Building up a Large Ontology from Wikipedia Japan with Infobox and Category Tree

Shinya Sakurai[1], Takuya Tejima[1] , Takeshi Morita[1],
Noriaki Izumi[2], and Takahira Yamaguchi[1]

[1] Keio University, 3-14-1 Hiyoshi, Kohokuku, Yokohama-shi, 223-8522 Japan
{s_saku, t_tejima, t_morita, yamaguti}@ae.keio.ac.jp
[2] National Institute of AIST, 1-16-13 Sotokanda, Chiyoda-ku, Tokyo 101-0021 Japan
n.izumi@aist.go.jp

Abstract. In this paper, we propose new methods for building up a large ontology from Wikipedia Japan with infobox and category tree. We call it Wikipedia Ontology. We applied several extraction techniques on Wikipedia Japan, and obtained the six types of data.

1 Introduction

Constructing general ontologies such as WordNet and EDR[1] took a lot of time and cost by hand. Therefore it is difficult to maintain the quality of vocabulary for up-to-date word, named entity etc. On the other hand, semi-structured information resources such as Wikipedia are being created on the present WWW with the emergence of Web2.0. Some studies focus on constructing large scale ontologies and their instances from Wikipedia using information resources such as infobox, external links, categories the article belongs to, and so on ([2-3]).

In this paper, we propose new methods for building up a large ontology from Wikipedia Japan with infobox and category tree, and we call it “Wikipedia Ontology”. We applied several extraction techniques on Wikipedia Japan, and obtained the following six types of data: word linkage, synonym, class-instance relationship, is-a relationship, infobox triple, and property and its domain. Especially, the novelty of our methods are extracting is-a relationship and domain of property using infobox template and category tree.

2 Building up Wikipedia Ontology

To build up Wikipedia Ontology, we applied several techniques on Wikipedia Japan as of May 2, 2008, and extracted the following six types of data.

We extracted word linkage using Wikipedia mining mentioned by Nakayama et al. [4]. We extracted synonym from 292,036 redirect links exist in Wikipedia. We

extracted class-instance relationships by scraping the 8,300 listing pages of Wikipedia in which various things names are enumerated.

We propose two methods to extract is-a relationships from Wikipedia. One of them uses two kinds of string matching (backward string matching and forward matched string eliminating) on the category names and the other uses infobox template and category tree. Since sub categories are made to classify articles newly made, they tend to be named in the form of compound term including superior categories. We focus on the relation among the infobox templates, categories to which the articles which have an infobox belong (infobox's categories), and category tree. We extracted is-a relationships by matching infobox templates name to category names. Figure 4 shows the overview, and the procedure of this method is shown following (1)－(4). (1) Extract the infobox templates, infobox's categories, and category tree from Wikipedia dump data. (2) Match infobox templates name to category names in the category tree. (3) Extract sub-categories of the matched categories at (2) and trim the categories other than infobox's categories from the sub-categories. (4) Define is-a relationships between the extracted categories at (2) and (3). (The extracted categories at (2) as super classes and the extracted categories at (3) as sub classes.) The structure “article name– attribute – value” in infobox is regarded as RDF triple. We extracted the structure as infobox triple. We extracted the attributes of infobox templates as properties and the infobox templates' name as the domain of the properties.

Table 1 shows the scale of the constructed Wikipedia Ontology.

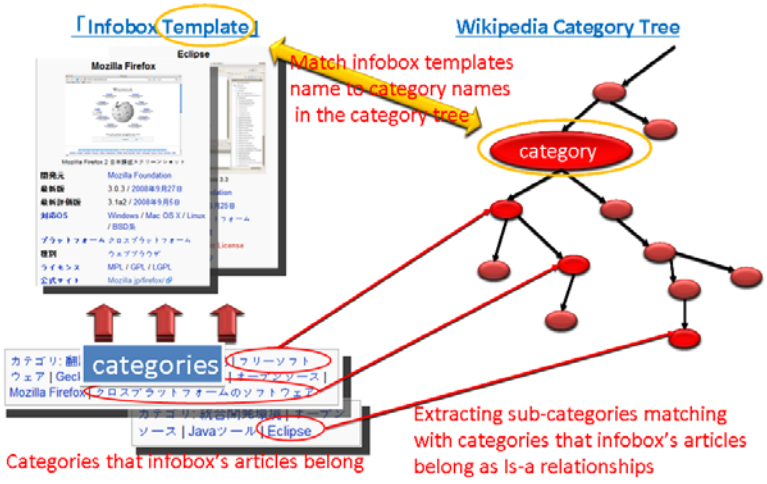

Fig. 1 Matching infobox template name to category name

Table 1. The scale of the constructed Wikipedia Ontology

# Word Linkage	1,030,444	**# Class-Instance Relationship**	331,535
# Synonym	292,036	**# Infobox Triple**	511,146
# Is-a Relationship	9,970	**# Domain of Property**	9,644

Additionally, we implemented a search application which enables to access the Wikipedia Ontology database. When a user inputs a class, the application outputs the statements about instances and is-a relationships of the class. Figure 2 shows the screenshot of the application.

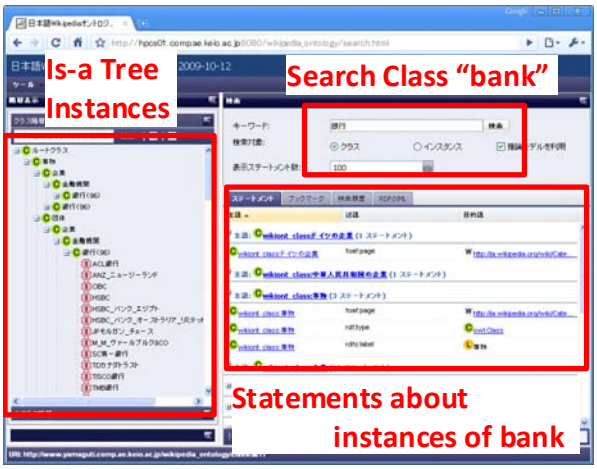

Fig. 2 Search application of Wikipedia Ontology

3 Future Work

As for future work, we should integrate existing upper ontology into Wikipedia Ontology to improve the quality of is-a relationships. Furthermore, we are considering linking up extracted instances as Linked Data to the database of DBpedia[3] and GeoNames etc.

References

1. National Institute of Information and Communications Technology: (EDR Electronic Dictionary Technical Guide) http://www2.nict.go.jp/r/r312/EDR/index.html
2. F. M. Suchanek, G. Kasneci, and G. Weikum.: Yago: A Core of Semantic Knowledge. In: Proc. of the 16th Int. Conference on WWW, ACM (2007) 697-706
3. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives.: DBpedia: a Nucleus for a Web of Open Data. In: Proc. of the ISWC+ASWC 2007, LNCS 4825 (2007) 722-735
4. Nakayama, K., Hara, T. and Nishio, S.: Wikipedia Mining for an Association Web Thesaurus Construction. In Proc. of the Int. Conference on Web Information Systems Engineering (WISE) (2007) 322-334