Computational Method and Experimental Postprint for Semantic Similarity of Linked Data Entities in Dunhuang Murals Thesaurus

Authors: Gao Jinsong, Fu Jiawei, Li Ke

Date: 2023-04-01T16:02:50+00:00

Abstract

[Purpose/Significance] With the rise of cultural heritage digitization and the humanities computing research paradigm, the demand for utilizing cultural heritage data resources has become increasingly prominent among humanities scholars participating in digital humanities research. Semantic fusion and interoperability of multi-source, heterogeneous cultural heritage information resources have become key issues in the construction of current digital humanities data infrastructure, and effective entity semantic similarity calculation methods have become an important means to achieve this goal. [Method/Process] Taking the Dunhuang murals thesaurus linked data as an example, and based on the analysis of the dataset's ontology model and data framework, this paper proposes an entity semantic similarity calculation method that combines multi-granularity matching and weighted operations according to its content distribution and structural characteristics. "Feitian" (flying apsaras) related entities in the Dunhuang murals thesaurus linked data are selected as experimental objects, and various existing entity semantic similarity calculation methods such as attribute features and edit distance are introduced for comparative experiments. [Results/Conclusion] Experimental results show that the entity semantic similarity calculation method based on multi-granularity matching proposed in this paper can better adapt to the content and structural characteristics of the Dunhuang murals thesaurus linked data, and demonstrates superior performance in terms of calculation result accuracy compared to similar methods. This represents another feasible approach for promoting data interconnection and knowledge sharing of heterogeneous humanities information resources under the background of digital humanities.

Full Text

A Method of Entity Semantic Similarity Calculation for Dunhuang Mural Thesaurus Linked Data with Experiment

Gao Jinsong[1], **Fu Jiawei**[1], **Li Ke**[2] [1]School of Information Management, Central China Normal University, Wuhan 430079 [2]Qingdao Hisense Hitachi Air Conditioning Marketing Co., Ltd., Qingdao 266510

Abstract: [Purpose/Significance] With the rise of cultural heritage digitization and humanities computing research paradigms, scholars in the humanities field have increasingly highlighted their need for utilizing cultural heritage data resources when participating in digital humanities research. The semantic integration and interoperability of multi-source, heterogeneous cultural heritage information resources have become critical issues in the construction of digital humanities data infrastructure, and effective entity semantic similarity calculation methods have emerged as essential means to achieve this goal. [Method/Process] Taking the Dunhuang Mural Thesaurus Linked Data as an example, this paper proposes an entity semantic similarity calculation method that combines multi-granularity matching with weighted operations based on analysis of its content distribution and structural characteristics. The study selects "Feitian" (flying apsaras) related entities in the Dunhuang Mural Thesaurus Linked Data as experimental objects and introduces multiple existing entity semantic similarity calculation methods, including attribute features and edit distance, for comparative experiments. [Result/Conclusion] Experimental results demonstrate that the proposed multi-granularity matching-based entity semantic similarity calculation method can better adapt to the content and structural characteristics of the Dunhuang Mural Thesaurus Linked Data, exhibiting superior performance in calculation accuracy compared to similar methods. This represents another feasible approach for promoting data interconnection and knowledge sharing of heterogeneous humanities information resources under the digital humanities framework.

Keywords: Dunhuang murals; linked data; multi-granularity; semantic similarity; entity similarity **Classification Number:** G254; TP391 **DOI:** 10.13266/j.issn.0252-3116.2021.08.010

With the rapid development of big data, machine learning technologies, and successful practices in cultural heritage digitization, Digital Humanities and Humanities Computing have become emerging research topics in the field of cultural heritage resource organization, attracting widespread attention from academia and industry. Digital Humanities and Humanities Computing introduce new thinking patterns for the digital preservation of cultural heritage under new technological conditions and enrich the ways traditional humanities scholars utilize cultural heritage data resources. In practice, this process is primarily achieved through the construction and publication of Linked Open Data, with representative cases including Europeana, MuseumFinland, Chinese Genealogy Linked Data, and Dunhuang Mural Thesaurus Linked Data.

To date, research on integrating cultural heritage information resources using linked data technologies has achieved initial progress domestically and internationally. Various cultural heritage preservation institutions, represented by museums, art galleries, and archives, have undertaken digitization initiatives based on physical collections, releasing substantial cultural heritage data resources online. This has significantly improved the data infrastructure for digital humanities research and satisfied humanities scholars' demands for fundamental data when participating in digital humanities research. After initially addressing data sourcing issues, the next stage of digital humanities data infrastructure construction should focus on high quality, broad domain coverage, and fine granularity. Against this backdrop of deepening research, promoting the aggregation and integration of multi-source, heterogeneous cultural heritage datasets has become a necessary step toward semantic, knowledge-based, and intelligent humanities information resource services, with effective semantic similarity calculation methods being one of the key technologies to accomplish this task.

This paper addresses the semantic integration and interoperability needs of cultural heritage domain data resources, proposes an entity semantic similarity calculation method based on multi-granularity matching, and uses "Feitian" related entities in the Dunhuang Mural Thesaurus Linked Data as a case study to explore the value and prospects of applying this method in the new stage of digital humanities infrastructure construction.

Related Research Status

The rapid rise of digital humanities and humanities computing paradigms has significantly increased humanities researchers' demand for high-quality, broad-domain, fine-grained data infrastructure. Existing research demonstrates that semantic web technologies such as ontologies, linked data, and knowledge graphs play crucial roles in transforming unstructured cultural heritage resources into structured semantic datasets. These technologies effectively support the structured integration of cultural heritage data resources across various themes, types, modalities, and unstructured formats to meet digital humanities research needs. Building upon these semantic datasets, achieving broader domain coverage, finer granularity, and higher quality multi-source integration and data interoperability has become a key concern in the new stage of digital humanities infrastructure construction. Consequently, developing effective dataset entity semantic similarity calculation methods has become critical to realizing this goal.

Entity semantic similarity calculation essentially quantifies the similarity relationship between a pair of named entities through specific numerical values. In recent years, with the increasing creation and publication of semantic datasets under the Linked Open Data standard, research on dataset semantic association discovery based on semantic similarity has grown domestically and internationally, yielding various entity semantic similarity calculation strategies. These

include similarity algorithms relying on external data such as domain ontologies and corpora, internal data-driven methods like association visualization and association rule mining, and data feature analysis-based approaches focusing on paths, attributes, and content that effectively support the quantification of semantic correlations between entities.

By horizontally comparing these entity semantic similarity calculation methods, we find that their primary differences lie in the choice of measurement granularity for semantic datasets. Granularity is a metric unit used to compare the coarseness of data, information, or knowledge, with its fineness depending on the depth of dataset refinement levels or the scale of partitioning patterns: deeper hierarchies and more patterns result in finer granularity, while shallower hierarchies and fewer patterns produce coarser granularity. Classifying mainstream dataset entity semantic similarity algorithms based on multi-granularity thinking reveals that coarse-grained methods primarily include data visualization tools such as sunburst charts, tree diagrams, and circular packing diagrams, as well as various entity similarity algorithms based on path distance. Medium-grained entity similarity calculation methods mainly include ontology-based attribute feature analysis and link predicate-based association rule discovery. Fine-grained entity similarity algorithms primarily achieve quantification of entity correlations in datasets by mining domain background knowledge or contextual information about entities.

As semantic linked data practices continue to deepen, the scale of entities contained in large-scale knowledge bases grows rapidly, while entity attribute features and annotation levels become increasingly refined. Entity semantic similarity calculation methods for semantic datasets are increasingly characterized by multi-granularity and multi-method integration. For example, Jia Limei et al. improved the accuracy of semantic similarity calculation by introducing dynamic weight-based semantic similarity algorithms and dynamic weighting mechanisms oriented toward attribute importance and value types on the basis of linked data attribute features. R. Meymandpour et al. proposed a context-based linked data similarity calculation strategy that comprehensively obtains attribute lists and value content of linked datasets through SPARQL queries and introduces corpus-based word vector models for semantic similarity calculation. Liu Xiaojuan et al. proposed an improved vector space model based on analysis of the implicit knowledge network characteristics of linked data, further enhancing linked data entity semantic similarity calculation accuracy through attribute weighting ideas. These studies also reflect that current dataset entity semantic similarity calculation methods are gradually shifting from algorithm technology-oriented to object requirement-oriented design thinking, placing greater emphasis on analyzing background knowledge of entity domains and model framework structures during method design. Through weighted operations, they integrate entity similarity calculation methods oriented toward different granularities, thereby ensuring and improving the accuracy and reliability of entity semantic similarity calculation results against the backdrop of rapid evolution in semantic dataset construction technologies.

Entity Semantic Similarity Calculation Method for Dunhuang Mural Thesaurus Linked Data

Basic Overview of Dunhuang Mural Thesaurus Linked Data

Dunhuang studies represent a special field in Chinese cultural heritage research, and Dunhuang murals are treasures of human cultural heritage with extremely high artistic and scientific research value. With the rise of cultural heritage digitization and digital humanities research, Dunhuang researchers have accumulated substantial firsthand information resources, providing important conditions for Dunhuang studies and the dissemination of Dunhuang murals. To explore the semantic information contained in Dunhuang mural resources and organize and express them effectively, domestic scholars have compiled the Dunhuang Mural Thesaurus and implemented its linked data publication using semantic web technologies. The core achievement of this research, the "Dunhuang Mural Thesaurus Linked Data," has become one of the representative practice cases in the field of cultural heritage semantic organization. The published Dunhuang Mural Thesaurus Linked Dataset contains over 4,500 semantic entities and more than 27,500 triples, covering 5 major facets, 25 secondary categories, and 3,896 controlled vocabulary terms of the Dunhuang Mural Thesaurus, providing effective data support for deep semantic annotation, semantic retrieval, knowledge organization, information association, and sharing of Dunhuang mural digital resources.

Semantic Description Granularity Analysis of Dunhuang Mural Thesaurus Linked Data

Linked data describes and organizes resources through RDF triples, where link predicates establish connections between head entities (Subjects) and tail entities (Objects) to describe attribute association relationships between different resources. In linked data publication practice, semantic entities describing specific resources are often composed of multiple triples. Due to different link predicates, the semantic description granularity of each triple within an entity often varies. In semantic similarity calculation processes, hierarchical relationships, logical relationships, and attribute parameters between entities all exert varying degrees of influence on similarity calculation results. Direct comparison of multiple triples with different granularities using a single calculation method often leads to semantic information loss and subsequent calculation errors. Therefore, in the entity semantic similarity calculation process for Dunhuang Mural Thesaurus Linked Data, it is necessary to reveal the semantic description granularity of different triple types by analyzing the composition of link predicates, thereby matching appropriate semantic similarity calculation methods for different granularity levels.

Ontology construction is a crucial component of linked data creation and publication. Ontology models define semantic relationships and hierarchical structures between resource entities by defining classes and class property relationships.

The Dunhuang Mural Thesaurus Ontology is built upon the logical structure of the Dunhuang Mural Thesaurus by reusing terminology elements from the GVP ontology, SKOS data model, and DCMI metadata standard. This ontology defines the hierarchical structure of the Dunhuang Mural Thesaurus Linked Data, providing a terminology framework for the semantic transformation and linked data publication of the Dunhuang Mural Thesaurus. In the entity semantic similarity calculation process for Dunhuang Mural Thesaurus Linked Data, comprehensive extraction of link predicates in Dunhuang mural linked data can be achieved through analysis of the Schema framework of the Dunhuang Mural Thesaurus Ontology, thereby effectively revealing their semantic description granularity. The attribute definitions of the Dunhuang Mural Thesaurus Ontology are shown in Table 1 . According to different description functions, they are divided into object properties and data properties. Object properties describe relationships between classes, with most object properties only describing relationships between concepts, such as exactMatch and related properties for describing identical or related relationships between concepts, and inScheme and hasTopConcept properties for describing inclusion relationships between concepts and thesauri or between thesauri and facets. The broader and narrower properties are defined as description media for multiple types of inter-class relationships, capable of describing both hierarchical relationships between concepts and hierarchical relationships between concepts and facets. Data properties describe the inherent characteristics of entities in various aspects, such as specific information including name, creation time, and creator, with attribute values mostly being short text character types, except for the scopeNote property which is specifically used to record background knowledge extracted from domain professional literature and has long text type values.

Based on the analysis of the Dunhuang Mural Thesaurus Ontology model and Schema framework, this paper divides the semantic description granularity of Dunhuang Mural Thesaurus Linked Data into three levels: (1) Coarse-grained level, composed of triples reflecting hierarchical structural relationships between different entities in Dunhuang Mural Thesaurus Linked Data, with corresponding link predicates including broader, narrower, and hasTopConcept properties reflecting conceptual hierarchical relationships, as well as exactMatch properties reflecting concept co-reference relationships. (2) Medium-grained level, composed of triples reflecting logical relationship information between entities in the thesaurus linked data, with corresponding link predicates including object properties such as type, inScheme, and related reflecting entity semantic relationships, as well as short text properties such as preLabel, created, creator, and rights reflecting entity inherent characteristics. (3) Fine-grained level, composed of triples annotating domain background information possessed by some entities in the thesaurus, with the corresponding link predicate being the long text property scopeNote.

Multi-Granularity Matching-Based Entity Semantic Similarity Calculation Model

Current research on entity semantic similarity calculation methods is gradually shifting from single technology orientation to requirement orientation oriented toward calculation object characteristics, placing greater emphasis on analyzing background knowledge of entity domains and model framework structures. By matching methods adapted to entity triples of different granularities, semantic similarity calculation is performed. Building upon these ideas and based on the analysis results of semantic description granularity of Dunhuang Mural Thesaurus Linked Data, this paper proposes an entity semantic similarity calculation model that combines multi-granularity matching with weighted operations. Its basic framework is shown in Figure 1 [Figure 1: see original paper].

First, access and obtain triple data of entities to be calculated, including head entities, tail entities, and link predicates, through the SPARQL query endpoint of Dunhuang Mural Thesaurus Linked Data. Second, match them with coarse, medium, and fine-grained modules in the model according to the semantic granularity level corresponding to the link predicates in the triples. Third, set corresponding calculation methods for each module's triples based on their content and structural characteristics and complete semantic similarity calculations. Finally, allocate weights according to the composition of link predicates in each module's triples and obtain the comprehensive semantic similarity of the entity pair through weighted operations.

Coarse-Grained Module Entity Similarity Calculation Method The coarse-grained module addresses triples used to describe entity hierarchical structural relationships in Dunhuang Mural Thesaurus Linked Data. The model employs a semantic similarity calculation method combining attribute co-reference and path distance.

(1) Attribute Co-Reference-Based Semantic Similarity Calculation. Before calculating semantic similarity between two entities in Dunhuang Mural Thesaurus Linked Data, it should first be determined whether they have equivalence attributes. If two entities have hierarchical properties indicating co-reference relationships such as owl:sameAS, rdfs:seeAlso, or skos:exactMatch, their semantic similarity should be determined as 1; otherwise, this part's similarity is 0, as shown in Formula (1).

(2) Path Distance-Based Semantic Similarity Calculation. As a semantic publication outcome of the Dunhuang Mural Thesaurus, the hierarchical nature between entities is an important characteristic of Dunhuang Mural Thesaurus Linked Data. Therefore, in the semantic similarity calculation process, the hierarchical relationship features between entities should be fully considered, introducing the semantic similarity calculation concept oriented toward concept relative depth. Path distance is a semantic similarity calculation method following this concept: the shorter the path dis-

tance between two entities, the higher their semantic similarity, as shown in Formula (2). Where length(x,y) represents the path length between entities x and y in the concept hierarchical structure tree (i.e., the number of hops from x to y), and α is a tuning parameter that can typically be set to 1.

Medium-Granularity Module Entity Similarity Calculation Method
The medium-granularity module addresses triples used to describe entity inherent attributes and related relationships in Dunhuang Mural Thesaurus Linked Data. The model employs a semantic similarity calculation method combining attribute features and edit distance.

(1) Attribute Feature-Based Semantic Similarity Calculation. In linked data, triples with object properties as link predicates can describe specific semantic relationships between head entities and tail entities. Therefore, the differences and similarities in object properties contained between different entities can effectively reflect their semantic relevance. The Tversky model is a typical method for calculating entity semantic similarity based on attribute features. This model quantifies semantic similarity according to the number of common attributes and different attributes shared by a pair of entities, using the calculation method shown in Formula (3). Where f(x y) represents the number of common attributes contained by entities X and Y, f(x-y) represents the number of attributes contained in entity x but not in entity y, and conversely, f(y-x) represents the number of attributes contained in entity y but not in entity x. α and β are tuning parameters reflecting the importance of entities X and Y, with default values of 1.

Based on the Tversky model, necessary improvements must be made in combination with the specific characteristics of Dunhuang Mural Thesaurus Linked Data. Although a pair of entities may share a certain common attribute, the tail entities corresponding to this attribute in their respective triples are not identical. To address this phenomenon, this model makes the following adjustment to Formula (3): for a group of attributes with the same link predicate (Predicate), only when the tail entities (Object) linked in the triples are also identical are they considered common attributes of the two entities; otherwise, they are treated as unique attributes of their respective entities and included in the denominator part of Formula (3).

(2) Edit Distance-Based Semantic Similarity Calculation. Edit distance is a typical method for semantic similarity calculation. In this model, it is primarily used to calculate the semantic similarity of short text attribute values such as skos:prefLabel and dc:created. This method employs a transformation concept to quantify the text similarity of attribute values between original and target entities, as shown in Formula (4). Where tc(x-y) represents the minimum number of transformations required to convert x to y, with operations including addition, insertion, replacement,

and deletion of attribute values, and max[|x|,|y|] represents the maximum character length of the two attribute values.

Fine-Granularity Module Entity Similarity Calculation Method The fine-granularity module addresses triples used to record entity-related background information in Dunhuang Mural Thesaurus Linked Data, primarily focusing on semantic similarity calculation of values for the long text property skos:scopeNote. Since long text attribute values often contain multiple sentences and paragraphs with complex text structures and high information capacity, the edit distance method for short text attribute values mentioned above is often not applicable. To address the semantic similarity calculation needs for long text information, this model employs a topic similarity calculation strategy combining topic identification with the Tversky model. First, text topic identification tools are used to extract a specified number of topic words from the long text attribute values of original and target entities respectively. Then, the numbers of common topic words and unique topic words are counted and substituted into the Tversky model to quantify their semantic similarity, as shown in Formula (5).

The above sections propose corresponding calculation methods for entity semantic similarity in coarse-granularity, medium-granularity, and fine-granularity modules of Dunhuang Mural Thesaurus Linked Data. In actual calculation processes, it is also necessary to reasonably set weight coefficients for each calculation method in the three granularity modules by analyzing the specific circumstances of calculation objects' content distribution and attribute characteristics, thereby obtaining the comprehensive semantic similarity of the entity pair. The calculation process is shown in Formula (6) (where α, β, and γ are weight coefficients for each module).

Entity Semantic Similarity Calculation Experiment for Dunhuang Mural Thesaurus Linked Data

Data Source

To verify the practical effectiveness of the proposed calculation method in entity semantic similarity calculation for Dunhuang Mural Thesaurus Linked Data, this paper uses "Feitian" related entities in the dataset as experimental objects and introduces multiple similar algorithms to conduct comparative experiments on semantic similarity calculation. Experimental data was obtained through SPARQL queries from the Dunhuang Mural Thesaurus Linked Data service platform: using the SPARQL query shown in Figure 2 [Figure 2: see original paper] at the platform's SPARQL Endpoint [27] to retrieve all entities with "飞天" (Feitian) in their skos:prefLabel property values, yielding 8 valid entities: Double Feitian ; Feitian Hair Bun ; Lotus Feitian Caisson Pattern ; Feitian Pattern ; Feitian ; Feitian Musician ; Central Plains Style Feitian ; Western Regions Style Feitian . The query results are shown in Figure 3 [Figure 3: see

original paper].

Experimental Process

Experimental Content The 8 retrieved linked data entities were paired to construct an 8$×$8 entity similarity matrix, generating 28 semantic similarity calculation tasks, as shown in Table 2 . The following sections use the multi-granularity matching-based entity similarity method, Tversky model-based attribute feature similarity method, and edit distance-based label text similarity method to calculate entity semantic similarity.

Multi-Granularity Matching-Based Entity Semantic Similarity Calculation Using the multi-granularity matching-based method for entity semantic similarity calculation requires prior analysis of the triple composition of entities to be calculated, allocating appropriate weight coefficients to each granularity module based on the distribution characteristics of triples at different granularity levels. The author downloaded the RDF documents of the aforementioned 8 entities through the data acquisition port provided by the Dunhuang Mural Thesaurus Linked Data service platform, classified the 55 link predicates they contained, and statistically analyzed the proportion of each attribute type, with results shown in Table 3 .

(1) In the coarse-granularity module, since none of the 8 entities involved in the experiment contained co-reference properties such as owl:sameAs, rdfs:seeAlso, or skos:exactMatch, the impact of entity equivalence on semantic similarity calculation could be disregarded in this experiment. As shown in Table 3, the 8 entities contained a total of 20 link predicates reflecting hierarchical attributes, accounting for the highest proportion among all attributes. This indicates that entity path distances in this dataset have a relatively high influence on semantic similarity calculation. Based on the proportion of link predicates, the weight coefficient for the coarse-granularity module was defined as 0.3636.

(2) In the medium-granularity module, the 8 entities contained 16 link predicates reflecting object attributes. Based on this proportion, the weight for the improved Tversky model algorithm in this module was set to 0.2909. Additionally, the number of link predicates reflecting short text attributes was also 16, so the weight for the edit distance similarity algorithm in this module was also set to 0.2909. Among them, date data annotated by the dc:created property needed to be converted to timestamp text before edit distance calculation.

(3) In the fine-granularity module, since only the entities Feitian , Lotus Feitian Caisson Pattern , and Double Feitian each contained one skos:scopeNote property among the 8 entities, domain background information of each entity had minimal impact on calculation results in this experiment. Therefore, based on its proportion, the weight coefficient

for the fine-granularity module was set to 0.0545.

In summary, based on analyzing the composition of link predicates in each module, the experimental scheme shown in Table 4 was defined, and semantic similarity calculations for coarse-granularity, medium-granularity, and fine-granularity modules were completed accordingly.

Taking task "T1: Sim(tema245, tema445)" as an example to illustrate the multi-granularity matching-based entity semantic similarity calculation process: In the coarse-granularity module, the path distance between Feitian and Feitian Pattern is 10. Substituting into Formula (2) yields a path distance similarity of 0.0909, which becomes 0.0331 after weighting. In the medium-granularity module, after calculation by the improved Tversky model, the attribute feature similarity between Feitian and Feitian Pattern is 0.4, becoming 0.1164 after weighting. Short text attributes yield a similarity of 0.6333 through edit distance (Formula (4)), becoming 0.1842 after weighting. In the fine-granularity module, since Feitian Pattern does not contain the skos:scopeNote property, the semantic similarity of the fine-granularity module between them is 0. In summary, the semantic similarity between Feitian and Feitian Pattern is 0.3336. The same method can be used to calculate the similarities of the other 27 entity pairs.

Tversky Model-Based Attribute Feature Similarity Calculation As shown in Formula (3), the classic Tversky model calculates semantic similarity between two entities based solely on the number of their common and different attributes, without considering specific attribute values. For example, in task T1, entities Feitian and Feitian Pattern have 5 triples with identical attributes. Feitian contains 2 unique attributes, while Feitian Pattern contains no unique attributes. Substituting into Formula (3) yields a similarity of 0.7143. The same algorithm can be used to complete the other 27 similarity calculation tasks.

Edit Distance-Based Label Text Similarity Calculation Edit distance-based text similarity calculation is a common method in large-scale linked data integration and interoperability practice. The basic idea is that properties in linked data entities such as dc:title and skos:prefLabel, which reflect title and label information, have values selected from natural language by data creators or publishers as representative and standardized terms. Semantic similarity calculation based on label edit distance can effectively balance calculation efficiency, result quality, and performance overhead. Label edit distance employs a transformation-based calculation concept: measuring semantic similarity through the ratio of minimum edit times to maximum character length for a group of entity attribute values. Using T1 as an example again: the skos:prefLabel property values of entities and are "飞天" (Feitian) and "飞天纹" (Feitian Pattern) respectively, with a minimum edit distance of 1 and maximum character length of 3. Substituting into Formula (4) yields a label edit distance similarity of 0.6667. The same algorithm is applied to the other 27 similarity calculation tasks.

Experimental Analysis

Using the three methods—classic Tversky model, label edit distance, and multi-granularity matching-based entity semantic similarity calculation—to complete the 28 calculation tasks in Table 1, the results are shown in Table 4 .

Taking the calculation results of T4, T7, and T23 in Table 3 (shown in Figure 4 [Figure 4: see original paper]) as examples to compare the effects of the three algorithms in "Feitian" related entity semantic similarity calculation. The basic profiles of these three tasks are as follows:

(1) T4: Sim(tema245, tema2551) calculates similarity between entities "Feitian " and "Central Plains Style Feitian ." In the Dunhuang Mural Thesaurus Linked Data, the former is the superordinate concept of the latter ( ), with a path distance of 1 between them, indicating high semantic relevance. Comparing the semantic similarity calculation results of different methods for T4 can highlight each method's sensitivity to entity path distance.

(2) T7: Sim(tema245, tema3655) calculates similarity between entities "Feitian " and "Feitian Hair Bun ." In the Dunhuang Mural Thesaurus Linked Data, the former is the subordinate concept of entity "Buddhist Deity " ( ), serving as a general term for a specific type of Buddhist figure. The latter is the subordinate concept of entity "Hairstyle " ( ), used to describe a styling style of mural figures. Although their label texts are similar, their path distance in the dataset is as high as 12, and their actual semantic relevance is low. Comparing calculation results of different methods for T7 can intuitively evaluate each method's ability to identify "error-prone" entities with similar label content but low semantic association.

(3) T23: Sim(tema2551, tema2552) calculates similarity between entities "Central Plains Style Feitian " and "Western Regions Style Feitian ." In the Dunhuang Mural Thesaurus Linked Data, both are subordinate concepts of entity "Feitian ," used to describe the image styles of "Feitian" imagery in different regional cultures. Although their path distance in the dataset (distance of 2) is greater than that of the two entities in task T4 (distance of 1), they have higher semantic similarity at the prior knowledge level. Therefore, T23 is suitable for comparing different calculation methods' ability to identify such implicitly highly relevant entities.

First, the calculation results based on the classic Tversky model are: T4: Sim(tema245, tema2551) = T7: Sim(tema245, tema3655) < T23: Sim(tema2551, tema2552). It can be observed that the Tversky model-based semantic similarity calculation results are generally high compared to the other two methods. The reason is that due to the domain scope and content structural characteristics of annotation objects, entities in the Dunhuang Mural

Thesaurus Linked Data generally exhibit the basic characteristic of having few attributes with high repetition rates. For the classic Tversky model oriented toward attribute features, these characteristics easily lead to high numerical values and insufficient discrimination in semantic similarity calculation results. Therefore, when applying the Tversky model for linked data semantic similarity calculation, necessary improvements should be made according to calculation requirements by adjusting the criteria for identifying common attributes between entities to avoid interference from these phenomena in calculation results.

Second, the calculation results based on label edit distance are: T4: Sim(tema245, tema2551) = T23: Sim(tema2551, tema2552) < T7: Sim(tema245, tema3655). It can be seen that this method yields relatively accurate results for T4 but has significant errors in semantic similarity calculation for T7 and T23. The reason is that this method directly uses entity label text content as the basis for semantic similarity evaluation. For special entities in cultural heritage linked data such as “Feitian ” and “Feitian Hair Bun ” (similar literal content but low semantic association) and “Central Plains Style Feitian ” and “Western Regions Style Feitian ” (dissimilar literal content but high semantic association), accurate semantic similarity calculation is often difficult.

Third, the calculation results based on multi-granularity matching are: T7: Sim(tema245, tema3655) < T4: Sim(tema245, tema2551) < T23: Sim(tema2551, tema2552), which is basically consistent with the prior knowledge descriptions of T4, T7, and T23 above. The reason is that the multi-granularity matching-based calculation method can reasonably divide dataset components into granularity levels according to content structural characteristics and select specific calculation methods adapted to each module. When performing semantic similarity calculation on Dunhuang Mural Thesaurus Linked Data with rich domain background knowledge and complex hierarchical structures, this method can achieve more accurate results compared to other single-approach-based calculation methods.

Comparing the calculation results of the three methods yields the following understanding: In using semantic similarity for semantic integration and interoperability of linked datasets in the cultural heritage domain, it is necessary to select targeted calculation methods based on thorough investigation and analysis of domain ontology models and linked data Schema frameworks, due to objective conditions such as complex background knowledge in related domains and blurred semantic boundaries between different entities. Meanwhile, the multi-dimensional nature of knowledge structures in the cultural heritage domain also makes single-strategy-based calculation methods difficult to fully satisfy all entities’ semantic similarity calculation needs within datasets. Against this background, entity semantic similarity calculation for cultural heritage domain linked data should follow these approaches: First, perform modular processing of linked datasets based on semantic description granularity analysis; second,

select appropriate semantic similarity calculation methods oriented toward the content and structural characteristics of different modules, and obtain optimal semantic similarity calculation results by reasonably setting weight coefficients for each module.

Conclusion

Against the backdrop of the rising humanities computing research paradigm and humanities scholars' needs for semantic integration and interoperability of cultural heritage domain datasets when participating in digital humanities research, this paper takes Dunhuang Mural Thesaurus Linked Data as an example and proposes a multi-granularity matching-based entity semantic similarity calculation method on the basis of dataset semantic description granularity analysis. This provides a feasible approach for data interconnection and knowledge sharing of heterogeneous humanities information resources under the digital humanities framework. In the experimental section, this paper uses "Feitian" related entities in Dunhuang Mural Thesaurus Linked Data as examples and conducts comparative experiments on semantic similarity calculation using the proposed multi-granularity matching method alongside currently representative attribute feature similarity and label edit distance similarity methods. Experimental results demonstrate that the multi-granularity matching-based entity semantic similarity calculation method can better adapt to the structural characteristics of Dunhuang Mural Thesaurus Linked Data, which features complex domain background knowledge and blurred entity semantic boundaries, achieving more accurate calculation results compared to the other two single-strategy-based semantic similarity algorithms. In future research, the calculation method proposed in this paper can be further applied to other linked datasets in the cultural heritage domain, and technical details such as weight allocation at different granularity levels and parameter settings in different algorithms can be adjusted and optimized through large-scale cross-dataset semantic similarity calculation experiments.

References

[1] Huang Shuiqing. Humanities Computing and Digital Humanities: Concepts, Issues, Paradigms, and Key Links[J]. Library Development, 2019(5): 68-78.

[2] Dunhuang Mural Thesaurus Linked Data Service Platform[EB/OL].[2021-01-06]. http://dh.whu.edu.cn/dhvocab/home.

[3] Zuo Dan, Ou Shiyan. Review of Research and Practice on Semantic Description and Organization of Humanities Information Resources[J]. Library Tribune, 2019, 39(8): 21-31.

[4] Hou Xilong, Tan Guoxin, Zhuang Wenjie, et al. Research on Knowledge Management of Intangible Cultural Heritage Based on Linked Data[J]. Journal of Library Science in China, 2019, 45(2): 88-108.

[5] Chen Tao, Liu Wei, Shan Rongrong, et al. Research on the Application of Knowledge Graphs in Digital Humanities[J]. Journal of Library Science in China, 2019, 45(6): 34-49.

[6] Xia Cuijuan, Zhang Lei. Application of Linked Data in Genealogy Digital Humanities Services[J]. Library Journal, 2016, 35(10): 26-34.

[7] Zhai Shanshan, Xu Xin, Xia Lixin, et al. Research on the Application of Semantic Publishing Technology in Digital Resource Sharing of Intangible Cultural Heritage[J]. Library and Information Service, 2017, 61(2): 23-31.

[8] Zeng Ziming, Zhou Zhi, Jiang Lin. Research on Knowledge Organization of Digital Humanities Visual Resources Based on Linked Data[J]. Information and Documentation Services, 2018(6): 6-12.

[9] Gong Zhen, Fan Bingbing. Research on Semantic Association Discovery Methods for Datasets[J]. Computer Applications and Software, 2018, 35(8): 83-86, 185.

[10] Zhang Zhe. Research on Linked Data Model Based on Semantic Similarity Analysis[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[11] Wang Zhongyi, Zhou Jie, Huang Jing. Creation and Publication of Multi-Granularity Linked Data for Digital Libraries[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(8): 885-896.

[12] PASSANT A. Measuring Semantic Distance on Linking Data and Using It for Resource Recommendations[C]//AAAI Spring Symposium: Linked Data Meets Artificial Intelligence. 2010: 93-98.

[13] HICKSON M, KARGAKIS Y, TZITZIKAS Y. Similarity-Based Browsing over Linked Open Data[EB/OL].[2021-04-03]. https://arxiv.org/pdf/1106.4176v1.pdf.

[14] TVERSKY A. Features of Similarity[J]. Readings in Cognitive Science, 1977, 84(4): 290-302.

[15] Deng Lanlan, Li Chunwang. Research on Similarity Calculation Methods for Linked Data Resource Sets[J]. Information Studies: Theory & Application, 2012, 35(5): 112-116.

[16] Sun Haixia, Qian Qing, Cheng Ying. Review of Ontology-Based Semantic Similarity Calculation Methods[J]. New Technology of Library and Information Service, 2010(1): 51-56.

[17] Jia Limei, Zheng Zhiyun, Li Dun, et al. Research on Semantic Similarity Algorithm for Linked Data Based on Dynamic Weight[J]. Computer Science, 2014, 41(8): 263-266, 273.

[18] MEYMANDPOUR R, DAVIS J. A Semantic Similarity Measure for Linked Data: An Information Content-Based Approach[J]. Knowledge-Based Systems, 2016, 109(19): 276-293.

[19] Liu Xiaojuan, Liu Qun. Research and Implementation of an Exploratory Retrieval System Based on Linked Data[J]. Journal of the China Society for Scientific and Technical Information, 2017, 61(5): 117-124.

[20] Zhang Libo, Sun Yihan, Luo Tiejian. A Semantic Similarity Calculation Method Based on Large-Scale Knowledge Base[J]. Journal of Computer Research and Development, 2017, 54(11): 2576-2585.

[21] Wang Xiaoguang, Hou Xilong, Cheng Hanghang, et al. Construction and Linked Data Publication of Dunhuang Mural Thesaurus[J]. Journal of Library Science in China, 2020, 46(4): 69-84.

[22] Introduction to Dunhuang Mural Thesaurus Project[EB/OL].[2021-01-06]. http://dh.whu.edu.cn/dhvocab/dhresource/html/intro.html.

[23] Ontology Model[EB/OL].[2021-01-06]. http://dh.whu.edu.cn/dhvocab/ontology.

[24] RADA R, MILI H. Development and Application of a Metric on Semantic Nets[J]. IEEE Transaction on System Man & Cybernetics, 1989, 19(1): 17-30.

[25] He Yuanxiang, Shi Baoming, Zhang Yong. Research on Ontology-Based Semantic Similarity Algorithm[J]. Computer Applications and Software, 2013, 30(11): 312-315.

[26] Deng Lanlan, Li Chunwang. Research on Web Data Association Creation Strategies[J]. New Technology of Library and Information Service, 2011(5): 1-6.

[27] Dunhuang Mural Thesaurus Linked Data Query[EB/OL].[2021-01-06]. http://dh.whu.edu.cn/dhvocab/sparql.

Author Contributions: Gao Jinsong: Paper guidance, writing, and revision; Fu Jiawei: Data collection, experimental design and operation, paper writing and revision; Li Ke: Experimental design and operation, paper writing.

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