User Generated Content Oriented Chinese Taxonomy Construction

Jinyang Li, Chengyu Wang, Xiaofeng He, Rong Zhang, Ming Gao*

Institute for Data Science and Engineering, Software Engineering Institute,
East China Normal University, Shanghai, China
{jinyangli,chengyuwang}@ecnu.cn,{xfhe,rzhang,mgao}@sei.ecnu.edu.cn

Abstract. The taxonomy is one of the basic components in knowledge graphs as it establishes types of classes and semantic relations among the classes. In this regard, taxonomy derivation is a vital task. Taxonomies are normally constructed either manually, or by language-dependent rules or patterns for type and relation extraction or inference. Existing work on building taxonomies for knowledge graphs is mostly in English language environment. In this paper, we propose a novel approach for large-scale Chinese taxonomy construction based on user generated content. We take Chinese Wikipedia as the data source, develop methods to extract classes and their relations mined from user tagged categories, and build up the taxonomy using a bottom-up strategy. The algorithms can be easily applied to other Wiki-style data sources. The experiments show that the constructed Chinese taxonomy achieves better results in both quality and quantity.

Keywords: knowledge graph, taxonomy, Wikipedia

1 Introduction

The past decades has witnessed advances in the construction of knowledge graphs in both academia and industrial circles. Projects such as Google Knowledge Graph, YAGO [1–3], DBpedia [4, 5] and Probase [6] have successfully built knowledge graphs by extracting entities and relations from Web data sources.

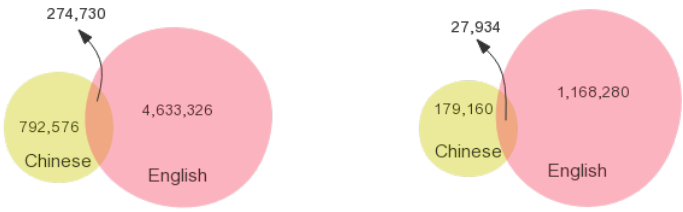
In a knowledge graph, the *taxonomy* is a basic component of the entire system as it specifies the sets of classes and entities, relations between classes and entities, and the topological structure of the knowledge graph. Currently, much research work has been devoted to creating taxonomies for knowledge graphs. Taxonomies are created manually or automatically. In projects such as ReadTheWeb [7] and DBpedia, classes and their hierarchical relations are predefined, while automatic approaches have also been proposed by exploiting the rich semantics in unstructured texts (in Probase) or semi-structured wikis (in YAGO and WikiTaxonomy[8]).

Taxonomy construction is a hot research problem. Other than the common challenges of taxonomy construction such as sparsity, incompleteness and

* Corresponding author

heterogeneity [6], taxonomy construction approaches are still highly language-dependent [9]. Therefore, taxonomy construction from Chinese Wikipedia is also challenging due to the following factors:

- **Lack of data sources:** The construction of taxonomy deeply relies on data sources. For example, Freebase[1] provides a knowledge repository, which is a backbone in Google Knowledge Graph. WordNet [10] contains rich hierarchical relations between entities and is served as a source for classes in YAGO. However, these counterparts in Chinese are not readily available.
- **Hard to obtain language patterns:** Extraction patterns in English language cannot be directly extended to other languages, such as Chinese. For example, there are no explicit singular/plural forms in Chinese nouns, while plural forms can be used to detect concepts in English [1, 8]. In [11], Chinese language patterns are designed to extract isA relations from plain text. But they can not be applied to user generated categories, which are short text rather than complete sentences.
- **Low coverage of cross-lingual approaches:** Although some cross-lingual methods have been proposed based on cross-lingual links between data sources in different languages [3, 12], the coverage is relatively low. Figure 1 shows the number of articles, categories and their overlaps for Chinese and English Wikipedias. We can find only 34.66% of articles and 15.60% of categories in Chinese Wikipedia are covered in the English version. Thus, cross-lingual approaches have limited power for entities and classes that are unique in certain languages.



(a) Number of articles (b) Number of categories

Fig. 1. Comparison of Chinese and English Wikipedias

In this paper, we construct a taxonomy from Chinese Wikipedia. We address the the problem in terms of two main ideas, namely: (a) detecting the isA relations from Wikipedia article titles and user tagged categories as accurate as possible, and (b) constructing a taxonomy in terms of all isA relations, where an isA relationship represents that one entity/class is a subclass of another. We summarize the main research contributions of this work as follows:

[1] http://www.freebase.com/

- We design a classification-based method to extract isA relations from Chinese Wikipedia titles and categories with an accuracy of over 95%. We also apply inference-based and mining-based approaches to generate isA relations that are not explicitly expressed in Wikipedia categories.
- We assemble these isA relations into a complete taxonomy. Specially, we construct a taxonomy from Wikipedia in a bottom-up manner via integrating these isA relations.
- We evaluate our constructed taxonomy in scale and accuracy measures. The experimental results show that our constructed taxonomy has high coverage and accuracy in entity space, class space and relation space.

The rest of the paper is organized as follows. Section 2 covers the related work. In section 3, we formulate the problem and present the stages of our approach for constructing taxonomy from Chinese Wikipedia. Section 4 and 5 elaborate the isA relation detection and taxonomy construction process, respectively. Section 6 covers our experimental studies on Chinese Wikipedia. Finally, we give the concluding remarks in Section 7.

2 Related Work

In recent years, a lot of research work has been focused on constructing taxonomies for knowledge graphs. In this section, we review various approaches towards the construction of a large-scale taxonomy. There are three ways of constructing taxonomy: manual approaches [4, 5], automatic approaches [8, 1–3, 6] and cross-lingual approaches [3, 12, 13].

Some knowledge graphs have a hand-craft, fixed taxonomy with fine quality, such as NELL and DBpedia. In NELL, categories are manually arranged into a hierarchical structure so that entities are extracted from texts and mapped to certain categories by coupled training [14, 7]. In DBpedia, there is a cross-lingual, universal taxonomy. Entities are mapped to the taxonomy by contributors of the project [4, 5]. The major drawback of manually constructed taxonomies is relatively low coverage, especially in newly emerged areas and specific domains.

Several projects leverage the rich semantic information in Wikipedia to derivate the taxonomy automatically. WikiTaxonomy [8] utilizes methods based on the connectivity of Wikipedia network and lexicon-syntactic features to classify isA and notIsA relations. In WordNet [10], concepts (synsets) are well organized by experts with clear semantic relations. YAGO [1–3] combines Wikipedia categories and WordNet by mapping Wikipedia categories to WordNet concepts. Currently the largest taxonomy is Probase [6]. Instead of extracting relations from Wikipedia, it takes natural languages from Web pages as input and generates isA pairs using Hearst patterns [15]. However, these approaches focus on English sources, and cannot be easily extended to Chinese sources.

The existing taxonomy can also be leveraged to construct a taxonomy in another language. In YAGO3 [3], Wikipedias in multiple languages are used to build one coherent knowledge base with the English version. Also, Wang et

al. [12, 13] studied the problem of cross-lingual taxonomy derivation from English and Chinese Wikipedias and proposed a cross-lingual knowledge validation method via Dynamic Adaptive Boosting. Although cross-lingual approaches are promising when multilingual links or knowledge exist. Due to the low coverage of cross-lingual information and the significant difference between Chinese and English, these methods can not be employed to construct taxonomies with a lot of language-specific knowledge.

3 Chinese Taxonomy Construction

Constructing a Chinese taxonomy is challenging. We briefly introduce our problem and provide a sketch of our approach in this section.

Problem Description Wikipedia is a large repository that can be modeled as a set of Wikipedia articles W. In our paper, each article is a 2-tuple $w = (e, C)$ where e is the title of the article, which is served as a candidate entity in our taxonomy and C is the set of user generated categories for e.

A taxonomy $T = (V, E)$ is a rooted, labeled tree where nodes V are entities or classes and edges E represent isA relations. Specifically, for each non-root $e \in V$, there exists a class c where (e, isA, c) holds.

However, it is a non-trivial task to identify (e, isA, c) from w because most categories express the semantic relatedness to the entity, or the topics or fields the entity belongs to. For example, in the article for *Jack Ma* in Chinese Wikipedia, categories include *1964 births*, *Alibaba Group*, *Business person in online retailing*, etc. Only *Business person in online retailing* is the suitable class for *Jack Ma*.

Besides, in the previous example, *1964 births* indicates that *Jack Ma* is a *person*. However, we do not know the isA relation between *Business person in online retailing* and *person*, while this relation is necessary to construct the high level structure of the taxonomy.

In our paper, we further divide isA relations into two types, namely *instanceOf* and *subclassOf*. The relations between entities and classes are called *instanceOf* relations. And *subclassOf* relations are used between classes. Our goal is to derive a large and accurate Chinese taxonomy T from W which contains *instanceOf* and *subclassOf* relations.

Overview of Our Approach Our approach consists of two key stages below.

Stage 1: Generate isA relations. As shown in Figure 2, Stage 1 generates isA relations from categories by a classification model, infers isA relations from rational categories and extends existing isA relations via rule mining.

Stage 2: Construct Chinese taxonomy. In Stage 2, we derive a tree to present the Chinese taxonomy. In this stage, we take each isA relation as a subtree and propose an algorithm to construct the taxonomy in a bottom-up manner via node merging, cycle removal and subtree merging.

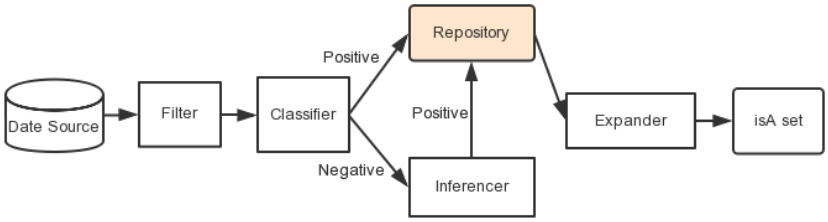


Fig. 2. The framework of generating isA relations

4 isA Relation Generation

In this section, we give a detailed description of our isA relation generation algorithm. The framework of our algorithm is shown in Figure 2. We first preprocess the Chinese Wikipedia pages by a filter to remove irrelevant pages, which will be discussed in Section 6. And then we train a classification model to detect isA relations. For the negative ones, we generate some efficient rules to infer isA relations. Finally, all the isA relations will be extended and we will obtain the whole isA relations set.

4.1 isA Relation Classification

Scoring function To distinguish isA relations from others, we carefully design a scoring function. Given an entity e, a category c, and two sets of features $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$, the function outputs a positive number for isA relation; negative otherwise, defined as follows:

$$y(e, c) = \boldsymbol{w_1} \cdot \boldsymbol{F_1}(e, c) + \boldsymbol{w_2} \cdot \boldsymbol{F_2}(c) + w_0 \tag{1}$$

where $\boldsymbol{F_1}(e, c)$ considers both information of e and c (called entity-dependent features) while $\boldsymbol{F_2}(c)$ (called entity-independent features) only takes the properties of c into account.

Feature sets We now briefly introduce our two feature sets. Features 1-4 are entity-dependent features while features 5-7 are entity-independent features.

Feature 1: Length of a category

The basic intuition is that if the length of a category is too long or short, it may be too general or too specific to describe the class of an entity.

Feature 2: POS tag

Usually a valid class is a noun or a noun phrase. We perform word segmentation and POS tagging on category names. We use the POS tag of the head word of the category as a feature.

Feature 3: Thematic category

As is described in [8], some categories, such as finance, politics, entertainment, etc. are thematics categories rather than conceptual classes. We have collected a set of themes in Chinese. We take whether a category or the head word of a category is a thematic word as a feature.

Feature 4: Language pattern

In English, a conceptual category is often in the form of *premodifier + head word + postmodifier* (see [1]). We have observed that in Chinese, the pattern is *premodifier + "de" + head word* where "de" is an auxiliary character "的" in Chinese. We then perform pattern matching on categories.

Feature 5: Common sequence of entity and category

In Chinese, entities and categories may have a common subsequence. For example, the category *political party* is a class for the entity *Labor Party*. We take the existence of the common sequence of a feature.

Feature 6: Head word matching

Similar to Feature 5, if the longest common sequence (LCS) of an entity and a category is the head word of the category, the category is likely to be a class.

Feature 7: Purity of a category

A category c is appeared on a set of articles with an entity set E_c. Intuitively, if most entities are person names, the category is likely to a class related to the concept *person*. We employ named entity recognition (NER) to tag entities. The purity of a category c is defined as:

$$purity(c) = max_{l \in L} \frac{|E_l \cap E_c|}{|E_c|} \tag{2}$$

where L is a collection of NE tags and E_l is the collections of entities that are labeled as l. Given a pre-defined threshold τ, we define whether $purity(c) > \tau$ as a feature.

Weights learning Features in $\boldsymbol{F_1}(e, c)$ and $\boldsymbol{F_2}(c)$ have different discriminative powers for the isA classification task, which are represented by weight vectors $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$. To learn the weights $\boldsymbol{w} = (\boldsymbol{w_1}, \boldsymbol{w_2}, w_0)$, we employ a sequential minimal optimization technique based on linear SVM. Given a training set of positive and negative samples, we optimize the objectives $\| \boldsymbol{w} \|^2 + C\Sigma_i \xi_i$ under the constraint:

$$y_i(\boldsymbol{w_1} \cdot \boldsymbol{F_1}(e, c) + \boldsymbol{w_2} \cdot \boldsymbol{F_2}(c) + w_0) \geq 1 - \xi_i \tag{3}$$

where $\xi_i \geq 0$ and C is the turning parameter.

4.2 isA Relation Inference

The classifier can be of help to extract *conceptual* classes from Wikipedia categories, however, a large number of *relational* categories can provide useful knowledge about an entity.

Rational categories can be leveraged to extract relations or properties. For example, we may extract a relation *(A, graduateFrom, PKU)* from the category *PKU graduates* in article *A*. We can also extract a property *(A, diedIn, 1964)* from the category *1964 deaths*. In the following, we elaborate how to infer isA relations from relational facts and properties.

Formally, a *relation* is a triple *(subj, predicate, obj)* that the subject and object are entities of certain classes. A *property* is also a triple but the object is a literal (such as string, numerical, date, etc.) rather than an entity.

For each relation R, given a relation predicate, a subject class SC, and an object class OC, then the referred rule is shown as follows:

$$(subj, R, obj) \Rightarrow (subj, isA, SC) \wedge (obj, isA, OC) \tag{4}$$

For each property P, given a property predicate and a subject class SC, the inferred rule is shown as follows:

$$(subj, P, obj) \Rightarrow (subj, isA, SC) \tag{5}$$

However, not all such rules can be used to generate isA relations. The reasons are twofold: firstly, the type inference rules may not be accurate. Secondly, the classes of subjects and objects can be very diverse. As a result, it is difficult to assign a simple class (i.e., SC or OC) for the rule. To ensure high accuracy, we only consider the top most frequent rules as isA relations. See Section 6 for detailed rules and their evaluation results.

4.3 isA Relation Expansion

Classes have different levels of abstraction. Some classes are high-level and cover a broad spectrum of entities, such as *person*, while others describe a domain specific region, such as *Chinese pop music composer*. In this part, our goal is to extract isA relations between classes of different levels. For example, given two isA relations *(A, isA, Chinese pop music composer)* and *(A, isA, person)*, isA relation *(Chinese pop music composer, isA, person)* should be extended to build the taxonomy.

In this section, we introduce our relation expansion technique in the framework of association rule mining. Given a class c, the *contribution* of the class is defined as the number of distinct entities that labeled by the class c:

$$contrib(c) = |\{e.subj | e.obj = c \wedge e \in E\}| \tag{6}$$

The *match* for two classes c_1 and c_2 is the number of matched entities:

$$match(c_1, c_2) = |\{e.subj | e.obj = c_1 \wedge e \in E\} \cap \{e.subj | e.obj = c_2 \wedge e \in E\}| \tag{7}$$

Then, the measure *confidence* between c_1 and c_2 corresponds to the ratio of the match and the contribution of c_1:

$$conf(c_1, c_2) = \frac{match(c_1, c_2)}{contrib(c_1)} \tag{8}$$

We can see $conf(c_1, c_2)$ determines whether class c_1 is a subclass of class c_2. When $conf(c_1, c_2)$ is no less than a pre-defined threshold, isA relation (c_1, isA, c_2) is formed. The higher the confidence score $conf(c_1, c_2)$ is, the more likely it is that class c_1 belongs to class c_2. Because when $conf(c_1, c_2)$ is closer to 1, more entities in c_1 also belong to the class c_2.

Other than isA relations generated from Wikipedia categories, this approach tries to extract the potential isA relations from a higher level of inter-articles.

5 Taxonomy Construction

The naive approach is to construct a graph via combining all relations detected in Stage 1. This approach is not effective because there exist many inconsistent or noisy relations after Stage 1. To avoid these drawbacks, we effectively construct the taxonomy in a bottom-up manner via incorporating three subtree operations. In summary, the process of taxonomy construction can be divided into three phases, namely, node merging, cycle removal and sub-tree merging.

Node merging Initially, isA relations are considered as sub-trees, $T(x)$ with root node x. We highlight two key operations to construct the sub-trees.

- **Horizontal merge** Given two isA relations (a, isA, x) and (b, isA, x) (i.e., two sub-trees), we join the sub-trees together to construct a new one when two child nodes a and b share a common parent node x.
- **Vertical merge** Given two isA relations (a, isA, b) and (b, isA, x) (i.e., two sub-trees $T(b)$ and $T(x)$), we extend the depth of $T(x)$ via adding child node a to node b in $T(x)$.

These two operations will be repeated until no sub-tree is generated or changed.

Cycle removal A taxonomy T can be viewed as a directed, acyclic graph. Unfortunately, cycles may be formed when we merge nodes vertically. For example, two isA relations *(a, isA, b)* and *(b, isA, a)* will be merged into a cycle. Thus we proposed an algorithm to remove cycles.

In the cycle removal algorithm, we first create a direct graph G from the set of isA relations S. Then, we utilize the DFS algorithm to check the connectivity of G. For each connected component *cc* in graph G, we check whether any edge exists in *cc* but not in the DFS tree produced by DFS. The edges are removed and no cycle exists as a result.

Sub-tree merging After node merging and cycle removal, sub-trees with a root of high level classes have been produced. However, these sub-trees are not inter-connected with each other. The manual effort in the whole taxonomy construction process is that we define several classes with high level of abstraction (e.g., *animal*, *event*, *organization*, etc.) and connect these sub-trees together. Finally, we assign a common root node to the sub-trees to build the complete taxonomy. And we label isA relations between classes and entities as *instanceOf*, others as *subClassOf*.

6 Experiments and Evaluation

6.1 Data Source

In this paper, our dataset is from Chinese Wikipedia dump[2] generated from 12 September 2014. In total, we extract 677,246 candidate entities for Chinese taxonomy construction. Every title of articles in Wikipedia dump is considered as a candidate entity. We clean up the data by the following steps:

[2] http://download.wikipedia.com/zhwiki/20140912/

1. Convert traditional Chinese characters to simplified Chinese;
2. Filter out pages without useful information;
3. Remove list pages, redirect pages, disambiguation pages, template pages and administrative pages, which do not contain candidate entities.

6.2 Taxonomy Analysis

Size and Accuracy In our taxonomy, there are in total 581,616 entities and 79,470 classes. Among these classes, 72,873 are extracted from Wikipedia categories, and the rest are classes of high level abstraction, generated from either inferring or mining approaches described in Section 4.

To evaluate the accuracy of extracted relations. We randomly select 2,000 relations from each set of relations (*instanceOf*, *subclassOf* and the whole *isA* relations) and manually label whether a relation is correct or not. We calculate the confidence interval of accuracy with significance level $\alpha = 0.05$. As shown in Table 1, the accuracy is over 95% for both *instanceOf* and *subclassOf* relations.

Relation type	Number	Accuracy	Samples
subClassOf	85,072	95.85% ± 2.16%	2000
instanceOf	1,233,291	97.80% ± 0.86%	2000
total	1,317,956	97.60% ± 0.71%	2000

Table 1. Size and accuracy of relations

Comparison It is not easy to compare our taxonomy with others, especially when they are for different languages. Because each taxonomy is a part of the knowledge graph and knowledge graphs are usually based on different data sources, structures and relations.

However, to show that our taxonomy contains unique knowledge that can not be captured by knowledge graphs in English, such as YAGO. We utilize the inter-language links in Wikipedia to map entities and categories from English to Chinese. If there is a hyperlink between an English and a Chinese article describing the same entity, the mapping can be formed. We perform the mapping process on Wikipedia categories in a similar fashion. We call the Chinese version of YAGO generated by mapping approach *YAGO-C* in this paper. Table 2

	Our taxonomy	YAGO-C	Coverage
Entity	581,616	274,730	47.15%
Class from Wikipedia categories	72,873	27,934	38.33%
High level class	6,597	-	11.70% (estimated)

Table 2. Comparison in size and coverage

shows the comparison results between YAGO-C and our taxonomy. Compare to 274,730 entities and 27,934 classes in YAGO-C, there are 581,616 entities and 79,470 classes in our taxonomy, which is much larger in size.

Except for classes and entities extracted from Widipedia, we also generate high level classes by isA inference and expansion. As YAGO combines Wikipedia and WordNet to construct a knowledge base, we analyze the coverage of high level classes in our taxonomy by language translation. We sample 1,000 high level classes randomly from our taxonomy and translate them into English. And we

find that only 117 high level classes are covered in WordNet. Generally, the coverage of entities, classes from Wikipedia categories and high levels classes is quite low in YAGO-C, with percentages of 47.15%, 38.33% and 11.70%, respectively.

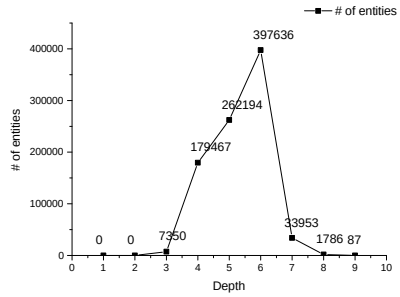


Fig. 3. Entity size distribution

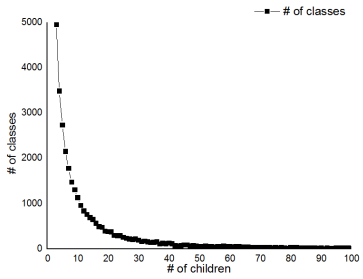


Fig. 4. Class size distribution

Topological Structure To understand the structure of our constructed taxonomy better, we evaluate the coverage of the taxonomy by observing the structure of the tree. We measure the depth of each leaf node and breadth of the taxonomy tree. We find that the depth ranges from 3 to 9, and the breadth ranges from 87 to 882,473.

We also evaluate the ability of the taxonomy on abstraction and expression. The depth of an entity in the tree indicates the ability of describing the entity. For example, given an entity *Lu Chen (magician)*, if the depth is 2 with the parent node *person*, we only know Lu Chen is a person. But if the depth is 5 with the path *living being, person, producer, Taiwanese television personality, Lu Chen (magician)*, we will know much more about Lu Chen.

In Figure 3, it shows that entities with the depth of 6 account for the majority of the entity set. It is normal that an entity may have different paths from it to the root, especially for people. We consider entity with multi-paths as different ones and that is why the size of entity set is larger than the entity space.

We also count the number of children for each class. As shown in Figure 4, the number of classes decreases rapidly as the number of children increases. When the number is more than about 50, the number of classes is very small. In fact, the classes with single child account for about 27.5%.

6.3 Performance Evaluation

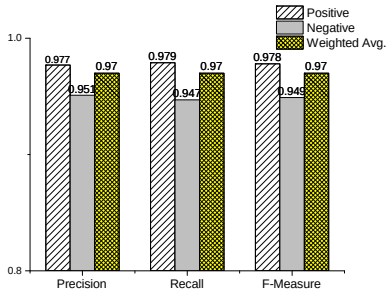
Classification To evaluate the performance of our classification model, We randomly select 4,600 (entity, category) pairs from the dataset and label them as positive (isA) or negative (notIsA). We randomly split 85% of the labeled data to train the classifier and test on the remainder.

We use precision, recall and F-measure to evaluate our linear SVM classifier. As as shown in Figure 5, the overall F-Measure is 97.0%, which proves the efficacy of our classification approach.

To further compare the contributions of features, we remove one feature and train the classifier with the rest at a time. As a result, we train another seven classifiers. To evaluation the contribution of each feature f, we define *Decrement in F-measure* as follows:

$$DF(f) = \frac{FM(F) - FM(F \setminus f)}{FM(F)} \tag{9}$$

where $FM(F)$ denotes the F-measure of the original classifier with feature set F and $FM(F \setminus f)$ denotes the F-measure of the classifier without feature f. From Figure 6, it is clearly observed that feature 3 is more discriminative with a higher DF score.



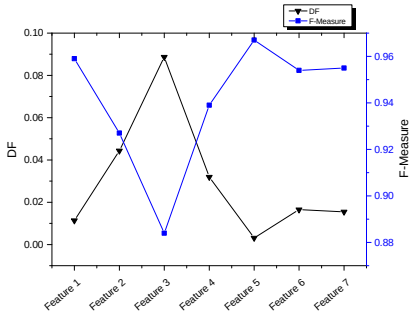


Fig. 5. Performance of linear SVM classifier **Fig. 6.** Evaluation of features and classifiers

isA Relation Inference And Expansion As discussed in Section 4.2, we use pattern matching to leverage the semantics of relational categories. In the imple-

Subject Class	Object Class	Regular Expression	Num of Extraction	Accuracy
city	province	(.*省)市镇	32,091	100%
political leader	position	(.*(委员|参议员|参政员|议员))	13,881	100%
person	-	(.*? \ d{1, 4}年)逝世	10,148	99%
person	-	(.*? \ d{1, 4}年)出生	4,801	99%
monarch	-	(.*?)(君主|国王)	3,649	100%

Table 3. Examples of inference rules

mentation, we use regular expressions to match Wikipedia categories to generate isA relations. In total, we design 70 regular expressions to match categories. Table 3 shows some of the regular expressions we use to perform inference. Note that when the object class does not exist, the rule can be leveraged to extract a property rather than a relation. We perform accuracy tests on each rule as well.

As discussed in section 4.3, we expand isA relations by calculating the *confidence*. In fact, 4,707 isA relations are generated from all the existing isA relations. We set confidence to be 0.05 to filter out noisy, incorrect isA relations. We extract 3,380 isA relations with the accuracy 88% and coverage 71.8%.

7 Conclusion

In this paper, we propose a hybrid method to construct a Chinese taxonomy from user generated content. We generate a large number of accurate isA relations via

directly classifying relations from categories, inferring relations from relational facts and properties, and extending existing relations by association rule mining. Furthermore, we construct the hierarchical structure to represent the taxonomy in a bottom-up manner. The experimental results illustrate that our Chinese taxonomy has a large scale and achieves a high accuracy.

However, Wikipedia is a rich knowledge repository that contains more than entities and categories. More isA relations can be mined from plain texts in Chinese. We will take effort to extract more isA relations and enlarge our taxonomy.

Acknowledgment

This work is partially supported by National Science Foundation of China (Grant No.61232002, 61402177 and 61332006).

References

1. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: WWW. (2007)
2. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. **194** 28–61
3. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A knowledge base from multilingual wikipedias. In: CIDR. (2015)
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: ISWC, ASWC. (2007) 722–735
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal **6**(2) (2015) 167–195
6. Wu, W., Li, H., Wang, H., Zhu, K.: Probase: A probabilistic taxonomy for text understanding. In: SIGMOD. (2012)
7. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI. (2010)
8. Ponzetto, S.P., Strube, M.: Deriving a large-scale taxonomy from wikipedia. In: AAAI. (2007) 1440–1445
9. Wang, C., Gao, M., He, X., Zhang, R.: Challenges in chinese knowledge graph construction. In: ICDEW. (2015) 59–61
10. Fellbaum, C., ed.: Wordnet, an Electronic Lexical Database. MIT Press (1998)
11. Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T.: Learning semantic hierarchies via word embeddings. In: ACL. (2014) 1199–1209
12. Wang, Z., Li, J., Li, S., Li, M., Tang, J., Zhang, K., Zhang, K.: Cross-lingual knowledge validation based taxonomy derivation from heterogeneous online wikis. In: AAAI. (2014) 180–186
13. Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., Tang, J.: Xlore: A large-scale english-chinese bilingual knowledge graph. In: ISWC. (2013) 121–124
14. Carlson, A., Betteridge, J., Wang, R.C., Jr., E.R.H., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: WSDM. (2010)
15. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: COLING. (1992) 539–545