Ontology-Based Semantic Analysis Methods for Research Topics: Postprint

Authors: Feng Jia, Zhang Yunqiu

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

Abstract

[Purpose/Significance] This study aims to analyze research topics at the deep semantic level. [Method/Process] An ontology-based semantic analysis method for research topics is proposed, developed along two dimensions of semantic types and semantic associations, and validated through empirical research using "medical informatics" as a case study. [Results/Conclusion] The results demonstrate that semantic type analysis can assist researchers in achieving deeper semantic understanding of research topic content; semantic association analysis examines the semantic relationships among research topics from a semantic perspective, enabling researchers analyzing topics in a given field to conduct comprehensive analyses of related topics and discover new research intersections.

Full Text

Preamble

Vol. 62 No. 7, April 2018
Research on Semantic Analysis Methods for Research Topics Based on Ontology
Feng Jia, Zhang Yunqiu
School of Public Health, Jilin University, Changchun 130021

Abstract

[Purpose/Significance] This study aims to analyze research topics at a deeper semantic level. **[Method/Process]** We propose a semantic analysis method for research topics based on ontology, examining both semantic types and semantic associations. In the empirical study, we validate the method using "medical informatics" as a case study. **[Result/Conclusion]** The results demonstrate that semantic type analysis can assist researchers in achieving a deeper semantic understanding of research topic content, while semantic association analysis examines the semantic relationships among topics from a conceptual perspective.

When analyzing research topics in a given field, this approach enables comprehensive analysis of related topics and discovery of new research intersections.

Keywords: research topic; semantic analysis; ontology
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.07.011

Introduction

A research topic typically refers to the main content discussed or investigated in an article. Research topics in a particular field reflect the direction of that field. Identifying research topics and understanding domain trends is of great significance to scientific researchers. In recent years, topic models have gained rapid popularity and been widely applied to topic extraction from various corpora, including academic literature, web texts, and social media resources. Topic modeling is a method for modeling latent topics in documents and can identify potential topics based on text corpora. Currently, topic modeling is extensively used in academic corpora to identify research topics. However, analysis of these topics often relies heavily on researchers' background knowledge, and due to variations in researchers' scientific literacy and knowledge backgrounds, the results tend to be highly subjective.

In terms of quantitative analysis, research topic analysis can be categorized into bibliometric methods and knowledge graph-based visualization methods. Bibliometric approaches typically analyze topics from perspectives such as temporal distribution, journal distribution, regional analysis, national distribution, and author distribution, combined with content analysis. For example, in 2014, Jing Fachong et al. used text mining methods to analyze and interpret the content of topics from the U.S. National Science Foundation. In 2009, Luan Chunjuan conducted a bibliometric analysis of papers and citations on international patentometrics published in *Scientometrics* between 1995 and 2007, mapping author co-citation networks, keyword co-occurrence networks, and author collaboration networks to visually represent key figures and research topics in international patentometrics research [2]. In 2013, Wei Xiaofeng employed knowledge graphs to analyze the evolution, hot topics, and frontiers of international information visualization research, conducting further analysis based on knowledge graphs [3]. In 2014, S.Y. Cheng used visualization techniques to identify and analyze research topics in the field of electronic government [4].

Currently, most research topic analyses rely on simple presentations of results, such as lists, matrices, or knowledge graphs. This study proposes mapping research topics to ontologies and leveraging the semantic types and structures of ontologies to conduct in-depth content-level analysis of research topics.

As a conceptual modeling tool that can describe information at the semantic and knowledge levels, an ontology provides a formal specification of a conceptual system. R. Studer et al. offer a clear definition: "An ontology is an explicit, formal, and shared specification of a conceptualization" [5]. "Explicit"

means that the types of concepts and their application constraints are clearly defined; "formal" means that the ontology can be processed by computers; and "shared" means that the ontology should represent a consensus on conceptualization within a community. Ontologies can be viewed as "abstractions and descriptions of domain knowledge norms, and methods for expressing, sharing, and reusing knowledge" [6].

The original purpose of ontology construction is to integrate relevant knowledge in a domain and provide a common understanding of domain concepts. By abstracting domain knowledge into a conceptual system represented by a vocabulary—including explicit definitions of each term, relationships between terms, and axiomatic knowledge statements within the domain—consensus can be achieved among domain experts, thus forming the ontology for that domain. This study aims to analyze research topics from semantic type and semantic association perspectives based on the conceptual structure of domain ontologies, seeking to conduct analysis and interpretation at a deeper semantic level.

2. Semantic Type Analysis of Research Topics

The semantic type of a vocabulary or concept can be understood as a conceptual attribute that describes and explains the concept. British linguist L. Geoffrey Leech introduced the concept of semantic types in his book *Semantics* [7]. Semantic types are distinguished based on language usage patterns from semantic and human communication perspectives. He categorized word meanings into seven types: conceptual meaning, connotative meaning, social meaning, affective meaning, reflective meaning, collective meaning, and thematic meaning. Semantic types abstract and generalize the framework, lexical units, and frame elements of a framework network, representing inherent, essential semantic features of semantic components that are independent of lexical context. Moreover, semantic types form a semantic type structure system through certain logical relationships, laying a solid foundation for ontology applications in natural language processing.

Current semantic type analysis methods mainly include semantic role labeling and ontology-based semantic type analysis. Semantic role labeling enables systematic analysis and interpretation of research content in scientific texts, enhancing the depth and accuracy of researchers' understanding. This method automatically labels semantic roles (such as agent, patient, time, and location) for predicates in sentences, focusing on shallow semantic analysis at the sentence level [8]. For example, in 2013, Zhang Zeyu et al. [9] proposed a semantic document annotation method based on ontology knowledge bases and WordNet semantic knowledge. However, in text mining of scientific literature, the focus is on analyzing semantic types of professional terms (nouns, verbs, etc.).

Ontology-based semantic type analysis maps words in text to concepts in an ontology and analyzes the semantic types of these concepts. An ontology is a complete structured conceptual system where each concept has associated se-

mantic types that describe and explain it. This forms the basis for semantic type analysis through ontologies. Ontology is a conceptual semantic representation method, with representative semantic dictionaries including WordNet and HowNet. Researchers have attempted to analyze semantic types of text based on ontologies; for instance, in 2007, Zhang Han et al. [10] used semantic types of concepts in UMLS to mine latent relationships between documents. Applying ontology to semantic type analysis provides theoretical support for text mining at the semantic level.

This study employs ontology-based semantic type analysis by mapping topic terms to concepts, transforming "topic word bags" into "concept word bags," and deeply mining the semantic types of concepts within these bags to enrich research topic analysis results.

3. Semantic Association Analysis of Research Topics

To further analyze semantic information of research topics, we propose using semantic distance to measure semantic similarity between topics. Semantic similarity reflects the conceptual and logical relationships between terms. From a semantic analysis perspective, this study analyzes the semantic relevance of research topics.

Semantic relationships between terms are embodied in domain ontologies. An ontology is a conceptual framework that provides a vocabulary to identify a set of concepts [11]. A domain ontology contains the conceptual structure of a domain, organizing concepts in a hierarchical structure. The basis for calculating conceptual semantic distance using a domain ontology is that two concepts have semantic relevance, meaning there exists a path between them in the ontology network. The conceptual relationship between terms—their relative positions in the ontology—can be measured by semantic distance. Therefore, by mapping co-occurring vocabulary to ontology terms through conceptual mapping, we can measure their conceptual relationships based on relative positions of terms in the ontology. Thus, ontology-based semantic distance can represent the intrinsic knowledge-based relationships between concepts.

Current research on semantic distance is relatively mature with abundant achievements. In addition to path length between terms, ontology-based semantic distance considers factors such as depth of the concept hierarchy tree and regional density of the concept hierarchy tree. For two terms with the same path length, those located at deeper levels of the hierarchy have greater semantic distance; for terms with the same path length, those in high-density regions of the concept hierarchy tree should have greater semantic distance than those in low-density regions.

This study selects semantic distance to measure the intrinsic knowledge-based relationships between different concepts. Semantic distance refers to the sum of edge weights on the shortest path between concepts in an ontology hierarchy tree [12], effectively representing conceptual similarity through geometric

measurement of relevance. Semantic distance is the most fundamental factor in measuring conceptual similarity, generally having greater impact than other factors [13].

Semantic distance is typically based on semantic dictionaries, which organize concepts in tree or network hierarchical structures, often represented as ontologies or thesauri. Using ontologies as an example, this method calculates similarity between two concepts based on their positions and distances in the ontology tree. Commonly used semantic distance algorithms include the Leacock-Chodorow method [14], Weighted Links method [15], and Wu and Palmer method [16].

This study selects the classic Leacock-Chodorow method to calculate semantic distance. The core idea of this algorithm is that concept similarity is related to the path length between concepts in the ontology hierarchy and the depth of the ontology hierarchy structure. The calculation formula is:

$$Sim(C_1, C_2) = -\log \frac{len(C_1, C_2)}{2 \times Depth}$$

In the above formula (1), $len(C_1, C_2)$ represents the shortest path length between concept terms C_1 and C_2 in the ontology tree, and $Depth$ represents the depth of the ontology tree.

The research approach for semantic association analysis of research topics is as follows: First, prepare data by obtaining topic word bags for research topics. Second, perform conceptual mapping of vocabulary in the word bags based on the domain ontology and 统计映射后的概念频次. Then, extract high-frequency concepts to construct a concept matrix, calculate semantic distances between concepts based on the domain ontology, and finally present and interpret the visualization results, as shown in [Figure 1: see original paper].

[Figure 1: see original paper] Semantic Distance Calculation Process

4. Empirical Analysis

Medical informatics is an emerging interdisciplinary field involving medicine, computer science, and information science. Currently, research results in medical informatics are continuously applied to clinical data analysis, drug management, disease modeling, and patient survival prognosis. Therefore, accurately identifying research topics in medical informatics helps strengthen strategic guidance for research management and provides important direction for researchers in this field. This study takes the field of "medical informatics" as an example, extracts topics from this domain, and conducts ontology-based topic analysis to verify the proposed method.

4.1 Corpus Construction

We selected the Web of Science Core Collection as the data source for the literature set. To comprehensively collect relevant literature in the medical informatics field, we used the "subject category" search function of the Web of Science Core Collection, which contains 252 subject categories, including "medical informatics." We collected literature from 2007 to 2016 under the "medical informatics" category, retrieving 35,981 records (downloaded on January 3, 2017). We then used the LDA model to extract research topics in this field, obtaining 19 major research topics in medical informatics, as listed in .

List of Research Topics in Medical Informatics

4.2 Semantic Type Analysis

Based on the research topics identified by the LDA method, we used MetaMap [17] to perform UMLS ontology mapping, converting topic terms that characterize research topics into knowledge concepts in the UMLS ontology. This process abstracts the semantics of these topic terms. After conceptual mapping of research topics in the medical informatics field, partial results are shown in .

We further 统计语义类型 to analyze the semantic types of different topics. lists the concepts and their semantic types for research topics in medical informatics. Based on the semantic type information in this field, the main categories can be divided into seven aspects:

1. **Conceptual category**: Including Conceptual Entity, Idea or Concept, Qualitative Concept, Quantitative Concept, Functional Concept, Spatial Concept, and Temporal Concept.
2. **Behavioral category**: Including Health Care Activity, Activity, Individual Behavior, Research Activity, Occupational Activity, Educational Activity, Mental Process, and Language.
3. **Population category**: Including Patient or Disabled Group, Population Group, and Professional or Occupational Group.
4. **Treatment and diagnosis category**: Including Therapeutic or Preventive Procedure, Finding, Clinical Attribute, and Diagnostic Procedure.
5. **Human function and phenomenon category**: Including Genetic Function, Organism Function, and Neoplastic Process.
6. **Material and device category**: Including Medical Device, Manufactured Object, and Research Device.
7. **Occupational category**: Including Occupation or Discipline and Biomedical Occupation or Discipline.

The semantic type matrix was visualized to clearly present the results. [Figure 2: see original paper] shows connections between topics and semantic types with a threshold greater than or equal to 2, where square nodes represent research topics, circular nodes represent semantic types, and the thickness of connecting lines represents association strength.

[Figure 2: see original paper] Semantic Type Graph

As shown in [Figure 2: see original paper], research topics in medical informatics share some common semantic types, such as intellectual product, health care activity, and functional concept. By analyzing topic content and semantic types, we can provide more information for interpreting research topics. For example, Topics 1 (tumor image analysis), 2 (medical applications of data mining algorithms), 4 (healthcare apps), 6 (clinical decision support), 18 (clinical knowledge semantic analysis), and 19 (medical data platform construction under big data background) primarily focus on "intellectual products" such as algorithms, models, standards, protocols, and technologies. Topic 5 (community health service research) mainly addresses "health care activities" such as telemedicine. Topic 7 (new medical models based on networks and computers) involves both "health care activities" like medical interventions and self-management, and "intellectual products" like technologies and methods in new medical models. Topic 15 (disease risk prediction) includes "body parts, organs, or components" such as heart and blood vessels, and "diseases or syndromes" such as coronary heart disease and heart failure, to analyze disease risks.

4.3 Semantic Association Analysis

For semantic distance calculation, we used the Leacock-Chodorow method based on the UMLS ontology to calculate semantic distances between concepts, utilizing the UMLS::Similarity [18] online system. To optimize visualization, we selected z-score [19] as the normalization method to obtain the semantic matrix.

shows the similarity matrix, where values represent similarity between different topics (higher values indicate greater similarity and closer semantic distance). A value of "1" indicates the same topic, while "0" indicates no semantic relevance between two topics.

To further analyze semantic distances between different topics, we visualized the semantic distance matrix, with results shown in [Figure 3: see original paper]. In [Figure 3: see original paper], topics are displayed with different numbers and sizes, where circle size represents relative topic scale and line thickness represents semantic association strength between topics.

[Figure 3: see original paper] Visualization Graph of Semantic Matrix

Through interpretation of the semantic association graph, researchers can comprehensively analyze different research topics. For example, in [Figure 3: see original paper], the two most strongly associated topics are Topic 17 (machine learning methods in healthcare) and Topic 4 (healthcare apps). By analyzing document-topic probability distributions and topic-word probability distributions, we can see from semantic content that healthcare apps monitor human ECG, EEG, and EMG signals through wearable devices, and combine data mining algorithms such as deep learning to achieve health data analysis and management.

lists the centrality measures of the research topic semantic matrix, which can be used to identify the research core and focus of the field and predict future research directions. According to , Topic 4 (healthcare apps) has the highest centrality in medical informatics research topics, indicating that this is currently the research focus of the field. Moreover, healthcare app research involves various research directions in medical informatics, such as Topic 17 (machine learning methods in healthcare), Topic 2 (data mining algorithms in medical field), Topic 13 (medical informatics methods and technology research), and Topic 9 (medical software development and application). Healthcare apps require integrating multiple methods and technologies from the medical informatics field, making the methods, technologies, and software development involved in healthcare app research the future direction of medical informatics.

Conclusion

This study proposes an ontology-based semantic analysis method for research topics from two dimensions: semantic type analysis and semantic association analysis. Semantic type analysis helps researchers achieve deeper semantic understanding of research topic content, while semantic association analysis examines the semantic relationships among topics, enabling comprehensive analysis of related topics and discovery of new research intersections.

This paper has made preliminary explorations in research topic analysis, focusing on concise analysis of semantic types and associations. Future work will further mine semantic information of research topics, integrating mature semantic analysis technologies to conduct more in-depth exploration of research topics in scientific literature, aiming to provide support for scientific innovation and decision-making.

References

[1] Jing Fachong, Li Chenying, Han Mingjie, et al. Topic analysis of emerging frontier projects in the Division of Biological Sciences of the U.S. NSF based on text mining [J]. Modern Information, 2014, 34(12): 107-112.

[2] Luan Chunjuan, Wang Xukun, Liu Zeyuan, et al. A quantitative analysis of international frontiers in patentometrics research [J]. Studies in Science of Science, 2008, 26(2): 334-338, 310.

[3] Wei Xiaofeng. Evolution, hotspots, and frontiers of international information visualization research based on knowledge graphs [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(5): 533-547.

[4] Cheng SY, Ding L. Mapping of electronic government: the trend of research fronts [C]//2014 seventh international joint conference on computational sciences and optimization. Piscataway: IEEE, 2014: 509-513. doi:10.1109/CSO.2014.100.

[5] Studer R, Benjamins VR, Fensel D. Knowledge engineering principles and methods [J]. Data and knowledge engineering, 1998, 25(1/2): 161-197.

[6] Liu Wei, Li Daling, Xia Cuijuan. Metadata and knowledge ontology [J]. Library Journal, 2004, 23(6): 50-54, 49.

[7] Leech G. Semantics [M]. Translated by Li Ruihua, Wang Tongfu, Yang Zijian, et al. Shanghai: Shanghai Foreign Language Education Press, 1987.

[8] Yang Xuanxuan, Zhang Lei. An information extraction model based on semantic roles and concept graphs [J]. Computer Applications, 2010, 30(2): 411-414.

[9] Zhang Zeyu, Li Li, Tan Feng, et al. Research on semantic-based document annotation methods [J]. Computer Engineering and Science, 2013, 35(9): 151-156.

[10] Zhang Han, Ren Zhiguo, Yu Qian, et al. Design and implementation of mining latent relationships between medical literature based on UMLS ontology [J]. New Technology of Library and Information Service, 2007, 2(9): 72-75.

[11] BoInoD, CornoF, PescarmonaF. Automatic learning of text-to-concept mappings exploiting WordNet-like lexical networks [C]//Proceedings of the 2005 ACM symposium on Applied computing. New York: ACM, 2005: 1639-1644. doi:10.1145/1066677.1067050.

[12] Xu Dezhi, Deng Chunhui, PASSIK. Research on concept semantic similarity based on SUMO [J]. Computer Applications, 2006, 26(1): 180-183.

[13] Tang Zhonglin. Research on ontology-based concept similarity calculation methods [D]. Wuhan: Wuhan University of Technology, 2013.

[14] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification [M]. Massachusetts: MIT Press, 1998: 265-283.

[15] Richardson R, Smeaton AF. Using WordNet in a knowledge-based approach to information retrieval [EB/OL]. [2017-06-15]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C38

[16] Wu Z, Palmer M. Verbs semantics and lexical selection [C]//Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 1994: 133-138. doi:10.3115/981732.981751.

[17] U.S. National Library of Medicine. Interactive MetaMap [EB/OL]. [2017-06-15]. http://ii.nlm.nih.gov/Interactive/UTS_{Required}/metamap.shtml.

[18] Pedersen T. UMLS::Similarity [EB/OL]. [2017-06-15]. http://atlas.ahc.umn.edu/cgi-bin/umls_{similarity}.cgi.

[19] CFOOL, MAFAUZY M. Does the use of mean or median Z-score of the thyroid volume indices provide a more precise description of the iodine deficiency disorder status of a population? [J]. European journal of endocrinology, 1999, 141(6): 557-560.

Author Contributions:
Feng Jia: Proposed research ideas and paper framework, conducted experiments and collected data, wrote and revised the paper.
Zhang Yunqiu: Determined the research topic, provided revision suggestions, and improved research content.

The final section regarding ProQuest and Taiwan Normal University is omitted as it constitutes promotional material unrelated to the academic content of the paper.

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