Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Knowledge extraction from Chinese wiki encyclopedias*

Zhi-chun WANG[1], Zhi-gang WANG[1], Juan-zi LI[1], Jeff Z. PAN[2]

([1]*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing 100084*, *China*)

([2]*Department of Computer Science*, *University of Aberdeen*, *Aberdeen AB24 3UE*, *UK*)

E-mail: {zcwang, wzhigang, ljz}@keg.cs.tsinghua.edu.cn; Jeff.z.pan@abdn.ac.uk

Received Aug. 10, 2011; Revision accepted Jan. 20, 2012; Crosschecked Feb. 27, 2012

Abstract: The vision of the Semantic Web is to build a 'Web of data' that enables machines to understand the semantics of information on the Web. The Linked Open Data (LOD) project encourages people and organizations to publish various open data sets as Resource Description Framework (RDF) on the Web, which promotes the development of the Semantic Web. Among various LOD datasets, DBpedia has proved a successful structured knowledge base, and has become the central interlinking-hub of the Web of data in English. However, in the Chinese language, there is little linked data published and linked to DBpedia. This hinders the structured knowledge sharing of both Chinese and cross-lingual resources. This paper deals with an approach for building a large-scale Chinese structured knowledge base from Chinese wiki resources, including Hudong and Baidu Baike. The proposed approach first builds an ontology based on the wiki category system and infoboxes, and then extracts instances from wiki articles. Using Hudong as our source, our approach builds an ontology containing 19 542 concepts and 2381 properties. 802 593 instances are extracted and described using the concepts and properties in the extracted ontology and 62 679 of them are linked to equivalent instances in DBpedia. As from Baidu Baike, our approach builds an ontology containing 299 concepts, 37 object properties, and 5590 data type properties. 1 319 703 instances are extracted from Baidu Baike, and 84 343 of them are linked to instances in DBpedia. We provide RDF dumps and SPARQL endpoint to access the established Chinese knowledge bases. The knowledge bases built using our approach can be used not only in Chinese linked data building, but also in many useful applications of large-scale knowledge bases, such as question-answering and semantic search.

Key words: Semantic Web, Linked Data, Ontology, Knowledge base
doi:10.1631/jzus.C1101008 **Document code:** A **CLC number:** TP311

1 Introduction

The Semantic Web is an extension of the current World Wide Web, where the semantics of data are accurately defined, and computers can directly and indirectly process the data on the Web (Berners-Lee, 1998; Shadbolt *et al*., 2006). To achieve the vision of the Semantic Web, World Wide Web Consortium (W3C, http://www.w3.org/) set up a Linked Open Data (LOD, http://linkeddata.org/) project to encourage people and organizations to publish various open data sets as Resource Description Framework (RDF) on the Web. LOD data sets are published following a set of rules outlined by Bizer *et al*. (2009b) and Heath and Bizer (2011), which are as follows: (1) Uniform resource identifiers (URIs) are used as names of things; (2) HTTP URIs are used so that people can look up those names; (3) When someone looks up a URI, it should provide useful information; (4) Links to other URIs should be included so that more things can be discovered. Founded in 2007, LOD has grown considerably since then. There were 295 datasets containing more than 31 billion RDF triples in the LOD project in September 2011. LOD datasets have been used in various domains, including DBLP (http://dblp.rkbexplorer.com/) in the domain of scientific publication, Myspace (http://dbtune.org/myspace/) in the domain of social networks, and LinkedMDB (http://linkedmdb.org/) and MusicBrainz

* Project supported by the National Natural Science Foundation of China (Nos. 661035004 and 60973102), the China Postdoctoral Science Foundation (No. 20110490390), and the THU-NUS Next Research Center

(http://dbtune.org/musicbrainz/) in the domain of entertainment. Besides these domain-dependent datasets, LOD also contains several large-scale cross-domain knowledge bases covering various things, including YAGO (Suchanek *et al*., 2007; 2008), DBpedia (Auer *et al*., 2007; Bizer *et al*., 2009a), and Freebase (Bollacker *et al*., 2008). These knowledge bases typically integrate information from different resources, and define consistent ontologies to describe structured information of various things. As an example, DBpedia extracts structural information from Wikipedia, provides approximately 1.2 billion RDF triples, and covers various domains including geographic information, people, companies, films, and music. Because of well-defined ontologies and wide coverage of things, knowledge bases such as DBpedia and YAGO have become the core of linked data (Bizer *et al*., 2009a). These knowledge bases have also been used in many applications such as music recommendation (Passant, 2010), tag disambiguation (García-Silva *et al*., 2009), and information extraction (Wu and Weld, 2007; 2008; Kasneci *et al*., 2008).

As various knowledge in different languages is used on the Web, multilingualism of the Semantic Web is evident. Currently, DBpedia provides several versions in non-English languages, including German, French, and Japanese. Because Chinese Wikipedia contains only 359 thousand articles, there is no Chinese DBpedia. As a result, LOD lacks a large-scale Chinese knowledge base. Also, in both DBpedia and YAGO, there is only upper-level ontology in English and no Chinese domain independent ontology for the linked data. These problems hinder the sharing of structured knowledge within both Chinese and cross-lingual resources in the Semantic Web.

In this paper, we propose an approach to build large-scale cross-domain Chinese knowledge bases from Chinese wiki resources. Chinese wikis, such as Hudong and Baidu Baike, have a large number of Chinese articles. Our approach takes these wikis as inputs to build Chinese knowledge bases. We make the following contributions:

1. We propose a method to extract an ontology from the category system and infobox schema of a Chinese wiki. Concepts and concept hierarchy are extracted from the category system by eliminating several inconsistent sub-category relations and too specific categories containing a small number of articles. Three kinds of properties are extracted from different parts of wiki articles, including general properties, infobox properties, and person-relation properties. Their domains and ranges are properly defined according to their associated concepts.

2. Based on the extracted ontology, structural information is extracted to define instances of the knowledge base. Instances are assigned to concepts according to the categories of their corresponding wiki articles. Instances are also linked to DBpedia instances by using cross-lingual links in Wikipedia.

3. Based on the proposed method, we build knowledge bases from Hudong and Baidu Baike, respectively. Based on Hudong, 52 404 concepts and 2381 properties are defined in the extracted ontology and 802 593 instances are extracted. Among these, 62 679 instances are linked to instances in DBpedia. From Baidu Baike, 299 concepts, 5627 properties, and 1 319 703 instances are extracted, 84 343 of which are linked to instances in DBpedia. Both RDF dumps and SPARQL endpoint are provided to access the newly built knowledge bases.

2 Preliminaries

This paper presents an approach for building Chinese knowledge bases from wiki encyclopedias. In this section we first introduce several well-known Chinese wiki encyclopedias, and then present some related definitions.

2.1 Chinese wiki encyclopedias

Currently, there are several large-scale Chinese wiki encyclopedias, including Chinese Wikipedia, Hudong, and Baidu Baike. Chinese Wikipedia was launched in 2002, and had 390 thousand articles in December 2011. Hudong was found in 2005, and had about 5.9 million articles in December 2011. Baidu Baike was found in 2006, and had more than 4.1 million articles in December 2011. All these wiki encyclopedias were built upon similar wiki software and therefore have similar structures. A wiki encyclopedia basically contains two groups of important elements, articles and categories. An article describes

a notable encyclopedic topic and contains rich information about that topic. Fig. 1 shows a snapshot of an article in Hudong.

Fig. 1 Snapshot of an article in Hudong

Typically, there are six elements in an article:

1. Title: each article has a unique title, which is at the top of the article page. Title represents the topic of the article.

2. Abstract: usually, the first paragraph in an article page summarizes the most important information in the article, and is often called abstract of the article.

3. Description: description of an article is a long text that describes detailed information in various sections of the article.

4. Links: similar to links on Web pages, links in an article refer to other articles within the wiki. They guide readers to the articles that provide related information.

5. Infobox: an infobox offers structured information about the article in a table format. It provides a set of subject-attribute-value triples summarizing the key aspects of the article.

6. Category tags: an article may have category tags that reflect the topic of the article. One article may have more than one category tag.

Besides article pages, wiki encyclopedias also contain category pages to group together articles on similar topics. A category page contains a list of articles that have been added to that particular category. There may also be a list of links to sub-categories of that category. All the categories and their sub-category relations constitute a tree-like network. Fig. 2 shows a category page in Hudong. It contains lists of super-categories, sub-categories, and articles belonging to that category.

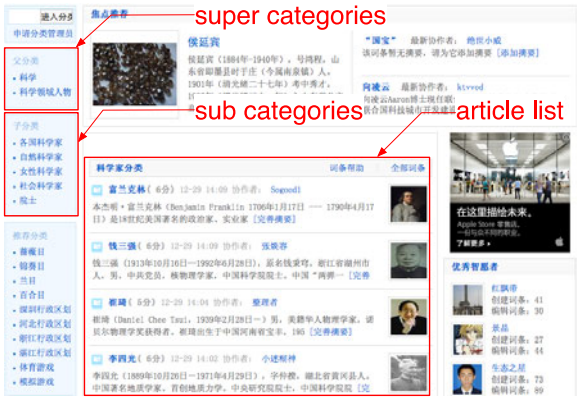

Fig. 2 A category page in Hudong

2.2 Related definitions

Here we present several definitions related to ontology and the knowledge base.

Definition 1 (Wiki encyclopedia) A wiki encyclopedia is a set of collaboratively written articles, which are organized in categories. It can be formally defined as a 3-tuple:

$$W =< A, C, L >,$$

where A denotes the set of articles, C the set of categories, and L the set of links in W.

As described in Section 2.1, an article contains several types of information. For each article $a_i \in A$, we represent it as a 6-tuple:

$$a_i =< T(a_i), G(a_i), D(a_i), L(a_i), I(a_i), C(a_i) >,$$

where $T(a_i)$, $G(a_i)$, $D(a_i)$, $L(a_i)$, $I(a_i)$, and $C(a_i)$ represent the title, abstract, description, links, infobox, and category tags of article a_i, respectively.

Definition 2 (Ontology) An ontology is a formal specification of a shared conceptualization, which

provides a vocabulary describing a domain of interest (Euzenat and Shvaiko, 2007). 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, and H^C and H^P represent the hierarchical relationships of concepts and properties, respectively.

Definition 3 (Ontology property) Ontology properties stating relationships from instances to data values are called data type properties, and properties describing relationships between instances are called object properties.

Definition 4 (Domain and range) In an ontology, the concepts that a property P describes are called the domain of property P, denoted as dom(P); the allowed concepts that the value of an object property P can be linked to are called the range of property P, denoted as rag(P).

Definition 5 (Knowledge base) Let I be a set of instances of concepts in ontology O. The ontology O and instances I constitute a knowledge base KB=$\{O, I\}$.

When extracting properties from wikis, we need to properly define their domains and ranges. Here we introduce the minimum general set (MGS), which is closely related to the computation of the domains and ranges of properties.

Definition 6 (Minimum general set) Given a set of concepts, $C=\{C_1, C_2, \ldots, C_n\}$, the MGS of C is a set of concepts C^g satisfying:

1. 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);

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

3. For each concept $C_i \in C^g$, $\neg\exists C_j \in C \backslash \{C_i\}$ such that $C_i \prec C_j$.

Given a set of concepts, C, Algorithm 1 shows how to obtain its MGS.

Algorithm 1 Minimum general set transformation

Input: A concept set $C=\{c_1, c_2, \ldots, c_n\}$.
Output: The minimum general set C^g of C.

Begin
  $C^g \leftarrow \varnothing$
  For each concept $c_i \in C$
    If $\neg\exists c_j \in C^g$ such that $c_i \prec c_j$
      $C^g \leftarrow C^g \cup \{c_i\}$
    End If
    For each concept $c_i \in C^g$
      If $c_j \prec c_i$
        $C^g \leftarrow C^g \backslash \{c_i\}$
      End If
    End For
  End For
  Return C^g
End

3 Ontology extraction

To build a knowledge base from a wiki encyclopedia, we first build an ontology to model the schema information of the extracted knowledge base. The ontology defines concepts and properties and their hierarchy relations. In this section, we present an approach for building an ontology based on the category system and infobox templates in a wiki encyclopedia.

3.1 Concept extraction

A concept in an 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, and benefit the information sharing and querying of the extracted knowledge base.

In wiki encyclopedias, the categories group similar articles and have super- and sub-category relations. Therefore, categories have very similar functions compared to ontology concepts, and we may define concepts and their relations based on the category system.

However, there are several problems with the wiki category system when transforming it into a concept hierarchy. First, there are some inconsistent sub-category links in the category system; some categories' sub-category may also be their super-category, or be the brother of their super-categories. As shown in Fig. 3, the sub-categories of 国家元首 (Head of State) contain a node 国家元首, which causes a circle in the category tree. Second, one

category may have several super-categories. In Fig. 3, the category 君主(Monarch) has two super-categories, 国家元首 and 领袖(Leader). Third, some categories are very specific and contain only one or two articles. These over-specific categories cannot represent a group of instances, and therefore is not suitable to be extracted as concepts.

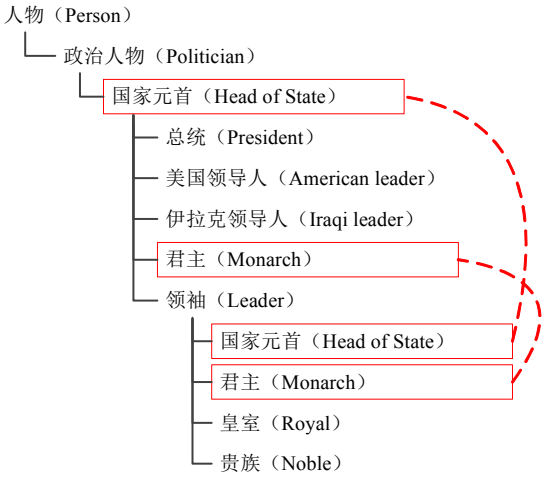

Fig. 3 A snapshot of Hudong's category tree to illustrate the inconsistency in the category system
The sub-categories of 国家元首 contain a node 国家元首; the category 君主 has two super-categories, 国家元首 and 领袖

To handle the above problems, we use the following methods to refine the category system before defining the concept hierarchy:

1. Delete the inconsistent sub-category relations. Enumerate all sub-category links in the category tree, and delete the links from a category on a lower level to categories on a higher level. By this step, the circles in the category tree are eliminated without destroying other relations between categories.

2. Delete multiple super-categories and keep the super-category closest to the root category. In this way, only the general definitions of categories are kept.

3. Delete over-specific categories that contain less than three entities.

After refining the category system of a wiki, 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 the wiki. Sub-category relations between categories are transformed into sub-concept relations between their corresponding concepts. All the defined concepts and hierarchical relations are recorded using the Web Ontology Language (OWL, http://www.w3.org/TR/owl-features/). Fig. 4 shows part of the concept hierarchy extracted from Hudong. All these concepts belong to the 人物 (Person) concept.

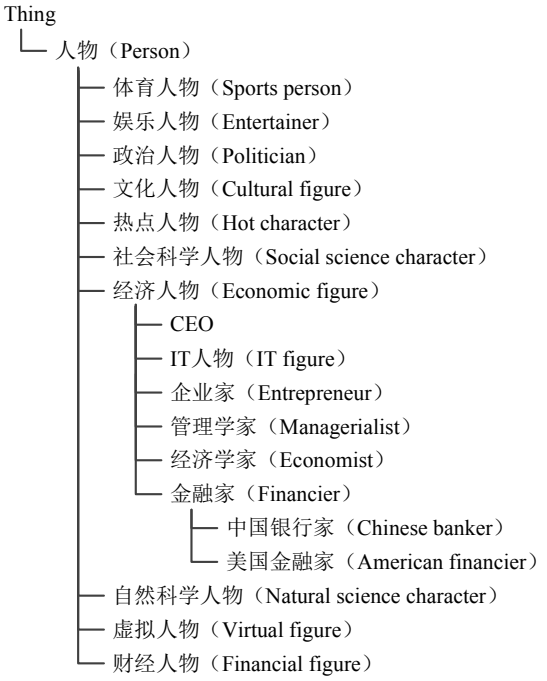

Fig. 4 Part of the concept hierarchy extracted from Hudong

3.2 Property extraction

Properties are used to describe the relationships between instances or from instances to data values. Properties can be divided into two types: data type properties, describing relations between instances of classes and RDF literals and eXtensible Markup Language (XML) schema data types; object properties, describing relations between instances of two classes. We define three kinds of properties according to three groups of information in wiki articles: general properties, infobox properties, and person-relation properties.

3.2.1 General properties

The general properties include label, abstract, and universal resource locator (URL). They are all data type 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 wiki page; the URL property gives the URL of the wiki page of an instance.

3.2.2 Infobox properties

Infobox properties are defined based on the attributes in the infoboxes, such as 姓名(name), 年龄(age), and 籍贯(native place) in a person's infobox. To determine the type of property, i.e., object or data type, the values of infobox attributes need to be analyzed first. If the values of an attribute are plain text, then this attribute can be defined as a data type property. For example, the attribute 姓名 can be defined as a data type property. If the values of an attribute contain links to other article pages, then this attribute is an object property. For example, the attribute 校长 (president) of a university is usually a link to an article page about a person; therefore, the attribute 校长 is defined as an object property.

3.2.3 Person-relation properties

In Hudong's articles, some pages that belong to the person category have a person relation graph describing the relations between the person and other persons. For example, Fig. 5 is an example of person relation of Yao Ming(姚明), a famous NBA Chinese player. The graph shows persons related to 姚明—his father, coach, daughter, and so on. We extract these relations between persons and define object properties based on these relations.

Each of the defined properties is assigned a unique URI. Here, we concatenate the namespace prefix http://CKB.org/ontology/ and the property's name as the URI of the defined property. For each defined property, we also specify its domain and range. For general properties, their domains are the most general concept 'thing', and their ranges are all defined as 'xsd:string'. For person-relation properties, because all these relations are between persons, the domains and ranges of person-relation properties are all set as the 人物(person) concept. The domains and ranges of infobox properties depend on categories that the infobox belongs to and the types of the attribute values. Here, we propose a method to determine the domains and ranges of infobox properties.

1. Domain

For each infobox property P, we enumerate all the wiki pages $W_P=\{w_1, w_2, \ldots, w_k\}$ where it appears, and record the category tags $T_P=\{t_1, t_2, \ldots, t_m\}$ in the wiki pages W_P. Let $D_P=\{C_1, C_2, \ldots, C_m\}$ be the set of defined concepts corresponding to categories $T_P=\{t_1, t_2, \ldots, t_m\}$. The MGS of D_P is defined as dom(P).

2. Range

For all the data type properties, their ranges are defined as 'xsd:string'.

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

Fig. 5 Person relation graph of 姚明(Yao Ming)

4 Instance extraction

In this section, we first introduce the process of defining instances based on wiki articles, and then present the method for linking those instances to DBpedia.

4.1 Defining instances from wiki articles

After an ontology is defined, article pages in the wiki encyclopedia are extracted as the instances of the ontology. A unique URI is assigned to each instance. The URI is the namespace prefix http://CKB.org/ontology/ connected to the instance name. Concept types are assigned to instances according to their category tags in the wiki article. There are three groups of properties for describing the 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 wiki page. For instances belonging to the 人物(person) concept, if there are person relation graphs in their pages, person-relation properties will be used to describe the relationships between this instance and other instances.

When extracting the values of object properties from infoboxes, the problem of missing links should be taken care of. A lot of property values are supposed to have links to other wiki pages, but sometimes they have only the instance name without links. For example, in the infobox of the 清华大学(Tsinghua University) article in Hudong, president of Tsinghua University is 顾秉林(Binglin Gu); however, the text 顾秉林 is not linked to the page of instance 顾秉林. 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 values of object properties are matched with the names of all the instances. 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 Linking instances to DBpedia

DBpedia is a large-scale structured knowledge base in English, and has become the central interlinking-hub of the Web of data. To make the extracted instances linked to other RDF datasets in the LOD, we propose methods for linking instances to DBpedia's instances. Since DBpedia is built on the Wikipedia, we extract 202 013 English-Chinese cross-lingual links in Wikipedia, and use them to help link Chinese instances to English instances. Specifically, instance links are established by the following steps:

1. Given a Chinese instance e, find the wiki article e' in Chinese Wikipedia with the same title.

2. Find whether there is a cross-lingual link between e' and an English article e'' in English Wikipedia. If e'' exists, get its URL.

3. Search the DBpedia URI of e'' by looking for the URL.

4. Declare URI(e) owl:sameAs URI(e'').

5 Extracted knowledge base

In this section, we demonstrate the results of the Chinese knowledge bases extracted from Hudong and Baidu Baike, respectively, and also introduce the SPARQL endpoint to access these knowledge bases.

5.1 Hudong knowledge base

We wrote a Web crawler that starts from the root of the category tree in Hudong and can download all the articles attached to the nodes in the category tree. We can download 687 thousand articles from Hudong. Although the number of extracted articles is relatively small as compared to the total number of articles in Hudong, these downloaded articles have higher quality than the remaining articles. These 687 thousand articles have rich information including infoboxes, categories, etc. Table 1 shows the number of articles in each upper category.

The extracted ontology contains 19 542 concepts, 2079 object properties, and 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), 生活(Life), 文化(Culture), 组织(Organization), 经济(Economics), 艺术(Art), 自然(Nature), 技术(Technology), 历史(History), 体育(Sport). As we see, the upper-level category 组织(Organization) is not contained in the Hudong

category system. The categories belonging to organizations appear in all the other upper-level categories. For example, the 经济组织(Economic Organization) category belongs to 经济(Economics), and the 科研机构(Scientific Organization) category belongs to 科学(Science). Because 组织(Organization) is an important concept, we manually aggregate all the related categories and build the 'Organization' concept in our ontology. Table 2 shows the numbers of concepts, associated properties, and hierarchy levels for each upper-level concept.

Table 1 Number of articles in Hudong's upper-level categories

Category	Number of Hudong articles	Percentage
Social	538 576	15.45%
Geography	520 869	14.94%
Science	471 083	13.52%
Person	111 899	3.21%
Culture	292 680	8.40%
Life	314 047	9.01%
Economics	211 229	6.06%
Art	261 794	7.51%
Nature	531 240	15.24%
Technology	143 537	4.12%
History	54 658	1.57%
Sport	33 657	0.97%
Total	3 485 269	100%

A total of 687 thousand articles are downloaded from Hudong. Since each Hudong article may occur in multiple categories, the total number of pages is much larger

Table 2 Hudong ontology information

Concept	Number of concepts	Number of related properties	Number of hierarchy levels
Social	13 515	1897	15
Geography	11 468	1482	18
Science	5044	964	19
Person	4345	2177	9
Life	3379	895	10
Culture	1947	963	10
Organization	1845	626	10
Economics	2346	594	10
Art	1536	816	10
Nature	7035	481	17
Technology	776	446	11
History	1826	672	10
Sport	694	588	8

Based on the extracted ontology, 802 593 instances are defined. These instances are described by various properties, resulting in 5 237 520 RDF triples. Table 3 shows the numbers of instances and RDF triples for each upper-level concept. Among these instances, 62 679 instances are linked to instances in DBpedia.

Table 3 Hudong instance information

Concept	Number of instances	Number of RDF triples
Social	326 774	2 447 922
Geography	311 952	1 999 392
Science	236 187	1 589 840
Person	144 254	1 153 841
Life	159 252	1 088 034
Culture	120 965	879 674
Organization	107 103	602 378
Economics	99 927	637 539
Art	98 219	726 341
Nature	94 672	1 043 903
Technology	51 822	294 569
History	36 979	271 885
Sport	18 701	177 989

5.2 Baidu knowledge base

We downloaded 1.3 million articles from Baidu Baike. Table 4 shows the number of articles in each upper-level category. By further observation we find that the category tree system of Baidu is relatively consistent compared with Hudong. Each Baidu article belongs to only one upper-level category.

Table 4 Number of articles in Baidu's upper-level categories

Category	Number of Baidu articles	Percentage
Social	133 969	10.15%
Geography	185 204	14.03%
Science	159 832	12.11%
Person	104 721	7.94%
Culture	266 902	20.22%
Life	214 302	16.24%
Economics	13 897	1.05%
Art	42 791	3.24%
Nature	71 369	5.41%
Technology	83 212	6.31%
History	21 356	1.62%
Sport	22 148	1.68%
Total	1 319 703	100%

The extracted ontology contains 1299 concepts, 37 object properties, and 5590 data type properties. There are 12 upper-level concepts in the ontology corresponding to the 12 categories in the Baidu category tree, including 社会(Social), 地理(Geography), 科学(Science), 人物(Person), 文化(Culture), 生活(Life), 经济(Economics), 艺术(Art), 自然(Nature), 技术(Technology), 历史(History), 体育(Sport). Table 5 shows the numbers of concepts, associated properties, and hierarchy levels for each upper-level concept.

Table 5 Baidu ontology information

Concept	Number of concepts	Number of related properties	Number of hierarchy levels
Social	89	819	3
Geography	133	1863	3
Science	157	601	3
Person	120	1109	3
Life	101	1722	3
Culture	131	834	3
Economics	30	212	3
Art	84	401	3
Nature	86	580	3
Technology	78	344	3
History	117	449	3
Sport	161	537	3

Unlike the Hudong knowledge base, each upper-level category of the Baidu knowledge base contains the same three levels. Besides, articles (instances) are attached only to the leaf concept. It is also found that the size of Baidu's category tree is much smaller than the Hudong one.

Based on the extracted ontology, 1 319 703 instances are defined. These instances are described by various properties, resulting in 4 590 144 RDF triples. 84 343 instances are linked to instances in DBpedia. Table 6 shows the numbers of instances and RDF triples for each upper-level concept. The number of instances in an upper-level concept is equal to the number of articles in the corresponding upper-level category in Baidu Baike.

5.3 SPARQL endpoint for knowledge bases

The extracted knowledge bases are recorded in RDF files. Then a SPARQL endpoint is set up for querying the information in knowledge bases. In the applications queries can be sent according to the SPARQL protocol to the endpoint to obtain the structured information of the instances. Fig. 6 shows the SPARQL query interface of our knowledge bases.

Table 6 Baidu instance information

Concept	Number of instances	Number of RDF triples
Social	133 969	445 396
Geography	185 204	665 287
Science	159 832	537 967
Person	104 721	386 266
Life	266 902	736 740
Culture	214 302	841 496
Economics	13 897	46 804
Art	42 791	142 953
Nature	71 369	286 191
Technology	83 212	250 934
History	21 356	77 474
Sport	22 148	90 972

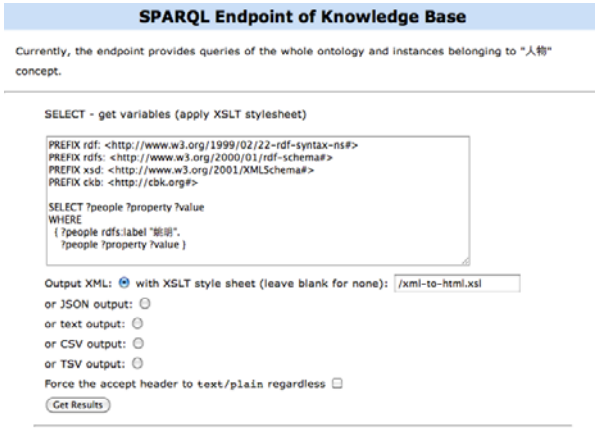

Fig. 6 The SPARQL query interface

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). After submitting this query, 38 triples are returned from the knowledge base. Table 7

Table 7 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(Ming Yao)”
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#慈善家(Philanthropist)>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#出生年月(Date of Birth)>	“1980 年 9 月 12 日”
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#英文名(English Name)>	“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(Ball Player)>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#身高(Height)>	“226 厘米(cm)”
<http://ckb.org/ontology#姚明>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://ckb.org/ontology#上海人(People in Shanghai)>
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#身材(Weight)>	“140 公斤(kg)”
<http://ckb.org/ontology#姚明>	<http://ckb.org/ontology#别名(Alias)>	“小巨人(Little Giant)”, “移动长城(The Moving Great Wall)”

Text in brackets, which is translation of Chinese text, does not appear in the original query result

shows part of the query results, including Yao Ming’s birth date, English name, height, etc.

Both the RDF files and SPARQL endpoint can be accessed from our project homepage http://keg.cs.tsinghua.edu.cn/project/ChineseKB/.

6 Related work

Traditional knowledge bases are built by experts, including WordNet (Fellbaum, 1998), EuroWordNet (Piek, 1997; Vossen, 1998), Cyc (Lenat, 1995; Matuszek *et al*., 2006), and SUMO (Niles and Pease, 2001; Pease and Niles, 2002). WordNet is a lexical knowledge base for the English language, created by linguists and computer engineers in Princeton University. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. In version 3.0, WordNet has 155 287 unique strings and 117 659 synsets. Similar to WordNet, EuroWordNet is also a lexical knowledge base. Besides English, EuroWordNet contains wordnets for Spanish, Dutch, and Italian. The Cyc knowledge base is a formalized representation of a vast quantity of fundamental human knowledge, created mostly by human experts. Cyc was started in 1984. After decades of development, Cyc has contained about five hundred thousand terms, including about fifteen thousand types of relations and about five million assertions relating to these terms. SUMO, first released in December 2000, defines a hierarchy of SUMO classes and related rules and relationships. It provides a foundation for middle-level and domain ontologies, and its purpose is to promote data interoperability, information retrieval, automated inference, and natural language processing. SUMO was designed to be relatively small, to keep the assertions and concepts understandable and easy to apply, and consists of approximately 4000 assertions (including over 800 rules) and 1000 concepts.

One of the advantages of traditional knowledge bases is their high quality and consistency. However, since building these knowledge bases by human costs a lot of time, their sizes are relatively small. As a huge amount of documents are available on the Web, much research has been done on building knowledge bases out of the information extracted from unstructured or semi-structured text (Maedche and Staab, 2001; Buitelaar *et al.*, 2005; Buitelaar and Cimiano, 2008). One representative of the ontology learning system is OntoLearn (Navigli *et al*., 2003; Navigli and Velardi, 2004), which can build domain ontologies from Web sites, and more generally, from the document corpus. OntoLearn extracts domain terminology from

available documents and then arranges domain terms in a hierarchy. However, unreliable source data often causes inaccuracy of the learned ontologies.

The development of wiki resources such as Wikipedia, attracts much research interest in building knowledge bases from Wikipedia.

YAGO (Suchanek *et al*., 2008) is a large ontology built on Wikipedia (http://www.wikipedia.org/) and WordNet (Fellbaum 1998). 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. To improve the quality of the concepts' hierarchy, YAGO links leaf categories of Wikipedia into the WordNet hierarchy. YAGO does not extract various properties in Wikipedia's infoboxes; instead, in this work we extract properties to describe the characteristics of the concepts.

DBpedia (Bizer *et al*., 2009a) is a knowledge base which extracts structured information from Wikipedia to make it available on the Web. DBpedia extracts entities from Wikipedia 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 the DBpedia ontology (http://wiki.dbpedia.org/Ontology), SKOS (http://www.w3.org/2004/02/skos/), YAGO (Suchanek *et al*., 2008), and UMBEL (http://www.umbel.org/). In this paper, we propose a framework to extract ontology and instances from Chinese wiki encyclopedias, and also instance links to DBpedia. These links provide the knowledge of English-Chinese languages that can be used in the application of the cross-lingual knowledge base.

Freebase (Bollacker *et al*., 2008) 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. Freebase extracts knowledge from Wikipedia as initial content for its database, which is then edited by Freebase users. Currently, there is no such kind of information in Chinese.

Ponzetto and Strube (2007) proposed an approach for deriving a large-scale taxonomy from Wikipedia. They took the category system in Wikipedia as a conceptual network, and created a subsumption hierarchy of concepts. To determine the isa relation between concepts, they used methods based on the connectivity of the network and on the application of lexico-syntactic patterns to Wikipedia articles. They focused mainly on building the subsumption relations between concepts and did not include the information of the instances or their infoboxes in the taxonomy.

Melo and Weikum (2010) 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 into 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 (Suchanek *et al*., 2008). Only categories denoting genuine classes are defined as classes. The subclass relations between classes are established using parent categories, category-WordNet subclass relationships, and WordNet hyponymy. Instances are extracted based on the infoboxes and categories in articles.

Most recently, Niu *et al*. (2011) have proposed a very similar approach, Zhishi.me, to publish large-scale Chinese semantic data. They extracted structural information of entities from Baidu Baike, Hudong, and Chinese Wikipedia, and used several strategies to link equivalent entities in different resources. Finally, they were able to publish five million distinct entities of semantic data. Compared with our knowledge bases, Zhishi.me contains more entities, but it does not define an ontology to describe the schema information of the published semantic data.

To summarize the related work, in this paper we present an approach to extract ontology and instances from Chinese wiki encyclopedias to build knowledge bases. Our approach first extracts an ontology for the knowledge base and then creates instances from wiki articles. Currently, there is no Chinese knowledge base like DBpedia, and only few Chinese data sets are linked to DBpedia. Our approach also establishes instance links to DBpedia. This lays the foundation for cross-lingual structured knowledge base sharing by integrating structured knowledge bases across existing wiki encyclopedias in different languages.

7 Conclusions and future work

This paper presents an approach for building Chinese knowledge bases from wiki resources. Our approach extracts an upper-level Chinese ontology based on the category system and infobox schema, and then extracts instances from wiki articles and establishes their links to DBpedia. Using the proposed approach, we build knowledge bases from Hudong and Baidu Baike, respectively. Our knowledge bases can be used as useful Chinese semantic resources for knowledge based applications. Currently, both RDF dump and SPARQL endpoint are provided to access the extracted knowledge bases.

As for 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 instance extraction. Currently, we build knowledge bases from Hudong and Baidu Baike. Another potential research is to integrate the information in different wiki resources and build a larger knowledge base covering much more instances.

References

Auer, S.R., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. DBpedia: a Nucleus for a Web of Open Data. Proc. 6th Int. Semantic Web Conf. and 2nd Asian Semantic Web Conf., p.722-735.

Berners-Lee, T., 1998. Semantic Web Road Map. Available from http://www.w3.org/DesignIssues/Semantic.html

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S.R., Becker, C., Cyganiak, R., Hellmann, S., 2009a. DBpedia—a crystallization point for the Web of data. *Web Semant.*, **7**(3):154-165. [doi:10.1016/j.websem.2009.07.002]

Bizer, C., Heath, T., Berners-Lee, T., 2009b. Linked data—the story so far. *Int. J. Semant. Web Inform. Syst.*, **5**(3):1-22. [doi:10.4018/jswis.2009081901]

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a Collaboratively Created Graph Database for Structuring Human Knowledge. Proc. ACM SIGMOD Int. Conf. on Management of Data, p.1247-1250. [doi:10.1145/1376616.1376746]

Buitelaar, P., Cimiano, P., 2008. Ontology Learning and Population: Bridging the Gap Between Text and Knowledge. Frontiers in Artificial Intelligence and Applications, **167**:45-69.

Buitelaar, P., Cimiano, P., Magnini, B., 2005. Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam.

Euzenat, J., Shvaiko, P., 2007. Ontology Matching. Springer-Verlag, Heidelberg (DE).

Fellbaum, C., 1998. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA.

García-Silva, A., Szomszor, Y., Alani, M.Y., Corcho, Ó.H.Y., 2009. Preliminary Results in Tag Disambiguation Using DBpedia. 1st Int. Workshop Collective Knowledge Capturing and Representation, p.41-44.

Heath, T., Bizer, C., 2011. Linked data: evolving the Web into a global data space. *Synth. Lect. Semant. Web Theory Technol.*, **1**(1):1-136. [doi:10.2200/S00334ED1V01Y201102WBE001]

Kasneci, G., Ramanath, M., Suchanek, F., Weikum, G., 2008. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Rec.*, **37**(4):41-47. [doi:10.1145/1519103.1519110]

Lenat, D.B., 1995. CYC: a large-scale investment in knowledge infrastructure. *ACM Commun.*, **38**(11):33-38. [doi:10.1145/219717.219745]

Maedche, A., Staab, S., 2001. Ontology learning for the Semantic Web. *IEEE Intell. Syst.*, **16**(2):72-79. [doi:10.1109/5254.920602]

Matuszek, C., Cabral, J., Witbrock, M., Deoliveira, J., 2006. An Introduction to the Syntax and Content of Cyc. AAAI Spring Symp., p.44-49.

Melo, G.D., Weikum, G., 2010. MENTA: Inducing Multilingual Taxonomies from Wikipedia. Proc. 19th ACM Int. Conf. on Information and Knowledge Management, p.1099-1108.

Navigli, R., Velardi, P., 2004. Learning domain ontologies from document warehouses and dedicated Web sites. *Comput. Ling.*, **30**(2):151-179. [doi:10.1162/089120104323093276]

Navigli, R., Velardi, P., Gangemi, A., 2003. Ontology learning and its application to automated terminology translation. *IEEE Intell. Syst. Their Appl.*, **18**(1):22-31. [doi:10.1109/MIS.2003.1179190]

Niles, I., Pease, A., 2001. Towards a Standard Upper Ontology. Proc. Int. Conf. on Formal Ontology in Information Systems, p.2-9.

Niu, X., Sun, X., Wang, H., Rong, S., Qi, G., Yu, Y., 2011. Zhishi.me—Weaving Chinese Linking Open Data. Proc. 10th Int. Conf. on the Semantic Web, **2**:205-220.

Passant, A., 2010. dbrec—Music Recommendations Using DBpedia. Proc. 9th Int. Semantic Web Conf., **2**:209-224.

Pease, A., Niles, I., 2002. IEEE standard upper ontology: a progress report. *Knowl. Eng. Rev.*, **17**(1):65-70. [doi:10.1017/S0269888902000395]

Piek, V., 1997. EuroWordNet: a Multilingual Database for Information Retrieval. Proc. Delos Workshop on Cross-Language Information Retrieval, p.5-7.

Ponzetto, S.P., Strube, M., 2007. Deriving a Large Scale Taxonomy from Wikipedia. Proc. 22nd National Conf. on Artificial Intelligence, **2**:1440-1445.

Shadbolt, N., Berners-Lee, T., Hall, W., 2006. The Semantic Web revisited. *IEEE Intell. Syst. Their Appl.*, **21**(3):96-101. [doi:10.1109/MIS.2006.62]

Suchanek, F.M., Kasneci, G., Weikum, G., 2007. Yago: a Core of Semantic Knowledge. Proc. 16th Int. Conf. on World Wide Web, p.697-706. [doi:10.1145/1242572.1242667]

Suchanek, F.M., Kasneci, G., Weikum, G., 2008. YAGO: a large ontology from Wikipedia and WordNet. *Web Semant.*, **6**(3):203-217. [doi:10.1016/j.websem.2008.06.001]

Vossen, P., 1998. Introduction to EuroWordNet. *Comput. Human.*, **32**(2/3):73-89. [doi:10.1023/A:1001175424222]

Wu, F., Weld, D.S., 2007. Autonomously Semantifying Wikipedia. Proc. 16th ACM Conf. on Information and Knowledge Management, p.41-50.

Wu, F., Weld, D.S., 2008. Automatically Refining the Wikipedia Infobox Ontology. Proc. 17th Int. Conf. on World Wide Web, p.635-644. [doi:10.1145/1367497.1367583]

Recommended reading

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., 2009. DBpedia—a crystallization point for the Web of data. *Web Semant.*, **7**(3):154-165. [doi:10.1016/j.websem.2009.07.002]

Suchanek, F.M., Kasneci, G., Weikum, G., 2008. YAGO: a large ontology from Wikipedia and WordNet. *Web Semant.*, **6**(3):203-217. [doi:10.1016/j.websem.2008.06.001]

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data—the story so far. *Int. J. Semant. Web Inf. Syst.*, **5**(3):1-22. [doi:10.4018/jswis.2009081901]

Wu, F., Weld, D.S., 2008. Automatically Refining the Wikipedia Infobox Ontology. Proc. 17th Int. Conf. on World Wide Web, p.635-644. [doi:10.1145/1367497.1367583]

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., Metakides, G., 2012. Internationalization of Linked Data: the case of the Greek DBpedia edition. *Web Semant.*, online 18 January 2012.