Postprint: An Ontology Alignment Framework for Chinese Ontology Schemas

Authors: Wang Ting, Gao Ying, Liu Jingwei

Date: 2017-11-08T00:00:00+00:00

Abstract

Abstract

Purpose: Existing ontology alignment methods often neglect the word-order sensitivity and polysemous semantic features of Chinese concepts. This paper proposes a large-scale Chinese ontology mapping model based on Tongyici Cilin and sequence alignment algorithms.

Method: We employ an improved Tongyici Cilin similarity algorithm to compute the semantic similarity of simple word tokens, and utilize an algorithm that integrates improved Tongyici Cilin with sequence alignment to measure semantic similarity between out-of-vocabulary words.

Results: Experimental results on association mapping using test corpora constructed from DBpedia (Chinese version), Baidu Baike, and Hudong Baike knowledge bases show that the model achieves average precision, recall, and comprehensive evaluation metrics of approximately 97.5%, 87.8%, and 92.1%, respectively.

Limitations: This model focuses solely on element-level similarity measurement of Chinese ontology concepts, and does not consider the influence of ontology properties and instances on concept equivalence relationships.

Conclusion: Evaluation results on large-scale open semantic datasets oriented toward Chinese web encyclopedias demonstrate that the overall performance of this model is significantly superior to existing algorithms.

Full Text

A Chinese Ontology Schema-Level Alignment Framework

Wang Ting, Gao Ying, Liu Jingwei
Information School, Capital University of Economics and Business, Beijing

100070, China

Abstract

[**Objective**] Existing ontology alignment methods often neglect the semantic characteristics of Chinese concepts, particularly their sensitivity to word order and polysemy. This paper proposes a large-scale Chinese ontology mapping model based on TongYiCiCiLin (TYCCL) and sequence alignment algorithms to address these issues. [**Methods**] The model employs an improved TYCCL-based similarity algorithm to compute semantic similarity between atomic concepts, and utilizes a hybrid algorithm combining improved TYCCL with sequence alignment to measure similarity between out-of-vocabulary terms. [**Results**] Experimental results on a test corpus constructed from DBpedia (Chinese version), Baidu Baike, and Hudong Baike demonstrate that the model achieves average precision, recall, and F-measure of approximately 97.5%, 87.8%, and 92.1%, respectively. [**Limitations**] The model focuses exclusively on element-level similarity measurement for Chinese ontology concepts, without considering the influence of ontology properties and instances on concept equivalence relationships. [**Conclusions**] Evaluation results on large-scale open semantic datasets derived from Chinese web encyclopedias confirm that the proposed model significantly outperforms existing algorithms.

Keywords: Chinese Linked Open Data; TongYiCiCiLin; Sequence Alignment; Ontology Mapping; Similarity Computing

Introduction

The vision of the Semantic Web is to establish a "Web of Data" that enables machines to comprehend semantic information on the Internet [1]. As a core element of the Semantic Web, an ontology provides a formal and standardized specification of shared concepts within a specific domain [2], forming the foundation for web-based knowledge sharing and semantic interoperability. Currently, research on Linked Open Data (LOD) [3] primarily focuses on the instance level [4-5]. However, heterogeneity among different ontologies hinders their reuse and sharing. Consequently, schema-level LOD construction serves as both the basis and prerequisite for linked data, making it an important research area [6].

Ontology mapping, a typical scenario in schema-level linked data construction, has been extensively studied. Its primary task is to discover semantic associations between concepts across heterogeneous ontologies or LOD datasets. With the rapid development of the Semantic Web, large-scale ontologies and knowledge bases described in Chinese have been increasingly constructed and shared. However, due to cultural and contextual factors, research on building large-scale Chinese linked data networks remains in its infancy, particularly lacking mature schema-level models for large-scale Chinese linked data. To address the challenges of semantic interoperability and sharing of Chinese ontologies within

the linked data environment, this paper proposes a novel large-scale Chinese ontology mapping model at the schema level.

Related Work

Researchers have proposed numerous mapping methods and systems. Melnik et al. [7] introduced Similarity Flooding, a structure-level ontology mapping algorithm that constructs a similarity propagation graph from the ontology' s concept hierarchy to propagate and refine similarities between concepts. Cohen et al. [8] analyzed several typical element-level similarity algorithms based on edit distance and token matching, and evaluated their performance. Giunchiglia et al. [9] proposed a linguistic approach that incorporates shared knowledge dictionaries such as WordNet [10] to discover semantic relationships through linguistic relations. Isaac et al. [11] presented an instance-level ontology mapping algorithm that measures concept similarity based on the number of shared instances. Nikolov et al. [12] developed KnoFuss, a workflow-based framework for linked data that leverages hierarchical relationships among concepts in ontology libraries to select optimal matching methods and parameters. Zhong et al. [13] introduced the RiMOM system, which employs multi-strategy mapping based on ontology instances, concept names, and structural features, and incorporates universal field theory principles to handle large-scale ontology mapping tasks. Jain et al. [6] released the BLOOMS system, which uses a bootstrapping approach with Wikipedia' s top-level category tree as a knowledge base for similarity computation to perform schema-level link construction in LOD environments. However, these systems are limited to handling ontology schema mapping tasks for English-language semantic datasets.

In recent years, increasing attention has been devoted to Chinese ontology and linked data construction. At the schema level (i.e., ontology mapping) of Chinese linked data network development, Li et al. [14] proposed an element-level concept similarity method based on HowNet [15] and implemented a Chinese ontology mapping system. However, this system overlooks the prevalent phenomena of word order sensitivity and polysemy in Chinese [16], limiting its applicability to large-scale ontology mapping tasks in linked data environments. Tian et al. [18] developed a Chinese word semantic similarity algorithm based on the Extended TongYiCiCiLin [17], but their approach did not address similarity computation for out-of-vocabulary terms, nor was it applied in real-world large-scale linked data network environments.

Additionally, numerous instance-level linked data systems have been developed. Silk [19-20] is a framework for discovering links between datasets that features a declarative language enabling users to configure links between two datasets, including link types and conditions, while supporting linking between remote and local datasets. Hassanzadeh et al. [21] proposed LinQL, a general and extensible framework that integrates various existing link discovery methods to help users select the most suitable approach for their datasets, with support for RDF data published from relational databases using tools like D2RQ or Virtuoso. Wang

et al. [5] extracted hierarchical relationships from Chinese encyclopedia classification systems (DMOZs) and obtained concept attributes and instances from Infobox-enabled web pages to construct two large-scale Chinese ontologies based on Baidu Baike and Hudong Baike, establishing coreference relationships with DBpedia through simple keyword matching strategies. Niu et al. [4] semantically integrated Baidu Baike [22], Hudong Baike [23], and Chinese Wikipedia [24-25] to develop Zhishi.me, an instance-level linked data application system for Chinese descriptions. To achieve knowledge sharing, reuse, and semantic interoperability in linked data environments, cross-lingual ontology linking and mapping becomes essential. Wang et al. [26] proposed a concept annotation method that enriches internal links using a small number of cross-lingual and internal link seeds, and employs a regression learning model to predict potential cross-lingual links between Chinese and English Wikipedia. However, these systems focus exclusively on constructing instance-level associations, lacking mechanisms for discovering and acquiring links at the ontology schema level.

In summary, currently few large-scale Chinese ontologies are published on the Web, and those that exist exhibit significant heterogeneity. Existing Chinese ontology mapping systems demonstrate low efficiency and limited usability when facing large-scale mapping tasks. Moreover, there remains a lack of large-scale ontology mapping systems specifically designed for Chinese language descriptions and adapted to LOD environments. To address these challenges, this paper proposes a novel Chinese ontology mapping model based on the Extended TongYiCiCiLin and sequence alignment principles. The model effectively resolves issues of word order sensitivity and polysemy in Chinese concept similarity computation. Experimental results on a large-scale ontology test set constructed from Chinese web encyclopedias demonstrate that the system achieves superior overall performance compared to previous work.

Formal Definitions

In the Extended TongYiCiCiLin (TYCCL), the collected vocabulary terms are referred to as simple word units. In Chinese ontology mapping systems, both simple word units and out-of-vocabulary terms correspond to ontology concepts. This paper designates simple word units as **Atomic Concepts (AC)** and collectively refers to out-of-vocabulary terms as **Component Concepts (CC)**, with the stipulation that a component concept consists of a linear arrangement of multiple atomic concepts.

Definition 1 (Ontology Mapping): Given two ontologies to be mapped, source ontology Os and target ontology Ot, for each concept Cs in Os, we aim to find a concept Ct in Ot with identical or similar semantics. The mapping function is defined as map: Os → Ot: For all Cs Os and Ct Ot, if sim(Cs, Ct) > t, then map(Cs) = Ct, where sim(Cs, Ct) represents the semantic similarity between Cs and Ct, and t is a threshold. When the semantic similarity between Cs and Ct exceeds t, the pair <Cs, Ct> is identified as an equivalent concept pair.

Definition 2 (Semantic Knowledge Base): This paper considers all vocabulary terms collected in the Extended TongYiCiCiLin and their semantic relationships as constituting a Semantic Knowledge Base (SKB), denoted as SKB_{TYCCL}. The set SKB_{TYCCL} is composed of atomic concepts, i.e., SKB_{TYCCL} = {AC_1, AC_2, ···, AC }, where N represents the total number of word units in the knowledge base.

Definition 3 (Component Concept): A component concept CC comprises an ordered sequence of atomic concepts. For all AC SKB_{TYCCL}, we introduce two-dimensional indices i and j, yielding the ordered sequence CC = [AC $_1$, AC $_2$, ···, AC], where j ≥ 1 and CC SKB_{TYCCL}. Here, j denotes the position of atomic concept AC in the ordered sequence CC . Specifically, for any atomic concept AC , we have AC = [AC].

Definition 4 (Concept Representation): For concepts Cs and Ct from ontologies Os and Ot respectively, we have Cs = CCs = [ACs_1, ACs_2, ···, ACs] and Ct = CCt = [ACt_1, ACt_2, ···, ACt], where m and n represent the lengths of the ordered sequences CCs and CCt corresponding to concepts Cs and Ct, with m, n ≥ 1.

4. Chinese Linked Data Mapping Based on TongYiCiCiLin and Sequence Alignment

The model comprises five main functional modules: ontology preprocessing, component concept segmentation, improved TYCCL similarity computation, scoring matrix construction, and component concept similarity calculation (which integrates improved TYCCL similarity computation and sequence alignment processing). The overall system framework is illustrated in Figure 1 [Figure 1: see original paper]. Based on the formal definitions above, we categorize and discuss various scenarios in the Chinese ontology concept mapping process.

For any two concepts Cs and Ct from the source ontology Os and target ontology Ot, three cases may arise during semantic similarity computation: (1) Both Cs and Ct are atomic concepts, i.e., Cs SKB_{TYCCL} and Ct SKB_{TYCCL}; (2) One of Cs and Ct is an atomic concept while the other is a component concept, i.e., Cs SKB_{TYCCL} or Ct SKB_{TYCCL}; (3) Both Cs and Ct are component concepts, i.e., Cs SKB_{TYCCL} and Ct SKB_{TYCCL}.

For case (1), we directly employ the "Improved TYCCL Similarity Computation" module to calculate semantic similarity between two atomic concepts.

For cases (2) and (3) involving component concept similarity computation, this paper adopts a multi-strategy fusion approach based on "Sequence Alignment Processing" and "Improved TYCCL Similarity Computation." Specifically, the "Component Concept Similarity Computation" module takes as input two word string sequences CCs and CCt to be mapped, along with their corresponding scoring matrix, which is collaboratively generated by the "Component Concept Segmentation" and "Scoring Matrix Construction" modules.

4.1 Improved TYCCL-Based Similarity Computation

TongYiCiCiLin encodes vocabulary terms and organizes them in a hierarchical structure resembling an inverted tree, where each node represents a concept. Chinese concept coreference identification can essentially be abstracted as Chinese synonym recognition and semantic similarity computation, making TongYiCiCiLin an optimal choice. This paper adopts the Extended TongYiCiCiLin from Harbin Institute of Technology as the commonsense knowledge base for Chinese ontology mapping relation extraction.

During experimentation, we observed that the traditional algorithm proposed by Tian et al. [18] overemphasizes semantic relatedness between concepts. Specifically, the parent-child relationships between vocabulary terms at different hierarchical levels in TongYiCiCiLin significantly interfere with the acquisition of equivalence relationships between ontology concepts. Since ontology mapping aims to discover equivalence rather than taxonomic relationships, this paper introduces a semantic adjustment factor and concept similarity weight coefficient to improve the traditional algorithm, making it suitable for Chinese ontology mapping tasks in LOD environments.

TongYiCiCiLin organizes word units in a hierarchical structure comprising five layers from top to bottom. Each layer has a corresponding encoding identifier, and the five-layer codes are arranged sequentially from left to right to form the word' s forest code. The implicit semantic relatedness between words increases with deeper hierarchical levels.

TongYiCiCiLin [27] is a Chinese thesaurus that encodes each word term. Table 1 explains the encoding format using the word "matter"(forest code: Ba01A02=) as an example.

Based on TongYiCiCiLin' s structural characteristics, we first parse the forest codes of concepts to be mapped, extracting sub-codes from layers 1 to 5, then compare them starting from layer 1. When sub-codes differ, the mapping pair receives a similarity weight corresponding to the layer at which the difference occurs. Differences appearing at deeper layers receive higher similarity weights, while differences at shallower layers (smaller code positions) indicate poorer semantic relatedness (lower similarity weights). Thus, the improved method simultaneously considers hierarchical factors in TongYiCiCiLin when computing similarity.

Additionally, the number of branch nodes at each layer affects similarity. The similarity computation method based on TongYiCiCiLin is given by Formula (1):

$$SIM_T(C_s, C_t) = \lambda \times \frac{L_i}{|L|} \times \frac{N_t - D + 1}{N_t}$$

Since ontology mapping tasks emphasize semantic similarity between concepts, we introduce a semantic adjustment factor λ to regulate the relationship be-

tween semantic relatedness and semantic similarity across different hierarchical levels, and to control the potential similarity between word units at different hierarchical branches, where λ (0,1). Larger λ values indicate greater possibility of similarity or equivalence between word units at different levels, and stronger influence of hierarchical semantic relatedness on the final concept similarity, and vice versa. Specifically, for Chinese ontology mapping tasks that prioritize semantic similarity between concepts, λ should not be set too high.

We define L = {1,2,3,4,5}, where for any L L, L represents the layer number at which sub-codes differ, and |L| denotes the number of elements in set L, which is constantly 5 in this system. The proposed concept similarity weight coefficient is $\lambda \times$ (L /|L|). N represents the total number of nodes at layer i for word units Cs and Ct, and D is the encoding distance between Cs and Ct. Specifically, when all five layers of codes for a concept pair are identical and the last character of the forest code is '=' , the similarity function SIM_T returns 1.0. The value range of SIM_T is clearly (0,1].

4.2 Component Concept Similarity Computation Based on Sequence Alignment

Numerous scholars have proposed solutions for computing similarity between Chinese component concepts. For instance, Li et al. [14] designed and implemented an element-level concept similarity method based on HowNet and developed a Chinese ontology mapping system. When handling out-of-vocabulary terms, this method traverses the atomic concept sequences corresponding to two component concepts to identify the mapping pairs with maximum similarity, then computes the similarity between the two component concepts based on these relatively maximal mapping pairs, as shown in Formula (2):

$$Sim(A,B) = \frac{\sum_{i=1}^{\max(m,n)} \max_i(B_{xy})}{\max(m,n)}$$

where B represents elements in the similarity matrix formed by known words from the segmentation of two terms, and max (B) denotes the i-th largest similarity value in the matrix. max(m,n) selects the larger of the row or column indices.

However, due to the prevalent characteristic of word order sensitivity in Chinese concepts, the aforementioned approach inevitably introduces errors in semantic similarity computation. For example, consider two component concepts from different ontologies: "historical theory" and "history of thought" . After segmentation, we obtain two ordered atomic concept sequences: [历史, 理论] and [思想, 史]. Using conventional methods for handling out-of-vocabulary terms yields the atomic concept mapping results shown in Figure 2 [Figure 2: see original paper]. Computing semantic similarity for each atomic concept mapping pair based on the Extended TongYiCiCiLin using Formula (1) and then applying Formula (2)

produces an element-level similarity of 1.0, which is clearly unreasonable. This erroneous result stems from neglecting the phenomena of word order sensitivity and polysemy prevalent in natural Chinese language.

Therefore, this paper proposes an improved concept semantic similarity computation method. Specifically, we introduce a global pairwise sequence alignment algorithm from bioinformatics for semantic similarity computation at the element level.

(1) Overview of Sequence Alignment Algorithms

In bioinformatics, pairwise sequence alignment involves arranging two DNA, RNA, or protein sequences to identify their similarities. Gap symbols may be inserted into sequences, with identical or similar symbols aligned in the same column. By comparing similar fragments and conserved sites between two sequences, the algorithm seeks potential molecular evolutionary relationships [28].

Alignment models fall into two categories: global alignment, which examines overall similarity between two complete sequences through full scanning and comparison; and local alignment, which focuses on specific fragments to compare similarity between sequence segments. Both can be solved using dynamic programming (DP) principles.

(2) Constructing the Dynamic Programming Scoring Matrix

A sequence is defined as a string composed of letter identifiers arranged according to specific rules.

Component Concept Segmentation: This system treats component concepts as word string sequences where each element is an atomic concept. We segment component concepts to obtain their corresponding word string sequences using ICTCLAS50 [29], developed by the Institute of Computing Technology, Chinese Academy of Sciences. The alphabet is defined as the semantic knowledge base of the Extended TongYiCiCiLin: SKB_{TYCCL}.

Scoring Matrix Construction: The two word string sequences to be aligned are represented as a scoring matrix M, with the sequences forming the two dimensions of the dynamic programming matrix. For concepts Cs and Ct from ontologies Os and Ot, row i of matrix M corresponds to atomic concept ACs in sequence CCs, and column j corresponds to atomic concept ACt in sequence CCt, where $i \leq m$ and $j \leq n$. The element at row i, column j is denoted as M .

Following dynamic programming principles, the two word string sequences are represented as rows and columns. If sequence CCs has length m and sequence CCt has length n, we form an $(m+1)\times(n+1)$ two-dimensional matrix with CCs as rows and CCt as columns. For example, segmenting the component concepts "Second Industrial Revolution" and "war criminals of World War II" yields two sequences: CCs = [第二, 次, 工业革命] and CCt = [第二, 次, 世界大战, 战犯].

(3) Optimal Recursive Solution Algorithm

Concept similarity computation for ontology mapping is abstracted as an alignment process between two word string sequences. Through a gap penalty function, the algorithm decides where to insert gap symbols '-' in the word string sequences to equalize their lengths, thereby establishing correspondences between atomic concepts or between atomic concepts and gap symbols. The essence of sequence alignment is to identify the optimal global pairing between two component concept sequences through a scoring strategy.

The Needleman-Wunsch algorithm, proposed in 1970 by Needleman and Wunsch, is a classic dynamic programming algorithm for global sequence similarity comparison, suitable for sequences with high overall macro-level similarity [30]. This paper primarily employs this algorithm and dynamic programming principles to recursively solve for the optimal alignment path in matrix M.

Algorithm 1: ConceptSimilarity(CCs, CCt) - **Input:** Scoring matrix M(i)(j) corresponding to component concepts CCs and CCt - **Output:** Matrix M (i)(j) containing the optimal alignment path

1. p ← -0.05 // Define constant p as the algorithm' s penalty factor, set to -0.05
2. For each i ← 1, 2, ···, m+1; j ← 1, 2, ···, n+1 // Initialize dynamic programming matrix
3. M(i)(n+1) ← p × (m-i+1)
4. M(m+1)(j) ← p × (n-j+1)
5. End for
6. For each i ← m, m-1, ···, 1
7. For each j ← n, n-1, ···, 1
8. M(i)(j) ← max(M(i+1)(j+1) + SIM_T(ACs , ACt), M(i)(j+1) + p, M(i+1)(j) + p) // Recursively compute cost for each matrix element
9. End for
10. End for
11. Backtrack to obtain matrix M (i)(j) containing the optimal alignment path
12. Return M (i)(j)

First, we set the sequence alignment penalty factor p = -0.05 and initialize column n+1 and row m+1 of the matrix using the rules: M(i)(n+1) = p × (m-i+1) and M(m+1)(j) = p × (n-j+1).

Second, we recursively solve for the remaining m×n elements in the scoring matrix based on the TYCCL similarity function SIM_T. The scoring function f is defined in Formula (3):

$$f(AC_{s_i}, AC_{t_j}) = \begin{cases} SIM_T(AC_{s_i}, AC_{t_j}), & \text{if } AC_{s_i} = AC_{t_j} \\ -1, & \text{if } AC_{s_i} \neq AC_{t_j} \end{cases}$$

Considering the prevalent word order sensitivity in Chinese component concepts, the recursion starts at the end of both component concept sequences, i.e., at

matrix element M . SIM_T is described in Formula (1). The recursive rule (gap penalty function) is given by Formula (4):

$$M(i)(j) = \max \begin{cases} M(i+1)(j+1) + f(AC_{s_i}, AC_{t_j}) \\ M(i)(j+1) + p \\ M(i+1)(j) + p \end{cases}$$

Finally, backtracking from element M to M_{11} yields the optimal alignment path. In the scoring matrix containing the optimal path, bold arrows indicate the optimal route. Specifically, the gap insertion strategy is: bold diagonal arrows pair the two atomic concepts at their tails; bold horizontal arrows insert a gap '-' before the atomic concept in sequence CCs at the corresponding row; bold vertical arrows insert a gap '-' before the atomic concept in sequence CCt at the corresponding column. If multiple optimal alignment paths exist, any one may be selected.

The complete element-level concept similarity algorithm based on global sequence alignment is presented in Algorithm 1. After inserting gap symbols '-', the two component concept term sequences have equal length, denoted as CC and CC , with length L. The final similarity between component concepts is computed based on the alignment results using scoring function f, as shown in Formula (5):

$$SIM_{NW}(CC'_s, CC'_t) = \frac{\sum_{k=1}^{L} f(AC_{s_k}, AC_{t_k})}{L}$$

5. Experiments

We invited four senior undergraduate students from the Information School of Capital University of Economics and Business to manually identify and annotate all objective equivalence relationships among top-level categories in DBpedia, Baidu Baike, and Hudong Baike. These annotations serve as reference correct mapping pairs for ontology mapping experiments, as shown in Tables 3 through 5 .

5.1 Data Sources

This study employs Chinese web-based open encyclopedia knowledge bases as experimental data sources. In addition to DBpedia (Chinese version), our system uses the HTMLParser toolkit to crawl and parse structured Infobox information from the open category pages and entry pages of Baidu Baike and Hudong Baike, based on references [5,31]. The extracted information is organized as Chinese triples to form large-scale Chinese open-domain knowledge bases ready for mapping. As shown in Table 2 , the ontology concept system is primarily constructed from the encyclopedia open category systems.

Table 2 provides information about the Chinese web encyclopedia knowledge bases, including statistics on Infobox triples, Infobox counts, and predicate numbers for Baidu Baike, Hudong Baike, and DBpedia 3.8 (Chinese version).

5.2 Evaluation Metrics

This study adopts precision, recall, and F-measure for evaluating Chinese concept equivalence identification as the final evaluation criteria:

$$Precision(P) = \frac{\text{Number of correctly mapped pairs}}{\text{Total number of mapped pairs}}$$

$$Recall(R) = \frac{\text{Number of correctly mapped pairs}}{\text{Total number of pairs in gold standard}}$$

$$F\text{-measure}(F1) = \frac{2 \times P \times R}{(P + R)}$$

5.3 Analysis of Sequence Alignment Results

After elaborating on the component concept similarity computation method based on sequence alignment, we re-examine the two similarity computation examples mentioned earlier.

Example 1: CCs = [思想, 史] and CCt = [历史, 理论]. Using Formula (2) yields Sim(CCs, CCt) = (1.0 + 1.0)/2 = 1.0. However, the sequence alignment result obtained by our algorithm is shown in Figure 3 [Figure 3: see original paper], with the corresponding scoring matrix in Figure 4 [Figure 4: see original paper]. The final component concept similarity should be SIM_{NW} = (-0.05 + 1.0 - 0.05)/3 = 0.3. This example mapping pair originates from the "history" subtask in the Hudong-DBpedia mapping task.

Example 2: CCs = [第二, 次, 工业革命] and CCt = [第二, 次, 世界大战, 战犯]. Using Formula (2) would incorrectly yield Sim(CCs, CCt) = 1.0 due to the polysemy of the atomic concept "次" (time/instance). Specifically, the word unit "次" has multiple encoding entries in the Extended TongYiCiCiLin, with entry Dn04B03= establishing "第二" (second) and "次" as equivalent word units. Consequently, Formula (2) produces four atomic concept mapping pairs with similarity 1.0: < 第二, 次 > = 1.0, < 第二, 第二 > = 1.0, < 次, 第二 > = 1.0, and < 次, 次 > = 1.0. Substituting into Formula (2) gives Sim(CCs, CCt) = (1.0 + 1.0 + 1.0 + 1.0)/4 = 1.0. In contrast, our sequence alignment algorithm yields a component concept similarity of SIM_{NW} = (1.0 + 1.0 + 0.18 - 0.05)/4 = 0.5325. The scoring matrix M (i)(j) containing the optimal matching path, obtained through Algorithm 1, is shown in Figure 5 [Figure 5: see original paper], with the corresponding optimal sequence matching result in Figure 6 [Figure 6: see original paper]. This example mapping pair originates from the "history" subtask in the Baidu-DBpedia mapping task.

These examples demonstrate that no equivalence relationship exists between the component concepts in either Example 1 or Example 2. Traditional methods erroneously assign extremely high similarity scores of 1.0, whereas Algorithm 1 produces more reasonable values. By accounting for word order sensitivity and polysemy in Chinese concepts, the global alignment algorithm based on Needleman-Wunsch effectively avoids erroneous mappings inherent in conventional methods such as [14]. Moreover, when mapping component concepts (i.e., out-of-vocabulary terms) whose atomic concept sequences share essentially the same semantic order, Algorithm 1 performs comparably to traditional methods. In summary, the element-level concept similarity algorithm based on global sequence alignment offers greater advantages and rationality for large-scale Chinese ontology mapping tasks.

5.4 Analysis of Large-Scale Chinese Ontology Mapping Results

Guided by the theoretical principles of the aforementioned algorithms, this study evaluates the performance of our prototype system using the three major Chinese web encyclopedia knowledge bases as data sources, targeting real-world large-scale linked data construction scenarios. Evaluation results from completing three major mapping tasks are presented in Tables 6 through 8 , showing precision, recall, and F1 values obtained using four typical similarity computation algorithms: (1) cross-lingual edit distance similarity algorithm [32], (2) traditional TYCCL-based Chinese word similarity algorithm [18], (3) ELOMC, a HowNet-based Chinese word similarity algorithm proposed by Li et al. [14], and (4) the comprehensive Chinese concept similarity computation algorithm proposed in this paper.

To ensure fairness, the similarity threshold for determining concept equivalence is uniformly set to t = 0.9.

Table 6 presents results for the Baidu-Hudong ontology mapping task. Our system' s average precision exceeds that of the traditional TYCCL algorithm and ELOMC by approximately 41% and 39%, respectively. Recall surpasses edit distance and traditional TYCCL algorithms by about 13% and 2% on average, remaining comparable to ELOMC. In terms of comprehensive F1 measure, our system outperforms edit distance, traditional TYCCL, and ELOMC algorithms by approximately 8%, 23%, and 20%, respectively.

Table 7 shows results for the Hudong-DBpedia ontology mapping task. For precision, our system exceeds edit distance, traditional TYCCL, and ELOMC algorithms by approximately 1%, 10%, and 11%, respectively. Recall surpasses edit distance and traditional TYCCL by about 6% and 1% on average, remaining comparable to ELOMC. Our system' s F1 measure exceeds edit distance, traditional TYCCL, and ELOMC algorithms by approximately 3%, 6%, and 6%, respectively.

Table 8 presents results for the Baidu-DBpedia ontology mapping task. Our system' s average precision exceeds traditional TYCCL and ELOMC algorithms

by approximately 39% and 43%, respectively. Recall surpasses edit distance, traditional TYCCL, and ELOMC algorithms by about 17%, 6%, and 3% on average. In comprehensive F1 measure, our system outperforms edit distance, traditional TYCCL, and ELOMC algorithms by approximately 8%, 26%, and 30%, respectively.

Examples of controversial synonym pairs include <nation, Chinese nation>, <criminal law, criminal>, <army, military>, and <Xinhai Revolution, revolution>. Such cases occur more frequently in the "society" sub-mapping task but rarely appear in other mapping tasks.

Macroscopically, our model achieves average precision, recall, and comprehensive evaluation metrics of approximately 97.5%, 87.8%, and 92.1%, respectively, across 18 sub-mapping tasks from the three major mapping tasks.

Although our model's precision is slightly lower than edit distance algorithm due to a few improperly classified synonym pairs in TongYiCiCiLin, edit distance algorithm mechanically compares literal similarity between concepts while completely ignoring semantic similarity, inevitably resulting in significantly lower recall across all mapping tasks. Our approach, which incorporates the semantic dictionary Extended TongYiCiCiLin with improvements to the traditional algorithm, achieves notably higher recall than edit distance algorithm. Consequently, our method significantly outperforms edit distance algorithm in final comprehensive F1 measure across all three mapping tasks.

In summary, our model achieves the best comprehensive evaluation metrics among comparable systems, with precision significantly higher than traditional TYCCL algorithm and ELOMC system, and recall higher than edit distance algorithm and traditional TYCCL similarity algorithm, remaining comparable to ELOMC system.

Conclusion

Currently, mature large-scale Chinese ontology mapping systems are lacking. This paper addresses schema matching challenges in linked data network construction by proposing a novel Chinese ontology mapping model that integrates TongYiCiCiLin with global sequence alignment algorithms. The system resolves usability issues in large-scale ontology mapping by focusing on the word order sensitivity and polysemy characteristics of existing large-scale Chinese ontologies for element-level mapping of component concepts. Future work will consider incorporating instance-level mapping parameters and concept definition similarity based on different Chinese ontology characteristics to further enhance the robustness and accuracy of Chinese mapping systems.

References

[1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web[J]. Scientific American, 2001, 284(5): 28-37.

[2] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Universiteit Twente, 2009.
[3] Bizer C, Heath T, Idehen K, et al. Linked Data on the Web[C]//Proceedings of the 17th International Conference on World Wide Web, Beijing, China. New York, USA: ACM, 2008: 1265-1266.
[4] Niu X, Sun X, Wang H, et al. Zhishi.me-Weaving Chinese Linking Open Data[C]//Proceedings of the 10th International Conference on the Semantic Web, Bonn, Germany. Heidelberg, Germany: Springer-Verlag Berlin, 2011: 205-220.
[5] Wang Z, Wang Z, Li J, et al. Knowledge Extraction from Chinese Wiki Encyclopedias[J]. Journal of Zhejiang University-Science C: Computer & Electronics, 2012, 13(4): 278-287.
[6] Jain P, Hitzler P, Sheth A P, et al. Ontology Alignment for Linked Open Data[C]//Proceedings of the 9th International Conference on the Semantic Web, Shanghai, China. Heidelberg, Germany: Springer-Verlag Berlin, 2010: 402-417.
[7] Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching[C]//Proceedings of the 18th International Conference on Data Engineering, San Jose, California, USA. Washington, USA: IEEE Computer Society, 2002: 117-128.
[8] Cohen W, Ravikumar P, Fienberg S. A Comparison of String Metrics for Matching Names and Records[C]//Proceedings of KDD Workshop on Data Cleaning and Object Consolidation. 2003, 3: 73-78.
[9] Giunchiglia F, Yatskevich M. Element Level Semantic Matching[C]//Proceedings of Meaning Coordination & Negotiation Workshop at ISWC. 2004.
[10] Stark M M, Riesenfeld R F. WordNet: An Electronic Lexical Database[C]//Proceedings of the 11th Eurographics Workshop on Rendering. MIT Press, 1998.
[11] Isaac A, Van Der Meij L, Schlobach S, et al. An Empirical Study of Instance-Based Ontology Matching[C]//Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference. Heidelberg, Germany: Springer-Verlag Berlin, 2007: 253-266.
[12] Nikolov A, Uren V, Motta E, et al. Integration of Semantically Annotated Data by the KnoFuss Architecture[C]//Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Heidelberg, Germany: Springer-Verlag Berlin, 2008: 265-274.
[13] Zhong Q, Li H, Li J, et al. A Gauss Function Based Approach for Unbalanced Ontology Matching[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, 2009: 669-680.
[14] Li Jia, Zhu Ming, Liu Chen, et al. Research and Implementation on Chinese Ontology Mapping[J]. Journal of Chinese Information Processing, 2007, 21(4): 27-33.
[15] Dong Zhendong, Dong Qiang, Hao Changling. Theoretical Findings of HowNet[J]. Journal of Chinese Information Processing, 2007, 21(4): 3-9.
[16] Lu Bingfu. Word Order Dominance and Its Cognitive Explanation[J]. Contemporary Linguistics, 2005, 7(1): 1-15.

[17] HIT-SCIR. TongYiCiCiLin (Extended Version)[EB/OL]. [2014-09-05]. http://ir.hit.edu.cn/demo/ltp/Sharing_{Plan}.htm.
[18] Tian Jiule, Zhao Wei. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System[J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 602-608.
[19] Volz J, Bizer C, Gaedke M, et al. Silk-A Link Discovery Framework for the Web of Data[C]//Proceedings of LDOW2009, Madrid, Spain. 2009.
[20] Volz J, Bizer C, Gaedke M, et al. Discovering and Maintaining Links on the Web of Data[C]//Proceedings of the International Semantic Web Conference. Springer Berlin Heidelberg, 2009: 650-665.
[21] Hassanzadeh O, Lim L, Kementsietsidis A, et al. A Declarative Framework for Semantic Link Discovery over Relational Data[C]//Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 1101-1102.
[22] Baidu Baike[EB/OL]. [2015-09-10]. http://baike.baidu.com/.
[23] Hudong[EB/OL]. [2015-09-17]. http://www.hudong.com/.
[24] DBpedia[EB/OL]. [2015-09-30]. http://wiki.dbpedia.org/.
[25] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A Crystallization Point for the Web of Data[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2009, 7(3): 154-165.
[26] Wang Z, Li J, Tang J. Boosting Cross-Lingual Knowledge Linking via Concept Annotation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2013.
[27] Mei Jiaju, Zhu Yiming, Gao Yunqi, et al. TongYiCiCiLin[M]. Shanghai: Shanghai Lexicographical Publishing House, 1983.
[28] Setubal J C, Meidanis J. Introduction to Computational Molecular Biology[M]. PWS Pub.Co., 1997.
[29] Institute of Computing Technology, Chinese Academy of Sciences. ICTCLAS[EB/OL]. [2013-01-03]. http://ictclas.org/.
[30] Needleman S B, Wunsch C D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins[J]. Journal of Molecular Biology, 1970, 48(3): 443-453.
[31] Wang T, Song J, Di R, et al. A Thesaurus and Online Encyclopedia Merging Method for Large Scale Domain-Ontology Automatic Construction[C]//Proceedings of the International Conference on Knowledge Science, Engineering and Management. Heidelberg, Germany: Springer-Verlag Berlin, 2013: 132-146.
[32] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals[J]. Soviet Physics Doklady, 1966, 10: 707-710.

Author Contributions

Wang Ting: Proposed research methodology and ideas, conducted system design and implementation, drafted the paper;
Gao Ying: Revised the paper;
Liu Jingwei: Participated in research design and data analysis.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is available at the journal's website: http://www.infotech.ac.cn.
[1] Wang T. HIT-IRLab-TongYiCiCiLin (Extended Version)*{{{full}}}*{{2005}}}.3.3.txt. Harbin Institute of Technology TongYiCiCiLin (Extended Version).
[2] Wang T. ICTCLAS50_{{{Windows}}*{{32}}}*{JNI}.rar. NLPIR Chinese Word Segmentation System.
[3] Wang T. Baidu-Ontology-Concepts.rar. Baidu Baike Top-level Category Tree—13 Major Category Ontology Concept Sets.
[4] Wang T. Hudong-Ontology-Concepts.rar. Hudong Baike Top-level Category Tree—13 Major Category Ontology Concept Sets.
[5] Wang T. DBpedia V3.8 zh-(Sub-ontology).rar. Chinese Wikipedia Top-level Category Tree—23 Major Category Ontology Concept Sets.

Received Date: 2016-08-18
Revised Date: 2016-10-22

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