Semantic Knowledge Base Construction from Domain-Specified Metadata

Jiyi LI†, Toshiyuki SHIMIZU†, and Masatoshi YOSHIKAWA†

† Department of Social Informatics, Graduate School of Informatics,
Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
E-mail: †jyli@db.soc.i.kyoto-u.ac.jp, ††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

Abstract Existing semantic knowledge bases such as WordNet and Yago contain the information of relations between entities. Although they are not domain-specific, they contain limited information and have limitations on the usage scopes and domains. For example, they do not hold the information about the domain-specific commonsense relations between concepts like "horse" and "farm" or "ocean" and "atmosphere" which intuitively have close relations on semantics in the domains of image description or earth observation data description. Such semantic knowledge are useful in the corresponding domains for various applications such as annotation, recommendation, search, suggestion and so on. Because in data collections, metadata which is used to describe data is widespread in various domains, in this paper we propose an approach to collect this kind of relations and construct knowledge bases for specific domains by mining knowledge of global structure and internal association in the metadata of data collections.

Key words Knowledge Base, Metadata, Domain-Specific, Relation Extraction

1. Introduction

Knowledge base (KB) has been an active research area in the past decades. It harvests and manages facts among entities as knowledge with representation that is readable by machines to provide knowledge-centric services like natural language processing assistant, question answering, and semantic search. There are many existing knowledge bases and each of them focuses on different types and scopes of information, for example, WordNet [1], [2] is a lexical database; YAGO [3], [4], [5] is a semantic knowledge base derived from Wikipedia(注1), WordNet and GeoNames(注2); DBpedia [6] extracts structured content from Wikipedia; Freebase [7] is an online collaborative knowledge base mainly by its community people; KnowItAll [14], [15], [16] project aims to automatically extract fact collections from Web resources.

These existing semantic knowledge bases such as WordNet and YAGO contain the information of relations between entities. Although they are not domain-specific, they contain limited information and have limitations on the usage scopes and domains. For example, they do not hold the information about the relations between the concepts "horse" and "farm" which intuitively have close relations on semantics in the domains of image description. Figure 2 shows more examples of domains and the relative concepts in the domains, including concepts "ocean" and "atmosphere" in the domain of earth observation data and concepts "knowledge harvesting" and "information extraction" in the domain of scholarly document.

Such kinds of relations between concepts have abstract or vague semantics. They seems to obviously and intuitively exist but actually are difficult to be represented exactly in words. For the example of "horse" and "farm", the semantic relations can be explained like "horse" lives on "farm" or "horse" and "farm" usually co-exist in same scenes and images. The relations are known by human beings but not easy to be acquired and understood by machine. It can be regarded as a kind of commonsense knowledge and we concentrate on it in this paper. Here "commonsense" is a domain-specific notion because it may be not commonsense for all people and may only be commonsense for the people with knowledge background of a specific domain; the relations exist and are used in specific domains. We can harvest these knowledge from the data collections in corresponding domains, for example, harvesting the relations like that between "horse" and "farm" from image datasets in image description domain.

Because in these data collections, metadata, which is data about data, is widespread applied in various domains and

(注1) : www.wikipedia.org
(注2) : www.geonames.org

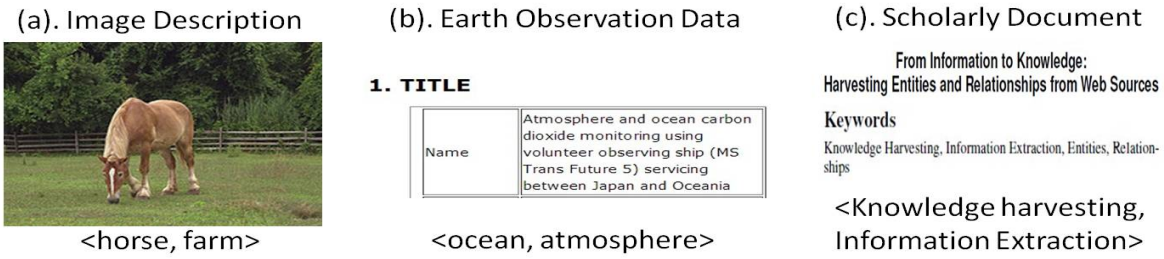


Figure 1 Example of Domains and Concepts with Relations

contain useful information for describing data content. We therefore extract the relation information from descriptive metadata. The descriptive metadata is diverse in format and content; some metadata like keywords are only list of words without syntactic information; some metadata like social tags are user generated and contain various noises; furthermore, the commonsense relations we extract are not easy to be represented with special patterns in natural language sentences. Therefore, we convert the textual description in metadata into bag of concepts. We propose a solution to construct specific knowledge bases for specific domains by mining the knowledge of global structure and internal association in the metadata of data collections.

The main contributions of this paper are twofolds. On one hand, for the communities of specific domains like image retrieval, earth observation and academic library, the commonsense concept relations we collect into the knowledge bases can be widely used for data management and information retrieval in these domains, such as, keyword annotation, keyword recommendation, keyword search, query suggestion and so on.

We do not focus on entities like that used in YAGO and relational facts between them. In contrast, we focus on concepts that used in WordNet and commonsense relations between them. Note that although the data we focus on is domain-specific, the approach we propose is not domain-specific. Without loss of generally, we do not use the information in the detailed content of data in a domain, such as image content, earth observation data, body of paper and so on. We only use the information of metadata in the format of bag of concepts. Actually our relation extraction approach is more general than existing approaches that leverage linguistics syntax information, because it is based on bag of concepts and is independent from the sentence structure and language, which also means our solution can be regarded as a multilingual solution.

Our approach is supervised and without expensive human involvement on manually relation labeling for both knowledge base construction and training set construction. The problem is that our solution need to make efforts on the precision and recall of the knowledge bases because of the lack of usable information.

On the other hand, for the community of knowledge base and semantic web, our work extends the research on the following aspects. The extension on various aspects is an important issue for the area of knowledge base.

- *Concepts and Relations:* Most existing Knowledge bases focus on harvesting relational facts between entities, for example, unary relations like "X IsA FootballPlayer" and binary relations like "X hasWonPrize Y". Our work focuses on harvesting commonsense relations between concepts in a specific domain, this kind of relations is not covered in existing knowledge bases.
- *Resources and Domains:* Some existing knowledge bases is constructed manually which their resources are human knowledge and references. Many recent work propose approaches to automatically collect the knowledge from text and web sources. There is no specific domain declared though there are limitations on the usage scopes. Our work harvests the knowledge from the metadata of data collections for a specific domain.
- *Relation Extraction Method:* Existing knowledge bases focus on relational facts and their relation extraction approaches leverage linguistics syntax in the text by using pattern-based gathering or constraint-based reasoning approaches. Our relation extraction approach do not refer linguistics syntax information and leverage bag of concepts by analyzing data characteristics, use heuristic rules and learn classifiers.
- *Applications:* Existing knowledge bases enable knowledge centric applications like natural language processing assistant, question answering, and semantic search for entities and relations, while our knowledge bases enhance data management and retrieval applications like keyword an-

notation, keyword recommendation, keyword search, query suggestion, and so on in specific domains.

The remainder of this paper is organized as follows. In Section 2 we review the related work. In Section 3, we propose our solution on semantic knowledge base construction for special domains with their metadata. In Section 4 we report and discuss the experimental results. We give a conclusion in Section 5.

2. Related Work

In this section, we brief review the related work. There are many existing knowledge bases, Ref. [8], [9] list some of them. Specially, we review two typical and widely used knowledge bases, WordNet and Yago, which are closely related to the motivation of our work.

WordNet [1] is an important lexicon in information retrieval and natural language processing. The core concept in WordNet is the synset. A synset contains one or more word senses and each word sense belongs to exactly one synset. In turn, each word sense has exactly one word that represents it lexically, and one word can be related to one or more word senses. WordNet follows taxonomy and has tree-like structure. To adopt WordNet in the semantic web research community, researchers convert it into RDF or OWL representation, e.g., Ref. [2].

YAGO [3], [4], [5] is a core of semantic knowledge in which the information is extracted from Wikipedia (e.g., categories, redirects, infoboxes), WordNet (e.g., synsets, hyponymy) and GeoNames, using rule-based and heuristic methods. It contains additional knowledge beyond WordNet on individuals like persons, organizations, products, and so on, with their relations. It has high coverage and contains more than 1 million entities and 5 millon facts. The empirical evaluation of fact correctness by the author shows that it has a high accuracy. It contains both hierarchical relations and non-taxonomic relations. It is compatible with RDF.

WordNet and YAGO do not explicitly contain the relations like "inherited hypernym" relation between $\langle horse, animal \rangle$ and "sister term" relation $\langle horse, zebra \rangle$. These implicit relations in Yago and Wordnet can be generated by additional inferences. YAGO contains additional information extracted from Wikipedia, but still doesn't contain the information we need. Some relations like that between $\langle horse, farm \rangle$ are neighter explicitly nor implicitly contained in WordNet and YAGO. Table 1 lists some examples that YAGO and WordNet contain or not.

For the approaches of relation extraction and knowledge base construction, WordNet [1], [2] is created manually. Freebase [7] is constructed by collaboratively by its community people. Many work make efforts on automatically harvesting the relational facts from web sources by using pattern-based gathering or constraint-based reasoning. Yago [3], [4], [5] and DBpedia [6] are rule-based and extract information from infobox of wikipedia. Some work [10], [11], [12], [13], [14], [15], [16], [17] leverage the linguistics syntax information in the text. Our relation extraction approach does not refer linguistics syntax information and leverages bag of concepts, We analyze data characteristics about global structure and internal association in the metadata of data collections, use heuristic rules and learn classifiers.

Table 1 Relations in Existing Semantic Knowledge Base

subject, object	YAGO	WordNet	Relation
horse, equid	subClassof	hypernym	direct hypernym
horse, animal	"indirect"	"indirect"	inherited hypernym
horse, zebra	"indirect"	"indirect"	sister term
horse, black	"no"	"no"	descriptive
horse, farm	"no"	"no"	coexist nouns

3. Semantic Knowledge Base Construction from Metadata

We propose our approach of semantic knowledge base construction in this section. Instead of relational facts between entities, we focus on domain-specific commonsense relations between concepts in this work. We first describe the preprocessing method on the data in the collection and then propose the relation extraction approach. After that, we introduce the post-processing method for arranging and managing the extracted relations and concepts in knowledge bases. Another issue about enrichment, extension and aggregation of the constructed knowledge bases is not yet covered in this paper and is one of the future work.

3.1 Definition and Formulation

Definition: Commonsense Relation R_c between two concepts w_i and w_j in a specific domain is that these two concepts subjectively has intuitive association for the people in this domain.

We use R_c instead of Commonsense Relation for short. This R_c in our work has the following properties.

- It is a binary relation: $w_i R_c w_j$.
- It is an abstract and hypernym relation: $R_c = \{r_c\}$.

R_c is proposed to handle the abstract or vague semantics that intuitively exist but may be difficult to be represented exactly in words. It can be interpreted to various detail relations r_c in different cases. Table 1 shows some examples of detailed cases and interpretations like "direct hypernym", "inherited hypernym", "sister term", "descriptive" and "coexist nouns". Specially, in our work, we focus on the relations that are not contained in existing knowledge bases such as that between "horse" and "farm" in Table 1.

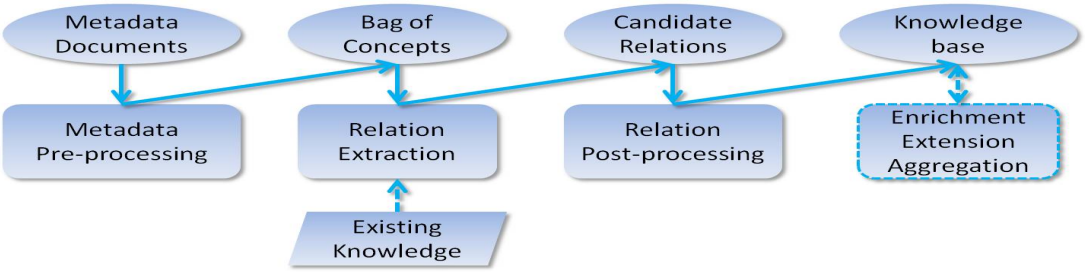


Figure 2 Overview of Our Solution

- It is a swappable relation: $w_i R_c w_j \Rightarrow w_j R_c w_i$. R_c has a definition like w_i "hasCommonsenseRelation" with w_j which is swappable. Although its sub relations like "inherited hypernym" are not non-swappable, e.g., "equid" is the "direct hypernym" of "horse", according to the definition of R_c, this example can be interpreted in the mean of that "horse" has R_c (or "direct hypernym") relation with "equid".
- It is a non-transitive relation: $w_i R_c w_j \wedge w_j R_c w_k \not\Rightarrow w_i R_c w_k$. The relation between any two concepts of w_i, w_j and w_k needs to be evaluated separately. In addition, because it is a non-transitive relation, an inference based on existing relations to generate new relations is not available.

We formulate the notions in our work as follows. In a specific domain, given a data collection $\mathcal{D}$ which is composed of data records, we denote each record as a document d_i. The data records can be in various forms, such as an image, a earth observation dataset, an academic paper, and so on. We denote the metadata of d_i as m_i. We convert the textual description in the metadata into bag of concepts. We denote each concept as w_x, the set of concepts in metadata d_i as $m_i = \{w_x | w_x \in m_i\}$, and the set of concepts used in this data collection as W. Our work is to extract the relations (predicate) p_{ij} between two concepts w_i and w_j: $t_{ij} = \langle w_x, p_{xy}, w_y, c_{xy} \rangle$. c_{xy} represents the confidence score of the relation, which is also used in YAGO [3]. We denote this quadruple t_{xy} as a term of "tuple" to distinguish the term of "fact" used in existing knowledge bases.

3.2 Data Pre-processing

In this sub-section, we describe the pre-processing method on the metadata of data collection. In this paper, we use social image public dataset NUS-WIDE in the domain of image description as the example for our solution. The detailed introduction of this dataset and domain is in the experimental section. The data in other domains like earth observation data and scholarly document are not yet collected and pre-processed. Without loss of generality, we use user generated social tags in English in the metadata of the images to extract the commonsense relations. The social tags of an image is a group of words. We can convert it to bag of concepts easily. For other kinds of textual information in metadata such as title or abstract which are in sentences, we can use natural language processing to split them into concepts. There are various natural language toolkits available for this task in each language. How to convert the plain text in a specific language to bags of concepts is not covered in this paper. Although the loss of information on syntax, sequence and so on, bag of words model has been proved effective and widely used in information retrieval for representing documents. We therefore use bag of concepts model in our work.

The purpose of data pre-processing is to prepare a list of candidate concepts, and then our approach can extract relations among them. After we convert the textual description in the metadata into bag of concepts, we filter these concepts with ontology-like vocabularies related to the specific domain. For example, we use WordNet for image description domain. The concepts with two low frequency are regarded as noises and have been deleted by the authors of NUS-WIDE. The concepts that are not contained in the noun synsets in WordNet are not included. The concepts with same homomorphy are merged. WordNet stores concepts like "horse", "farm" and so on, and therefore is proper to provide a concept list for this example. Although the concepts in this example based on WordNet are only some words like "horse", the concepts can also be some phrases like "information extraction" based on the vocabularies and the domains. For the domain of earth observation data, this vocabulary can be GCMD Science Keywords [18]; for the domain of scholarly document, this vocabulary can be constructed from the category tree of Wikipedia.

3.3 Relation Extraction

To reduce or avoid expensive human involvement on manually relation labeling of knowledge base construction, we propose a solution to automatically extract relations from metadata in data collections. Our fundamental idea is that

Table 2 Concept Co-occurrence v^1_{xy} of Selected Concepts

	horse	building	black	farm	road	animal	zebra
horse	2399	17	68	100	20	187	17
building		6995	189	43	190	30	4
black			7407	59	141	285	77
farm				2236	74	175	6
road					3004	62	7
animal						9553	133
zebra							663

Table 3 Symmetric Relative Co-occurrence v^2_{xy} *1000 of Selected Concepts

	horse	building	black	farm	road	animal	zebra
horse	1000	1.81	6.98	22.05	3.72	15.89	5.58
building		1000	13.30	4.68	19.37	1.82	0.52
black			1000	6.16	13.73	17.09	9.63
farm				1000	14.32	15.07	2.07
road					1000	4.96	1.91
animal						1000	13.19
zebra							1000

Table 4 Google Distance exp(-v^4_{xy}) * 10 of Selected Concepts

	horse	building	black	farm	road	animal	zebra
horse	10	2.795	3.703	5.153	3.460	4.347	4.388
building		10	3.662	3.456	4.485	2.063	2.886
black			10	3.648	4.144	3.764	4.676
farm				10	4.617	4.340	3.733
road					10	3.262	3.646
animal						10	4.909
zebra							10

when two concepts are used in same metadata, it is possible that they have commonsense relations. However, it always does not true. Although many concepts are used in the metadata of data records, only some of them have relations. We need to consider the global structure and internal association of a concept pair in the data collection to evaluate the possibility of their relation. Therefore we come to a series of relation measures that may reflect concept relations.

- *Concept Co-occurrence:* We denote the the frequency of word w_x in the data collections as $f(w_x)$, a word multiple appears in the metadata of a data record is computed as once. The number of times that two concepts are used for same data record is defined as $f(w_x, w_y)$. The relation measure is

$$v^1_{xy} = f(w_x, w_y).$$

- *Relative Co-occurrence:* Concept co-occurrence is the most intuitive measure and shows the global structure of the concept pairs in the data collection. Some other related co-occurrence measures that consider internal association of concept pairs can be used. For example, a symmetric measure is defined as

$$v^2_{xy} = \frac{f(w_x, w_y)}{g(w_x, w_y)},$$

which is also known as Jaccard similarity coefficient. $g(w_x, w_y)$ is the number of times that w_x or w_y is used for a data record. An asymmetric measure is defined as

$$v^3_{xy} = \frac{f(w_x, w_y)}{f(w_x)} \frac{f(w_x, w_y)}{f(w_y)},$$

which keeps the feature vector of c^3_{xy} and c^3_{yx} same. Another form of this asymmetric measure can be defined as

$$v^{3'}_{xy} = \frac{1}{2}(\frac{f(w_x, w_y)}{f(w_x)} + \frac{f(w_x, w_y)}{f(w_y)}).$$

- *Google Distance:* There is a semantic similarity measure proposed based on the hits number of Google search engine for a given set of keywords. It can also be adapted to used in our scenario.

$$v^4_{xy} = \frac{max\{logf(w_x), logf(w_y)\} - logf(w_x, w_y)}{logM - min\{logf(w_x), logf(w_y)\}}.$$

M is the total number of the data records. A Monotonically increasing form of this measure is $v^{4'}_{xy} = exp(-v^4_{xy})$, in which the concept pairs with higher similarity have higher value.

- *Associated Word Distribution:* We also propose an original measure that refers the distribution of the other concepts that co-occur with a given concept pair. This measure is computation consuming (n^3 matrix, n is the number of uniqure concepts) and therefore is not used currently. It can be used for classifying sub relations in future work. For example, we can use light-computation measures to filter out the candidates with relations and then use this heavy-computation measure to classify them into sub relations.

We make an investigation with some samples to check whether such relations measures is possible to evaluate the relations. We first check the measure of concept co-occurrence. Table 2 shows the co-occurrence of selected concepts to the images in the NUS-WIDE [21] dataset. We compute the number of times that two concepts are labeled to same images. "horse" and "buildings" with no relation has low co-occurrence value; "horse" and "farm" with a commonsense relation has high co-occurrence value. It shows that maybe concept co-occurrence can be used for evaluating the relations. A naive relation extraction method is to set an empirical threshold, e.g., $\theta^1 = 30$, and select the concept pairs with concept co-occurrence higher than this threshold. Note that although there are some exsiting text analysis tools can visualize the co-occurrence between words with graph, word co-occurrence is not equivalent to commonsense relation.

However, "horse" and "zebra" which has a commonsense relation has same value on this measure with "horse" and "buildings". It shows that only concept co-occurrence is not enough for evaluating the relations. We then investigate other measures. Table 3 and Table 4 show that in

the measures of symmetric relative co-occurrence and Google distance the pair of "horse" and "buildings" and the pair of "horse" and "zebra" are distinguishable on the commonsense relation. Based on above investigations, we draw an assumptions that these measures can be used for evaluating the relations. We therefore construct feature vectors $v_{xy} = \{h(v_{xy}^k)\}$ for concept pairs with a selection of these relation measures, and then leverage machine learning technologies to classify the concept pairs into a class of commonsense relation or not. h is some transformations on the basic measures, e.g., the monotonically increasing form of Google distance.

To reduce or avoid expensive human involvement on manually relation labeling for training set construction. We propose a solution to automatically generate the training set for learning the model. The idea is to collect the learning samples from available existing knowledge bases. In our example on image description, we use the same ontology-like resources with data pre-processing, i.e., WordNet. We assume that various sub-commonsense relations have similar characteristics on the above mentioned measures, specially the relations labeled in WordNet and the relations that are not contained in WordNet have similar distributions on these measures. It means that we can use the relations in WordNet that are manually labeled and have high precision and confidence as training samples to learning the model and threshold for the commonsense relations. We therefore scan all candidate concept pairs in the concept space W. We check whether these concept pairs have a selected list of relation types explicitly or implicitly labeled in WordNet and add the positive ones into the training set.

There are two problems in this automatic solution. One is the noises in the training set; the other is the lack of negative samples. For the first problem, although WordNet is manually constructed. It only has high precision on its own relation types. Not all relation types labeled by WordNet can be regarded as commonsense relation; for a give relation type labeled by WordNet, not all concept pairs labeled in WordNet have commonsense relations. For example, not all concepts pairs with "inherited hypernym" relation in WordNet have commonsense relations; many concepts share same hypernym and therefore have "sister term" relations in WordNet do not have commonsense relations. Actually, some concept pairs with relations labeled by WordNet have 0 frequency in the data collection and therefore have 0 values on all measures. To solve this problem, we set empirical rules on each selected relation type to reduce the noises. In this example, we select the relation types of "direct hypernym" and "sister term" in WordNet to generate training samples. For "sister term" relation, the maximum depth of the concepts in the synset tree of WordNet should be higher than a threshold α_1, e.g., $\alpha_1 = 10$. The frequency of concept pairs should be higher than a threshold α_2, e.g., $\alpha_2 = 5$. α_k is defined as thresholds for generating training set from the resources.

For the second problem, the concept pairs that do not have relations labeled in WordNet cannot be regarded as negative samples. To manually construct negative samples with the same order of magnitude of positive samples is expensive. How to automatically select negative samples is a problem. We address it as future work. Currently our solution is to learn the model without negative samples. Some machine learning technologies are proposed for such case. We use one class SVM [20] for learning the model to extract the relations.

3.4 Post-processing

After we extract the rough tuples of relations, we serialize the concept pair matrix, and store the list of concept pairs with commonsense relation into a file to construct the knowledge base. It also can be converted to a graph representation or stored in RDF format. Because our commonsense relation is non-transitive relation, we do not need to carry out an inference based on existing detected tuples to generate new tuples is not available. In the future work, we will also add methods to improve the precision in the post-processing.

The predicate p_{xy} in a tuple of two concept is defined as "hasCommonsenseRelation". To assign the confidence score c_{xy} for each tuple of concept pairs. If the machine learning method can generate probability value on the positive class for each concept pair, we can use it as the confidence score. If it cannot return probabilistic value and can only return class labels, e.g., the implement of one class SVM used this paper, we use the correctness ratio in the testing set for all concept pairs instead.

There are some candidate solutions for improving the precision and recall of the generated knowledge base, for example, we can check and add the tuples manually; we also can add more documents or data collections. In this paper, we introduce an approach that how to construct the knowledge base based on one data collection. This knowledge base generated from only one data collection has limited information to used for the whole specific domain. We can enrich, extend this knowledge base with more documents, and we can also integrate multiple knowledge bases into one knowledge base. This important issue is not yet covered in this paper and is one of the future work.

4. Experiment

In this experimental section, we first introduce the candidate domains we select and the candidate available datasets in these domains. We then describe the details of the experimental settings on the example domain and dataset. We report and discuss the experimental results at last.

4.1 Dataset Resources

We consider several candidate domains to collect the the domain-specified metadata for experiments. They are social image, earth observation data and scholarly document.

a) Social Image

Social images are from the huge image collections on social networking services websites.The metadata of social images contains exchangeable image file format (Exif) information, user generated textual information, social information such as owner, favorites and groups, and so on. We concentrate on the social tags which are labeled by users for a brief description to the images to collect concept relations. The constructed knowledge base can be used for image annotation and image retrieval. Some public datasets in which the images are gathered from social image hosting websites like Flickr are available, for example, NUS-WIDE [21] with 269,648 images and MIR Flickr [22] with 25,000 images.

b) Earth Observation Data

The domain of earth observation science gather huge amount of data on earth physics, chemistry and biology. The metadata of earth observation data are to describe the dataset on the information of content, investigator, publication, and so on. There are many data centers around the world that publish various data, and the candidate data centers or services we consider to collect data for our work are the Data Integration and Analysis System (DIAS) project (注3) which holds hundreds of datasets and Pangaea(注4) which provides rich API interfaces to access their data.

c) Scholarly Document

Scholarly document collections with millions of publications in various scientific areas have entered big data era. The metadata of scholarly document, e.g., academic papers, contain authors information, paper content information, publication information, and so on. For the candidate datasets, CiteSeer is a public digital library mainly in the areas of computer science and provide OAI interface (注5) to harvest the records in it. Mendeley is reference management application and provides the public Mendeley's DataTEL dataset (注6) for a challenge.

4.2 Experimental Settings

As we have introduced in Section 3, we use the social tag metadata of the data collection NUS-WIDE in the image description domain as the example for our work. NUS-WIDE has deleted the concepts with too low frequency. After the pre-processing stage with homomorphy merge and noun verification on the candidate concept set, there are 4031 unified concepts in the concept space W at last.

In the stage of relation extraction, the feature vector of concept pairs w_x and w_y is $v_{xy} = \{v^1_{xy}, v^2_{xy}, v^3_{xy}, v^{4'}_{xy}\}$. For training set construction, we select the relations of "direct hypernym" and "sister term" in WordNet. Because the exist of noisy samples, for the filtering thresholds of generating training set, we set some strict thresholds to ensure the precision of the training set. The maximum depth of concepts for "sister term" should be higher than 10 ($\alpha_1 = 10$); the frequency of concept pairs should be higher than 5 ($\alpha_2 = 5$). Note that we do not set too large frequency threshold because the positive testing concept pairs with low frequency may be rejected, and we don't need to consider the recall of the training examples. There are 2181 positive samples in the training set for the 4031 unified concepts, which means 0.54 relations for each concept in average. To learn the model, the implement of one class SVM [20] we use on our automatically generated positive samples is LibSVM [19]. The "nu" parameter of one class SVM is set to 0.2 because it reaches good results on the cross validation on training set and does not overfit the training set.

The size of a knowledge base is huge and manually evaluating the whole knowledge base is impractical. We refer the empirical evaluation of relation precision which is used by Ref. [3] for evaluation. We sample a subset of the concept space W and corresponding concept pairs to construct a testing set for evaluation by human judgment. We use the 24 concepts from the image annotation ground truth in MIR Flickr dataset [22]. Because the target applications of our knowledge base are data management and retrieval applications like keyword annotation and recommendation, these 24 concept are regarded typical and common concepts in image annotation. In addition, we also use additional 6 concepts from Table 2. There are 30 concepts in total, and we will evaluate the relations of 435 concept pairs.

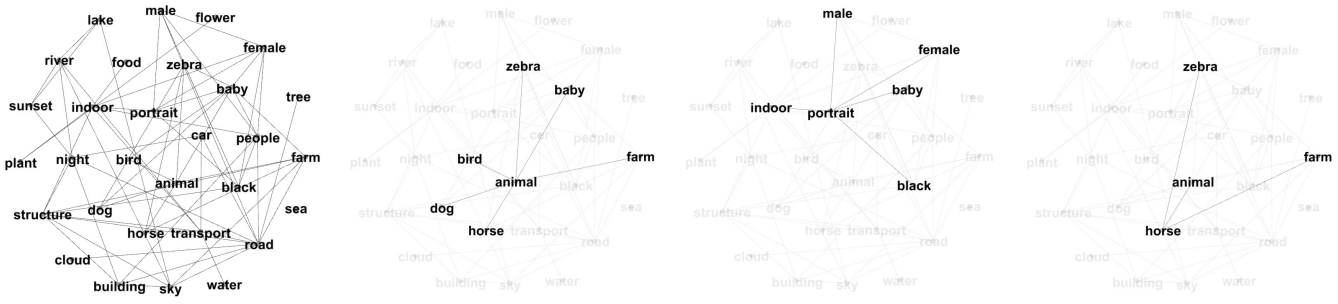
The baseline method we compare with is setting empirical threshold θ^1 to the measure of concept co-occurrence. According to Table 2, we set $\theta^1 = 30$.

4.3 Experimental Results

We evaluate the precision of the constructed knowledge base. We define this metric as Knowledge Base Precision (kbPre). Generally, for many existing knowledge bases like YAGO, this metric of evaluating the precision is enough. They focus more on precision and less on recall because there is no boundary on the knowledge, and list the huge number of relation records in the knowledge base. On the balance between precision and recall, precision is more fundamental for knowledge bases.

The kbPre of our generated knowledge base is $51/66 \approx 77.3\%$. It is higher than that of baseline $100/269 \approx 37.2\%$.

(注3) : DIAS: http://www.editoria.u-tokyo.ac.jp/projects/dias/.
(注4) : Pangaea: http://www.pangaea.de/.
(注5) : CiteSeerX Data: http://csxstatic.ist.psu.edu/about/data.
(注6) : Mendeley's DataTEL: http://dev.mendeley.com/datachallenge/.



(a) (b) (c) (d)

Figure 3 Sample Results

Figure 3 shows sample results with a graph representation. Concepts are nodes, and relations are edges. Figure 3.a shows the relations in the concepts of the testing set. Figure 3.b, 3.c and 3.d show all extracted relations of three concepts. It shows that the knowledge base generated by our method has a good precision. Furthermore, on the whole concept space with 4,031 concepts, our approach extracts 845,560 tuples from all 8,122,465 concept pairs and uses these tuples to construct the knowledge base.

5. Conclusion

In this paper, we discuss how to construct knowledge base from metadata in a specific domain. We focus on collecting the domain specific commonsense relations. The constructed knowledge bases can be used for annotation, recommendation, search, suggestion and so on in the specific domains. Our work is a significant extension on the covered type of concepts and relations, domains and data resources, and applications in the research community of knowledge base.

References

[1] C. Fellbaum, editor. "WordNet: An Electronic Lexical Database", MIT Press, 1998.

[2] Wordnet 3.0 in RDF, available at `http://semanticweb.cs.vu.nl/lod/wn30/`.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum. "YAGO - A Core of Semantic Knowledge", WWW'07, pp. 697-706, 2007.

[4] J. Hoffarta, F. M. Suchanekb, K. Berberich, E.L. Kelham, G.D. Melo, G. Weikum. "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages", WWW'11, pp. 229-232, 2011.

[5] J. Hoffarta, F. M. Suchanekb, K. Berbericha, and G. Weikum. "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia", Artificial Intelligence, Vol. 194, pp. 28-61, Jan. 2013.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. "DBpedia: a Nucleus for a Web of Open Data", ISWC'07/ASWC'07, pp. 722-735, 2007.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge", SIGMOD'08, pp. 1247-1250, 2008.

[8] F. Suchanek, and G. Weikum. "Knowledge Harvesting from Text and Web Sources", ICDE'13, pp. 1250-1253, 2013.

[9] F. Suchanek, and G. Weikum. "Knowledge Harvesting in the Big-Data Era", SIGMOD'13, pp. 933-938, 2013.

[10] N. Nakashole, M. Theobald, and G. Weikum. "Scalable Knowledge Harvesting with High Precision and High Recall", WSDM'11, pp. 227-236, 2011.

[11] N. Nakashole, G. Weikum, and F. Suchanek. "Discovering and Exploring Relations on the Web", VLDB Endowment, Vol. 5, No. 12, pp. 1982-1985, 2012.

[12] N. Nakashole, G. Weikum, and F. Suchanek. "PATTY A Taxonomy of Relational Patterns with Semantic Types", EMNLP-CoNLL'12, pp. 1135-1145, 2012.

[13] N. Nakashole, G. Weikum, and F. Suchanek. "Discovering Semantic Relations from the Web and Organizing Them with PATTY", SIGMOD Record, Vol. 42, No. 2, pp. 29-34, Jun 2013.

[14] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. "Unsupervised Named-Entity Extraction from the Web: an Experimental Study", Artificial Intelligence, Vol. 165, Iss. 1, pp. 91-134, 2005.

[15] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. "Open Information Extraction from the Web", IJCAI'07, pp. 2670–2676, 2007.

[16] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. "Open Information Extraction: the Second Generation", IJCAI'11, pp. 3-10, 2011.

[17] G. Weikum, and M. Theobald. "From Information to Knowledge: Harvesting Entities and Relationships from Web Sources", PODS'10, pp. 65-76, 2010.

[18] L.M. Olsen et al.. "NASA/Global Change Master Directory (GCMD) Earth Science Keywords", Version 8.0.0.0.0, 2013.

[19] C.C. Chang and C.J. Lin. "LIBSVM : a library for support vector machines", ACM TIST, Vol. 2, Iss. 3, No. 27, 2011.

[20] B. Schlkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. "Estimating the support of a high-dimensional distribution", Neural Computation, Vol. 13, Iss. 7, pp. 1443-1471, 2001.

[21] T.S. Chua, J.H. Tang, R.C. Hong, H.J. Li, Z.P. Luo, and Y.T. Zheng. "NUS-WIDE: A Real-World Web Image Database from National University of Singapore", CIVR'09, 2009.

[22] M.J. Huiskes, and M.S. Lew. "The MIR Flickr Retrieval Evaluation", MIR'08, pp. 39-43, 2008.