SKOS-based Multi-granular Semantic Annotation Method for Academic Journal Text Resources (Postprint)

Authors: Xia Lixin, Zheng Lu, Zhang Yuchen, Zhai Shanshan, Sun Jingqiong

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

Abstract

[Purpose/Significance] To address the two persistent challenges in semantic annotation of academic journal text resources—the difficulty of constructing general-purpose ontologies and the limitation to single-granularity annotation—this paper proposes a multi-granularity semantic annotation method for academic journals based on SKOS, thereby further advancing the applied development of semantic annotation and better meeting users' multi-granularity academic information needs. [Method/Process] Building upon a SKOS description of the "Chinese Thesaurus," this study takes academic journal text resources as its object to implement multi-granularity semantic annotation and verifies the feasibility of the method through empirical research. [Results/Conclusion] The implementation of multi-granularity semantic annotation for academic journal text resources using SKOS demonstrates certain advantages over annotation results in current academic retrieval systems across four dimensions: "recall," "precision," "internal feature retrieval entry points," and "retrieval result feedback format."

Full Text

Abstract

[**Purpose/Significance**] Semantic annotation of academic journal text resources faces two major challenges: the difficulty of constructing a universal ontology and the limitation of single-granularity annotation. To address these issues, this paper proposes a multi-granularity semantic annotation method for academic journals based on SKOS, which advances the application development of semantic annotation and better satisfies users' multi-granular academic information needs. [**Method/Process**] Building upon the SKOS description of the *Chinese Thesaurus*, this study takes academic journal text

resources as the annotation object to implement multi-granularity semantic annotation and validates the feasibility of the method through empirical research. [**Result/Conclusion**] Using SKOS to achieve multi-granularity semantic annotation of academic journal text resources demonstrates certain advantages over current annotation results in academic retrieval systems across four dimensions: recall, precision, internal feature retrieval access points, and retrieval result feedback forms.

Keywords: semantic annotation; multi-granularity; SKOS thesaurus; academic journal
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.09.015

Academic journals serve as important platforms for presenting and exchanging scholarly achievements. With continuous digital development, massive amounts of academic information both meet users' research needs and create severe information overload, making it difficult for users to efficiently extract required academic information [1]. Meanwhile, differing standards and specifications across academic journals from various sources create obstacles for dissemination and sharing. Semantic annotation offers a new knowledge organization method for academic journals. By tagging original data with semantic information, content becomes understandable not only to humans but also to machines [2], significantly improving the efficiency of academic journal retrieval and utilization.

Currently, ontology-based semantic annotation methods are most widely used. However, direct construction of a domain ontology often consumes substantial resources and requires assistance from domain experts. Different ontologies frequently suffer from heterogeneity issues, making them difficult to generalize and integrate. Therefore, constructing a truly universal ontology has become a bottleneck. Coarse-grained annotation methods prevent the vast amount of information contained within academic journals from being further retrieved, filtered, and extracted. To address this, some scholars have attempted fine-grained annotation of academic journals, extending annotation units to the smallest knowledge units. However, both coarse-grained and fine-grained annotation methods only achieve single-granularity knowledge organization, while users' academic information needs often exhibit multi-granularity characteristics. Consequently, single-granularity annotation cannot satisfy users' multi-granular academic information needs, making multi-granularity annotation essential for meeting these demands.

To further advance the application development of semantic annotation and better satisfy users' multi-granular academic information needs, this paper explores implementation approaches for multi-granularity semantic annotation of academic journal text resources based on the SKOS resource description framework and related technologies.

Related Research

SKOS Research Status

SKOS (Simple Knowledge Organization System), published by the World Wide Web Consortium (W3C) in 2004, provides a solution for knowledge management and semantic processing of controlled vocabularies [3]. Using SKOS-ified controlled vocabularies for semantic annotation has become a potential solution to ontology construction challenges. Since the release of the SKOS resource description framework, research has primarily focused on the SKOS transformation of controlled vocabularies. While numerous successful examples of English controlled vocabularies have been SKOS-ified—with the W3C SKOS official website's Datasets page sharing up to 39 SKOS-ified controlled vocabularies [4]—unfortunately, none are Chinese controlled vocabularies. For Chinese controlled vocabularies, scholars' research has concentrated on the SKOS conversion of the *Chinese Thesaurus* or *Chinese Classification Thesaurus*. Fan Wei proposed using SKOS to construct machine-understandable knowledge organization systems, with case studies using thesauri [5]. Jia Junzhi provided a clear description of SKOS content and structure for the *Chinese Thesaurus*, completing a SKOS description demonstration [6]. Additionally, Zhang Shinan et al. designed a SKOS description scheme for the *Library Classification of the Chinese Academy of Sciences* [7].

Although certain achievements have been made in SKOS-ifying controlled vocabularies, they have not yet been widely applied, and related research remains fragmented. J. Pastor-Sanchez et al. compared SKOS schemes with other controlled vocabulary representation schemes in the semantic web, ultimately concluding that SKOS is the optimal semantic description scheme for thesauri [8]. Wang Qian et al. discussed macro-level knowledge organization models for the semantic web using SKOS, enhancing knowledge description capabilities through extensions of SKOS classes and properties [9]. Xiong Taichun analyzed the feasibility of using SKOS for indexing information resources in network environments [10].

In summary, current research on the SKOS resource description framework primarily focuses on semantic description of controlled vocabularies, with abundant achievements in SKOS description of controlled vocabularies but relatively insufficient follow-up application research, particularly lacking practical outcomes.

Academic Journal Semantic Annotation Research Status

Zhu Jiaxian et al. define semantic annotation as "using ontology technology to annotate semantic information such as concepts, attributes, and relations in information resources as machine-understandable metadata, achieving association between annotation information and resources" [11]. Currently, semantic annotation methods for academic journal text resources are mainly implemented based on ontology technology. At the theoretical level, Wei Moji et al. proposed a semantic annotation method for discipline-specific documents based

on domain ontologies [12]; Leng Fuhai et al. combined semantic annotation technology, rule extraction technology, and regular expression technology to propose a method for extracting academic information such as specific theories involved in academic literature [13]; the General Architecture for Text Engineering (GATE) developed by the University of Sheffield represents a prominent multi-ontology-based semantic annotation method, though its limitation lies in the lack of mapping associations between multiple ontologies, hindering interconnection and interoperability [14]. Many scholars have also proposed more specific ontology-based optimization methods for academic journal text resource semantic annotation, but most remain in the experimental stage. At the practical level, the DBpedia project is a typical representative of text resource semantic annotation, extracting structured information from Wikipedia entries and linking datasets to Wikipedia, ultimately publishing data as linked data [15]. Subsequently, numerous scholars have conducted practical attempts at semantic annotation of academic text resources based on this project. For example, F. Norberto et al. proposed a collaborative semantic annotation framework based on DBpedia that fully leverages the advantages of manual semantic annotation, integrates basic user operations with semantic annotation operations, and reduces the burden on non-expert annotators [16]; Tang Yijie et al. combined the Chinese Academy of Sciences Integrated Information Platform (CASIIP) with the DBpedia dataset, using DBpedia's information resource description and organization forms to semantically annotate data information in CASIIP, achieving semantic extension of the CASIIP platform [17].

In summary, research on semantic annotation of academic text resources has accumulated certain achievements, but studies primarily focus on entire documents or resource collections as objects. Structurally, annotation has not penetrated into document chapters; content-wise, it has not involved knowledge units that characterize document content features. In other words, semantic annotation research on academic text resources has not considered finer-grained semantic annotation schemes [18], and lacks a universal ontology that makes semantic description of academic text resources more structured and standardized. These two major issues directly affect users' multi-granular academic information needs—the deeper and more detailed the needs, the less effectively users can locate required information resources. Therefore, based on existing research, this study achieves multi-granularity semantic annotation of academic journal text resources through SKOS-ification of the *Chinese Thesaurus*.

SKOS Application Analysis in Multi-Granularity Semantic Annotation of Academic Journal Text Resources

Advantages of Multi-Granularity Semantic Annotation

Multi-granularity semantic annotation essentially involves dividing document content into granularity layers and performing semantic annotation on each layer separately, forming hierarchical and structured annotation results. Coarse

granularity provides comprehensive description of a theme; medium granularity describes a specific aspect of a theme; fine granularity describes a specific problem. Multi-granularity annotation reveals both overall summaries and in-depth descriptions of specific document content. Compared with commonly used single-granularity annotation (either coarse or fine), the rich annotation results provide greater support for both higher-level knowledge organization forms and user retrieval feedback.

Advantages of SKOS for Thesaurus Description

Currently, many thesaurus description schemes exist based on XML and RDF, with some alternative methods such as topic maps. Compared with other thesaurus representation schemes, SKOS offers distinct advantages as shown in Table 1 .

Table 1 Advantages of SKOS Over Other Thesaurus Description Schemes

Thesaurus Description Scheme	Relative Advantage of SKOS
XML Vocabularies (ZTHES, MESH)	Uses RDF for semantic integration at the description level
Concept Maps, Topic Maps (XTM)	Uses OWL for semantic integration at the logical level
Other RDF Vocabularies (LIMBER, CERES, ILRT)	More flexible and standardized concept description changes
OWL Ontology	Simpler description and maintenance tasks

In the *Chinese Thesaurus*, both descriptors and non-descriptors are described as lexical labels of SKOS concepts. If the SKOS standard language lacks properties corresponding to lexical attributes, the SKOS standard language can be customized and extended with new properties, with the extended portion called SKOSEX language. This paper uses some SKOSEX language for attribute description in depicting the *Chinese Thesaurus*. In addition to concept and semantic relationship descriptions, some properties are established for thesaurus creation, such as addition time, term frequency, term type, and edit count, which need not be displayed to users and can therefore be ignored. Ultimately, terms in the *Chinese Thesaurus* are described using the properties shown in Table 2 .

Table 2 Term Description Properties for the Chinese Thesaurus

Chinese Thesaurus Term Relationship	Corresponding SKOS Concept Property
Class number in *Chinese Library Classification*	skos:broadMatch
Category number in category index	skos:broadMatch

Chinese Thesaurus Term Relationship	Corresponding SKOS Concept Property
Pinyin representation of descriptor	skos:prefLabel xml:lang="zh-latn"
English translation of descriptor	skos:prefLabel xml:lang="en"
Chinese label of descriptor	skos:prefLabel xml:lang="zh"
Abbreviation of descriptor (optional label)	skosex:abbreviation
Equivalence relationship between two descriptors	skos:exactMatch
Non-descriptor synonymous with descriptor	skos:altLabel xml:lang="zh"
Top term of word family	skos:broader
Broader term	skos:narrower
Narrower term	skos:related
Related term	skos:topBroader
Word family leader	skos:leadBroader
Component concept forming compound concept	skosex:coordinationOf
Component concept generating compound concept	skosex:coordinationOf
Compound concept formed by single descriptor	skosex:coordinatedTo
Some notation symbol corresponding to descriptor	skos:notation
User evaluation note	skos:note
Historical note	skos:historyNote
Evaluation note	skosex:evaluationNote

Where xmlns:skos represents the SKOS standard language defined by W3C, and xmlns:skosex represents the extension language of SKOS. The language code for the xml:lang attribute is defined by the IETF BCP 47 standard.

A partial SKOS language description example for the descriptor "information retrieval" from the *Chinese Thesaurus* is shown in Figure 1 [Figure 1: see original paper].

Framework Design for Multi-Granularity Semantic Annotation of Academic Journal Text Resources Based on SKOS

The research process is illustrated in Figure 2 [Figure 2: see original paper]. Specifically: (1) Perform thesaurus semantic description, including revelation of semantic relationships; (2) Design the multi-granularity annotation process for academic journal text resources; (3) Combine both to implement SKOS-based multi-granularity semantic annotation of academic journal text resources; (4) Develop an implementation scheme for the SKOS-based multi-granularity semantic annotation method, using journal articles as examples for annotation to verify feasibility and evaluate annotation results.

The framework can be divided into three main components:

1. **Thesaurus to SKOS Conversion**: The thesaurus serves as the information retrieval language selected for use in the annotation process. Traditional thesaurus concepts and semantic relationships must be expressed using SKOS standard description language. The SKOS-ified thesaurus

is the fundamental tool for semantic annotation. This paper selects the SKOS description results of the *Chinese Thesaurus* as the annotation tool for the multi-granularity semantic annotation method. The SKOS description scheme for the *Chinese Thesaurus* was explained in Section 3.3.

2. **Multi-Granularity Processing and Annotation Term Selection for Academic Journal Text Resources**: Before semantic annotation, three basic processing steps are required for annotation objects. First, construct a multi-granularity hierarchical structure for annotation objects by dividing academic journal content into hierarchical tree structures according to different granularity units. Second, perform multi-granularity segmentation of academic journal content based on the granularity division results. Finally, conduct importance calculation for multi-granularity annotation candidate terms on the segmentation results—i.e., construct annotation term evaluation metrics, calculate scores for each term, and select current granularity annotation terms based on score rankings.
3. **SKOS-Based Multi-Granularity Annotation and Result Generation for Academic Journal Text Resources**: After selecting appropriate annotation terms through calculation, use SKOS-ified thesaurus to perform semantic annotation of academic journal text resources and organize results through XML-structured documents, preserving the structural hierarchy of multi-granularity annotation results. The semantic annotation process requires further concept description and semantic relationship revelation according to needs, including description of non-descriptors among annotation terms.

Granularity Division of Academic Journal Text Resources

Through analysis of academic journal structural characteristics, this paper divides the annotation granularity levels of academic journal text resources as shown in Figure 3 [Figure 3: see original paper].

Based on the thematic structural features of academic journal text resources, this paper divides annotation granularity in the annotation process as follows: (1) Coarse granularity: full-text content of academic journals; (2) Medium granularity: chapter and section units of academic journals; (3) Fine granularity: natural paragraphs of academic journals.

In academic journal text resources, coarser-grained content contains finer-grained content, and there often exists a hierarchical relationship between them during annotation. Therefore, this paper selects a bottom-up annotation direction—first annotating natural paragraphs (fine granularity), then sections/chapters (medium granularity), and finally full-text units (coarse granularity).

Multi-Granularity Segmentation of Academic Journal Text Resources

This paper uses the Chinese word segmentation system NLPIR developed by the Chinese Academy of Sciences for segmentation. The dictionaries used by NLPIR mainly include the core dictionary (coreDict), word association dictionary (BigramDict), person name dictionary (nr), translated person name dictionary (tr), and location name dictionary (ns). The multi-granularity segmentation process for academic journals proceeds bottom-up, starting from the paragraph level and segmenting each granularity unit sequentially. The segmentation process is illustrated in Figure 4 [Figure 4: see original paper].

The multi-granularity segmentation process proceeds as follows: First, take fine-grained unit text "paragraphs" and use NLPIR for segmentation according to the workflow shown in Figure 4. After segmentation, store paragraph segmentation results separately with proper identifiers. Then take medium-grained unit text "sections" and repeat the segmentation workflow, storing section segmentation results separately with identifiers. The analysis process for medium-grained text "chapters" is the same as for "sections." Finally, take coarse-grained text "full-text" and complete segmentation following the same workflow for storage and identification. Multi-granularity segmentation aims to capture subtle differences in segmentation algorithms and results caused by different text object bases, such as differences in "new word discovery" results due to varying granularity texts. Therefore, the sum of segmentation results from all "paragraphs" under a chapter often does not completely equal the one-time segmentation result of that chapter text. Multi-granularity segmentation can comprehensively grasp the full picture of current granularity unit text, obtaining the most appropriate segmentation results to express current granularity text content, preparing for the next step of annotation term selection.

Selection of Multi-Granularity Annotation Terms for Academic Journal Text Resources

Before calculating annotation term selection, preprocessing of multi-granularity segmentation results is necessary, with the most important step being stopword removal. Stopword removal eliminates the impact of meaningless high-frequency words on annotation results. After stopword removal, candidate terms with annotation significance are obtained. Finally, importance scoring of candidate terms yields annotation terms. This method uses the metrics shown in Figure 5 [Figure 5: see original paper] for candidate term importance scoring:

Figure 5 [Figure 5: see original paper] Importance Calculation Metrics for Multi-Granularity Annotation Candidate Terms of Academic Journal Text Resources

1. **TF-IDF Value**: In different granularity levels, the TF-IDF value of the same candidate term should be recalculated according to granularity units. The TF value refers to the frequency of the candidate term in the current annotation granularity unit content, calculated as:

$$TF = \frac{\text{Current term frequency in granularity text}}{\text{Total segmentation count in current text}}$$

Similarly, IDF values differ across granularity levels. In this multi-granularity semantic annotation method, IDF calculation is also defined differently based on granularity units.

- Fine granularity (paragraph-based): The IDF value for a specific term in a paragraph is:

$$IDF_i = \log \frac{\text{Total paragraphs in the academic journal}}{\text{Number of paragraphs containing the term}}$$

- Medium granularity (chapter-based): The IDF value for a specific term in a chapter is:

$$IDF_j = \log \frac{\text{Total chapters in the academic journal}}{\text{Number of chapters containing the term}}$$

- Coarse granularity (document-based): The IDF value for a specific term in an academic journal document is:

$$IDF_k = \log \frac{\text{Total documents in the retrieval system}}{\text{Number of documents containing the term}}$$

After calculating TF and IDF values for each candidate term in respective annotation granularity units, multiply them to obtain the absolute TF-IDF value. Then use the "min-max" normalization method to map all TF-IDF values to the interval [0,1] for easier comparison.

2. **Position**: The position metric uses direct assignment based on sub-metrics. If a candidate term appears in three important positions—"text title," "keywords," or "section title"—each sub-metric is assigned a value of 1. These three sub-metrics do not interfere with each other. If a candidate term appears in two or more positions simultaneously, values can be assigned separately and then multiplied by weights in the calculation formula.

3. **Part of Speech**: Part-of-speech filtering during annotation can quickly eliminate numerous terms without annotation significance. This method sets a "noun part-of-speech" importance bonus in the term importance evaluation metrics. This sub-metric also uses direct assignment: noun candidate terms receive a value of 1, non-noun candidate terms receive 0. When calculating, the part-of-speech used in the current granularity unit text is primary. If multiple part-of-speech forms appear in the current granularity unit text, noun part-of-speech alone can assign a value of 1.

4. **Inter-term Relationships**: The inter-term relationship in candidate term importance evaluation refers to the relationship between candidate terms in coarser granularity units and annotation terms in finer granularity units, which can be divided into three categories: "equivalence relationship," "hierarchical relationship," and "associative relationship." This method uses the *Chinese Thesaurus* as the information retrieval language for annotation. The relationships in the thesaurus correspond to the three evaluation sub-metrics as follows: "equivalence relationship" corresponds to "Use (Y)" and "Used for (D)" in the thesaurus; "hierarchical relationship" corresponds to "Broader term (S)," "Narrower term (F)," and "Top term (Z)"; "associative relationship" corresponds to "Related term (C)."

This importance evaluation metric also uses direct assignment. When an annotation candidate term in a coarser granularity unit has one of these relationships with a specific annotation term in a finer granularity unit, the corresponding sub-metric is assigned a value of 1. For a given candidate term and a specific annotation term, only one most appropriate relationship type can be selected to avoid duplicate assignment. However, for the entire network of lexical relationships where the candidate term resides, the candidate term can have multiple types of relationships with different annotation terms, but each relationship type is assigned only once without cumulative duplication. Notably, in the finest granularity annotation, all candidate terms receive a score of 0 for this metric since no finer-grained annotation results exist.

Let variable TI represent the TF-IDF value of the current candidate term in the current annotation granularity text (the same candidate term has different TF-IDF values in different annotation granularity texts and should be recalculated for each granularity). Let variables DT, KW, CT represent whether the current candidate term appears in the text title, keywords, or section title positions (assigned value 1 if present, 0 otherwise). Let variable N represent whether the current candidate term is a noun (assigned value 1 if noun, 0 otherwise; when a term has multiple part-of-speech forms, the part-of-speech used in the current granularity unit text is primary; if multiple part-of-speech forms appear in the current granularity unit text, noun part-of-speech alone can assign value 1). Let variables E, G, R represent whether the current candidate term has equivalence, hierarchical, or associative relationships with annotation terms in the previous granularity level (assigned value 1 if such relationship exists, 0 otherwise; only one most appropriate inter-term relationship can be selected for a specific term without duplicate assignment with other relationships; however, for the entire network of lexical relationships, the candidate term can possess all types of relationships, but these are not cumulatively assigned). The importance score of candidate term i in the current annotation granularity text can then be calculated using formula (5):

$$\text{Score}_i = \frac{TI_i - MIN(TI)}{MAX(TI) - MIN(TI)} + (DT_i + KW_i + CT_i) \cdot w_{pos} + N_i \cdot w_{noun} + (E_i + G_i + R_i) \cdot w_{rel}$$

Formula (5) calculates the comprehensive importance score of each candidate term in the current annotation granularity text across four aspects: "TF-IDF," "position," "part of speech," and "inter-term relationship." Sorting scores in descending order allows selection of an appropriate number of candidate terms as annotation terms for the current granularity text based on annotation needs. Following the multi-granularity division structure of academic journals, annotation term selection for each granularity text is completed sequentially from bottom to top, ultimately forming multi-granularity annotation results corresponding to the hierarchical structure of academic journals.

Representation of Multi-Granularity Annotation Results for Academic Journal Text Resources

SKOS is an RDF-based description language whose basic format adopts XML. Therefore, this paper continues to use XML for describing multi-granularity annotation results of academic journal text resources. This approach avoids conflict with the SKOS resource description framework and allows direct nested usage, while XML's extensibility through custom tags conveniently defines multi-granularity hierarchical levels of annotation results, preserving structural relationships between results.

XML (Extensible Markup Language) allows users to define tags for identifying structured document content. This paper uses three tag sets—`<document></document>`, `<chapter></chapter>`, and `<paragraph></paragraph>`—to identify coarse, medium, and fine granularity annotation levels respectively. The document structure of multi-granularity annotation results is illustrated in Figure 6 [Figure 6: see original paper].

Empirical Study of Multi-Granularity Semantic Annotation for Academic Journal Text Resources Based on SKOS

Basic Experimental Setup

The empirical study selected a theoretical research article titled "Data Governance: Development Opportunities for Library Undertakings" by scholar Gu Liping, published in the 5th issue of 2016 of *Journal of Library Science in China* [19]. This article has two notable characteristics:

1. **Single author**: Single-author papers ensure consistency of viewpoints and systematic discussion of issues throughout the academic journal, with coherent internal logical structure that better reflects a complete thought process on an academic problem by an independent mind, making inter-granularity relationships more prominent.
2. **Rigorous structure**: The article takes "data governance as a development opportunity for library undertakings" as its thesis, discussing the theme from four sub-aspects: "data acquisition governance," "data sharing governance," "data reuse governance," and "data value-added gover-

nance," forming a typical "general-specific-general" structure. The rigorous internal structure makes granularity division clearer, themes more obvious across granularities, differences significant among parallel granularity units at the same level, and relationships close between granularity units at different levels.

Therefore, this article is both representative of academic journals and well-suited for this method. This paper uses this text as the annotation object for an empirical study of SKOS-based multi-granularity semantic annotation of academic journals to verify the method's feasibility and annotation effectiveness.

Granularity Division of Experimental Object

Analyzing the organizational structure of this academic journal article and performing granularity division based on its internal logical structure, this study represents the granularity division as a tree diagram according to the annotation framework constructed in Section 4.1, as shown in Figure 7 [Figure 7: see original paper].

After granularity division, each granularity unit requires unique numbering for subsequent annotation work. This method adopts a composite numbering approach in the hierarchical structure, assigning each granularity unit an "ABC" code where A is the full document number, B is the chapter number in document A, and C is the paragraph number in chapter B, progressing layer by layer to form a unique number for each granularity unit. The specific representation of A, B, and C can be set according to actual needs. In this empirical study, A, B, and C all use two-digit decimal numbers based on structural analysis of the article. For example, the position of the second paragraph in chapter one is represented as "010102."

Specifically, "title" positions are represented by "00"—for instance, the title position of chapter two is "010200." Full-text titles, keywords, and other positions not belonging to any chapter (i.e., empty at a certain division level) are padded with "00," such as the full-text title position "010000" and keywords position "010001."

The hierarchical composite numbering approach assigns a unique number to each granularity unit and facilitates reading and annotation of any granularity unit content in subsequent annotation work. For example, when annotating medium granularity, only the "B" position in the numbering needs to be distinguished.

Multi-Granularity Segmentation of Experimental Object

After completing granularity division of the academic journal article, each granularity unit must be segmented. This study uses the NLPIR segmentation tool to segment content of each granularity unit. The tool provides multiple segmentation methods and available part-of-speech tag sets. After preliminary comparison of segmentation results using test texts, this empirical study selects

its maximum matching segmentation method and ICTPOS first-level part-of-speech tagging set.

After segmenting the academic journal article, stopwords must be removed from segmentation results to obtain annotation candidate terms, assign position information to each term, and export in EXCEL format for the next annotation term selection calculation. The stopword list uses the extended version published by Harbin Institute of Technology. After stopword removal, partial annotation candidate terms are shown in Table 3 .

Table 3 Partial Annotation Candidate Terms (Coarse Granularity)

[Table content would be preserved here with original data]

Multi-Granularity Annotation Term Selection Calculation for Experimental Object

According to the annotation candidate term importance evaluation metrics, calculate each metric score for each candidate term in each granularity unit text using the methods in Section 4.3, ultimately obtaining the total importance score for each annotation candidate term. Notably, in this empirical study, since it is difficult to obtain all documents in the retrieval system, the TF-IDF value in coarse granularity is temporarily replaced by TF value for calculating candidate term importance.

Calculate importance scores for each annotation candidate term sequentially from fine to coarse granularity levels. After sorting scores in descending order, select an appropriate number of candidate terms as annotation terms based on needs. Considering that the number of annotation terms should match text length, fine-grained text units generally select the highest-scoring term as the annotation term, medium-grained text units select the top three highest-scoring terms, and coarse-grained text units select the top five highest-scoring terms.

Taking medium-grained annotation units as an example, partial candidate term importance calculation results are shown in Table 4 .

Table 4 Partial Candidate Term Importance Score Calculation Results

[Table content would be preserved here with original data]

From the data in Table 4, the annotation terms for chapter one at medium granularity are “economy,” “data governance,” and “data sharing.”

Multi-Granularity Annotation Results of Experimental Object

After completing importance score calculation for annotation candidate terms at each granularity unit, select an appropriate number of candidate terms as annotation terms for each granularity unit. In this study, fine-grained text units select the highest-scoring term, medium-grained text units select the top

three highest-scoring terms, and coarse-grained text units select the top five highest-scoring terms. The annotation results for each granularity unit of this academic journal are shown in Table 5 .

Table 5 Multi-Granularity Annotation Term Selection Results

[Table content would be preserved here with original data]

Using SKOS vocabulary to describe these annotation terms and organizing them with XML-structured documents yields the final XML annotation document for the academic journal. A partial annotation document example is shown in Figure 8 [Figure 8: see original paper].

Figure 8 [Figure 8: see original paper] Partial XML Document Example of Multi-Granularity Semantic Annotation for Academic Journal Article Based on SKOS

Experimental Result Evaluation

Since no multi-granularity annotation applications currently exist, this empirical study can only use the annotation results of this article in current retrieval systems as the control group. As the two annotation results cannot completely correspond, this evaluation process conducts qualitative analysis of retrieval effects for comparative assessment.

This evaluation borrows the concepts of "recall" and "precision" to theoretically analyze and compare multi-granularity semantic annotation effects with currently used annotation results under the same retrieval expressions, identifying potential impacts on retrieval system recall and precision rates. Additionally, internal feature retrieval access points provided to users based on annotation results and retrieval feedback forms also constitute important functions affecting user retrieval results and utilization effectiveness, serving as reference evaluation metrics for multi-granularity annotation results.

Using the article's retrieval status in CNKI as the reference group, this study comparatively analyzes the retrieval performance of SKOS-based multi-granularity annotation results to evaluate the method's effectiveness. Specific comparison results are shown in Table 6 .

Table 6 Retrieval Performance Evaluation of Multi-Granularity Annotation Results for Academic Journal Articles Based on SKOS

Evaluation Dimension	CNKI Control Group	Multi-Granularity Annotation
Internal Feature Retrieval Access Points	Title, keywords, abstract, full-text	Full-text theme, chapter theme, paragraph theme
Retrieval Result Feedback Form	Document only	Document, chapter, paragraph

Evaluation Dimension	CNKI Control Group	Multi-Granularity Annotation
Impact on Recall	Depends on user query expansion skills	Automatic expansion through semantic relationships
Impact on Precision	Literal matching	More accurate content retrieval

1. **Internal Feature Retrieval Access Points**: Current CNKI retrieval systems generally provide four internal feature access points: "title, keywords, abstract, full-text" with direct natural language matching. CNKI's "theme" access point is a combination of "title, keywords, abstract" and thus not considered a separate access point. Multi-granularity semantic annotation provides three granularity-level content access points: full-text theme, chapter theme, and paragraph theme. The two systems offer different retrieval access point systems that complement each other. While difficult to determine which is superior, comparing roughly equivalent access points—CNKI's title access point versus coarse-grained access point—reveals that title matching retrieves the article using core terms from the title "Data Governance: Development Opportunities for Library Undertakings," such as "data governance," "library," "library undertakings," and "development opportunities." In multi-granularity annotation retrieval, using corresponding coarse-grained annotation terms "data governance," "data sharing," "data reuse," "data value-added," and "open data" as search terms can also retrieve the article. The former uses title core terms directly as the article's theme, while the latter obtains the article's theme through bottom-up multi-granularity annotation, achieving expansion of core themes and richer thematic description of the full text, though losing limiting themes like "library undertakings." This result has both advantages and disadvantages; in actual retrieval processes, lost thematic information can be supplemented through medium and fine-grained annotation results.

2. **Retrieval Result Feedback Form**: Currently, CNKI can only feedback document-level (coarse-grained) content to retrieval users. Multi-granularity semantic annotation results can organize documents in different hierarchical structures, presenting retrieval results of different granularity sizes simultaneously in user retrieval feedback—complete documents, chapters related to retrieval themes, or even specific paragraphs related to retrieval themes. Users can select or filter based on personalized information needs, directly retaining retrieval results of a certain granularity.

3. **Impact on Recall**: Currently, document annotation results in CNKI cannot directly improve retrieval recall; recall assurance mainly depends on retrieval users' skills in query construction, such as expansion with syn-

onyms and related terms. SKOS-based multi-granularity semantic annotation results already possess inter-term relationships from the thesaurus, enabling retrieval range expansion through equivalence, hierarchical, and associative relationships without increasing users' cognitive burden.

4. **Impact on Precision**: Although difficult to determine the impact of the two annotation results on retrieval system precision, and although precision is conceptually inverse to recall (precision often decreases when recall significantly improves), considering only theme-relevant content in retrieval results, multi-granularity semantic annotation processes chapter and paragraph information within documents rigorously. Compared with simple literal full-text matching, content retrieval is clearly more accurate.

However, SKOS-based multi-granularity semantic annotation for academic journals also inevitably brings disadvantages. Complete reliance on the thesaurus for annotation inevitably creates mismatches between natural language and controlled vocabulary. Thesauri only present simple inter-concept relationships based on disciplines, unable to reveal personalized and rich conceptual connections, resulting in far fewer usable semantic relationships than professionally constructed domain ontologies, sometimes even increasing irrelevant associative relationships and causing annotation result redundancy. Additionally, multi-granularity annotation results are significantly more numerous than single-granularity results, inevitably making information processing, storage, and matching calculations in retrieval systems more complex and creating new problems, thereby increasing system burden.

Conclusion

This study proposes a SKOS-based multi-granularity semantic annotation method for academic journal text resources and conducts empirical research based on semantic annotation theories and SKOS-related technologies. The method offers two main advantages: (1) SKOS is currently the optimal scheme for thesaurus description. Compared with RDF and OWL languages, SKOS provides more flexible and standardized description of concepts and relationships with simpler maintenance operations, enabling multi-level retrieval and automatic retrieval expansion based on thesauri. (2) Multi-granularity semantic annotation can satisfy users' needs for knowledge units of different granularities, richly revealing semantic relationships between different documents at the same granularity level and between different granularity levels within the same document.

This study is only a preliminary attempt at multi-granularity semantic annotation of academic journal text resources using SKOS-ified thesauri, with several issues warranting further research: The method uses only a single thesaurus as the annotation tool; future work could attempt using multiple thesauri for semantic annotation, involving issues such as heterogeneity between different thesauri, mapping of concepts and semantic relationships across different sys-

tems, and priority issues in thesaurus selection during annotation. However, using multiple thesauri for semantic annotation can obviously enrich annotation content and solve current problems where some concepts and relationships are difficult to describe. Furthermore, future applications of SKOS-based multi-granularity semantic annotation for academic journal text cannot be separated from the development of related tools, systems, and platforms. This study is merely an attempt, and the empirical stage only tested the method on a small sample. Future work must develop relevant tool systems based on further method refinement to promote practical application.

References

[1] Yu Yiwen, Chen Aiping, Zhao Huixiang. Preliminary exploration of academic journal development based on semantic web[J]. Chinese Journal of Scientific and Technical Periodicals, 2013, 24(5): 954-956.

[2] Qiu Junping, Mou Nan, Lou Wen, et al. Analysis of research progress on semantic annotation at home and abroad[J]. Information Studies: Theory & Application, 2014, 37(5): 12-16.

[3] W3C. Introduction to SKOS[EB/OL]. [2017-08-03]. https://www.w3.org/2004/02/skos/intro.html.

[4] W3C. Datasets[EB/OL]. [2017-08-03]. https://www.w3.org/2001/sw/wiki/SKOS/Datasets.

[5] Fan Wei. Case study of thesaurus in semantic web environment: Using SKOS to construct machine-understandable knowledge organization systems[J]. Information Science, 2006(7): 1073-1077.

[6] Jia Junzhi. Simple knowledge organization system and Chinese thesaurus[J]. Journal of Library Science in China, 2008, 34(1): 75-78.

[7] Zhang Shinan, Song Wen. Design of SKOS description scheme for Library Classification of Chinese Academy of Sciences[J]. New Technology of Library and Information Service, 2010, 26(6): 7-11.

[8] PASTOR-SANCHEZ J, MARTINEZ J, RODRIGUEZ-MUNOZ V. Advantages of thesaurus representation using the simple knowledge organization system (SKOS) compared with proposed alternatives[J]. Information research: an international electronic journal, 2009, 14(4): 422-432.

[9] Wang Qian, Tao Lan, Wang Bizuo. Knowledge organization model based on SKOS in semantic Web[J]. Computer Engineering and Design, 2007, 28(6): 1441-1443.

[10] Xiong Taichun. Subject indexing of network information resources based on SKOS[J]. Library Science Research, 2009(7): 63-66.

[11] Zhu Jiaxian, Bai Weihua, Li Jigui. Research on multi-granularity semantic annotation of Web resources and its application technology[J]. Computer Science, 2011, 38(8): 83-87.

[12] Wei Moji, Yu Tao. Domain ontology-based semantic annotation method for professional documents[J]. Journal of Computer Applications, 2011, 31(8): 2138-2142.

[13] Leng Fuhai, Bai Rujiang, Zhu Qingsong. Research on hybrid semantic information extraction method for scientific and technical literature[J]. Library and Information Service, 2013, 57(11): 112-119.

[14] CUNNINGHAM H, MAYNARD D, BONTCHEVA K, et al. A framework and graphical development environment for robust NLP tools and applications[C]//Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. DBLP, 2002: 168-175.

[15] DBpedia. DBpedia home[EB/OL]. [2017-08-05]. http://wiki.dbpedia.org/.

[16] FERNANDEZ N, FISTEUS J A, FUENTES D, et al. A Wikipedia-based framework for collaborative semantic annotation[J]. International journal on artificial intelligence tools, 2011, 20(5): 847-867.

[17] Tang Yijie, Zhang Min, Ding Xiaoqin. Research on semantic implementation method of integrated information platform based on linked data[J]. Journal of Modern Information, 2016, 36(6): 66-73.

[18] Xu Xukan, Fang Daowei, Jiang Xun, et al. Research on knowledge granular representation and standardization in knowledge organization[J]. Documentation, Information & Knowledge, 2014(6): 101-106.

[19] Gu Liping. Data governance: Development opportunities for library undertakings[J]. Journal of Library Science in China, 2016, 42(5): 40-56.

Author Contributions

Xia Lixin: Responsible for topic proposal and idea generation, provided revision suggestions for paper writing.
Zheng Lu: Responsible for paper framework design, literature collection and writing, paper revision.
Zhang Yuchen: Responsible for literature collection and paper writing.
Zhai Shanshan: Participated in paper framework design and paper revision.
Sun Jingqiong: Participated in paper framework design and manuscript proofreading.

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