Postprint: An Ontology Alignment Framework for Chinese Ontology Schemas

Authors: Wang Ting, Gao Ying, Jingwei Liu

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

Abstract

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

Method: An improved Tongyici Cilin similarity algorithm is adopted to calculate the semantic similarity of simple tokens. Furthermore, an algorithm that integrates improved Tongyici Cilin with sequence alignment is utilized to measure semantic similarity between out-of-vocabulary words.

Results: Experiments 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 exclusively on element-level similarity measurement of Chinese ontology concepts, without considering 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 model's overall performance is significantly superior to existing algorithms.

Full Text

ChinaXiv Partner Journal, Issue 2, 2017

A Chinese Ontology Schema Alignment Framework*

Wang Ting, Gao Ying, Liu Jingwei (School of Information, Capital University of Economics and Business, Beijing 100070, China)

Abstract

[**Objective**] Existing ontology alignment methods often overlook the semantic characteristics of Chinese concepts, particularly their sensitivity to word order and prevalence of polysemy. This paper proposes a large-scale Chinese ontology mapping model based on TongYiCiCiLin (TYCCL) and sequence alignment algorithms. [**Methods**] The model employs an improved TYCCL-based similarity algorithm to compute semantic similarity between atomic concepts. For out-of-vocabulary terms, it integrates the improved TYCCL algorithm with sequence alignment to measure semantic similarity. [**Results**] Experiments on a test corpus constructed from DBpedia (Chinese version), Baidu Baike, and Hudong Baike demonstrate that the model achieves average precision, recall, and F1-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 and does not consider 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

Classification Number: G353.1

Introduction

The vision of the Semantic Web is to create 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 normalized specification of shared concepts within a specific domain [2], forming the foundation for web-based knowledge sharing and semantic interoperability. Current research on Linked Open Data (LOD) [3] primarily focuses on the instance level [4-5]. However, due to heterogeneity among different ontologies, reusing and sharing ontologies remains challenging. Consequently, schema-level linked data construction research is equally important as a prerequisite for LOD [6].

Ontology Mapping, a typical scenario in schema-level linked data construction, has been extensively studied. Its objective is to discover semantic associations between concepts in heterogeneous ontologies or LOD datasets. With the rapid development of the semantic web, large-scale Chinese ontologies and knowledge bases are increasingly being constructed and shared. However, due to cultural and contextual factors, research on building large-scale Chinese linked data networks remains in its early stages, and mature schema-level models for large-scale Chinese linked data are lacking. To address semantic interoperability and sharing challenges for Chinese ontologies in linked data networks, 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 similarity propagation graphs from concept hierarchies and propagates similarity scores between concepts. Cohen et al. [8] analyzed several typical element-level similarity algorithms based on edit distance and token matching, evaluating their performance. Giunchiglia et al. [9] proposed a linguistic approach incorporating shared knowledge dictionaries such as WordNet [10] to discover semantic relationships. 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 that leverages hierarchical relationships 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, incorporating universal field theory to handle large-scale ontology mapping tasks. Jain et al. [6] presented the BLOOMS system, which uses Wikipedia's top-level category tree as a knowledge base for similarity computation in LOD-based schema-level linking. However, these systems are designed exclusively for English-language semantic datasets.

In recent years, scholars have increasingly focused on Chinese ontology and linked data construction. At the schema level (i.e., ontology mapping), 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 neglects 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] proposed a Chinese word semantic similarity algorithm based on the Extended TongYiCiCiLin [17], but it does not address similarity computation for out-of-vocabulary terms and has not been applied in actual large-scale linked data networks.

Additionally, several instance-level linked data systems exist. Silk [19-20] is a framework for linking datasets across different sources, featuring a declarative language that allows users to configure links between datasets, including link

types and conditions, and supports linking remote datasets with local ones. Hassanzadeh et al. [21] provided a general and extensible framework called LinQL that integrates various existing link discovery methods to help users select appropriate approaches for dataset linking, including RDF data published from relational databases via D2RQ or Virtuoso. Wang et al. [5] extracted hierarchical relationships from Chinese encyclopedia category systems and obtained concept attributes and instances from Infobox-enabled web pages to construct two large-scale Chinese ontology libraries based on Baidu Baike and Hudong Baike, establishing simple keyword-matching-based coreference relationships with DBpedia. Niu et al. [4] semantically integrated Baidu Baike [22], Hudong Baike [23], and Chinese Wikipedia [24-25] to develop Zhishi.me, a Chinese instance-level linked data application system. For knowledge sharing, reuse, and semantic interoperability across linked data networks, cross-lingual ontology linking and mapping become 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, then employs a regression learning model to predict potential cross-lingual links between Chinese and English Wikipedia. However, these systems focus solely on instance-level relationships without addressing schema-level link discovery.

In summary, few large-scale Chinese ontologies are currently published on the Web, and significant heterogeneity exists among them. Existing Chinese ontology mapping systems exhibit low efficiency and limited usability for large-scale tasks, and no large-scale Chinese ontology mapping system specifically designed for LOD environments exists. This paper proposes a new Chinese ontology mapping model based on the Extended TongYiCiCiLin and sequence alignment principles. The model effectively addresses word order sensitivity and polysemy in Chinese concept similarity computation, demonstrating superior overall performance compared to previous work on large-scale ontology test sets constructed from Chinese web encyclopedias.

Problem Definitions

In the Extended TongYiCiCiLin (TYCCL), the vocabulary included are termed simple word elements. In Chinese ontology mapping systems, both simple word elements and out-of-vocabulary terms correspond to ontology concepts. We refer to simple word elements as **Atomic Concepts (AC)** and out-of-vocabulary terms as **Component Concepts (CC)**, where a component concept consists of a linear arrangement of multiple atomic concepts.

Definition 1 (Ontology Mapping): Given two ontologies to be mapped, O_s and O_t, for each concept C_s in O_s, find a concept C_t in O_t with identical or similar semantics. The mapping function is defined as $map : O_s \to O_t$:

For $\forall C_s \in O_s, \forall C_t \in O_t$, if $sim(C_s, C_t) > t$, then $map(C_s) = C_t$, where $sim(C_s, C_t)$ represents the similarity between C_s and C_t, and t is a threshold. When the semantic similarity exceeds t, the pair $\langle C_s, C_t \rangle$ is considered an equiv-

alent concept mapping.

Definition 2 (Semantic Knowledge Base): The complete set of words included in the Extended TongYiCiCiLin and their semantic relationships constitute a Semantic Knowledge Base, denoted as SKB_{TYCCL}. This set comprises atomic concepts: $SKB_{TYCCL} = \{AC_1, AC_2, \ldots, AC_N\}$, where N is the total number of word elements included in the knowledge base.

Definition 3 (Component Concept): A component concept CC_i consists of an ordered sequence of atomic concepts. For $\forall AC_i \in SKB_{TYCCL}$, we introduce two-dimensional indices i and j, yielding the ordered sequence $CC_i = [AC_{i1}, AC_{i2}, \ldots, AC_{ij}]$, where $j \geq 1$ and $CC_i \notin SKB_{TYCCL}$. Here, j represents the position of atomic concept AC_i in the ordered sequence CC_i. Notably, for all atomic concepts AC_i, we have $AC_i = [AC_i]$.

Definition 4 (Concept Representation): For concepts C_s and C_t from ontologies O_s and O_t, respectively, we have $C_s = CC_s = [AC_{s1}, AC_{s2}, \ldots, AC_{sm}]$ and $C_t = CC_t = [AC_{t1}, AC_{t2}, \ldots, AC_{tn}]$, where m and n denote the lengths of the ordered sequences CC_s and CC_t, with $m, n \geq 1$.

4. A Chinese Ontology Mapping Model Based on TongYiCiCiLin and Sequence Alignment

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

[Figure 1: see original paper] shows the Chinese ontology mapping model based on TongYiCiCiLin and sequence alignment. For any two concepts C_s and C_t from source ontology O_s and target ontology O_t, three cases arise during semantic similarity computation:

1. Both C_s and C_t are atomic concepts: $C_s \in SKB_{TYCCL}$ and $C_t \in SKB_{TYCCL}$.
2. One concept is atomic while the other is a component concept: $C_s \notin SKB_{TYCCL}$ or $C_t \notin SKB_{TYCCL}$.
3. Both concepts are component concepts: $C_s \notin SKB_{TYCCL}$ and $C_t \notin SKB_{TYCCL}$.

For case (1), we directly apply the "Improved TYCCL Similarity Computation" module to calculate semantic similarity between two atomic concepts. For cases (2) and (3), we employ a multi-strategy fusion approach combining "Sequence Alignment Processing" and "Improved TYCCL Similarity Computation." The "Component Concept Similarity Computation" module receives as input two word sequence strings CC_s and CC_t and their corresponding scoring matrix,

which is collaboratively generated by the "Component Concept Segmentation" and "Scoring Matrix Construction" modules.

4.1 Improved TongYiCiCiLin Similarity Computation

TongYiCiCiLin is a Chinese thesaurus that encodes each word and organizes them in an inverted tree structure, where each node represents a concept. Chinese concept coreference identification can be abstracted as Chinese synonym recognition and semantic similarity computation, making TYCCL an optimal resource. This study adopts the Harbin Institute of Technology's Extended TongYiCiCiLin as the commonsense knowledge base for Chinese ontology mapping relationship extraction.

During experimentation, we observed that the traditional algorithm proposed by Tian et al. [18] overemphasizes semantic relatedness between concepts—specifically, the hierarchical parent-child relationships in TYCCL interfere significantly with extracting equivalence relationships between ontology concepts. Since ontology mapping aims to discover equivalence rather than taxonomic relationships, we introduce a semantic adjustment factor and concept similarity weight coefficient to adapt the traditional algorithm for Chinese ontology mapping tasks in LOD environments.

TYCCL organizes word elements in a hierarchical structure with five layers. Each layer has a corresponding code identifier, and the five-layer codes are arranged sequentially to form a word's TYCCL code. Using the word "matter" (TYCCL code: Ba01A02=) as an example, explains the encoding format.

shows the TYCCL code example with sub-codes B, a, 01, A, 02, and "=", representing major category, medium category, minor category, word group, atomic word group, and synonym/equivalent/isolated relationships across layers 1-5, respectively.

Based on TYCCL's structural characteristics, we first parse the TYCCL codes of concepts to be mapped, extracting sub-codes from layers 1 to 5, then compare them starting from layer 1. If sub-codes differ at deeper layers, higher similarity weights are assigned; differences at shallower layers indicate poorer semantic relatedness (lower weights). This improved method simultaneously considers the impact of hierarchical factors on similarity computation results, with branch node counts at each layer also influencing similarity.

We present the TYCCL-based similarity computation method in 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 emphasizes semantic similarity between concepts, we introduce a semantic adjustment factor λ to regulate the relationship between

semantic relatedness and similarity across different hierarchical levels and control the potential similarity between word elements at different branch levels, where $\lambda \in (0, 1)$. Larger λ values indicate greater possibility of similarity or equivalence between word elements at different levels and stronger influence of hierarchical semantic relatedness on final concept similarity. For Chinese ontology mapping tasks, which prioritize semantic similarity, λ should not be set too high.

We define $L = \{1, 2, 3, 4, 5\}$, where for $\forall L_i \in L$, L_i represents the layer number where sub-codes differ, and $|L|$ denotes the number of elements in set L (constantly 5 in this system). The proposed concept similarity weight coefficient is $\lambda \times (L_i/|L|)$. N_t represents the total number of nodes at layer i for word elements C_s and C_t, and D is the code distance between C_s and C_t. Notably, when all five layers of codes are identical and the last character of the TYCCL code is "=", the similarity function SIM_T returns 1.0. The function's range is $(0, 1]$.

4.2 Sequence Alignment-Based Component Concept Similarity Computation

Many scholars have proposed solutions for Chinese component concept similarity computation. For example, Li et al. [14] designed an element-level concept similarity method based on HowNet. When handling out-of-vocabulary terms, this method traverses the atomic concept sequences of two component concepts to identify the atomic concept mapping pair with maximum similarity, then computes the overall similarity between component concepts using formula (2):

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

where B_{xy} represents elements in the similarity matrix formed by known words from the split vocabulary, $\max_i(B_{xy})$ denotes the i-th largest similarity value in the matrix, and $\max(m, n)$ takes the larger of the row or column indices.

However, Chinese concepts universally exhibit "word order sensitivity," making the above approach prone to semantic similarity computation errors. For instance, consider two component concepts from different ontologies: "historical theory" and "history of thought". After segmentation, we obtain the ordered sequences [历史, 理论] and [思想, 史]. Using conventional out-of-vocabulary processing methods yields the atomic concept mapping results shown in [Figure 2: see original paper]. Based on the Extended TYCCL and formula (1), the element-level similarity between these concepts is computed as 1.0 using formula (2), which is completely unreasonable. This error stems from neglecting the prevalent "word order sensitivity" and "polysemy" characteristics of Chinese natural language.

[Figure 2: see original paper] illustrates the incorrect matching result. Therefore, we propose an improved semantic similarity computation method that incorporates global pairwise sequence alignment algorithms from bioinformatics for element-level similarity calculation.

(1) Overview of Sequence Alignment Algorithms

In bioinformatics, pairwise sequence alignment arranges two DNA, RNA, or protein sequences to identify similarities, allowing gap insertion so that identical or similar symbols align in columns. By comparing similar fragments and conserved sites, potential molecular evolutionary relationships can be identified [28]. Alignment models fall into two categories: global alignment, which examines overall similarity between sequences through full scanning and comparison, and local alignment, which focuses on specific fragments. Both can be solved using dynamic programming.

(2) Constructing the Dynamic Programming Scoring Matrix

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

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

Scoring Matrix Construction: The two word sequences to be aligned are represented as a scoring matrix M. For concepts C_s and C_t from ontologies O_s and O_t, row i of matrix M corresponds to atomic concept AC_{si} in sequence CC_s, and column j corresponds to atomic concept AC_{tj} in sequence CC_t, where $i \leq m$, $j \leq n$. The element at row i, column j is denoted M_{ij}.

Following dynamic programming principles, the two sequences are represented as rows and columns. If sequence CC_s has length m and CC_t has length n, they form an $(m+1) \times (n+1)$ matrix with CC_s as rows and CC_t as columns. For example, component concepts "Second Industrial Revolution" and "World War II war criminals" yield sequences $CC_s = [\ ,\ ,\ \ ]$ and $CC_t = [\ ,\ ,\ \ ,\ ]$ after segmentation.

(3) Optimal Recursive Solution Algorithm

Concept similarity computation is abstracted as an alignment process between two word sequences. Through a gap penalty function, gaps "-" are inserted at appropriate positions to equalize sequence lengths, establishing correspondences between atomic concepts or between atomic concepts and gaps. Sequence alignment essentially identifies the optimal global pairing between two component concept sequences using a scoring strategy.

The Needleman-Wunsch algorithm, proposed in 1970, is a classic dynamic programming algorithm for global sequence similarity comparison, suitable for se-

quences with high macro-level similarity [30]. This work primarily uses this algorithm and dynamic programming principles to recursively solve for the optimal alignment path in matrix M.

Algorithm 1: ConceptSimilarity(CC_s, CC_t) - **Input:** Scoring matrix $M(i)(j)$ for component concepts CC_s and CC_t - **Output:** Matrix $M'(i)(j)$ containing the optimal alignment path

```
p ← -0.05 // Define constant p as penalty factor, equal to -0.05
for each i ← 1, 2, ⋯, m+1; j ← 1, 2, ⋯, n+1 // Initialize dynamic programming matrix
    M(i)(n+1) ← p×(m-i+1)
    M(i+1)(j) ← p×(n-j+1)
end for
for each i ← m, m-1, ⋯, 1
    for each j ← n, n-1, ⋯, 1
        M(i)(j) ← max(M(i+1)(j+1) + SIM_T(AC_si, AC_tj), M(i)(j+1) + p, M(i+1)(j) + p)
          // Recursively compute cost for each matrix element
    end for
end for
  Backtrack to obtain matrix M'(i)(j) containing the optimal alignment path
  return M'(i)(j)
```

First, the penalty factor $p = -0.05$ is defined, and the $(n+1)$-th column and $(m+1)$-th row of the matrix are initialized using $M(i)(n+1) = p \times (m-i+1)$ and $M(m+1)(j) = p \times (n-j+1)$. Second, the remaining $m \times n$ elements are recursively solved using the TYCCL similarity function SIM_T. The scoring function f is defined in formula (3):

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

Considering word order sensitivity in Chinese component concepts, recursion begins at the end of both sequences (element M_{mn}). The recursive rule (gap penalty function) is given in formula (4):

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

Finally, backtracking from M_{mn} to M_{11} yields the optimal alignment path. In the resulting matrix, bold arrows indicate the optimal path: diagonal bold arrows pair corresponding atomic concepts, horizontal bold arrows insert a gap "-" before the atomic concept in sequence CC_s, and vertical bold arrows insert a gap before the atomic concept in CC_t. If multiple optimal paths exist, any one may be selected.

After gap insertion, the two sequences have equal length, denoted CC'_s and CC'_t with length L. The final similarity between component concepts is computed using formula (5):

$$SIM_{NW}(CC'_s, CC'_t) = \frac{\sum_{i=1}^{L} f(AC_{si}, AC_{ti})}{L}$$

5. Experiments

5.1 Data Sources

This study uses Chinese web-based open encyclopedia knowledge bases as experimental data sources. In addition to DBpedia (Chinese version), we crawled and parsed open category pages and entry pages from Baidu Baike and Hudong Baike using the HTMLParser toolkit, following the methodology in [5,31]. The extracted Infobox structured information was organized as Chinese triples to form large-scale Chinese open-domain knowledge bases for mapping. As shown in , the ontology concept system primarily consists of encyclopedia open category systems.

presents statistics for the Chinese web encyclopedia knowledge bases: Baidu Baike contains 21,152 Infobox Chinese triples with 2.30% Infobox frequency; Hudong Baike contains 10.10% frequency; and DBpedia 3.8 (Chinese version) contains 19.74% frequency with predicates in Infoboxes.

5.2 Evaluation Metrics

We adopt Precision, Recall, and F1-measure as evaluation criteria for Chinese concept equivalence identification:

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

$$Recall(R) = \frac{\text{Number of correctly mapped pairs output}}{\text{Total number of mapping pairs in ground truth}}$$

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

Four senior undergraduate students from the School of Information at Capital University of Economics and Business manually identified and annotated Chinese concept equivalence relationships in the top-level category trees of DBpedia, Baidu Baike, and Hudong Baike. These annotations serve as ground truth mapping pairs for the ontology mapping experiments, with statistics provided in through .

shows reference mapping counts for the Baidu-Hudong mapping task across top-level categories. and present similar statistics for Hudong-DBpedia and Baidu-DBpedia mapping tasks, respectively.

5.3 Sequence Alignment Results Analysis

After presenting the sequence alignment-based similarity computation method, we re-evaluate previous examples:

Example 1: CC_s = [,], CC_t = [,]. Using formula (2) yields $Sim(CC_s, CC_t) = (1.0 + 1.0)/2 = 1.0$. However, the sequence alignment algorithm produces the alignment shown in [Figure 3: see original paper] with the scoring matrix in [Figure 4: see original paper], resulting in $SIM_{NW}(CC'_s, CC'_t) = (-0.05 + 1.0 - 0.05)/3 = 0.3$. This example mapping pair originates from the "History" subtask in the Hudong-DBpedia mapping task.

[Figure 3: see original paper] shows the correct alignment result, while [Figure 4: see original paper] displays the scoring matrix for Example 1.

Example 2: CC_s = [, ,], CC_t = [, , ,]. Using formula (2) incorrectly yields $Sim(CC_s, CC_t) = 1.0$ due to polysemy of "次". In TYCCL, "次" has multiple encoding entries, including "Dn04B03=" which identifies "第二" and "次" as equivalent. This produces four atomic concept mappings with similarity 1.0: ⟨第二, 次⟩, ⟨第二, 第二⟩, ⟨次, 第二⟩, and ⟨次, 次⟩, resulting in $Sim(CC_s, CC_t) = (1.0 + 1.0 + 1.0 + 1.0)/4 = 1.0$. The sequence alignment algorithm yields $SIM_{NW}(CC'_s, CC'_t) = (1.0+1.0+0.18-0.05)/4 = 0.5325$. The optimal scoring matrix $M'(i)(j)$ is shown in [Figure 5: see original paper], with the optimal sequence matching result in [Figure 6: see original paper]. This example originates from the "History" subtask in the Baidu-DBpedia mapping task.

[Figure 5: see original paper] shows the scoring matrix for Example 2, and [Figure 6: see original paper] displays the sequence matching result.

These examples demonstrate that traditional methods incorrectly assign similarity 1.0 to non-equivalent component concept pairs. In contrast, Algorithm 1 produces more reasonable similarity values. By addressing word order sensitivity and polysemy through the Needleman-Wunsch global alignment algorithm, our approach effectively avoids mapping errors inherent in conventional methods like [14]. When atomic concept sequences in component concepts share similar semantic order, Algorithm 1's performance aligns with traditional methods. Overall, the global sequence alignment-based element-level similarity algorithm offers superior advantages and rationality for large-scale Chinese ontology mapping tasks.

5.4 Large-Scale Chinese Ontology Mapping Results

Guided by the theoretical framework, we evaluated our prototype system's performance using three major Chinese web encyclopedia knowledge bases in a

large-scale linked data construction scenario. Evaluation results for three mapping tasks are presented in through , comparing four typical similarity algorithms: (1) cross-lingual edit distance [32], (2) traditional TYCCL-based similarity [18], (3) Li et al.'s HowNet-based ELOMC algorithm [14], and (4) our proposed comprehensive Chinese concept similarity algorithm. For fairness, the equivalence threshold was uniformly set to $t = 0.9$.

shows Baidu-Hudong mapping results. Our system's average precision is approximately 41% and 39% higher than traditional TYCCL and ELOMC algorithms, respectively. Recall exceeds edit distance and traditional TYCCL by about 13% and 2%, respectively, while matching ELOMC. The F1-measure surpasses edit distance, traditional TYCCL, and ELOMC by approximately 8%, 23%, and 20%, respectively.

presents Hudong-DBpedia mapping results. Precision exceeds edit distance, traditional TYCCL, and ELOMC by about 1%, 10%, and 11%, respectively. Recall surpasses edit distance and traditional TYCCL by about 6% and 1%, matching ELOMC. The F1-measure exceeds the three comparison algorithms by approximately 3%, 6%, and 6%, respectively.

shows Baidu-DBpedia mapping results. Precision is approximately 39% and 43% higher than traditional TYCCL and ELOMC, respectively. Recall exceeds edit distance, traditional TYCCL, and ELOMC by about 17%, 6%, and 3%, respectively. The F1-measure surpasses the three algorithms by approximately 8%, 26%, and 30%, respectively.

In Baidu-Hudong and Baidu-DBpedia tasks, our precision is slightly lower than edit distance due to controversial or misclassified synonym pairs in TYCCL (e.g., ⟨民族, 中华民族⟩, ⟨刑法, 刑事⟩, ⟨军队, 军事⟩, ⟨辛亥革命, 革命⟩). These appear frequently in the "Society" subtask but rarely in other tasks. Edit distance mechanically compares literal similarity without considering semantic similarity, resulting in significantly lower recall across all tasks. Our method's integration of TYCCL and improvements to the traditional algorithm yields higher recall than edit distance and traditional TYCCL, matching ELOMC's performance.

Overall, across 18 subtasks in three mapping tasks, our model achieves average precision, recall, and F1-measure of approximately 97.5%, 87.8%, and 92.1%, respectively. While precision is slightly lower than edit distance due to occasional misclassifications in TYCCL, our method's F1-measure consistently outperforms all comparison systems across three mapping tasks.

Conclusion

With limited mature large-scale Chinese ontology mapping systems available, this paper addresses schema matching challenges in linked data network construction by proposing a novel Chinese ontology mapping model integrating TongYiCiCiLin and global sequence alignment algorithms. The system resolves usability issues in large-scale ontology mapping by targeting the "word order

sensitivity" and "polysemy" characteristics of Chinese ontologies for element-level mapping. Future work will incorporate instance-level and concept definition similarity mapping parameters to further enhance system robustness and accuracy.

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 the Semantic Web and 2nd Asian Conference on 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. TongYicCiCiLin (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 Intelligent. 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/. (NLPIR Chinese Word Segmentation System[EB/OL]. [2013-01-03]. http://ictclas.nlpir.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 Contribution Statement

Wang Ting: Proposed research methodology and system design, implemented the system, and drafted the manuscript; Gao Ying: Revised the manuscript; 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 in the online version of the journal at 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 Ontology Concept Sets. [4] Wang T. Hudong-Ontology-Concepts.rar. Hudong Baike Top-Level Category Tree—13 Major Ontology Concept Sets. [5] Wang T. DBpedia V3.8 zh-(Sub-Ontology).rar. Chinese Wikipedia Top-Level Category Tree—23 Major Ontology Concept Sets.

NISO Publishes Updated ResourceSync Framework Specification

The National Information Standards Organization (NISO) recently announced the official publication of an updated ResourceSync framework specification (ANSI/NISO Z39.99-2017). Approved by the American National Standards Institute (ANSI), this version 1.1 improves a web standard detailing server capabilities for enabling third-party systems to synchronize with evolving resources. Such synchronization is crucial in contemporary environments where web-based content and its metadata change continuously.

Initially released in 2014 as the ResourceSync "core" specification, ANSI/NISO Z39.99 provides easily implementable server functions for remote systems to maintain synchronization with evolving resources. It describes how servers should declare supported facilities and includes extensive examples and use cases. The recent revision clarifies issues such as confusion between a resource's last modification date and the notification date of its modification.

"Web resources and collections of web resources continuously evolve, and in many cases applications that want to make use of these resources need to be confident that the data they use is up to date," said Herbert Van de Sompel, co-chair of the ResourceSync Working Group. "Our revision of the ResourceSync core specification strengthens a standard that can meet resource discovery and synchronization needs between different systems in scholarly communication, cultural heritage, and education. ResourceSync is designed to be highly modular, based on HTTP and the Sitemap protocol to ensure easy implementation in many applications, including but not limited to timely sharing of data from different types of repositories. Additionally, related optional specifications pro-

vide extensions to the ANSI/NISO ResourceSync core, including support for synchronizing information archives and push-based change notifications."

The ResourceSync specification and video tutorials for using ANSI/NISO Z39.99-2017 are available on the NISO website at http://www.niso.org/workrooms/resourcesync/.

(Compiled from: http://www.niso.org/news/pr/view?item_key=96962d7722cc13a1e20c40e2ca3c2ca8ca80359d

(Journal Correspondence)

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