifying their strengths and weaknesses, we can provide guidance for subsequent efforts to promote the further development of semantic enhancement of scientific papers. [Method/Process] Beginning with the concept of semantic enhancement, this study focuses on analyzing the core objectives, implementation paths, and key issues of semantic enhancement of scientific papers. Subsequently, it examines theoretical and practical achievements in semantic enhancement of both main text and paratext content in scientific papers, and presents a comparative analysis focusing on the three stages of the semantic enhancement pathway: semantic annotation, semantic organization, and visual presentation. [Results/Conclusion] The research further summarizes the features of current semantic enhancement of scientific papers and puts forward four recommendations for its future development and research.

Full Text

Preamble

The Survey and Trend Analysis of Semantic Enhancement for Scientific Papers

Song Ningyuan[1], Pei Lei[1], Wang Chunying[2]

[1] School of Information Management, Nanjing University, Nanjing 210023
[2] School of Information Management, Zhengzhou University, Zhengzhou 450001

Abstract

[**Purpose/Significance**] As scientific communication systems migrate to electronic media, the traditional content organization and presentation methods of scientific papers have introduced numerous drawbacks. Semantic enhancement of scientific papers can innovate the organization and presentation of scientific paper content and is key to solving these problems. It has received attention from research institutions and academic publishers, yielding a series of theoretical and practical achievements. Systematically reviewing and summarizing these achievements to identify their strengths and weaknesses can provide guidance for promoting the further development of semantic enhancement for scientific papers. [**Method/Process**] Starting from the concept of semantic enhancement, this paper focuses on analyzing the core objectives, implementation paths, and key issues of semantic enhancement for scientific papers. Subsequently, it reviews theoretical and practical achievements in semantic enhancement of both primary text and paratext content in scientific papers, and conducts a comparative analysis around three stages in the semantic enhancement path: semantic annotation, semantic organization, and visual presentation. [**Result/Conclusion**] The study further summarizes the characteristics of current semantic enhancement of scientific papers and proposes four suggestions for future development and research in this area.

Keywords: scientific papers, semantic enhancement, semantic annotation, semantic organization, visualization

Classification Number: G255
DOI: 10.13266/j.issn.0252-3116.2021.01.013

1. Introduction

Scientific papers serve as carriers of scientific communication with specific structures and formats. With the digital transformation of journals, electronic journals have become mainstream. As publication volumes increase and average reading time per article decreases, leveraging emerging digital technologies to innovate the content organization and presentation methods of scientific papers to improve readers' reading and communication efficiency has gradually attracted academic attention [1-2]. By employing semantic technologies such as ontologies, RDF, and linked data to accurately represent the semantic functions of scientific paper content, fine-grained knowledge fragments can be visualized, associated, and disseminated [3]. This series of technologies and methods has opened the path to semantic enhancement (Semantic Enhancement or Enrichment) for scientific papers.

Semantic enhancement is a value-added editorial processing activity aimed at

content that can increase the value of digital content assets. Currently, semantic enhancement of scientific papers has received widespread attention from academia and industry. D. Shotton et al. conducted a series of semantic publishing experiments [4], exploring methods and paths for semantic enhancement of papers. The Royal Society of Chemistry [5], Elsevier [6], *Nature* magazine, as well as Microsoft, Google, and some cultural heritage institutions have also carried out semantic enhancement experiments on academic publications and web resources. Through semantic enhancement, innovating the organization and presentation of scientific paper content can improve the utilization efficiency of scientific paper resources, excavate the potential value of scientific paper content, achieve interconnectivity between paper content and knowledge, and promote the transformation of scientific and technological information work toward intelligent services.

After decades of attempts and exploration, numerous research and practical achievements have been made in semantic enhancement for scientific papers. Systematically reviewing these achievements can help clarify the methods and implementation paths of semantic enhancement for scientific papers and identify future research directions, priorities, and trends, providing directional guidance for the reuse of scientific paper content and value.

2. Connotation, Requirements, and Implementation of Semantic Enhancement for Scientific Papers

2.1 Conceptual Development of Semantic Enhancement for Scientific Papers

Semantic enhancement is an emerging concept in information resource management and scientific paper publishing that accompanies the development of computer text processing technology and the semantic web, aiming to address the insufficient semantic revelation and encoding representation in existing electronic documents. Currently, scientific papers are mainly in HTML and PDF formats. Due to limitations in document encoding schemes, these documents generally lack semantic markup, making it difficult for search engines to understand the semantic features and functions of content fragments and elements in these documents. Consequently, readers also struggle to retrieve and utilize fine-grained fragments and knowledge points in documents, making semantic enhancement an unavoidable step in data resource transformation and semantic web construction.

In understanding the connotation of semantic enhancement, V. Damjanovic believes that semantic enhancement is related to various activities such as semantic retrieval, semantic organization, semantic annotation, and semantic analysis and knowledge discovery [7]. In the Europeana semantic enhancement framework, the basic stages of semantic enhancement are defined as analysis, association, and enhancement [8]. The IFLA LRM model defines five information tasks for users in a semantic enhancement environment: find, identify, select,

obtain, and explore [9]. The SURF Foundation report states that semantic enhancement of scientific papers primarily integrates research data, supplementary materials, data records, and published publications to extend and expand traditional paper content [10]. M. Hoogerwerf believes that semantic enhancement of scientific papers is object-based information integration, where objects refer to various multimedia elements and text blocks such as videos, user comments, and databases [11], which have significant relationships between them. L. Breure et al. believe that semantic enhancement of scientific papers should support both linear and non-linear reading under a complete semantic metadata system [12].

Based on the above perspectives, implementation stages, and focus areas from different studies, this paper argues that semantic enhancement for scientific papers aims primarily to improve user reading efficiency and knowledge acquisition effectiveness. It comprehensively utilizes various semantic and visualization technologies to conduct a series of structured, semantic, associative, and visual processing of scientific papers. The main stages of semantic enhancement include semantic annotation, semantic association, and visual presentation.

2.2 Core Objectives of Semantic Enhancement for Scientific Papers

The primary purpose of semantic enhancement for scientific papers is to fully reveal the potential knowledge contained within scientific papers, innovate the organization and presentation of scientific paper content, and improve users' reading efficiency and effectiveness. This involves constructing trustworthy, contextualized, associated, cognizable, predictable, and usable smart datasets through semantic enhancement, achieving the transformation and upgrading from traditional literature resources to smart data to fully mine the potential knowledge contained in scientific papers.

Around these core objectives, semantic enhancement for scientific papers has multiple application scenarios: knowledge discovery, semantic publishing, and strategic reading. In the field of knowledge discovery, rich semantic content data of scientific papers enables analysis from different perspectives, achieving advanced applications such as knowledge extraction, retrieval, and discovery. In semantic publishing, semantic enhancement enables the transition of publishing objects from article-level scientific papers to fine-grained statements. In strategic reading, revealing semantic features of scientific paper content at different granularities helps locate the most valuable information for users.

2.3 Implementation Methods and Key Issues

2.3.1 Implementation Methods Scientific papers are complex knowledge systems composed of large amounts of primary text and paratext content. Paratext mainly includes bibliographic information, abstracts, citations, and references, while primary text refers to the scientific paper content containing substantial knowledge. Paratext primarily serves to assist in understanding and

explaining the primary text. Semantic enhancement of scientific papers aims to semantically represent both primary text and paratext content, innovate content organization and presentation methods, and generate enhanced papers that improve user reading effectiveness. Therefore, the implementation path generally includes three stages: semantic annotation, semantic organization, and content visualization, as shown in Figure 1 [Figure 1: see original paper].

Figure 1. Implementation Methods for Semantic Enhancement of Scientific Papers

(1) **Semantic Annotation.** Semantic annotation refers to associating entities in scientific papers with concepts in knowledge organization tools such as ontologies and thesauri. It uses concepts, properties, and relationships defined in ontologies to reveal semantic features of scientific papers, generate semantically described content with semantic labels (semantic content), and achieve machine readability. Semantic annotation is a crucial process for transforming scientific papers from document-centric to entity-centric [13].

(2) **Semantic Organization.** Based on semantic annotation, semantic organization achieves association and organization of generated semantic content with semantic labels. This process involves designing organizational models, ontology interoperability, and ontology mapping, integrating multiple ontologies and metadata sets. The result is the generation of interconnected rich semantic scientific paper content datasets.

(3) **Content Visualization.** Content visualization comprehensively utilizes various computer vision technologies to graphically and multimedia-present scientific paper content datasets, generating enhanced papers suitable for users to improve content perception and promote knowledge acquisition efficiency.

2.3.2 Key Issues On the path of "semantic annotation - semantic organization - content visualization," the following key issues must be addressed to better achieve semantic enhancement of scientific papers:

(1) **Multi-dimensional Semantic Description of Primary Text Content.** Scientific papers contain large amounts of unstructured content data. Comprehensive representation of this data can reveal organizational patterns and basic structures, enabling the transition from document-level to fine-grained content-level, which is key to promoting further development of semantic enhancement.

(2) **Semantic Association, Organization, and Publication of Multi-source Data.** Semantic enhancement requires innovating scientific paper organization models, which necessitates fully associating and organizing multi-source data (bibliographic information, citation information, and content data) after semantic annotation, including designing semantic or-

ganization models and selecting organization and publication tools. The key lies in accurately describing and normatively defining the logical structure, semantic relationships, and citation relationships of scientific papers to construct applicable organizational models for different application scenarios.

(3) **Visual Presentation of Semantic Content Data.** Visual and interactive presentation of content data is the main way to improve users' content understanding efficiency. In addition to visualizing words and concepts, it is particularly necessary to consider how to accurately represent rich semantic content data such as logical structures and argumentation methods of scientific papers using graphics.

3. Analysis of Semantic Enhancement Paths for Scientific Papers

This section reviews different approaches to semantic enhancement for paratext and primary text content.

3.1 Semantic Enhancement of Paratext Content

3.1.1 Semantic Association and Organization of Bibliographic Information Bibliographic information of scientific papers includes article titles, author information, abstracts, keywords, project and funding information, etc., with clear formats that can be described using metadata sets such as Dublin Core. The main method for semantic enhancement of bibliographic information is designing bibliographic ontologies for semantic description and using multiple ontologies collaboratively to achieve semantic integration of multi-source information such as scientists, papers, conferences, and journals.

Semantic description of bibliographic information is represented by the Bibliographic Ontology Specification (BIBO) and FRBR-aligned Bibliographic Ontology (FaBiO) [14]. BIBO defines 69 elements, primarily focusing on document type definitions. FaBiO integrates classifications of works, expressions, manifestations, and items from the FRBR framework while including descriptions of creators and creation groups, forming an integrated ontology [15]. Based on semantic description of single-document bibliographic information, the VIVO ontology system [16] integrates BIBO, FOAF, DC and other ontologies and metadata sets, adding numerous semantic relationships to construct a semantic model of scientists' information exchange.

3.1.2 Semantic Enhancement of Abstracts Abstracts summarize the main content of scientific papers and contain rich information, thus having various enhancement methods. Yu Qichen et al. [17] summarized abstracts with different semantic enhancement approaches: structured abstracts add semantic elements (background, purpose, method, results, discussion, etc.) to single-paragraph abstracts to clarify structure and enrich semantics, helping

users quickly grasp key points; video abstracts and graphical abstracts use charts, audio-visual materials combined with text for multimedia and visual expression; structured digital abstracts focus on machine understandability and link to external knowledge bases through entity linking; highlight abstracts reveal the most important assertions and statements in papers with high intelligence value.

3.1.3 Semantic Description of Citation Functions Citation and reference information typically includes authors, titles, journals, and publishers of cited literature. Additionally, citation information links cited and citing documents, forming citation relationships that contain semantic attributes such as citation sentiment and context, which are also key concerns for semantic enhancement.

Current semantic enhancement of citations and references is mainly achieved through constructing relevant ontologies, with representative ones including the Citation Typing Ontology (CiTO) [18] and Citation Counting and Context Characterization Ontology (C4O) [19]. CiTO uses RDF to represent citation relationships and defines their semantic attributes. In CiTO, citation semantics are mainly defined by rhetorical relationships (citation sentiment: positive, neutral, negative) and factual relationships (functions such as citing data or methods). C4O primarily defines citation positions and contexts of the same reference in different documents and associates with Google Scholar to describe total citation counts. Both CiTO and C4O have strong extensibility and can be associated with FOAF ontology and Dublin Core metadata set to represent author information of cited literature, and can collaborate with Document Elements Ontology (DEO) and Document Component Ontology (DoCO) to represent fine-grained citation contexts.

3.1.4 Association and Publication of Paratext Content In terms of semantic organization and linked publication, paratext content semantic enhancement is relatively mature, forming a certain number and scale of open datasets and knowledge graphs.

In dataset construction and publication, the OpenCitations dataset is most representative. Built through crowdsourcing, it includes structured information of conference papers, book chapters, and journal articles, using CiTO, FaBiO and other ontologies for semantic description [20], and uses RDF language to link and open literature data indexed by CrossRef and ORCID.

Knowledge graphs are widely adopted by publishing and research institutions to associate and publish bibliographic and reference information. Springer Nature launched the SciGraph [21] project in 2015, achieving cross-modal semantic aggregation of multi-source heterogeneous data through data fusion, knowledge discovery, and content computing based on knowledge organization. Tsinghua University' s AMiner [22] scientific knowledge graph analyzes and mines scientific big data including scientific literature, experts, and academic activities

to provide knowledge services such as semantic search, analysis, and evaluation. Microsoft' s Academic Graph (MAG) [23] is a heterogeneous knowledge graph built through intelligent analysis of web academic entities and their relationships. Additionally, MAG and AMiner collaborated to build the Open Academic Graph (OAG) [24], achieving nearly 65 million linking relationships to support research on author collaboration networks and academic topic mining. Shanghai Jiao Tong University' s AceKG [25] academic knowledge graph contains over 100 million academic entities and 2.2 billion triples, providing rich attribute information for each entity to support diverse academic big data mining projects.

3.2 Semantic Enhancement of Primary Text Content

Primary text content includes concept entities (words, phrases), statements and propositions (sentences), content components (multiple sentences), and logical structures (formed by relationships between components). Current research has conducted theoretical and practical explorations of semantic enhancement at different levels and granularities.

3.2.1 Extraction and Representation of Conceptual Entities in Scientific Papers In conceptual entity extraction, comprehensive use of domain ontologies and natural language processing technologies including named entity recognition has achieved extraction and semantic representation of conceptual entities, such as mining and constructing micro-concept maps [26], extracting academic concept attributes [27], and extracting key terms [28]. In visual representation of conceptual entities, existing research and practice mainly present core concepts of scientific papers through tag clouds and tag trees.

3.2.2 Description and Linked Publication of Scientific Statements Scientific statements are the foundation of scientific paper content and the direct representation of conceptual entities' states and attributes. The most representative statement representation model is Nanopublication.

Nanopublication is a model of "the smallest publishable unit with scientific significance and machine readability" based on "scientific statements" [29]. This model contains core scientific statements and related contexts, facilitating knowledge processing at the scientific claim level, such as integration, querying, and reasoning of scientific claims. Generally, nanopublications consist of content and functional components. The content component is based on concept triples, treating each meaningful triple as a scientific statement. A scientific statement and its provenance information constitute the most basic nanopublication. Additionally, publication information (including attribution, integration time, citation status, etc.) and supporting information provide explanatory functions. Currently, the nanopublication model has been widely used in biomedical and digital humanities projects, forming a certain scale of nanopublication datasets.

3.2.3 Semantic Representation of Content Components and Logical Structures In addition to describing and enhancing statements as basic units, some research approaches scientific paper content from a discourse analysis perspective, proposing concepts for content components. According to different interpretation perspectives, they mainly focus on four directions:

(1) Rhetorical and Functional Components. Centering on the scientific investigation process and based on the Scientific Experiment Ontology (EXPO) and CISp, M. Liakata proposed the CoreSC (Core Scientific Concept) model [31], which classifies statements in scientific texts according to different stages of scientific experiments: hypothesis, motivation, goal, objective, background, method, experiment, model, observation, result, and conclusion. This model defines the scientific experimental process in detail but has insufficient semantic representation capability for large amounts of argumentative text. A. De Waard believes scientific papers are knowledge constructions around specific scientific goals and proposed the ABCDE model in 2006 [32]. This model describes scientific papers from five aspects: annotation, background, contribution, discussion, and entities, covering not only document content (background, contribution, discussion) but also metadata (annotation) and entity-level information (entities), though with coarse granularity and limited expressiveness. Focusing on document content modules, De Waard further proposed the Discourse Segment model, more finely revealing knowledge units in scientific papers, including seven categories: fact, hypothesis, goal, method, result, impact, and problem [33]. L. Zhang defined six information use tasks in scientific paper contexts based on users' functional needs during reading: learning background knowledge, referencing facts, referencing arguments, referencing methods, and following frontier research [34]. Additionally, L. Zhang et al. proposed a conceptual model containing 41 functional units by combining research space theory and genre analysis [34].

(2) Argumentative Structure. Semantic description of argumentative structure generally includes definitions of argumentative components and relationships. For argumentative components, S. Teufel proposed the Argumentative Zoning (AZ) model [35], which defines different content components in scientific texts as categories such as aim, contrast, basis, text, and background. Subsequently, Teufel expanded this theory by combining citation functions, author sentiment tendencies, and textual rhetorical functions to propose the more fine-grained Argument Zoning II framework [36], defining 14 different rhetorical components with new types including comparison (CoDI), pointing out flaws, contradictory views, support, and usage, making it more suitable for scientific papers. N. L. Green studied the representation of scientific paper structures in the biomedical domain [37], proposing an argumentation framework including hypothesis, conclusion, and background knowledge to represent argumentative structures [38-40].

For argumentative relationships, the more mature project is the Scholarly Ontologies Project [41]. In this project, S. J. Buckingham Shum et al. proposed

decomposing scientific papers into basic discourse knowledge units and, based on cognitive relationship theories, defined argumentative relationships including: causal, problem-related, similarity, generic, support/challenge, and classification relationships. Each relationship includes explicit polarity (positive or negative) and specific weights. This research has spawned a series of tools for annotating and visualizing argumentative relationships [42-43].

(3) Contextual Information Description. Contextual information reveals the existence status of scientific paper content components. P. Thompson designed the EventMine-MK annotation framework for contextual information in biomedical scientific papers, representing scientific paper contextual information through knowledge types, credibility levels, polarity, sources, degrees, and other dimensions [44]. Additionally, P. Thompson et al. designed a contextual information annotation framework for news events, adding dimensions such as information source, voice, and subjectivity based on polarity, time, and genre [45]. A. De Waard et al. proposed a contextual information representation model including three dimensions: certainty level, foundation, and source [46]. The certainty level dimension indicates the credibility of statements; the foundation dimension indicates the existence status of statements and propositions; the source dimension represents provenance information. Claim Framework [47], proposed by C. Blake, takes claims as the main description object, arguing that claim components include not only knowledge entities such as subject and object but also modal elements such as change, direction, modification, and foundation.

These three representation models provide different definitions of contextual information: EventMine-MK targets event-type knowledge, beneficial for event knowledge representation and mining; De Waard' s model focuses on multi-dimensional representation of statements; Blake' s framework constructs existence relationships between content components and conceptual entities, emphasizing the representation of logical relationships between entities.

(4) Scientific Paper Content Ontologies. Current research uses ontologies to normatively define content components and their relationships, designing and developing numerous scientific paper content ontologies.

Coarse-grained rhetorical ontologies include SALT [48] and Ontology of Rhetorical Block (ORB) [49], which macroscopically define the rhetorical structure of scientific paper content. Fine-grained rhetorical ontologies are represented by Discourse Elements Ontology [50] and Document Components Ontology (DoCO) [51], which meticulously define content components. In addition to rhetorical component ontologies, Peroni et al. proposed the Argument Model Ontology (AMO) [52], defining six argument elements including claim, evidence, support, challenge, qualifier, and warrant, as well as argumentative relationships such as support and question.

With deepening research on semantic enhancement of scientific paper content, the development of scientific paper content ontologies shows two trends: First, based on semantic modeling of scientific paper content, researchers strive to

develop more expressive and comprehensive ontologies. For example, Wang Xiaoguang et al. designed and developed a Functional Units Ontology (FUO) integrating contextual information based on functional unit theory and conducted preliminary deep annotation experiments [53]. Wang Xiaoguang et al. also reviewed argumentation ontologies and, referencing DEO, DoCO and other ontologies, further improved the definition of scientific evidence and designed the Scientific Argumentation Ontology (SAO) [54]. Second, some ontologies focus more on specific parts of scientific paper content, such as those for scientific conclusions [55] and scientific paper events [56-57], aiming for more complete definitions of specific knowledge.

3.2.4 Fine-grained Content Organization Models for Scientific Papers

Fine-grained content organization models for scientific papers represent a new direction for semantic enhancement of unstructured scientific paper content data, involving structural enhancement through association and reorganization after structuring and semanticizing unstructured information.

Among many content organization models, the most representative is the Micropublication model designed by T. Clark et al. [58]. This model distinguishes statements, assertions, data, methods, and other sentences with different semantics and functions, and clarifies argumentative relationships between sentences. Similarly, C. Bølling et al. [59] proposed the Semantic Evidence (SEE) representation method and model. SEE also provides a knowledge aggregation approach based on evidence, connecting scientific claims, evidence, and related materials, methods, hypotheses, reasoning, and other external knowledge bases on specific topics to form an interconnected and machine-readable representation. Other models, such as the Research Object Suite model [60], aim to provide a structured container that encapsulates research data with corresponding research methods and related metadata to form a suite around specific topics.

Essentially, scientific paper content organization models provide a new document representation method. Both micropublication and semantic evidence models split scientific literature into various argument units and then reorganize them according to argument structures, representing both the logical structure of scientific paper content and the association of content components. The Research Object Suite model associates research methods, experimental processes, and scientific data in research papers, representing both the scientific experimental process and providing clear attributes and background for scientific data.

4. Comparative Analysis of Semantic Enhancement for Scientific Papers

Based on the above review of theoretical exploration and practice in semantic enhancement of different components of scientific papers, this paper conducted a comprehensive analysis of semantic enhancement paths and implementation for different content data, as shown in Table 1 .

Table 1. Comparative Analysis of Semantic Enhancement Paths for Scientific Papers

Scientific Paper Components	Semantic Enhancement Path	Ontology/Model	Visualization Presentation
Paratext Content	Bibliographic Information	BIBO, FaBiO, VIVO, SciGraph, MAG	Rich media abstracts (image abstracts, video abstracts)
	Abstracts	Structured abstracts, SDA, highlight abstracts	SciGraph, MAG, OpenCitations, AMiner
	Citations & References	CiTO, C4O	Nanopublications
Primary Text Content	Conceptual Entities	Domain ontologies, NER	Tag clouds, tag trees
	Statement Description	Nanopublications	Micropublications, semantic evidence, research object suites
	Content Components & Logic	CoreSC, ABCDE, Discourse Segment, AMO, DoCO, SALT, EventMine-MK, Claim Framework	Micropublications, semantic evidence models
	Argumentative Structure	AMO, ScholoOnto, SALT	
	Contextual Information	EventMine-MK, De Waard' s model, Claim Framework	

As shown in Table 1, practice in semantic enhancement of paratext content is

relatively rich. For citation and reference information, existing research has achieved semantic description of citation functions, situations, and contexts, and constructed corresponding datasets. Meanwhile, through ontologies and knowledge graphs, bibliographic and reference information can be associated and published.

Semantic enhancement of primary text content is mainly theoretical exploration. In conceptual entity extraction and representation, significant progress has been made in extracting and representing conceptual entities using domain ontologies. In statement description and linked publication, nanopublication dataset construction is also steadily advancing. For semantic representation and association of content components and logical structures, although corresponding ontologies and organization models exist, large-scale datasets have not yet been constructed due to technical issues in the semantic annotation process.

Overall, research and practical achievements in semantic enhancement of scientific papers mainly concentrate on the semantic description and annotation stage. The importance of ontologies in the semantic enhancement process has gradually become prominent, while semantic organization models and knowledge graphs have also been produced. However, content visualization presentation research is still clearly insufficient.

5. Future Trends in Semantic Enhancement Research for Scientific Papers

Focusing on the core objectives and key issues of semantic enhancement for scientific papers, this paper proposes that future work and exploration can develop in the following directions:

(1) **Semantic Integration and Interoperability of Multi-dimensional, Multi-source Data.** Through the design, development, and application of different knowledge graphs, existing research has conducted semantic enhancement of scientific paper bibliographic and reference information. However, such knowledge graphs rarely involve association with primary text content data (statements, content components, and logical structures). How to fill the gaps in existing scientific paper knowledge graphs based on semantic representation of scientific paper content and achieve association between knowledge graphs and content semantic organization models will be fundamental to promoting semantic enhancement of scientific papers.

(2) **Visualization of Rich Semantic Content Data.** Using rich visualization methods to improve the perceptibility of scientific paper content is key to semantic enhancement. Current research on visualizing scientific papers is still scarce, with visualization methods mainly limited to tag clouds and tag trees, and visualization objects primarily based on words. How to completely, efficiently, and accurately visualize semantic content data such as argumentative structures and key information requires deeper exploration from both theoretical and practical perspectives.

(3) **Domain-specific Semantic Enhancement for Scientific Papers.** The complexity of scientific papers lies not only in their rich knowledge content but also in their subjection to domain research paradigms, methods, and writing norms. Existing content ontologies and organization methods mainly target the biomedical domain. How to apply existing achievements to humanities, social sciences, or other natural science domains requires building more expressive and comprehensive content semantic representation models tailored to domain characteristics and developing domain-specific semantic enhancement approaches.

(4) **Semantic Enhancement Oriented to Scientific Paper Reading Behavior.** The ultimate goal of semantic enhancement for scientific papers is to help researchers quickly acquire the substantial knowledge contained in papers. Therefore, it is necessary to conduct in-depth research on users' reading tasks, strategies, objects, and patterns by domain. Current basic theoretical research and practical exploration mostly start from analyzing textual features of scientific papers. Both semantic description and organization are based on text analysis and logical analysis, without fully considering users' reading characteristics and usage patterns. Therefore, future research on semantic enhancement of scientific papers needs to strengthen understanding of user reading behavior.

References

[1] RENEAR A H, CAROLE L P, Strategic reading, ontologies, and the future of scientific publishing[J]. Science, 2009, 325(5492): 828-832.
[2] SHOTTON D. Semantic publishing: the coming revolution in scientific journal publishing[J]. Learned publishing, 2009, 22(2): 85-94.
[3] SHOTTON D. The five stars of online journal articles: a framework for article evaluation[EB/OL]. [2020-12-20]. https://purl.pt/302/dlib/january12/shotton/olshotton.html.
[4] SHOTTON D, PORTWIN K, KLYNE G, et al. Adventures in semantic publishing: exemplar semantic enhancements of a research article[J]. PLoS computational biology, 2009, 5(4): e1000361.
[5] WENG Yanqin, LI Yuan, PENG Xilin. Research on the semantic publishing model of scientific journals by the Royal Society of Chemistry (RSC)[J]. Chinese Journal of Scientific and Technical Periodicals, 2013, 24(5): 825-829.
[6] WENG Yanqin, PENG Xilin. Research on Elsevier' s semantic publishing model[J]. Chinese Journal of Scientific and Technical Periodicals, 2014, 25(10): 1256-1261.
[7] KURZ T, DAMJANOVIC V, GUNTNERS G, et al. Semantic enhancement for media asset management systems[J]. Multimedia tools & applications, 2014, 70(2): 949-975.
[8] Europeana semantic enrichment[EB/OL]. [2020-02-23]. https://pro.europeana.eu/page/europeana-semantic-enrichment.
[9] ZENG M L. Semantic enrichment for enhancing LAM data and supporting digital humanities. review article[J]. El profesional de la informacion, 2019,

28(1): 1-35.
[10] WOUTERSEN-WINDHOUWER S, BRANDSMA R, HOGENAAR A, et al. Enhanced publications: linking publications and research data in digital repositories[M]. Amsterdam: Amsterdam University Press, 2009.
[11] HOGERWERF M. Durable enhanced publications[EB/OL]. [2021-01-03]. https://www.researchgate.net/publication/242732066_{{{Durable}}}{{{Enhanced}}}{Publications}.
[12] BREURE L, VOORBIJ H, HOGERWERF M. Rich internet publications: show what you tell[J]. Journal of digital information, 2010, 12(1): 1.
[13] PRASAD A R D, GIUNCHIGLIA F, DEVIKA P M. DERA: from document centric to entity centric knowledge modeling[C]// SLAVICA A, GNOLI C. Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017. Würzburg: Ergon Verlag, 2017: 169-179.
[14] Bibliographic ontology specification[EB/OL]. [2021-01-02]. http://www.bibliontology.com.
[15] PERONI S, SHOTTON D. FaBiO and CiTO: ontologies for describing bibliographic resources and citations[J]. Web semantics: science, services and agents on the World Wide Web, 2012, 17(17): 33-43.
[16] ZHANG Yanxia, QI Fei, BI Qiang. Research on semantic interconnection applications of linked data: a case study of VIVO[J]. Library and Information Service, 2013, 57(17): 17-21.
[17] YU Qichen, WANG Xiaoguang. Investigation on semantic enhancement forms of scientific paper abstracts[J]. Digital Library Forum, 2017(8): 8-15.
[18] CICCARESE P, SHOTTON D, PERONI S, et al. CiTO+SWAN: the Web semantics of bibliographic records, citations, evidence and discourse relationships[J]. Semantic Web, 2014, 5(4): 295-311.
[19] Citation counting and context characterization ontology[EB/OL]. [2019-12-20]. http://purl.org/spar/c4o.
[20] OpenCitation[EB/OL]. [2020-12-05]. http://opencitations.net/.
[21] SciGraph[EB/OL]. [2020-05-15]. http://www.springernature.com/cn/researchers/scigraph.
[22] Aminer[EB/OL]. [2020-05-15]. https://www.aminer.cn.
[23] Microsoft academic graph[EB/OL]. [2020-05-15]. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.
[24] Open academic graph[EB/OL]. [2020-05-15]. https://www.openacademic.ai/oag/.
[25] WANG R, YAN Y, WANG J, et al. AceKG: a large-scale knowledge graph for academic data mining[C]// Proceedings of the 27th ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2018: 1487-1491.
[26] REN Haiying, SHI Tong. Construction of micro-concept maps for scientific papers and mining of research ideas[J]. Library and Information Service, 2016, 60(4): 115-124.
[27] DING Junjun, ZHENG Yanning, HUA Bolin. Rule-based academic concept attribute extraction[J]. Information Studies: Theory & Application, 2011, 34(12): 10-14, 33.
[28] LE Xiaoqiu, ZHANG Fan, HE Yuanbiao. Research on key term extraction methods from academic paper outlines[J]. New Technology of Library and Information Service, 2014, 30(3): 73-79.

[29] WU Sizhu, LI Feng, ZHANG Zhixiong. Research on semantic representation and publishing models of knowledge resources: a case study of Nanopublication[J]. Journal of Library Science in China, 2013(4): 102-109.
[30] KING R D, ROWLAND J, OLIVER S G, et al. The automation of science[J]. Science, 2009, 324(5923): 85-89.
[31] LIAKATA M, SAHA S, DOBNIK S, et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics, 2012, 28(7): 991-1000.
[32] DE WAARD A, TEL G. The ABCDE format enabling semantic conference proceedings[EB/OL]. [2021-01-01]. https://www.researchgate.net/publication/220706582_{{{The}}}{{ABCDE
[33] DE WAARD A, BUITELAAR P, EIGNER T. Identifying the epistemic value of discourse segments in biology texts[C]// Proceedings of the eighth international conference on computational semantics. Stroudsburg: Association for Computational Linguistics, 2009: 351-354.
[34] ZHANG L, KOPAK R, FREUND L, et al. A taxonomy of functional units for information use of scholarly journal articles[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1): 1-10.
[35] TEUFEL S. Argumentative zoning: information extraction from scientific texts[D]. Edinburgh: University of Edinburgh, 1999.
[36] TEUFEL S. The structure of scientific articles: applications to citation indexing and summarization[M]. Stanford, CA: CSLI Publications (CSLI Studies in Computational Linguistics), 2010.
[37] GREEN N L. Representation of argumentation in text with rhetorical structure theory[J]. Argumentation, 2010, 24(2): 181-196.
[38] GREEN N. Identifying argumentation schemes in genetics research articles[C]// Proceedings of the 2nd workshop on argumentation mining. Denver: Association for Computational Linguistics, 2015: 12-21.
[39] GREEN N. Argumentation mining in scientific discourse[C]// Proceedings of the 18th workshop on computational models of natural argument. London: Association for Computational Linguistics, 2017: 7-13.
[40] GREEN N. Implementing argumentation schemes as logic programs[C]// Proceedings of the 16th Workshop on computational models of natural argument. New York: Association for Computational Linguistics, 2017: 1-7.
[41] SHUM S B, MOTTA E, DOMINGUE J. ScholOnto: an ontology-based digital library server for research documents and discourse[J]. International journal on digital libraries, 2000, 3(3): 237-248.
[42] BUCKINGHAM SHUM S J, UREN V, LI G, et al. Modeling naturalistic argumentation in research literatures: representation and interaction design issues[J]. International journal of intelligent systems, 2007, 22(1): 17-47.
[43] UREN V, BUCKINGHAM SHUM S, BACHLER M, et al. Sense-making tools for understanding research literatures: design, implementation and user evaluation[J]. International journal of human-computer studies, 2006, 64(5): 420-445.
[44] THOMPSON P, NAWAZ R, MCNAUGHT J, et al. Enriching a biomedical event corpus with meta-knowledge annotation[J]. BMC bioinformatics, 2011, 12(1). doi:10.1186/1471-2105-12-393.

[45] ANANIADOU S, THOMPSON P, NAWAZ R. Enhancing search: events and their discourse context[C]// GELBUKH A. Proceedings of the 14th international conference on computational linguistics and intelligent text processing. Berlin: Springer-Verlag, 2013: 318-334.
[46] DE WAARD A, MAAT H P. Epistemic modality and knowledge attribution in scientific discourse: a taxonomy of types and overview of features[C]// Proceedings of the workshop on detecting structure in scholarly discourse. Stroudsburg: Association for Computational Linguistics, 2012: 47-55.
[47] BLAKE C. Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles[J]. Journal of biomedical informatics, 2010, 43(2): 173-189.
[48] TUDOR G, SIEGFRIED H, KNUD M, et al. SALT - Semantically annotated LaTeX for scientific publications[C]// Proceedings of the 4th European semantic Web on the semantic Web: research and applications. Berlin: Springer-Verlag, 2007: 518-532.
[49] MA Yumeng, ZHU Zhongming. Analysis of scientific discourse rhetorical block ontology standards and their applications[J]. Journal of Intelligence, 2012(10): 112-116.
[50] The discourse element ontology[EB/OL]. [2020-05-15]. http://purl.org/spar/deo.
[51] CONSTANTIN A, PERONI S, PETTIFER S, et al. The document components ontology (DoCO)[J]. Semantic Web, 2016, 7(2): 167-181.
[52] The argument model ontology[EB/OL]. [2020-10-23]. http://www.essepuntato.it/2011/02/argumentmodel.
[53] WANG Xiaoguang, LI Menglin, SONG Ningyuan. Design of scientific paper functional unit ontology and indexing application experiment[J]. Journal of Library Science in China, 2018, 236(4): 75-90.
[54] WANG Xiaoguang, ZHOU Huimin, SONG Ningyuan. Design of scientific argumentation ontology and annotation experiment[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(9): 885-895.
[55] FATTHALLA S, VAHDATI S, AUER S, et al. SemSur: a core ontology for the semantic representation of research findings[C]// Proceedings of the 14th international conference on semantic systems. Vienna: Elsevier B.V., 2018: 151-162.
[56] JEONG S, KIM H G. SEDE: an ontology for scholarly event description[C]// Proceedings of the ACM international conference on information and knowledge management. New York: Association for Computing Machinery, 2018: 227-232.
[57] FATTHALLA S, VAHDATI S, LANGE C, et al. SEO: a scientific events data model[EB/OL]. [2020-11-12]. https://www.researchgate.net/publication/336594094_{SEO}A_{{{Scienti
[58] CLARK T, CICCARISE P, GOBLE C. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications[J]. Journal of biomedical semantics, 2014, 5(1): 28.
[59] BOLLING C, WEIDLICH M, HOLZHUETTER H G. SEE: structured representation of scientific evidence in the biomedical domain using semantic Web techniques[J]. Journal of biomedical semantics, 2014, 5(S1): S1.
[60] HETTNE K M, DHARURI H, ZHAO J, et al. Structuring research methods and data with the research object model: genomics workflows as a

case study[J]. Journal of biomedical semantics, 2014, 5(1): 41.

Author Contributions

Song Ningyuan: Framework design, paper writing;
Pei Lei: Research direction determination, providing directional suggestions;
Wang Chunying: Paper revision and finalization.

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

Source: ChinaXiv —Machine translation. Verify with original.