Design and Practice of a Semantic Enrichment Framework for Scientific Literature Retrieval Systems

Authors: Xie Jing, Wang Jingdong, Wu Zhenxin, Zhang Zhixiong, Wang Ying, Ye Zhifei

Date: 2017-08-21T00:00:00+00:00

Abstract

[Purpose/Significance] This paper aims to enhance the service functions and effectiveness of scientific literature retrieval systems by employing technical methods such as data mining, semantic recognition, and knowledge relationship computation, thereby enabling the presentation of richer knowledge-based semantic information and revealing more knowledge points and knowledge relationships to users. [Method/Process] This study applies data mining and relationship computation tools including semrap and clausIE to identify and extract semantic objects from scientific literature, analyze, compute, and construct semantic relationships, and designs and establishes a multi-dimensional semantic index tree based on the obtained semantic objects and relationships, thereby creating a new data organization and presentation model. [Results/Conclusion] A semantic enrichment retrieval demonstration system was developed, which fully reveals semantic information within scientific literature retrieval systems, providing users with enhanced navigation, association, discovery, and revelation at the knowledge content level, while also analyzing the advantages and disadvantages of the designed model.

Full Text

Preamble

Design of a Semantic Enrichment Framework for Scientific Literature Retrieval Systems

Xie Jing[1], Wang Jingdong[1], Wu Zhenxin[1], Zhang Zhixiong[1], Wang Ying[1], Ye Zhifei[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190)

Abstract

[**Objective/Significance**] This paper aims to enhance the service functions and effectiveness of scientific literature retrieval systems by employing data mining, semantic recognition, and knowledge relationship computation techniques. The goal is to present richer knowledge-based semantic information, revealing more knowledge points and relationships to users. [**Methods/Process**] This study utilizes the SemRep and ClausIE data mining and relationship computation tools to identify and extract semantic objects from scientific literature, analyze and construct semantic relationships, and design a multi-dimensional semantic index tree based on these objects and relationships, proposing a new data organization and presentation model. [**Results/Conclusions**] A semantic enrichment retrieval demonstration system was developed to fully reveal semantic information in scientific literature retrieval systems, providing users with enhanced navigation, association, discovery, and disclosure at the knowledge content level. The paper also analyzes the advantages and limitations of the proposed design model.

Keywords: Semantic Enrichment; Semantic Knowledge Organization; Multi-dimensional Index; Semantic Relationship Presentation
Classification Number: TP391

With the continuous development and application of data mining, semantic recognition, and knowledge relationship computation technologies in scientific literature, users increasingly expect retrieval systems to present richer semantic content and reveal more knowledge points and relationships. The design objective of the semantic enrichment framework is to disclose various types of semantic knowledge objects and their rich interconnections within scientific literature retrieval systems. Building upon data mining and semantic recognition technologies, this framework moves beyond existing keyword-oriented retrieval systems to reorganize semantically enriched data and present semantic knowledge and association information.

In the experimental design of this semantic enrichment framework, we selected article collections from PubMed [1] in the medical domain covering two topics —Migraine Disorder and Heart Diseases—from the past two years as the demonstration system' s test dataset. We employed the mature data mining and analysis tools SemRep and ClausIE as foundational instruments and designed a multi-dimensional semantic data organization index model to develop a retrieval demonstration system for exploring semantic enrichment in scientific literature retrieval.

1 Semantic Enrichment Overall Framework Design

[Figure 1: see original paper]

As shown in Figure 1, the semantic enrichment of scientific literature retrieval systems primarily involves two components:

(1) Semantic Annotation: This involves deep indexing of knowledge objects appearing in the literature. First, keywords are extracted from the documents, and the types of knowledge objects are identified (i.e., what they are). Second, the associative relationships among these knowledge objects are computed. Semantic annotation 主要包括: literature clue indexing, content semantic indexing, semantic relationship extraction, and syntactic relationship extraction.

- **Literature Clue Indexing:** This includes not only metadata indexing such as author, publisher, and publication year, but also segmenting abstract text data into sentences and paragraphs with independent content meaning, such as identifying research objectives, methods, tools, and results within the literature.
- **Content Semantic Indexing:** This involves indexing content elements in academic papers including problems, theories, methods, technical means, tools, models, and conclusions.
- **Semantic Relationship Extraction:** Based on knowledge objects indexed within the same sentence, possible semantic relationships between each pair of objects are queried using the UMLS and STKOS knowledge organization systems [2], and these discovered relationships are recorded as SPO triples [3].
- **Syntactic Relationship Extraction:** Through syntactic relationship computation [5], long sentences are split into shorter sentences and clauses, within which subject-predicate-object relationships are identified and recorded as SPO triples. A single sentence may be decomposed into multiple SPO triples.

(2) Semantic Indexing: This involves constructing a multi-dimensional semantic index system based on the extracted and annotated content, organizing semantic knowledge organically to facilitate use by the semantic retrieval platform.

- **Document Indexing:** The document index layer indexes metadata descriptions including article titles, authors, and publication dates. The sentence and paragraph layer index segments abstracts into paragraphs and sentences for indexing.
- **Knowledge Object Indexing:** Terms and entities identified and indexed from text are collectively referred to as knowledge objects [6]. According to the design requirements of the semantic enrichment demonstration system, knowledge object indexing comprises two parts: knowledge object indexing and knowledge object attribute indexing. Knowledge object indexing converts user-input keywords into semantic objects, serving a normalization function. Knowledge object attribute indexing enables retrieval and classification display of various attributes of knowledge objects.
- **Knowledge Object Relationship Indexing:** Semantic relationships and syntactic relationships computed during semantic annotation are collectively referred to as knowledge object relationships. Both are expressed

as SPO triples and indexed to enable retrieval and relationship revelation of knowledge object semantic relationships.

2 Semantic Annotation Function Design and Implementation

[Figure 2: see original paper]

Semantic annotation was performed on the titles and abstracts of selected literature data. Referencing UMLS and STKOS, semantic objects in the medical domain were categorized into 15 major categories and 134 subcategories for using SemRep and MetaMap tools to extract important semantic objects from the literature. MetaMap and ClausIE were used to compute and identify semantic relationships. The workflow is shown in Figure 2, where the left process represents semantic object annotation of text content through SemRep. The annotated experimental data undergoes normalization mapping and correction according to the selected 15 major categories and 134 subcategories. Annotation tasks include:

(1) **In-depth Content Object Annotation:** Based on large-scale scientific literature (200,000 records) and the STKOS ontology and UMLS thesaurus, SemRep and MetaMap tools were used to extract important content and objects across 15 major categories and 134 subcategories.

(2) **Knowledge Object Annotation:** After annotation and normalization, 4,935 knowledge objects were obtained (200,000 unnormalized objects).

2.2 Data Semantic Organization and Standardization

As shown in Figure 2, the right process represents the computation, organization, and standardization of object relationships. MetaMap tools compute and identify standardized semantic relationships, categorizing semantic object relationships into 30 standardized relationships. ClausIE tools identify syntactic tree relationships. Semantic relationship data identified by both MetaMap and ClausIE are merged and integrated, with experimental data normalized and corrected according to the 30 standardized relationships selected from MetaMap.

Completed data organization and standardization tasks include:

(1) **Semantic Relationship Annotation Within Documents:** Using SemRep and MetaMap tools to extract 30 types of semantic relationships and paragraph relationships from scientific literature, 挖掘知识对象之间潜在的语义关系.

(2) **Syntactic Relationship Annotation of Document Content:** Using ClausIE tools to extract syntactic relationships (SPO) from scientific literature, discovering potential associative relationships among knowledge objects (keywords and terms).

(3) **Integration of Semantic and Syntactic Relationship Annotation:** From experimental data of 1,116 document abstracts, 50,204 SPO relationships were extracted, including 41,590 semantic relationships and 8,614 syntactic relationships.

2.3 Key Problem Solutions

(1) Mapping Annotation Content to MeSH Thesaurus

SemRep processing results appear as follows:

```
SE|00000000||tx|1|entity|C0006142|Malignant neoplasm of breast|neop|Breast cancer|1000|1|13
```

The meaning is as follows:

SemRep Tag	Entity	C0004142	Malignant neoplasm of breast	Breast cancer
Data Field Meaning	SemRep tool marker	Text source marker	Entity position in text	MeSH thesaurus term code

The red field (e.g., "neop"in the example) represents abbreviations for 134 subcategory semantic relationships. This study has collected the English full names, English abbreviations, and Chinese names corresponding to the 134 subcategories and 15 major categories in the MeSH thesaurus. By associating through the red field, a mapping relationship is established between text-identified terms and the MeSH thesaurus, solving the correspondence problem between SemRep processing results and the 15 major categories and 134 subcategories.

(2) Correspondence Between ClausIE-Extracted Subjects (S), Predicates (P) and UMLS Thesaurus

ClausIE extracts triples based on syntactic relationships, so the extracted entities cannot completely match those from SemRep. Meanwhile, SemRep only extracts semantic verbs, with other verbs being ignored. For the first case, fuzzy matching is considered to ensure entity correspondence. For the second case, verbs are extracted from MetaMap for matching, ensuring the standardization and consistency of experimental data.

3.1 Semantic Index Basic Functions

[Figure 3: see original paper]

The semantic indexing design aims to reveal semantic objects and their multiple interrelationships, moving beyond current single-dimensional indexing approaches. Multiple index trees work collaboratively to present semantic content from multiple dimensions.

As shown in Figure 3, semantic indexing centers on knowledge objects and follows the user workflow. Starting from retrieval keywords, knowledge object indexes perform semantic recognition and disambiguation of input terms. Then, through knowledge object relationship indexes, the knowledge network is traversed to navigate and filter required associated knowledge. Bridge indexes locate the sentences and paragraphs containing the knowledge objects. Finally, document indexes query and display literature information containing relevant knowledge content. Based on these four steps, the index is divided into four functional components:

(1) Knowledge Object Index - **Knowledge Object Index:** Indexes full names, abbreviations, and aliases of knowledge objects, converting user-input keywords into relevant knowledge objects to achieve semantic retrieval transformation. - **Knowledge Object Attribute Index:** Retrieves and displays various attributes of knowledge objects, identifies keywords with semantic conflicts, and implements semantic disambiguation.

(2) Semantic Relationship Index - **Knowledge Object Semantic Relationship Index:** Indexes semantic relationships among knowledge objects appearing in text (semantic relationships are standardized associations from UMLS or STKOS), enabling retrieval and analytical display of semantic relationships. - **Knowledge Object Syntactic Relationship Index:** Indexes syntactic relationships among knowledge objects appearing in text (syntactic relationships are unstandardized associations from NLP parsing), used to distinguish retrieval and analytical display of semantic versus syntactic relationships.

(3) Bridge Index - **Object-Document Relationship Index:** Maps knowledge objects to their locations in documents and reveals co-occurrence relationships of semantic knowledge objects. - **Object-Paragraph Relationship Index:** Maps knowledge objects to their paragraph locations and reveals co-paragraph relationships of semantic knowledge objects. - **Object-Sentence Relationship Index:** Maps knowledge objects to their sentence locations and reveals co-sentence relationships of semantic knowledge objects.

(4) Document Index - **Metadata Index:** Indexes document metadata including title, author, publication year, etc., for displaying basic literature information. - **Document Content Index:** Indexes article abstracts (or full text) for content display and highlighting of relevant knowledge objects and relationships.

The experiment implemented indexing of 1,116 documents, 4,023 paragraphs, 7,684 sentences, and 4,935 standardized knowledge objects, with 50,204 knowledge relationships indexed.

3.2 Key Problems and Solutions

(1) Mapping Input Keywords to Standardized Knowledge Objects

During the experiment, input keywords might not perfectly match indexed

knowledge objects, preventing accurate mapping. Additionally, a single keyword might have multiple meanings, causing semantic ambiguity and preventing clear mapping to specific knowledge objects.

For the first issue, this study employs fuzzy matching, selecting the knowledge object with the highest matching score and displaying the top 5 matches to users for correction and semantic recognition. For the second issue, users are presented with different meanings of the knowledge object to select from, achieving semantic disambiguation. Future work may consider using user behavior context for intelligent semantic disambiguation.

(2) Statistical Revelation of Knowledge Object Associations

Knowledge object relationships are indexed as S-P-O triples in Apache Solr. To facilitate relationship analysis, redundant fields are added to the triple index. The indexing method establishes facets on subjects (S) with (PO) pairs and facets on objects (O) with (SP) pairs. Using Solr' s faceting and frequency statistics, when retrieving knowledge objects, facets on (PO) and (SP) can reveal the most frequent syntactic and semantic relationships in retrieval results, helping users discover potential knowledge associations.

4 Data Organization for Semantic Enrichment Experimental System

[Figure 4: see original paper]

To implement the semantic enrichment retrieval demonstration platform, the system organizes data across four dimensions, as shown in Figure 4. The first dimension is the abstract layer, revealing article titles, authors, publication dates, and other metadata. The second dimension is the sentence and paragraph layer, segmenting articles into paragraphs and sentences for expression revelation. The third dimension is the fact layer, representing semantic segmentation of sentences to reveal grammatical and syntactic relationships computed from knowledge objects. The fourth dimension is the knowledge object layer, revealing knowledge objects (terms and entities) identified in text and their attributes.

From a bottom-up perspective, the third and fourth dimensions decompose scientific literature into knowledge objects and their associations, forming a scientific knowledge network for semantic querying and associative navigation. The first and second dimensions, combined with document paragraphs and sentences, locate where knowledge exists in scientific literature, facilitating detailed associative reading for users.

5 Semantic Enrichment Demonstration Platform

The semantic enrichment demonstration platform is designed around users' knowledge application needs. The typical user retrieval workflow involves in-

putting keywords, displaying knowledge relationships, navigating to deeper specific knowledge points through associations, and viewing the specific articles containing the knowledge. The platform implements four functions:

(1) **Digital Object Semantic Recognition and Retrieval:** Identifies semantic objects from user-input keywords and performs semantic retrieval.
(2) **Retrieval Result Knowledge Relationship Revelation:** Displays surrounding knowledge relationship networks for retrieval content, revealing the full knowledge landscape.
(3) **Semantic Relationship Association Navigation:** Navigates to deeper specific knowledge points based on semantic relationships, filtering more precise retrieval results through semantic associative navigation.
(4) **Semantic Reading of Specific Articles:** Views specific articles containing the knowledge, with highlighted display of knowledge points and relationships to assist reading.

5.1 Digital Object Semantic Recognition and Retrieval Implementation

[Figure 5: see original paper]

The demonstration system can identify corresponding semantic objects based on user-input keywords and display relevant explanations. As shown in Figure 5, when the retrieval keyword "headache" is entered, the system identifies semantic objects related to headache, which belong to the category of "Signs or Symptoms." It also provides encyclopedia entries and related images about Headache.

Compared with traditional literature retrieval, this function' s advantage lies in standardizing user input, transforming fuzzy keyword matching into semantic object retrieval with semantic features, thereby making semantic enrichment retrieval more precise. The semantic recognition function can also indicate the type (or category) of semantic objects, assisting users in semantic disambiguation and avoiding semantic deviations common in traditional keyword retrieval.

5.2 Knowledge Relationship Revelation in Retrieval Results

[Figure 7: see original paper]

The knowledge relationship revelation function graphically displays involved knowledge objects, semantic relationships between knowledge elements, and article fragments (sentences, paragraphs, etc.) containing the knowledge, as shown in Figure 7. These knowledge elements and associations are presented as graph nodes and edges, with different colored nodes representing different types of knowledge objects and edges representing semantic relationships. Users can discover needed knowledge through click-associate-navigate interactions.

The demonstration system clearly displays hit knowledge relationships and their locations in articles, sentences, and paragraphs. Revealing important sentences

and knowledge relationships greatly helps researchers judge whether content meets their retrieval needs. This study argues that using relevant knowledge objects, facts, sentences, and paragraphs for retrieval is more helpful for precise semantic knowledge discovery than full-text retrieval. Users can click on semantic knowledge objects, sentences, or paragraphs to view the full text through associative links.

5.3 Semantic Relationship Association Navigation

[Figure 6: see original paper]

The semantic association navigation function matches semantic objects based on retrieval input and statistically identifies co-occurring, co-paragraph, and co-sentence semantic objects in retrieval result documents. It enables navigational browsing of associated semantic objects, helping researchers discover valuable content from potential associations and filter relevant literature. As shown in the upper part of Figure 6, when querying "Headache," co-occurrence relationships, co-sentence relationships, and co-paragraph relationships reveal "Migraine Disorders" and "Clinical Research," which may provide insights for researchers.

Similarly, SPO semantic relationships and syntactic relationships are revealed through faceting, using predicate-object (knowledge object) facet statistics to reveal potential semantic and syntactic relationships. As shown in the lower part of Figure 6, searching "Headache" can reveal articles on deep professional knowledge such as "Process of Child" treatment and "Process of Adolescent" treatment, as well as research papers on related therapeutic drugs (e.g., "followed epilepsy"), providing researchers with clear knowledge relationship inspiration and guidance.

The demonstration system' s semantic association and navigation functions, based on statistical data revelation, help discover implicit knowledge associations and potential new knowledge relationships, enabling exploration of new research points in interdisciplinary fields. This expands researchers' thinking and assists scientific and technological innovation.

5.4 Semantic Assisted Reading for Single Documents

[Figure 7: see original paper]

As shown in Figure 7, the semantic assisted reading function highlights and displays computed semantic objects and knowledge relationships when viewing a single document. The tree list on the left displays semantic knowledge objects identified in the document, grouped by type and marked with different colors. The central main section shows the document' s abstract information. When a specific type of knowledge object is selected, its occurrences in the abstract are highlighted in the corresponding color for easy reference. The right side displays computed semantic and syntactic relationships from the document, also allowing users to view specific locations in the text.

This semantic assisted reading approach helps users directly view the most important knowledge points, locate their positions, and guides readers to prioritize reading paragraphs and sentences containing key knowledge, thereby improving reading efficiency for full-text content.

Conclusion

Based on a medical domain dataset from PubMed and employing mature data mining and knowledge relationship computation tools, this study proposes a design model for a semantic enrichment framework and demonstrates its advantages and feasibility through a demonstration system. Overall, this research significantly improves semantic literature retrieval effectiveness in several aspects:

(1) **Semantic recognition technology** transforms fuzzy keyword matching into knowledge object retrieval with semantic features, improving retrieval precision. It assists users in semantic disambiguation, avoiding semantic deviations common in traditional keyword retrieval.

(2) **Using knowledge objects, factual relationships, and sentences** for more precise semantic knowledge retrieval better helps researchers judge content relevance. Semantic association locates the actual knowledge within full-text documents and paragraphs.

(3) **Semantic association navigation functions** reveal potential associated knowledge through statistical methods, helping researchers discover new knowledge, explore new research points in interdisciplinary fields, expand research thinking, and assist scientific and technological innovation.

(4) **Semantic assisted reading** highlights important knowledge points, guiding readers to prioritize key paragraphs and sentences, thereby improving full-text reading efficiency.

During the experiment, several issues and limitations were identified for future improvement:

(1) SPO triple relationships obtained through ClausIE syntactic analysis are unstandardized. While this study has used domain dictionaries to standardize subjects (S) and predicates (P) secondarily, predicate (P) standardization remains incomplete. The unstandardized predicates in experimental data are somewhat chaotic, impacting association navigation discovery. Future work will construct predicate standardization thesauri or develop predicate semantic recognition methods for improvement.

(2) Computed semantic relationships between knowledge objects frequently associate with broad hypernyms. Consequently, broad hypernyms appear frequently in knowledge relationship revelation, but most are not helpful for domain researchers. Future work will compute knowledge object weights using TF/IDF methods to filter out frequently occurring yet overly

broad hypernyms, thereby improving knowledge association navigation effectiveness.

References

[1] Pubmed [EB/OL]. [2015-10]. http://www.ncbi.nlm.nih.gov/pubmed
[2] Tan S, Zheng L. Methodology framework of knowledge organization system for scientific & technological literature. Library & Information. 2013;1:2-7.
[3] Semrep [EB/OL]. [2015-10]. https://semrep.nlm.nih.gov/
[4] Rindflesch, T.C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6):462-477.
[5] L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In Proceedings of the International World Wide Web Conference (WWW), 2013.

Author Contributions Statement:

Xie Jing: Designed the semantic enrichment retrieval model and demonstration system framework, developed the system, and primary author of the paper.
Wang Jingdong: Performed data semantic annotation and semantic relationship computation, author of the semantic annotation chapter.
Wu Zhenxin: Project coordination and management, organized paper structure, revised paper versions.
Zhang Zhixiong: Designed and guided the multi-dimensional indexing and semantic retrieval approach.
Wang Ying: Designed data organization and graphical display solutions for the demonstration system.
Ye Zhifei: Developed graphical display modules for the demonstration system.

Author E-mail: xiej@mail.las.ac.cn

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