Building a Large Scale Knowledge Base from Chinese Wiki Encyclopedia

Zhichun Wang[1], Zhigang Wang[1], Juanzi Li[1], Jeff Z. Pan[2]

[1]Department of Computer Science and Technology, Tsinghua University
[1]{zcwang, wzhigang, ljz}@keg.cs.tsinghua.edu.cn
[2]Department of Computer Science, The University of Aberdeen
[2]Jeff.z.pan@abdn.ac.uk

Abstract: DBpedia has been proved to be a successful structured knowledge base, and large scale Semantic Web data has been built by using DBpedia as the central interlinking-hubs of the Web of Data in English. But in Chinese, due to the heavily imbalance in size (no more than one tenth) between English and Chinese in Wikipedia, there are few Chinese linked data are published and linked to DBpedia, which hinders the structured knowledge sharing both within Chinese resources and cross-lingual resources. This paper aims at building large scale Chinese structured knowledge base from Hudong, which is one of the largest Chinese Wiki Encyclopedia websites. In this paper, an upper-level ontology schema in Chinese is first learned based on the category system and Infobox information in Hudong. Totally, there are 19542 concepts are inferred, which are organized in hierarchy with maximally 20 levels. 2381 properties with domain and range information are learned according to the attributes in the Hudong Infoboxes. Then, 802593 instances are extracted and described using the concepts and properties in the learned ontology. These extracted instances cover a wide range of things, including persons, organizations, places and so on. Among all the instances, 62679 of them are linked to identical instances in DBpedia. Moreover, the paper provides RDF dump or SPARQL to access the established Chinese knowledge base. The general upper-level ontology and wide coverage makes the knowledge base a valuable Chinese semantic resource. It not only can be used in Chinese linked data building, the fundamental work for building multi lingual knowledge base across heterogeneous resources of different languages, but also can largely facilitate many useful applications of large-scale knowledge base such as knowledge question-answering and semantic search.

Keywords: Semantic Web, Linked Data, Ontology, Knowledge base

1 Introduction

The vision of Semantic Web is to build a "web of data" that enables machines to understand the semantics of information on the Web [1]. In order to achieving the goal of Semantic Web, datasets in various domains have been published and

interlinked on the Web, such as DBLP[1] in the domain of scientific publication, Myspace[2] in the domain of social networks, Linked MDB[3] and Music Brains[4] in the domain of entertainment. Besides these domain dependent datasets, several large-scale domain independent knowledge bases covering various things have also been proposed, including YAGO [2, 3], DBpedia [4, 5] and Freebase [6]. These knowledge bases typically integrate information from different resources, and provide structured information of various objects. Take DBpedia as an example, it extracts structural information from Wikipedia, provides approximately 1.2 billion triples of information and covers various domains such as geographic information, people, companies, films, music, etc. Because of well-defined ontologies and wide coverage of things, DBpedia and YAGO have the works as the core of linked data [7], and have been used in applications such as music recommendation [8], tag disambiguation [9], information extraction [10, 11].

With various knowledge of different languages are used on the web, multilinguality of semantic web is increasingly evident. Currently, DBpedia provides several versions in non-English languages, including German, French and Japanese etc. However, there is not a Chinese knowledge base with wide coverage in the cloud of linked data. This is mainly because English Wikipedia has 359 thousand articles with inter-language links to Chinese Wikipedia that is only no more than one tenth comparing to 3.64 million English articles in Wikipedia. Also, both DBpedia and YAGO only builds upper-level ontology in English, there is not a Chinese domain independent ontology for the linked data. These problems hinder the structured knowledge sharing both within Chinese resources and cross-lingual resources in Semantic Web.

In this paper, we aim at building a large-scale domain independent Chinese knowledge base on an online Chinese Wiki encyclopedia website, Hudong. As a Chinese encyclopedia, Hudong has much more articles than Chinese Wikipedia; the categories in Hudong are organized in a hierarchy like a tree, which is more suitable for building concept taxonomy. Based on our method, the extracted knowledge base consists of Chinese ontological schema and covers a large number of instances. Specifically, our work makes the following contributions:

1) We propose a method to learn an ontology from the category system and Infobox schema in Hudong. Concepts and its hierarchy are extracted from the category system by eliminating inconsistent relations and too specific categories. Properties are extracted from the Infoboxes, and their domains and ranges are properly defined according to their associated concepts. Based on the proposed method, 52404 concepts are extracted and organized in hierarchy with maximally 20 levels are extracted from Hudong. At the same time, 2381 properties with domain and range are learned to describe various relations between different concepts;
2) Based on the extracted ontology, 802,593 instances are extracted from Hudong. Three kinds of properties including *General-properties*, *Infobox-*

[1] http://dblp.rkbexplorer.com/
[2] http://dbtune.org/myspace/
[3] http://linkedmdb.org/
[4] http://dbtune.org/musicbrainz/

properties and *Person-Relation-properties* are used to describe various attributes of entities and their relationships, resulting in more than 5.2 million RDF triples. And among which, 62679 entities are linked to their identical entities in DBpedia to make our knowledge base linked with others.

3) Both RDF dump and SPARQL endpoint are provided to access our knowledge base.

The rest of this paper is organized as follows: Section 2 introduces the Hudong which is the one of the most largest Chinese Wiki encyclopedia websites, Section 3 presents our approach of ontology extraction from Hudong, Section 4 describes how the entities are extracted; Section 5 shows the results of established knowledge base; Section 6 gives the conclusion and future work.

2 Preliminary

This section first gives some related definitions, and then briefly introduces the Hudong encyclopedia.

2.1 Related Definitions

An knowledge base consist of a ontology which model the schema information, and a set of instance defined and described under the ontology constitutes the main information in the knowledge base. We formally introduce some related notions as follows.

Definition 1: An ontology is a formal specification of a shared conceptualization, which provides a vocabulary describing a domain of interest [12]. An ontology can be described as a 4-tuple:

$$O = \{C, P, H^C, H^P\}$$

where C and P are the sets of concepts and properties, respectively. H^C and H^P represents the hierarchical relationships of concepts and properties, respectively.

Definition 2: Let I be a set of instances of concepts in ontology O, the ontology O together with instances I constitute a Knowledge Base $KB = \{O, I\}$.

Definition 3: In an ontology O, properties stating relationships from instances to data values are called datatype properties; properties describing relationships between instances are called object properties.

Definition 4: In an ontology O, the concepts that a property P describes, are called the domain of property P, denote as $dom(P)$; the allowed concepts that the value of an object property P can linked to, are called the range of property P, denote as $rag(P)$.

Definition 5. Given a set of concepts $C = \{C_1, C_2, ..., C_n\}$, the Minimum General Set (MGS) of C is a set of concepts C^g that satisfies:

a. For each concept $C_i \in C$, $C_i \in C^g$ or $\exists C_i^{'} \in C^g, C_i \prec C_i^{'}$ ($A \prec B$ means A is a sub-concept of B);

b. For each concept $C_i \in C^g$, $C_i \in C$;

c. For each concept $C_i \in C^g$, $\neg\exists C_j \in C \setminus \{C_i\}$ that $C_i \prec C_j$.

The MGS in Definition 5 will be used to derive the domains and ranges of properties in the following Section. Given a set of concept C, Algorithm 1 shows how to transform it to its Minimum General Set.

Algorithm 1. *Minimum General Set* Transformation

Input:

- A concept set $C = \{c_1, c_2, ..., c_n\}$

Output:

- The *Minimum General Set* C^g of C

Begin:

$C^g \leftarrow \varnothing$;
For each concept $c_i \in C$
If $\neg\exists c_j \in C^g$ that $c_i \prec c_j$
$C^g \leftarrow C^g \bigcup \{c_i\}$
EndIf
For each concept $c_j \in C^g$
If $c_j \prec c_i$
$C^g \leftarrow C^g \setminus \{c_i\}$
EndIf
EndFor
EndFor
Return C^g
End

2.2 Hudong

Our knowledge base is built on Hudong[5], one of the worlds's largest Chinese encyclopedia website. This section gives a brief introduction to Hudong. Hudong is found in 2005, and it has more than 5 million pages created by 3 million users in May 2011. Basic entries of Hudong are article pages; each describes a specific concept or thing. Typically, each article page contains the following elements:

- Title: Every article in Hudong has a unique title, which denotes the name of the article's subject. We call the title of the article entities in this paper. It is can be words or phrase.
- Content: Content of an article is a long text which describes information in various aspect of the article's subject.
- Links: An article consists of a hypertext document with hyperlinks to other pages within or outside Hudong. The hyperlinks guide readers to the pages that

[5] http://www.hudong.com/

provide related information about the article's subject.

- Infobox: Infobox offers structured information about the article's subject in table format. It summarizes the key aspects of the article's subject by attribute-value pairs.
- Category: An article may have category tags that reflect the topic of the article's subject. One article may have several or none category tags.

There is a classification tree in Hudong to organize its articles. Nodes in the classification tree are categories, and articles are associated with their corresponding categories. There are 12 upper categories under the root of classification tree, including 社会 (Social), 地理 (Geography), 科学 (Science), 人物 (Person), 文化 (Culture), 经济 (Economics), 艺术 (Art), 自然 (Nature), 技术 (Technology), 历史 (History), 体育 (Sport), 生活 (Life). We define the 13th categories of 组织（Organization）in our defined categories which play an important role in knowledge base.

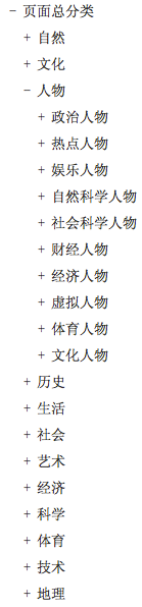

Fig1. Classification Tree in Hudong

3 Ontology extraction

We first build an upper level ontology to model the schema information of the extracted knowledge base. This section presents our approach to automatically build the ontology from the category system and Infobox templates in Hudong.

3.1 Concept extraction

A concept in ontology defines a group of instances that belong to the same type and share some common properties. Concepts can be organized in a hierarchy by specifying the *subclass-of* relation between them. The concepts and their hierarchy comprise the backbone of the ontology, which benefit the sharing and querying information of the extracted entities.

Here, we explore hudong's category system and transformed it to a taxonomy of concepts. Hudong's category system uses a classification tree to organize its articles. Articles describing the same type of things are grouped into one category, and categories have sub-categories and super-categories. For each category, there is a page in Hudong lists its sub-categories, super-categories, and articles belong to it. The categories' names and their hierarchical relations are built by collaborate editions of large number of users. In general, most hierarchical category relationships defined in the classification tree of Hudong is consistent and high qualified.

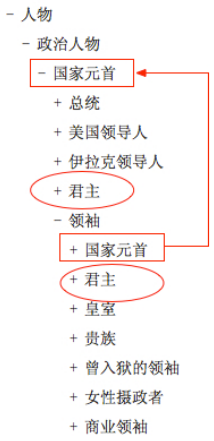

Fig. 2. A snap of the classification tree to illustrate the inconsistency in Hudong category system

We summarize three problems with Hudong categories for defining a concept hierarchy. First, there are some inconsistent sub-class links in the tree; some concepts' sub-classes may also be the super-class of it, or be the brother of its super-classes. Such as shown in Fig 2, the sub-categories of 国家元首 (Head of State) contains a node 国家元首 (Head of State), which causes a circle in the tree. Second, one category may have several super-categories. Such as shown in Fig 2, the category 君主 (Sovereign) has two sup-categories, 国家元首 (Head of State) and 领袖 (Leader). Third, some categories are too specific that only contains one or two articles, these over specific categories cannot represent a group of instances, therefore is not suitable to be extracted as concepts. In order to build a concept hierarchy, we first use the following methods to refine the category system of Hudong:

(1) Delete the inconsistent sub-category relations. Enumerate all sub-category links in the classification tree, delete the links from a category on lower level to categories on higher level. By this step, the circles in the classification tree are eliminated without destroying other category relations.
(2) Delete multiple super-categories, keep the super-category closest to the root category. In this way, only the general definitions of categories are kept.
(3) Delete specific categories that contain less than two entities.

After refining the category system of Hudong, we define concepts and concepts' hierarchy based on the refined category system. For each category, we define a concept and assign a unique URI to it. The URI of a concept is created by concatenating the namespace prefix *http://CKB.org/ontology/* and the name of the category. The hierarchy of concepts is extracted from the sub-category links in Hudong. If a concept's corresponding category has sub-categories, then concepts corresponding to these sub-categories are specified as its sub-concepts. All the defined concepts and hierarchical relations are recorded using the OWL[6] language. Fig 3 is a snap of our extracted concept hierarchy, all these concepts belongs to the 人物 (Person) concept.

Fig 3. A snap of the concept hierarchy extracted from Hudong.

3.2 Property extraction

Properties are used to describe the relationships between instances or from instances to data values. Properties in this paper are divided into two types: datatype properties, relations between instances of classes and RDF literals and XML Schema datatypes; object properties, relations between instances of two classes. We define three groups of properties for Hudong entities: *general-properties*, *Infobox-properties*, and *person-relation-properties*.

(1) *General-properties*

[6] http://www.w3.org/TR/owl-features/

The *general-properties* include label, abstract and url, they are all datatype properties. These properties describe basic information of instances. The label property specifies the name of an instance; the abstract property represents the first paragraph of the text in the instance's Hudong page; the url property gives the url of the Hudong page of an instance.

(2) *Infobox-properties*

Infobox-properties are defined based on the attributes in the Infobox, such as 姓名 (name), 年龄 (age), 籍贯(native place) in a person's Infobox. All the attributes are defined as properties with a unique URI. Here, we concatenate the namespace prefix *http://CKB.org/ontology/* and the attribute's name as the URI of the defined property. In order to determine the type of properties, i.e. object and datatype, the values of Infobox attributes need to be processed first. If the values of an attribute are plain texts, then this attribute can be defined as a datatype property. For example, the attribute 姓名 (name) can be defined as a datatype property. If the values of an attribute contain links to other entities' pages, then this attribute is an object property. For example, the attribute 校长 (president) of a university is usually a link to a person; therefore the attribute president is defined as a object property, its range is concept 人物 (person).

For each defined property, we also specify its domain and range. For the three general properties, their domain is the most general concept "*Thing*", and their range is defined as "*xsd:string*"[7]. The domains and ranges of Infobox properties are determined by as follows.

a. Domain

For each *Infobox-property* P, we enumerate all the wiki pages $W_p = \{w_1, w_2, ..., w_k\}$ that it appears; record the category tags $T_p = \{t_1, t_2, ..., t_m\}$ in the wiki pages W_p. Let $D_p = \{C_1, C_2, ..., C_m\}$ be the set of defined concepts corresponding to categories $T_p = \{t_1, t_2, ..., t_m\}$, the MGS of D_p is defined as $dom(P)$.

b. Range

For all the datatype properties, their ranges are defined as "*xsd:string*."

For each object property P, enumerate all the wiki pages $W_p = \{w_1, w_2, ..., w_k\}$ that it appears; record all the wiki pages $W_{pl} = \{w_1', w_2', ..., w_m'\}$ that the values of the property link to; enumerate pages in W_{pl} and record the category tags $T_p = \{t_1, t_2, ..., t_n\}$ in these wiki pages. Let $R_p = \{C_1, C_2, ..., C_n\}$ be the set of defined concepts corresponding to categories $T_p = \{t_1, t_2, ..., t_n\}$, the MGS of R_p is defined as $rag(P)$.

(3) *Person relation properties*

In most Hudong pages belong to person category, there is usually a person relation graph describing the relations between the person and other persons. For example, Fig. 4 is an example of person relation of Yao Ming (姚明). The graph shows other persons that related to 姚明 (Yao Ming), including his father, coach, daughter and so on. We extract these relations between persons and define object properties from them.

Because all these relations are between persons, the domains and ranges of person relation properties are all set as 人物 (person) concept.

4 Instance Extraction

4.1 Extract instances and descriptions

After the ontology being defined, entities in Hudong are extracted as the instances in the ontology. A unique URI is assigned to each instance, which its namespace prefix *http://CKB.org/ontology/* is connected with the instance's name. Concept types are assigned to instances according to their category tags in Hudong. There are three groups of properties to describe information of instances. *General-properties* including title, abstract and url are extracted for every instance. *Infobox-properties* are extracted if there is an Infobox in the instance's Hudong page. For instances belonging to 人物 (person) concept, if there are person relation graphs in their pages, *person-relation* properties will be used to describe the relationships between the instance and other instances.

Fig. 4 Person relation graph of 姚明(Yao Ming)

When extracting the values of object properties from the Infoboxes, we have to handle the problem of missing links. A lot of properties' values should be supposed to

have links to other entities, but sometimes they only have instance' name without links. For example, the president of Tsinghua University is "顾秉林" (Binglin Gu), the text "顾秉林"(Binglin Gu)" is not linked to the page of instance "顾秉林"(Binglin Gu)". Therefore, we have to find these missing links so that we can use object properties to establish RDF links between them. Here we use the method of name matching to add the missing links. The value of object property is matched with the names of all the entities. If there is an exactly matched name with the property value, then the property value is replaced with the link to the matched instance.

4.2 Link entities to DBpedia entities

In order to make our knowledge base linked with other linked data, we also create the *owl:sameAs* links with DBpedia. Indentical instances' URIs are found by the following method:

(1) Given an instance e extracted from Hudong, find the entity e' in Chinese Wikipedia with the same title.
(2) Find whether there is an inter-language links between e' and an entity e'' in English Wikipedia; if e'' is exist, get its url.
(3) Search the DBpedia URI of e'' by looking for the url.
(4) Declare URI(e) *owl:sameAs* URL(e'').

5 Results

5.1 Dataset

We wrote a web crawler that starts from the root of the classification tree in Hudong, and downloads all the articles attached to the nodes in the classification tree. Finally, we are able to download 687 thousand articles. Although the number of extracted articles is relative small comparing to the total number of articles in Hudong, these downloaded articles have high quality than the rest of articles. These 687 thousand articles usually have rich information including Infoboxes, categories, etc. Table 1 shows the number of articles in each upper category. It should be explained that each Hudong article may appear in multiple upper categories, therefore the total number of pages of 12 categories is much larger than 687 thousand.

5.2 Extracted Knowledge Base

The extracted ontology contains 19542 concepts, 2079 object properties, 302 data type properties. There are 13 upper level concepts in the ontology corresponding to the 13 categories in Hudong, including 社会 (Social), 地理 (Geography), 科学 (Science), 人物 (Person), 文化 (Culture), 组织 (Organization), 经济

(Economics), 艺术 (Art), 自然 (Nature), 技术 (Technology), 历史 (History), 体育 (Sport), 生活 (Life). As we noticed, there is not a upper level category 组织 (Organization) in Hudong category system. The categories belong to organizations appear in all the other upper level categories, such as 经济组织 (Economic Organization) category belongs to 经济 (Economics), 科研机构 (Scientific Organization) belongs to 科学 (Science), etc. Because organization is an important concept, we manually aggregate all the related categories and build "Organization" concept in our ontology. Table 2 shows the number of concepts, associated properties and hierarchy levels for each upper level concept.

Table 1. Number of articles in Hudong's upper categories

Category	#Hudong articles	Percentage
社会 (Social)	538576	15.45%
地理 (Geography)	520869	14.94%
科学 (Science)	471083	13.52%
人物 (Person)	111899	3.21%
文化 (Culture)	292680	8.40%
生活 (Life)	314047	9.01%
经济 (Economics)	211229	6.06%
艺术 (Art)	261794	7.51%
自然 (Nature)	531240	15.24%
技术 (Technology)	143537	4.12%
历史 (History)	54658	1.57%
体育 (Sport)	33657	0.97%
Total	3485269	100%

Table 2. Ontology information

Concept	#Concepts	#Related properties	#Hierarchy Levels
社会 (Social)	13515	1897	15
地理 (Geography)	11468	1482	18
科学 (Science)	5044	964	19
人物 (Person)	4345	2177	9
生活 (Life)	3379	895	10
文化 (Culture)	1947	963	10
组织 (Organization)	1845	626	10
经济 (Economics)	2346	594	10
艺术 (Art)	1536	816	10
自然 (Nature)	7035	481	17
技术 (Technology)	776	446	11
历史 (History)	1826	672	10
体育 (Sport)	694	588	8
文化 (Culture)	1947	963	10

Based on the extracted ontology, 802593 instances are defined. These instances are described by various properties resulting in 5237520 RDF triples. Table 3 shows the number of instances and RDF triples for each upper level concept.

Our knowledge base is recorded in an RDF file, and we also provide a SPARQL endpoint for querying the knowledge base. Applications can send queries by the SPARQL protocol to endpoint to get instances' structured information. Fig 5 shows the SPARQL query interface of our knowledge base. There is a sample query as follows:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ckb: <http://cbk.org#>

SELECT ?people ?property ?value
WHERE
  { ?people rdfs:label "姚明".
    ?people ?property ?value }
```

This query looks up information about a person 姚明 (Yao Ming), a famous NBA Chinese player. After submitting this query, 38 triples are returned from the knowledge base. Table 4 shows part of the query results, including Yao Ming's birth data, English name, height, etc.

Both the RDF file and SPARQL endpoint can be assessed in our project's homepage: *http://keg.cs.tsinghua.edu.cn/project/ChineseKB/*.

Table 3. Instances Information

	#Instances	#RDF tripples
社会 (Social)	326774	2447922
地理 (Geography)	311952	1999392
科学 (Science)	236187	1589840
人物 (Person)	144254	1153841
生活 (Life)	159252	1088034
文化 (Culture)	120965	879674
组织 (Organization)	107103	602378
经济 (Economics)	99927	637539
艺术 (Art)	98219	726341
自然 (Nature)	94672	1043903
技术 (Technology)	51822	294569
历史 (History)	36979	271885
体育 (Sport)	18701	177989

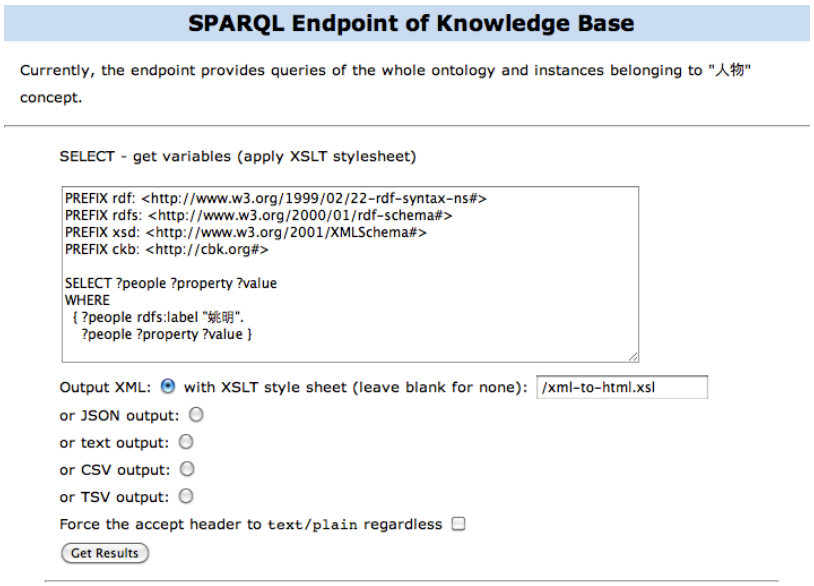

Fig. 5. The SPARQL query interface of our knowledge base

Table 4. Sample query results from the SPAQRL endpoint

people	property	value
<http://ckb.org/ontology#姚明>	<http://www.w3.org/2000/01/rdf-schema#label>	"姚明"
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#慈善家>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#出生年月>	"1980 年 9 月 12 日"
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#英文名>	"Yao Ming"
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#球类运动员>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#身高>	"226 厘米"
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#上海人>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#身材>	"140 公斤"
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#别名>	"小巨人、移动长城"
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#男演员>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#重要事件>	"2009.7，收购上海篮球队，成为新老板 2009 年带领火箭队进入季后赛第二轮 1998 年入选中国篮球明星队"

5 Related Work

In this section, we review some work related to our paper.

YAGO [2] is a large ontology built based on Wikipedia[7] and WordNet [13]. It extracts more than 1.7 million entities and 14 relationships from Wikipedia. The category system and the redirect pages are used to establish a hierarchy of concepts. In order to improve the quality of the concepts' hierarchy, YAGO links leaf categories of Wikipedia into the WordNet hierarchy. Different from our work, YAGO does not extract various properties in the Wikipedia's Infoboxes, instead, we extract 2079 properties to describe the characteristics of the concepts in this paper.

DBpedia [4] is a knowledge base which extracts structured information from Wikipedia and to make this information available on the Web. DBpeida extracts entities from Wikipeida and describes entities by a set of general properties and a set of Infobox-specific properties. The extracted entities are also mapped into four classification schemata, including DBpedia ontology[8], SKOS[9], YAGO [2] and UMBEL[10]. In our paper, we propose a framework to extract schema ontology and entities information according to the features of Chinese wiki Encyclopedia Hudong, and also generate a Chinese structured knowledge base, and 62679 of them are linked to DBpedia. These links provide the knowledge of English-Chinese languages that can be used in the application of cross-lingual knowledge base.

Freebase [6] is an open repository of structured data of almost 22 million entities. Users of Freebase can edit the data in a similar way as they edit Wikipedia articles. Free-base extracts knowledge from Wikipedia as initial content for their database, which is then edited by Freebase users. In Chinese, currently there is no such kind of the information.

Ponzetto et.al [14] proposed an approach for deriving a large-scale taxonomy from Wikipeida. They took the category system in Wikipedia as a conceptual network, and created subsumption hierarchy of concepts. In order to determine the *isa* relation between concepts, they used methods based on the connectivity of the network and on applying lexico-syntactic patterns to Wikipedia articles. They mainly focused on building the subsumption relations between concepts and did not include the instances and their Infoboxes' information in the taxonomy.

Melo et.al [15] explored the multilingual nature of Wikipedia, and built a large multilingual entity taxonomy MENTA, which describes 5.4 million entities in various languages. They integrated entities from all editions of Wikipedia and WordNet to a single coherent taxonomic class hierarchy. Categories are extracted as candidates of classes; categories denoting genuine classes and topic labels are distinguished by the singular/plural heuristic proposed for YAGO [2]. Only categories denoting genuine

[7] http://www.wikipedia.org/
[8] http://wiki.dbpedia.org/Ontology
[9] http://www.w3.org/2004/02/skos/
[10] http://www.umbel.org/

classes are defined as classes. The *subclass* relations between classes are established by making use of parent categories, category-WordNet subclass relationships and WordNet Hyponymy. Instances are extracted based on the Infoboxes and categories in articles.

To summarize the related work, in this paper we propose a framework to extract schema ontology of knowledge base and generate the structured knowledge base from one of the largest wiki Encyclopedia website-Hudong. Currently, there is no RDF DBpedia like knowledge base in Chinese and few Chinese data sets are linked to DBpedia. Besides, we provide links to DBpedia, which lay the foundation for cross lingual structured knowledge base sharing by integrating structured knowledge base across existing wiki Encyclopedia websites of different languages.

6 Conclusion

This paper presents a Chinese knowledge base built from a Chinese Wiki websites, Hudong. An upper level Chinese ontology is first built based on the category system and Infobox schema of Hudong. Then more than 800 thousand entities in various domains are extracted and classified according to the defined ontology. Structured information of entities is described by the defined properties. As the development of semantic techniques and applications, our knowledge base can be used as a useful Chinese semantic resource. Currently, both RDF dump and SPARQL endpoints are provided to access our knowledge base.

As our future work, we will concentrate on improving the quality of the Chinese structured knowledge base including the refinement of the schema ontology and the process of the entities learning. We want to build links from other data sets such as in news domain and academic domain to this structured knowledge base to build Chinese linked data. Also, linking Chinese structured knowledge with DBpedia or other knowledge bases of different languages could fulfill structured knowledge sharing across heterogeneous knowledge bases of different languages.

7 Acknowledgement

The work is supported by the National Natural Science Foundation of China (No. 6 61035004, 60973102), the National Basic Research Program of China (973 Program) (No. 2007CB310803), the China Postdoctoral Science Foundation (No. 20110490390), it is also supported by THU-NUS Next research center.

Reference

[1]. Berners-Lee, T. *Semantic Web Road map*. 1998; Available from: http://www.w3.org/DesignIssues/Semantic.html.

[2]. Suchanek, F.M., G. Kasneci, and G. Weikum, *YAGO: A Large Ontology from Wikipedia and WordNet*. Web Semantics: Science, Services and Agents on the World Wide Web, 2008. **6**(3): p. 203-217.

[3]. Suchanek, F.M., G. Kasneci, and G. Weikum, *Yago: a core of semantic knowledge*, in *Proceedings of the 16th international conference on World Wide Web*2007, ACM: Banff, Alberta, Canada. p. 697-706.

[4]. Bizer, C., et al., *DBpedia - A crystallization point for the Web of Data*. Web Semantics: Science, Services and Agents on the World Wide Web, 2009. **7**(3): p. 154-165.

[5]. Auer, S.r., et al., *DBpedia: A Nucleus for a Web of Open Data The Semantic Web*, in *The Semantic Web*, K. Aberer, et al., Editors. 2007, Springer Berlin / Heidelberg. p. 722-735.

[6]. Bollacker, K., et al., *Freebase: a collaboratively created graph database for structuring human knowledge*, in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*2008, ACM: Vancouver, Canada. p. 1247-1250.

[7]. Bizer, C., T. Heath, and T. Berners-Lee, *Linked Data - the story so far*. International Journal on Semantic Web and Information Systems, 2009. **5**(3).

[8]. Passant, A., *dbrec - Music Recommendations Using DBpedia*, in *9th International Semantic Web Conference (ISWC2010)*2010.

[9]. García-Silva, et al., *Preliminary Results in Tag Disambiguation using DBpedia*, in *First International Workshop Collective Knowledge Capturing and Representation CKCaR09*2009: Redondo Beach, California, USA.

[10]. Wu, F. and D.S. Weld, *Automatically refining the wikipedia infobox ontology*, in *Proceeding of the 17th international conference on World Wide Web*2008, ACM: Beijing, China. p. 635-644.

[11]. Kasneci, G., et al., *The YAGO-NAGA approach to knowledge discovery*. SIGMOD Record, 2008.

[12]. Euzenat, J. and P. Shvaiko, *Ontology Matching*2007: Springer-Verlag,Heidelberg(DE).

[13]. Fellbaum, C., *WordNet: An Electronic Lexical Database*. WordNet: An Electornic Lexical Database, ed. C. Fellbaum1998: MIT Press.

[14]. Ponzetto, S.P. and M. Strube, *Deriving a large scale taxonomy from Wikipedia*, in *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*2007, AAAI Press: Vancouver, British Columbia, Canada. p. 1440-1445.

[15]. Melo, G.d. and G. Weikum, *MENTA: inducing multilingual taxonomies from wikipedia*, in *Proceedings of the 19th ACM international conference on Information and knowledge management*2010, ACM: Toronto, ON, Canada. p. 1099-1108.