Taxonomy Induction from Chinese Encyclopedias by Combinatorial Optimization

Weiming Lu(✉), Renjie Lou, Hao Dai, Zhenyu Zhang, Shansong Yang, and Baogang Wei

College of Computer Science and Technology, Zhejiang University, Hangzhou, China
luwm@zju.edu.cn

Abstract. Taxonomy is an important component in knowledge bases, and it is an urgent, meaningful but challenging task for Chinese taxonomy construction. In this paper, we propose a taxonomy induction approach from a Chinese encyclopedia by using combinatorial optimizations. At first, **subclass-of** relations are derived by validating the relation between two categories. Then, integer programming optimizations are applied to find out **instance-of** relations from encyclopedia articles by considering the constrains among categories. The experimental results show that our approach can construct a practicable taxonomy from Chinese encyclopedias.

1 Introduction

Nowadays, knowledge bases such as DBPedia [1], YAGO [2], Freebase [3] and BabelNet [4] are becoming increasingly crucial in many fields of Artificial Intelligence. Taxonomy is the backbone of a knowledge base, which is a conceptual network including **subclass-of** relations between concepts and **instance-of** relations between concepts and instances.

DBpedia taxonomy was manually created from the most commonly used infobox templates in Wikipedia, which forms a shallow subsumption hierarchy. YAGO has a deep subsumption hierarchy, which takes the hyponymy relations of WordNet [5]. However, manual construction of taxonomies is extremely laborious, time consuming, and costly, which makes the scope and the scale of the taxonomies limited.

Recently, several researchers try to construct the taxonomies automatically. Pattern based methods are widely used to mine hypernym-hyponym relations from text corpora [6–9], but these methods often suffer from low recall and precision because of the noisy text and the low quality of patterns. Other works [10–12] mined the taxonomies from Wikipedia, which mainly learned the hypernym-hyponym relations from Wikipedia's category system. But the coverage is still limited, since they only used categories in the Wikipedia and ignored the numerous articles.

However, the knowledge bases in Chinese is quite limited currently. For example, DBpedia has become the central hub and reference point in the Web of data

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 299–312, 2015.
DOI: 10.1007/978-3-319-25207-0_25

in English, but the coverage of Chinese knowledge in DBpedia is restricted since Wikipedia has only 824,693 articles. Therefore, the construction of Chinese taxonomy is an urgent, meaningful but challenging task. Several works began to construct the Chinese taxonomies, where [13–16] were mainly based on Chinese encyclopedias (Hudong and Baidu Baike), while [17, 18] identified the hypernym-hyponym relation from free text.

In this paper, we also try to address the Chinese taxonomy induction problem based on encyclopedias. In addition to mine the taxonomy from the category system of encyclopedias, we also mine the hypernym-hyponym relations from articles to form a more wide-coverage and fine-grained taxonomy. The core idea of our approach is that we leverage the taxonomy learned from category system to improve the coverage of the hypernyms for articles while guarantee the high precision.

The rest of the paper is organized as follows. Section 2 discusses the related works about taxonomy induction. Section 3 describes our approach in detail. Then, we evaluate our approach on real-world datasets in Section 4, and conclude the paper in Section 5.

2 Related Work

Previous works about taxonomy induction mainly focus on pattern-based methods and encyclopedia-based methods.

Hearst [6] is a pioneer of using manually constructed lexical patterns like "X such as Y" to harvest hyponym-hypernym relations. In order to address the limitation of small number of hand-crafted patterns, LASER [7] iteratively discovered new patterns through a sequential pattern mining framework by taking in a set of seed Hearst patterns. Probase [19] can take advantage of the existing knowledge they already learned to discover more syntactic patterns, and then to acquire more knowledge. Other patterns like word-class lattices [8, 9] and dependency path [20] can also be learned to extract hypernym-hyponym relations.

Wikipedia is very useful in automatic taxonomy construction. WikiTaxonomy [10] was generated by automatically assigning is-a and not-is-a labels to the relations between categories in Wikipedia. However, it only used a set of lightweight heuristics for the assignment. Furthermore, WordNet was used in [11] to restructure the WikiTaxonomy. On the other hand, KOG [12] modeled subsumption detection task as a binary classification problem.

Machine-learning and purely distributional approaches also contributed to the task of hypernym discovery. [21] proposed an approach to distinguish hypernyms and co-hyponyms by a linear support vector machine with distributional features. [22] proposed a supervised approach with the selective distributional inclusion hypothesis for hypernymy detection. Probabilistic models were used to incorporate multiple evidences from hyponym and coordination cues for semantic taxonomy induction [23–25]. TAXIFY [26] was proposed to learn a taxonomy from a domain-specific corpus. It first uses Hearst patterns to collect initial set of *is-a* relations, and then improves the recall and precision by a clustering-based inference procedure and incorrect edges detection. [27] considered the hypernym extraction problem as a sequential classification task, which combined linguistic, definitional and graph-based information.

Recently, several works started to construct the Chinese taxonomies. Zhishi.me [13] focused on the infobox information extraction and Chinese LOD construction, but it did not construct the taxonomy between concepts. [14] built an ontology based on the category system and infobox templates in Hudong and Baidu Baike. However, they only used some simple heuristic methods to refine the category system, which would bring many wrong sub-concept relations to the ontology since sub-category relations between categories are not strict sub-concept relations. For instance, in the category system of Hudong Baike, 海洋*(Ocean)* contains sub-categories 海底隧道*(channel tunnel)* and 填海工程*(reclamation)*, 计算机病毒*(Computer Virus)* has super-category 计算机安全*(Computer Security)*, and 昆虫*(insert)* contains sub-category 昆虫学*(insectology)*. They all would bring wrong sub-concept relations to the ontology. XLore [15] utilized a classification-based method to correctly semantify the wikis' category systems. [16] extracted candidate hypernyms from multiple sources, and applied a statistical ranking model to select correct hypernyms. [17,18] identified the hypernym-hyponym relation by using the word-embedding-based semantic projections between words and their hypernyms. Zheshi.schema [28] extracted semantic relations between categories from a large number of popular Chinese social Web sites. However, they all don't utilize the relatedness between hypernyms when inducing the hypernym-hyponym relations.

Multilingual information could also be used to reinforce the performance of taxonomy induction. For example, a cross-lingual knowledge validation based model was proposed in [29] to iteratively reinforce the performance of taxonomy derivation. MENTA [30] induced multilingual taxonomies from all editions of Wikipedia and WordNet.

The closely-related previous works are WiBi [31] and [16]. WiBi presented an approach to create an integrated taxonomy of Wikipedia pages and categories. In their work, the category taxonomy and page taxonomy could be enriched mutually, but they only used statistics information for taxonomy induction. [16] extracted candidate hypernyms from multiple sources, and applied a statistical ranking model to select correct hypernyms. However, they directly used the ranking models such as Support Vector Machine and Logistic Regression on features of candidate hypernyms without considering the relations among candidate hypernyms. Our approach mutually considers the linguistic features, structural features and the taxonomy have been learned into a unified learning model for taxonomy induction. In other words, our approach also leveraged the learned relations among candidate hypernyms to enhance the taxonomy induction.

3 The Proposed Approach

In this section, we formally defined our taxonomy induction problem at first, and then addressed this problem by a two step approach including **subclass-of** relation induction and **instance-of** relation induction.

3.1 Problem Formulation

In encyclopedia, each article can be considered as an instance, and then can be represented as a 5-tuple $a = \{L(a), A(a), C(a), P(a), T(a)\}$, where $L(a)$ is the title of article a, $A(a)$ is the set of linked articles of a, $C(a)$ is the catalog of a, $P(a)$ is the set of properties of the article's infobox, and $T(a)$ is the set of category tags of a. Figure 1 shows a sample article from Chinese encyclopedias.

Fig. 1. Sample Article Page from Chinese Encyclopedias

Then, our taxonomy induction task for Chinese encyclopedias is illustrated in Figure 2.

From the figure, we observed that many category tags of an article are hypernyms of the article, and these tags may have **subclass-of** relation among them. Taking *magpie* as an example in Figure 2, *Songbird*, *Ornamental bird*, *Animal*,

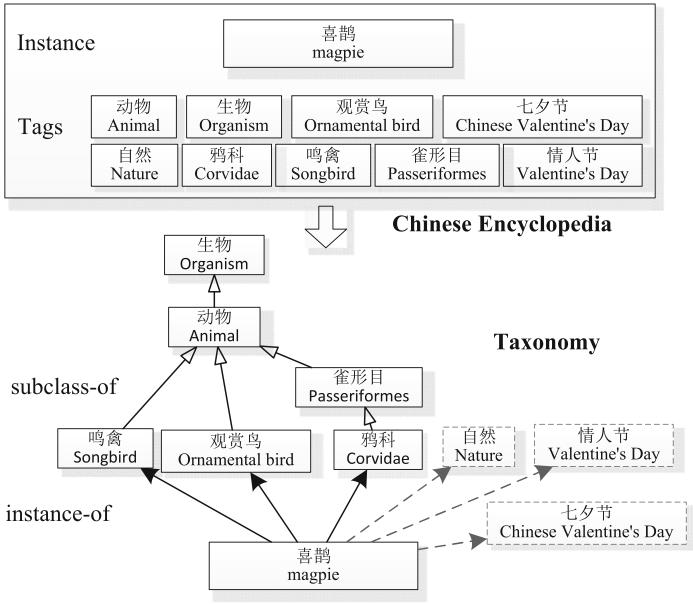


Fig. 2. The taxonomy induction task from Chinese Encyclopedia

etc. are hypernyms of *magpie*, and *Songbird* has **subclass-of** relation with *Animal*. However, there are also noisy tags, such as *Valentine's Day*, which only has related topic with *magpie*, so we should filter them out.

These category tags in the encyclopedia form a category system to provide navigational links to all articles in a hierarchical way, However, they do not form a real subsumption hierarchy, which includes *isa* and *notis* relations between categories. If the hypernym-hyponym relations between category tags can be mined, it could be very helpful for **instance-of** relation induction. For instance, if *Songbird* and *Ornamental bird* are selected as hypernyms of *magpie*, then *Animal* and *Organism* should also be the hypernyms of *magpie*, while *Valentine's Day* is not likely to be the hypernym for *magpie*, since *Animal* and *Organism* are the super classes of *Songbird* and *Ornamental bird*, and *Valentine's Day* and *Songbird* can not be the hypernyms of *magpie* simultaneously.

Therefore, our taxonomy induction problem are defined as follows. Given an encyclopedia with articles $\mathcal{A}$ and categories $\mathcal{C}$, we would like to induce the **subclass-of** relations between categories in $\mathcal{C}$, and then leverage the relatedness of categories to induce the **instance-of** relation between category tags and articles.

3.2 Subclass-of Relation Induction

The **subclass-of** relation induction problem is treated as a binary classification problem: given two category tags $c_i \in \mathcal{C}$ and $c_j \in \mathcal{C}$, we try to train a classifier to predict whether c_i and c_j have the **subclass-of** relation. we use linguistic features and structural features as in [29].

Linguistic Features.

1. ***Head matching feature***. This feature represents whether two labels have a common head or not, which is obtained by suffix matching. Usually, it implies the existence of **instance-of** or **subclass-of** relations, e.g. 中国演员*(Chinese Actor)* is a subclass of 演员*(Actor)*, and 牡丹鹦鹉*(Peony parrot)* is an instance of 鹦鹉*(parrot)*.
2. ***Modifier matching feature***. This feature represents whether one label is the other's modifier or not, which is obtained by prefix matching. It indicates that **instance-of** or **subclass-of** relation probably doesn't exist, e.g. 食品安全*(Food safety)* and 食品*(Food)* do not have **instance-of** and **subclass-of** relations.

Structural Features. Distributional Inclusion Hypothesis [22] states that more specific terms appear in a subset of the distributional contexts in which more general terms appear. Here, we use the set of articles which connect to the category as the distributional context of a category, and use Normalized Google Distance [32] to calculate the diversity between two categories. The structural features for category $c_i \in \mathcal{C}$ and $c_j \in \mathcal{C}$ are listed in Table 1. Then, a simple function $f(x) = \frac{1}{1+x}$ is used to normalize the structural features to [0,1].

After we obtain a six-dimensional features for two categories, we train a SVM classifier to predict the validity of **subclass-of** relation between categories.

Table 1. The structural features for categories

Feature	Calculation	Comment
Article Diversity	$d_a(c_i, c_j) = \frac{\max(\lvert A(c_i)\rvert, \lvert A(c_j)\rvert - \log(\lvert A(c_i) \cap A(c_j)\rvert))}{\log(\lvert A\rvert) - \log(\min(\lvert A(c_i)\rvert, \lvert A(c_j)\rvert))}$	$A(c)$ is the set of articles in the category c, and $\lvert A\rvert$ is the total number of articles
Property Diversity	$d_p(c_i, c_j) = \frac{\max(\lvert P(c_i)\rvert, \lvert P(c_j)\rvert - \log(\lvert P(c_i) \cap P(c_j)\rvert))}{\log(\lvert P\rvert) - \log(\min(\lvert P(c_i)\rvert, \lvert P(c_j)\rvert))}$	$P(c)$ is the set of properties of $A(c)$, and $\lvert P\rvert$ is the total number of properties.
Category Diversity	$d_t(c_i, c_j) = \frac{\max(\lvert T(c_i)\rvert, \lvert T(c_j)\rvert - \log(\lvert T(c_i) \cap T(c_j)\rvert))}{\log(\lvert T\rvert) - \log(\min(\lvert T(c_i)\rvert, \lvert T(c_j)\rvert))}$	$T(c)$ is the set of category tags of $A(c)$, and $\lvert T\rvert$ is the total number of category tags.
Catalog Diversity	$d_c(c_i, c_j) = \frac{\max(\lvert C(c_i)\rvert, \lvert C(c_j)\rvert - \log(\lvert C(c_i) \cap C(c_j)\rvert))}{\log(\lvert C\rvert) - \log(\min(\lvert C(c_i)\rvert, \lvert C(c_j)\rvert))}$	$C(c)$ is the set of words in catalogs of $A(c)$, and $\lvert C\rvert$ is the total number of words in catalogs.

3.3 Instance-of Relation Induction

In the **instance-of** relation induction task, we collect hypernym candidates for each instance at first, and then select the correct hypernyms with the help of the relatedness among hypernym candidates.

In addition to the category tags, we also collect the head words of the title of the article a as the hypernym candidates, denoted as $\mathcal{H}_a$. For instance, 鹦鹉*(parrot)* is the head word of 牡丹鹦鹉*(Peony parrot)*, so it is also collected as a hypernym candidate. Then, we train a classifier with pairs $\langle a, h \in \mathcal{H}_a \rangle$, with features similar to those for **subclass-of** relation classification by replacing c_i,c_j with a,h for the structural features. That is to say, besides the linguistic features, the structural features $d_a(a,h)$,$d_p(a,h)$, $d_t(a,h)$ and $d_c(a,h)$ are calculated by Normalized Google Distance, and then a SVM classifier was also trained on the normalized features.

The probability output could be regarded as the coherence between the article a and the hypernym candidate h, which is denoted by $coh(a,h)$. Although $coh(a,h)$ could be used to predict the correct hypernyms, however, it ignores the relatedness among hypernym candidates. Here, we denote the results of the **subclass-of** relation induction by two sets: $Subset = \{\langle c_i, c_j \rangle | c_i \textit{ subclass-of } c_j\}$ and $Mutex = \{\langle c_i, c_j \rangle | c_i \textit{ not-subclass-of } c_j \textit{ and } c_j \textit{ not-subclass-of } c_i\}$, and the probability output of the classifier in section 3.2 is regarded as the coherence between two categories, denoted as $coh(c_i, c_j)$.

Then, we exploit the Integer Programming method to introduce the knowledge from category tags into the **instance-of** relation induction. Let $y_i \in \{0,1\}$ be an indicator variable specifying whether $h_i \in \mathcal{H}_a$ is a hypernym of a, then the optimization problem can be formulated by a Mixed Integer Programming (MIP) optimization and Integer Quadratic Programming (IQP) optimization, where MIP utilizes the $Subset$ and $Mutex$ constrains, while IQP relaxes the $Mutex$ constrain, since $Mutex$ is too restrict. This is because we simply consider $Mutex = \mathcal{C} \times \mathcal{C} - Subset$ in our implementation.

1. **MIP** by utilizing the *Subset* and *Mutex* constraints.

$$
\begin{aligned}
\textbf{maximize} \quad & \sum_{h_j \in \mathcal{H}_a} y_j \cdot coh(a, h_j) - \sum_{\langle h_j, h_k \rangle \in Subset} \zeta_{jk} - \sum_{\langle h_j, h_k \rangle \in Mubset} \xi_{jk} \\
\textbf{s.t.} \quad & \forall h_j, h_k \in \mathcal{H}_a, y_j \in \{0,1\}, \zeta_{jk}, \xi_{jk} \geq 0 \\
& y_k \geq y_j - \zeta_{jk}, \langle h_j, h_k \rangle \in Subset \\
& y_j + y_k \leq 1 + \xi_{jk}, \langle h_j, h_k \rangle \in Mutex
\end{aligned}
$$

Here, ζ_{jk} is the penalty for violation of $Subset$ constraint. $y_k \geq y_j - \zeta_{jk}, \langle h_j, h_k \rangle \in Subset$ means if h_j is a subclass of h_k, and when $y_j = 1$, then y_k should be 1 for a smaller ζ_{jk}. That is to say, if h_j is selected as hypernym, then h_k who is the super class of h_j should also be selected as a hypernym.

ξ_{jk} is the penalty for violation of $Mutex$ constraint. $y_j + y_k \leq 1 + \xi_{jk}, \langle h_j, h_k \rangle \in Mutex$ means if h_j and h_k are mutually exclusive to be super classes of an instance, then $y_j = y_k = 1$ would make a penalty $\xi_{jk} \geq 1$. That is to say, h_j and

h_k can not be selected as hypernyms simultaneously. If they are both selected, then the penalty $\xi_{jk} \geq 1$.

2. **IQP** by relaxing the *Mutex* constraint with the penalty for small $coh(h_k, h_j)$.

$$\begin{aligned}
\textbf{maximize} \quad & \sum_{h_j \in \mathcal{H}_a} [| \mathcal{H}_a | \cdot y_j \cdot coh(a, h_j) + \\
& \lambda \cdot \sum_{h_k \in \mathcal{H}_a} y_k \cdot (1 - y_j) \cdot (1 - coh(h_k, h_j))] \\
\textbf{s.t.} \quad & \forall h_j \in \mathcal{H}_a, y_j \in \{0, 1\}, \\
& \forall \langle h_j, h_k \rangle \in Subset, y_k - y_j \geq 0
\end{aligned}$$

where $\sum_{h_k \in \mathcal{H}_a} y_k \cdot (1 - y_j) \cdot (1 - coh(h_k, h_j))$ indicates that when $coh(h_k, h_j)$ is small, then if h_k is selected as a hypernym, then h_j is encouraged not to be selected. This makes $\sum_{h_k \in \mathcal{H}_a} y_k \cdot (1 - y_j) \cdot (1 - coh(h_k, h_j)) = 1 - coh(h_k, h_j) > 0$. λ is the tradeoff parameter for precision and recall, and larger λ could get higher precision but lower recall. The experiments in Section 4.3 shows this phenomenon.

We solve the above optimization problem by using *IBM CPLEX* optimizer[1].

4 Experimental Evaluation

4.1 Experiment Settings

Data Set. *Hudong Baike*[2] and *Baidu Baike*[3] are the two largest collaboratively Chinese encyclopedias. In order to evaluate our approach, we crawled about 3.3M articles with 14963 categories from *Hudong Baike*, and 5.3M articles with 4639 categories from *Baidu Baike* in Jan 2015. We observed that *Hudong Baike* generally has more tags than *Baidu Baike*, but has more noise. For instance, 喜鹊*(magpie)* has tags of 鸦科*(Corvidae)*, 鸟类*(Bird)*, 雀形目*(Passeriformes)*, 民俗*(Folklore)* and 动物*(Animal* in *Baidu Baike*, while *Hudong Baike* has other more tags such as 鸣禽*(Songbird)*, 观赏鸟*(Ornamental Bird)*, 情人节*(Valentine's Day)* and 七夕节*(Chinese Valentine's Day)*. Obviously, *Hudong Baike* has more useful tags such as 鸣禽*(Songbird)* and 观赏鸟*(Ornamental Bird)*, but also introduces some noisy tags such as 情人节*(Valentine's Day)* and 七夕节*(Chinese Valentine's Day)* as the hypernyms of 喜鹊*(magpie)*. In statistics, *Hudong Baike* has about 6 tags with precision of 52.55% in average, while *Baidu Baike* has about 4 tags with precision of 56.21% in average.

From these data, we respectively formed two datasets $Hudong_D$ and $Baidu_D$ with human annotation as the ground truth. For each dataset, we randomly sampled 1400 pairs of *(category,sub-category)* and 1400 pairs of *(article,category tag)* for training classifiers, and then used the classifiers to validate all pairs in

[1] http://www-01.ibm.com/software/commerce/optimization/cplexoptimizer/
[2] http://www.baike.com
[3] http://www.baike.baidu.com

the *Hudong* and *Baidu Baike*. For each instance, we further used combinatorial optimizations to determine the **instance-of** relations. For testing, we randomly sampled 1000 articles with more than 3 tags as the testing data for **instance-of** relation induction, and sampled 500 categories with its direct sub-categories as the testing data for **subclass-of** relation induction.

Additionally,we also asked annotators to rate the category tags for each article according to the *specificity*. Larger *specificity* indicates the hypernym is more specific for the instance. For example, Table 2 shows the tags of *magpie* with the corresponding rates, which can be denoted by $\{h_1 : r_1, h_2 : r_2, ..., h_n : r_n\}$, where r_i is the rate for candidate h_i according to the specificity. In the table, we could find *Nature*, *Valentine's Day* and *Chinese Valentine's Day* are not the hypernyms of *magpie*, and *Songbird* and *Ornamental Bird* are more specific hypernyms than *Animal* and *Organism*.

Table 2. A user annotated example for **instance-of** relation

instance	category tags and its rates
magpie	Organism:1, Animal:2, Passeriformes:3, Songbird:4, Ornamental bird:4, Corvidae:4, Nature:0, Valentine's Day:0, Chinese Valentine's Day:0

Evaluation Metrics. We used precision, recall, F1 and accuracy to evaluate the **subclass-of** and **instance-of** relation induction. The pairs which are predicted to be the correct hypernym-hyponym relations by models are denoted by M_t, and the others are denoted by M_f. By comparing with the human labeled data sets H_t (the pairs with correct hypernym-hyponym relations) and H_f (the pairs with incorrect hypernym-hyponym relations), *precision*, *recall*, *F1* and *accuracy* could be defined as: $Prec = |M_t \cap H_t|/|M_t|$, $Rec = |M_t \cap H_t|/|H_t|$, $F_1 = 2 \cdot Prec \cdot Rec/(Prec + Rec)$, and $Acc = (|M_t \cap H_t| + |M_f \cap H_f|)/(|H_t| + |H_f|)$. Taking into the consideration of hypernym *specificity*, we also defined the weighted version of precision, recall, F1 and accuracy: $Prec^w = \sum_{x \in M_t \cap H_t}(x.r + 1)/\sum_{x \in M_t}(x.r + 1)$, $Rec^w = \sum_{x \in M_t \cap H_t}(x.r + 1)/\sum_{x \in H_t}(x.r + 1)$, $F_1^w = 2 \cdot Prec^w \cdot Rec^w/(Prec^w + Rec^w)$ and $Acc^w = (\sum_{x \in M_t \cap H_t}(x.r + 1) + \sum_{x \in M_f \cap H_f}(x.r + 1))/\sum_{x \in H_t \cup H_f}(x.r + 1)$.

In addition, we used Normalized Discounted Cumulative Gain(NDCG) [33] to evaluate our approach when we treated the hypernym prediction results as a ranked list.

Concretely speaking, given an instance a, its annotated list $\{h_1 : r_1, h_2 : r_2, ..., h_n : r_n | r_i \geq r_j, i < j\}$ can be used to calculate the IDCG (Ideal Discounted cumulative gain) $= r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2(i)}$. The comparative methods re-rank the list according to y_i and $coh(a, h_i)$. That is to say, the tags with $y_i = 1$ and larger $coh(a, h_i)$ are ranked in the front, like $\{h_{i_1} : r_{i_1}, h_{i_2} : r_{i_2}, ..., h_{i_n} : r_{i_n} | y_{i_1} = ... = y_{i_t} = 1, y_{i_{t+1}} = ...y_{i_n} = 0, coh(a, h_{i_m}) \geq coh(a, h_{i_n}), m < n\}$. Then, its DCG is calculated by $r_{i_1} + \sum_{j=2}^{n} \frac{r_{i_j}}{\log_2(j)}$. So $NDCG = \frac{DCG}{IDCG}$. For instance, if the tags of *magpie* in Table 2 is ranked as a list of *Songbird, Ornamental bird, Nature, Passeriformes, Corvidae, Organism, Animal, Valentine's Day, Chinese Valentine's Day*, then $IDCG = 4 + 4/\log_2(2) + 4/\log_2(3) + 3/\log_2(4) + 2/\log_2(5) +$

$1/\log_2(6) = 13.27$, $DCG = 4 + 4/\log_2(2) + 0/\log_2(3) + 3/\log_2(4) + 4/\log_2(5) + 1/\log_2(6) + 2/\log_2(7) = 12.32$, so $NDCG = 12.32/13.27 = 0.928$.

4.2 Performance Evaluation

In our experiments, we found that the precisions of **subclass-of** relation classification in *Hudong Baike* and *Baidu Baike* are 97.56% and 98.11% respectively, which are high enough to be the knowledge for **instance-of** relation induction.

Table 3 shows the average results of the **instance-of** relation induction, where we also evaluated the performance of hypermyn candidates only from article titles, and used *specificity* as weights for performance evaluation.

From the table, we observe that although *MIP* can achieve the highest precision, but the recall is very low. This is because *MIP* has very strict constraints, especially for the *Mutex* constrains. *IQP* achieves the best performance with F1 measurement. In addition, article's title includes high-quality hypernym candidates because of the high Prec^h, Rec^h, Acc^h and F1^h. Table 4 shows an specific example for the **instance-of** relations for *Scarlet Macaw*.

From the table, we find (1) *SVM* can not verify the relations between entity and high-level classes well, because of the great diversity of the high-level classes. But low-level classes have higher rate, so the NDCG of *SVM* is greater than *MIP* in Table 3. (2) *IQP* can select more low-level classes than *MIP*, since the derived **subclass-of** relations can not cover every valid relations, so the *Mutex* could be noisy, if we just take $Mutex = \{pairs \notin Subset\}$. In this example, since the **subclass-of** relation between 鹦鹉*(Parrot)* and 攀禽*(Zygodactyl)* is not obtained, this makes them can not be simultaneously selected as hypernyms in *MIP*.

Table 3. The performance of **instance-of** relation induction, where $*^h$ means the measurements are only considering the hypernym candidates from the head of title, $*^w$ means it is the weighted version of the measurements

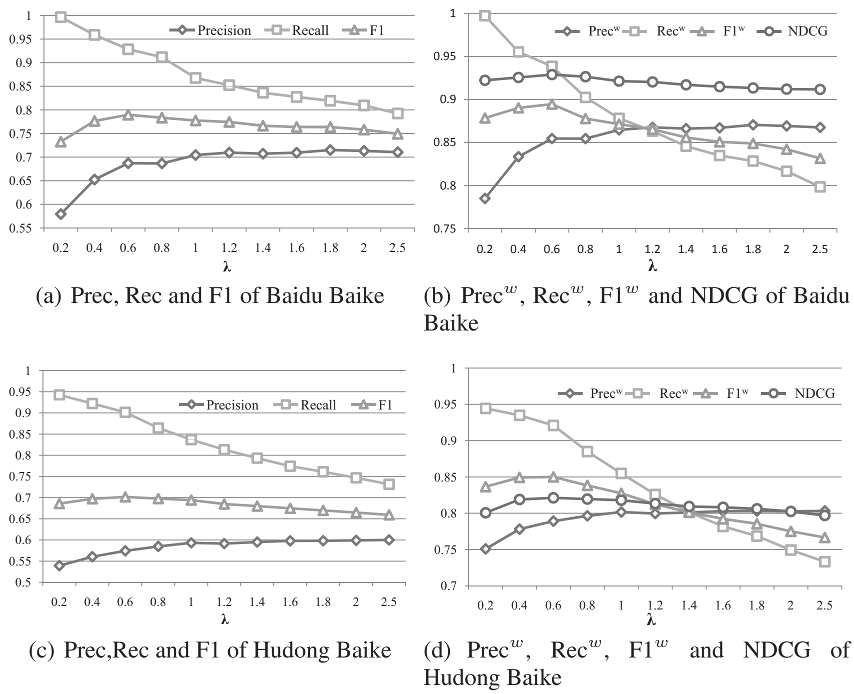
	Baidu Baike			*Hudong Baike*		
methods	SVM	MIP	IQP	SVM	MIP	IQP
Prec	72.85%	80.41%	68.70%	54.62%	59.62%	57.44%
Rec	68.16%	38.56%	92.84%	54.93%	24.10%	90.13%
Acc	67.79%	60.14%	72.20%	52.18%	51.38%	59.58%
F1	70.42%	52.12%	**78.97%**	54.78%	34.32%	**70.16%**
Prec^h	99.01%	100.00%	98.34%	83.08%	93.10%	80.22%
Rec^h	81.70%	54.47%	96.34%	50.00%	25.00%	67.60%
Acc^h	81.78%	56.59%	94.96%	53.57%	40.71%	62.14%
F1^h	89.53%	70.53%	**97.33%**	62.43%	39.42%	**73.37%**
Prec^w	87.78%	91.24%	85.46%	77.43%	79.76%	78.90%
Rec^w	68.87%	36.90%	93.86%	57.69%	23.71%	92.10%
Acc^w	68.52%	48.47%	92.90%	55.56%	38.13%	75.56%
F1^w	77.19%	52.55%	**89.46%**	66.12%	36.55%	**85.00%**
NDCG	92.12%	92.10%	**92.89%**	81.98%	80.19%	**82.12%**

Table 4. An example of *instance-of* relations, where the underline indicates the selected hypernyms for the instance

	instance-of relations for 五彩金刚鹦鹉 (Scarlet Macaw)
SVM	鹦形目(Psittaciformes), 鸟类(Birds), 动物(Animal), 攀禽(Zygodactyl), 自然(Nature), 鹦鹉(Parrot), 生物(Organism)
MIP	鹦形目(Psittaciformes), 鸟类(Birds), 动物(Animal), 攀禽(Zygodactyl), 自然(Nature), 鹦鹉(Parrot), 生物(Organism)
IQP	鹦形目(Psittaciformes), 鸟类(Birds), 动物(Animal), 攀禽(Zygodactyl), 自然(Nature), 鹦鹉(Parrot), 生物(Organism)

4.3 Parameter Setting

Parameter λ in *IQP* influences the performance of **instance-of** relation induction. According to the objective function of *IQP*, larger λ could result in larger precision but lower recall. We solved the *IQP* with different λ, and the results are shown in Figure 3.



(a) Prec, Rec and F1 of Baidu Baike

(b) Prec^w, Rec^w, F1^w and NDCG of Baidu Baike

(c) Prec,Rec and F1 of Hudong Baike

(d) Prec^w, Rec^w, F1^w and NDCG of Hudong Baike

Fig. 3. Performance with different parameter λ for Baidu Baike and Hudong Baike

From the figure, we can see that recall decreases and precision increases when λ increases. Therefore, we selected $\lambda = 0.6$ for both *Hudong Baike* and *Baidu Baike* as it can reach the best $F1$.

5 Conclusion and Future work

In this paper, we propose a taxonomy induction approach from a Chinese encyclopedia by combinatorial optimizations. In future, we plan to take the semantic similarity among articles into account, since similar articles are most likely to take the similar categories as their hypernyms.

Acknowledgments. This work is supported by the Fundamental Research Funds for the Central Universities(2014QNA5008), Chinese Knowledge Center of Engineering Science and Technology(CKCEST), and the National Natural Science Foundation of China (No.61103099).

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia-a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web **7**(3), 154–165 (2009)
2. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706. ACM (2007)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM (2008)
4. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
5. Fellbaum, C.: Wordnet: An electronic database (1998)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, vol. 2, pp. 539–545. Association for Computational Linguistics (1992)
7. Li, T., Chubak, P., Lakshmanan, L.V.S., Pottinger, R.: Efficient extraction of ontologies from domain specific text corpora. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1537–1541. ACM (2012)
8. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1318–1327. Association for Computational Linguistics (2010)
9. Navigli, R., Velardi, P., Faralli, S.: A graph-based algorithm for inducing lexical taxonomies from scratch. In: IJCAI, pp. 1872–1877 (2011)
10. Ponzetto, S.P., Strube, M.: Wikitaxonomy: a large scale knowledge resource. In: ECAI (2008)

11. Ponzetto, S.P., Navigli, R.: Large-scale taxonomy mapping for restructuring and integrating wikipedia. In: IJCAI (2009)
12. Wu, F., Weld, D.S.: Automatically refining the wikipedia infobox ontology. In: WWW. ACM (2008)
13. 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, Part II. LNCS, vol. 7032, pp. 205–220. Springer, Heidelberg (2011)
14. Wang, Z., Wang, Z., Li, J., Pan, J.Z.: Knowledge extraction from chinese wiki encyclopedias. Journal of Zhejiang University SCIENCE C **13**(4), 268–280 (2012)
15. 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: International Semantic Web Conference (Posters & Demos), vol. 1035, pp. 121–124 (2013)
16. Fu, R., Qin, B., Liu, T.: Exploiting multiple sources for open-domain hypernym discovery. In: EMNLP, pp. 1224–1234 (2013)
17. Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T.: Learning semantic hierarchies via word embeddings. In: Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1 (2014)
18. Ruiji, F., Guo, J., Qin, B., Che, W., Wang, H., Liu, T.: Learning semantic hierarchies: A continuous vector space approach. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(3), 461–471 (2015)
19. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: SIGMOD. ACM (2012)
20. Snow, R., Jurafsky, D., Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. Advances in Neural Information Processing Systems 17 (2004)
21. Weeds, J., Clarke, D., Reffin, J., Weir, D., Keller, B.: Learning to distinguish hypernyms and co-hyponyms. In: Proceedings of COLING, pp. 2249–2259 (2014)
22. Roller, S., Erk, K., Boleda, G.: Inclusive yet selective: supervised distributional hypernymy detection. In: COLING (2014)
23. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 801–808. Association for Computational Linguistics (2006)
24. Zhang, F., Shi, S., Liu, J., Sun, S., Lin, C.-Y.: Nonlinear evidence fusion and propagation for hyponymy relation mining. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1159–1168. Association for Computational Linguistics (2011)
25. Bansal, M., Burkett, D., de Melo, G., Klein, D.: Structured learning for taxonomy induction with belief propagation, pp. 1041–1051 (2014)
26. Alfarone, D., Davis, J.: Unsupervised learning of an is-a taxonomy from a limited domain-specific corpus. CW Reports (2014)
27. Espinosa-Anke, L., Ronzano, F., Saggion, H.: Hypernym extraction: combining machine-learning and dependency grammar. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9041, pp. 372–383. Springer, Heidelberg (2015)
28. Wang, H., Wu, T., Qi, G., Ruan, T.: On publishing chinese linked open schema. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 293–308. Springer, Heidelberg (2014)

29. 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: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
30. de Melo, G., Weikum, G.: Menta: Inducing multilingual taxonomies from wikipedia. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1099–1108. ACM (2010)
31. Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two is bigger (and better) than one: the wikipedia bitaxonomy project. In: ACL, pp. 945–955 (2014)
32. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering **19**(3), 370–383 (2007)
33. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)