Clinga: Bringing Chinese Physical and Human Geography in Linked Open Data

Wei Hu(✉), Haoxuan Li, Zequn Sun, Xinqi Qian, Lingkun Xue, Ermei Cao, and Yuzhong Qu

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{whu,yzqu}@nju.edu.cn, hxli.nju@gmail.com, zqsun.nju@gmail.com, xqqian.nju@gmail.com, lkxue.nju@gmail.com, emcao.nju@gmail.com

Abstract. While the geographical domain has long been involved as an important part of the Linked Data, the small amount of Chinese linked geographical data impedes the integration and sharing of both Chinese and cross-lingual knowledge. In this paper, we contribute to the development of a new Chinese linked geographical dataset named Clinga, by obtaining data from the largest Chinese wiki encyclopedia. We manually design a new geography ontology to categorize a wide range of physical and human geographical entities, and carry out an automatic discovery of links to existing knowledge bases. The resulted dataset contains over half million Chinese geographical entities and is open access.

Resource Type: Dataset and ontology
Permanent URL: http://w3id.org/clinga

1 Introduction

To embrace the vision of the Semantic Web, significant efforts have been devoted towards creating linked data for the geographical domain, such as GeoNames[1], GeoLink[2], GeoWordNet [3] and LinkedGeoData [7], in addition to a few multilingual, open-domain knowledge bases like DBpedia [5], Freebase [2], YAGO [4] and Wikidata [8] involving geographical data as well. Although they are being widely used by many semantic applications, the amount of Chinese geographical data is relatively limited and mainly exist in their labels. For example, according to our analysis of GeoNames 2015, only 4.6 % of its geographical features involve Chinese names, even for the features in China, this proportion is still 63 %. Despite two early works, Zhishi.me [6] and XLore [9], which extracted knowledge from Chinese wiki encyclopedias, to the best of our knowledge, there is no prior work on building Chinese linked geographical data.

[1] http://www.geonames.org/ontology.
[2] http://www.geolink.org.

© Springer International Publishing AG 2016
P. Groth et al. (Eds.): ISWC 2016, Part II, LNCS 9982, pp. 104–112, 2016.
DOI: 10.1007/978-3-319-46547-0_11

In this paper, we present a new **C**hinese **lin**ked **g**eographical d**a**taset, Clinga[3], which contains generally but not exclusively a large number of geographical entities in China (e.g. cities) and their relations (e.g. has-capital). Compared with existing geographical datasets, Clinga has three distinguished features:

- We obtain our data from Baidu Baike[4], the largest collaboratively-built Chinese wiki encyclopedia. Both structural data and textual description (e.g. in main text) of an article are extracted and translated to RDF using our structure ontology, to achieve the completeness of the article at our best effort.
- We manually design a physical and human geography ontology to categorize various geographical entities. Following the Chinese naming conventions, we combine the type-based heuristic rules and an SVM classifier to obtain good categorization accuracy.
- We link Clinga to existing knowledge bases like DBpedia and GeoNames. An automatic discovery of entity links is conducted based on bilingual (Chinese and English) labels and manually-defined ontology mappings.

Clinga is expected to be useful not only as a complementary source for location-based semantic applications such as DBpedia Mobile [1], but also as our primary knowledge base for answering geographical questions in the national matriculation examination of China (called GaoKao), under the support of the National High-tech R&D Program of China.

2 Development Methods

Figure 1 shows the general steps to develop the Clinga dataset. To achieve a satisfactory qualitative and quantitative result, the methodology that we follow is largely automatic, with a little amount of human intervention in critical parts.

We choose Baidu Baike rather than the Chinese Wikipedia[5] as the main data source due to its larger scale and richer content, especially about geography and contemporary people in China. At the time of writing, Baidu Baike has 13 million articles and is 15 times more than the Chinese Wikipedia. For another Chinese wiki encyclopedia, Hudong Baike[6], it significantly overlaps with Baidu Baike but has less influence in the Chinese language community[7]. It worth noting that the proposed methods can be smoothly applied to these wiki encyclopedias.

Technically, our data extraction process is similar to DBpedia, but having geographical data in Baidu Baike is not a trivial task. First, the category structure of Baidu Baike is often incorrect and inconsistent. For example, "Administrative division" and "Administrative region" categories co-exist

[3] Clinga is publicly available at http://w3id.org/clinga, under the Creative Commons BY-NC 4.0 license. It is also registered on https://datahub.io. Documentation and online services including SPARQL endpoint and keyword search are accessible at http://ws.nju.edu.cn/clinga.

[4] http://baike.baidu.com.

[5] http://zh.wikipedia.org/.

[6] http://www.baike.com/.

[7] https://strategy.wikimedia.org/wiki/Case_studies/Baidu_and_Hudong.

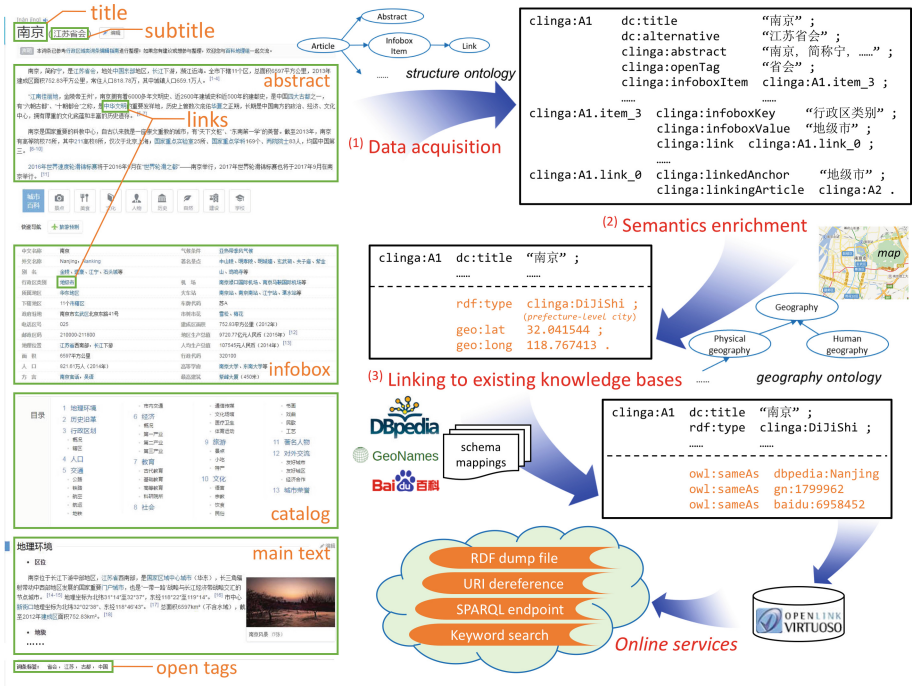

Fig. 1. Left part is a typical article in Baidu Baike, and right part is the methodological steps to develop Clinga, including: (1) data acquisition, (2) semantics enrichment, and (3) linking to existing knowledge bases.

with a very similar meaning; "Ancient place name" appears redundantly under "Place name". Second, mismatches frequently happen between categories and articles. For instance, several articles about people are assigned to "Administrative division"; hundreds of mountains are not categorized under "Mountain". These problems prevent us from directly using the category structure of Baidu Baike to obtain its geographical data. Besides, the main text is often causal and complex.

Therefore, we exhaustively crawl all the articles in Baidu Baike, and design a geography ontology to identify physical and human geographical entities. We also conduct an automatic discovery of links to knowledge bases such as DBpedia and GeoNames. The details of our methods are described below.

2.1 Data Acquisition

For an article in Baidu Baike (see the left part of Fig. 1), a unique and sequential ID is generated for entity URI (e.g. http://ws.nju.edu.cn/clinga/A1). Then, the following items are extracted and translated using our structure ontology.

Title and subtitle. An article has exactly one title and optionally one subtitle, which are converted as the values of dc:title and dc:alternative, respectively.

Infobox. An infobox presents the structural facets of an article using key-value pairs, and keys are later converted to properties in our geography ontology. A compound structure is used to model a key, a value, and links in the value.

Abstract. The first few paragraphs before infobox form the abstract of an article, which is described as the value of clinga:abstract.

Catalog and main text. An article organizes its main text by separated sections, and section titles form the catalog of the article. The correspondence between the title and paragraphs of a section is reserved, which is especially useful for geographical question answering, e.g. searching a related section and performing natural language understanding. However, the text in each paragraph is not RDFized yet.

Open tag. An article may have a few plain-text tags indicating its topics, which are converted as the values of clinga:openTag. Note that these tags are much more casual than the categories in the Chinese Wikipedia.

Link. Links in an article refer to other articles within Baidu Baike or webpages outside, which are described using clinga:link.

Furthermore, two other kinds of information are particularly considered.

Redirect. Baidu Baike uses redirects to handle the synonym problem. Two articles with a redirect relation are presented using clinga:redirects.

Disambiguation. Baidu Baike does not offer disambiguation pages, but rather puts all meanings together in a polysemy list. This relation is presented using clinga:disambiguates.

2.2 Semantics Enrichment

Before developing our own ontology, we investigate some existing geographical ontologies. The ontology in GeoNames has no class hierarchy, but uses feature codes for categorization. DBpedia and XLore automatically extracted their ontologies from categories, but Baike categories are incomplete and inconsistent to be reused directly. By discussing with the professors at the School of Geographic and Oceanographic Sciences in our university and referencing the DBpedia ontology and textbooks, we design a new geography ontology involving two main branches, physical geography and human geography, to categorize the geographical entities in Clinga. See Fig. 2 for an excerpt of our geography ontology.

We introduce a two-step method for entity categorization. In the first step, we define 213 heuristic rules to extract the candidate entity set for each most-specific type (i.e. the "leaf" class at the class hierarchy). The rationale is based on the Chinese naming conventions and distinguished infobox keys. For example, the names of mountains in Chinese usually end with the same Chinese character "山". With these rules, we generate candidate entities for each type.

However, the heuristics-based categorization may also contain wrongly-typed entities. For instance, the last character of many Chinese people's names is also "山". So in the second step, we adopt a machine learning method to filter these errors.

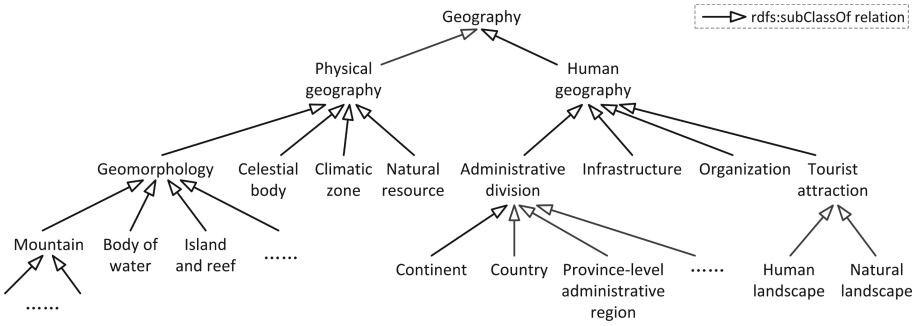

Fig. 2. Excerpt of the geography ontology. The actual version contains 130 classes with preferred labels in Chinese and alternative labels in both Chinese and English, where 35 leaf classes are newly involved as compared with GeoNames.

Specifically, an SVM classifier is employed with the RBF (Radial Basis Function) kernel[8], which performs better than the linear kernel based on human observation on some sample data. Article titles, subtitles, keys in infoboxes, first sentences of abstracts and open tags are pre-processed by word segmentation, stop word removal and TF-IDF scoring, and form the feature vectors for SVM. By manually creating a training set for each type, we find that the SVM-based filtering performs very well. More details are provided in Sect. 3.

After identifying all geographical entities for Clinga, we extract all the properties from their infobox keys. We follow the idea of DBpedia to distinguish raw properties and mapping-based properties. In total, there are 41,827 raw properties in all the infoboxes, in which 124 properties are mandated by Baidu Baike when creating articles, while the rest of them are user-customized and not validated. By considering the infobox values and manual refinement, 54 properties are identified as object properties and 70 are datatype properties.

To ensure the coverage of the Clinga dataset, we also use a gazetteer about administrative divisions and landforms in China. Furthermore, coordinates are complemented (if not exists in infoboxes) by querying Baidu Map.

2.3 Linking to Existing Knowledge Bases

Following the Linked Data principles, we connect Clinga to DBpedia and GeoNames to promote the integration and sharing of both Chinese and cross-lingual geographical knowledge. DBpedia has become the hub of the Linked Data, and GeoNames is one of the most famous geo-spatial database, but their schemas are completely different. To overcome the heterogeneity and achieve good accuracy, we first manually exploit 118 mappings among their classes with our best effort. Two class mappings between Clinga, DBpedia and GeoNames are below:

[8] We use scikit-learn 0.17.1 (via LIBSVM) with default parameters.

$$\text{clinga:JiChang} =_{\mathbf{M}} \text{dbpedia:Airport} =_{\mathbf{M}} (\text{gn:featureCode value 'S.AIRP'}),$$

where $=_{\mathbf{M}}$ denotes the exact match relation between two classes.

Then, we use these class mappings to guide the entity linking process. Let $\mathbf{C}$ be the Clinga dataset and $\mathbf{D}$ be DBpedia or GeoNames. The links between them, denoted by $link(\mathbf{C}, \mathbf{D})$, are defined as the set of entity pairs having compatible types and similarities larger than $\eta \in [0, 1)$:

$$link(\mathbf{C}, \mathbf{D}) = \{(c, d) \in \mathbf{C} \times \mathbf{D} \mid type(c) \simeq_{\mathbf{M}} type(d), sim(c, d) > \eta\}, \quad (1)$$

where $\simeq_{\mathbf{M}}$ denotes the compatible relation derived from the class mappings and subclass relations. $sim()$ calculates similarities based on name comparison:

$$sim(c, d) = \begin{cases} 1 & \exists a \in alias(c)\, \exists b \in alias(d)\ (a = b) \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where $alias()$ involves various aliases, such as official names and English names. Besides, Traditional Chinese names are translated to Simplified Chinese.

For the entities holding the disambiguation relation, an *ad hoc* post filtering method is used by considering subtitles and coordinates, which keeps only one link with highest similarity for each entity.

3 Dataset Statistics

In this section, we report the statistical data of Clinga and a preliminary evaluation on its accuracy. Overall, we obtained 13,068,395 articles from Baidu Baike, and the uncompressed raw data is about 180 GB. From the articles, we identified more than 624 thousand geographical entities and generated around 73 million RDF triples (see Table 1). As compared with the Chinese portion of GeoNames 2015, Clinga contains a larger number of Chinese geographical entities and much more triples. Moreover, XLore reported in 2012 [9] that it obtained 185 thousand Chinese geographical entities using the original "Geography" category of Baidu Baike. We argue that this smaller quantity is caused by the increase of articles in recently years and the mismatch between categories and articles.

Table 2 lists the numbers of entities and links w.r.t. eight upper-level classes. We can see that the numbers vary between classes. The three classes with most entities are "Infrastructure", "Administrative division" and "Organization". Except "Administrative division", the other two have little overlap with DBpedia and GeoNames, indicating the complementarity of the three datasets.

To evaluate the accuracy of entity categorization, for each class we manually set up a training set containing 500 examples (the ratio of positive and negative examples is 5:1), and conducted 5-fold cross-validation. Classes with less than 500 entities were not counted. The F1-score in average is 0.94 and the standard deviation is 0.01. This good result was achieved because of the significant difference between entities in geography and other classes (e.g. people and movies).

Table 1. General statistics and comparison with GeoNames and XLore

	No. of entities	No. of RDF triples
Clinga	624,391	73,326,425
GeoNames (gn:alternateName lang 'zh')	467,679	546,824
XLore ("Geography" category of Baidu Baike)	185,204	665,287

Last accessed date of GeoNames RDF dump is April 21, 2015.

Table 2. Class-based statistics and entity links to DBpedia and GeoNames

Classes	No. of entities	No. of links to	
		DBpedia	GeoNames
Geomorphology	41,999	26,914	33,076
Celestial body	365	272	0
Climatic zone	28	17	0
Natural resource	1,324	128	56
Administrative division	128,697	50,694	63,820
Infrastructure	356,937	38,968	12,311
Organization	88,848	2,789	250
Tourist attraction	8,801	1,497	1,567

Note that an entity can have multiple types, e.g. Mount Huang can be both a mountain and a tourist attraction.

However, for entities not included using our heuristic rules, they would be missed by the entity categorization algorithm, which is a limitation of our method.

For entity linking, we randomly chose 100 entities for each class and manually judged the correctness of 17,381 entity links in total. The precision of entity links is 0.81. We observed that the precision on "Street" and "Administrative village" is not good, because there exist a number of articles with exactly the same title but no more information for disambiguation.

4 Related Work

The vast increase of data sources containing geographical information is bound up with the diversity of geo-spatial applications such as location-based services, among which GeoNames (see Footnote 1) is perhaps the most famous geo-spatial database collecting data like place names in many languages from many sources. GeoWordNet [3] is a semantic and linguistic resource developed by the full integration of GeoNames with WordNet and MultiWordNet Italian portion. LinkedGeoData [7] transformed and represented OpenStreetMap using RDF, and linked itself with DBpedia, GeoNames and others. GeoLink (see Footnote 2) aims to improve data retrieval, reuse, and integration of seven geoscience data repositories with ontologies. Moreover, GeoLinkedData.es[9] and Ordnance Survey[10]

[9] http://geo.linkeddata.es.
[10] http://data.ordnancesurvey.co.uk/datasets/os-linked-data.

created linked geographical data for Spain and Great Britain, respectively. Currently, Chinese data in these datasets are still sparse and often limited to names.

Zhishi.me [6] and XLore [9] are two early works on extracting open-domain knowledge from the Chinese Wikipedia, Hudong Baike and Baidu Baike. Their difference is that Zhishi.me did not provide an ontology to describe the crawled data, while XLore automatically built an ontology using categories and infobox keys. By focusing on the geographical domain, we manually constructed a geography ontology having a more consistent category structure and covering more physical and human geographical topics. Also, the amount of entities and RDF triples in the Clinga dataset is larger, due to our exhaustive ways of entity extraction and content RDFization.

5 Conclusion and Future Work

Clinga is our ongoing effort to make Chinese geographical data easily findable, accessible, interoperable and reusable. We obtained the data from Baidu Baike, and built a geography ontology to categorize physical and human geographical entities, which we further linked to DBpedia and GeoNames. At present, Clinga is freely accessible through RDF dump, URI dereference, SPARQL endpoint and keyword search. The challenges for future work include a continual improvement of the Clinga's quality, and the study of new knowledge extraction and integration methods, to better support semantic geographical applications.

Acknowledgments. This work is supported by the National High-tech R&D Program of China (No. 2015AA015406) and the National Natural Science Foundation of China (Nos. 61370019, 61223003 and 61321491).

References

1. Becker, C., Bizer, C.: Exploring the geospatial semantic web with DBpedia mobile. J. Web Seman. **7**(4), 278–286 (2009)
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD 2008, pp. 1247–1250. ACM (2008)
3. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.: GeoWordNet: a resource for geo-spatial applications. In: Aroyo, L., Antoniou, G., Hyvönen, E., Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6088, pp. 121–136. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13486-9_9
4. Hoffart, J., Suchanek, F., Berberich, K., Weikum, K.: YAGO2: a spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. **194**, 28–61 (2013)
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Seman. Web J. **6**(2), 167–195 (2015)
6. Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y.: Zhishi.me - weaving chinese linking open data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011. LNCS, vol. 7032, pp. 205–220. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25093-4_14

7. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: LinkedGeoData: a core for a web of spatial open data. Seman. Web J. **3**(4), 333–354 (2012)
8. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
9. Wang, Z., Wang, Z., Li, J., Pan, J.Z.: Knowledge extraction from chinese wiki encyclopedias. J. Zhejiang Univ. Sci. C **13**(4), 268–280 (2012)