Wikipedia Category Ontology: A Framework for Utilization of the Wikipedia Category Structure by Knowledge Engineers ⋆

Masaharu Yoshioka[1,2,3,4] and Takanori Nakagawa[1⋆⋆⋆]

[1] Graduate School of Information Science and Technology, Hokkaido University N-14 W-9, Kita-ku, Sapporo 060-0814, Japan

[2] Faculty of Information Science and Technology, Hokkaido University

[3] Global Station for Big Date and Cybersecurity, Global Institution for Collaborative Research and Education, Hokkaido University

[4] Center for Advance Intelligence Project, RIKEN Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi,Chuo-ku, Tokyo 103-0027, Japan

Abstract. Wikipedia categories are intended to group together pages on similar topics and are organized in a hierarchical structure. Since the editorial policy for Wikipedia categories differs from policies used by knowledge engineers, various types of relationship exist in its category structure. In this paper, we propose a novel framework called "Wikipedia category ontology" (WCO) that aims to act as a basis for interpreting the Wikipedia category structure and is based on a classification of category and relationship types. WCO enables particular Wikipedia category substructures to be extracted, including class–subclass hierarchies, class–instance, and sets of diffused categories for a particular category. WCO is available online in the form of linked open data at `http://wcontology.org/`.

1 Introduction

Wikipedia[5] is a free online encyclopedia that anyone can edit and that contains a huge number of articles. A characteristic of Wikipedia is that its articles are organized in a semistructured format and researchers have aimed to extract structural information that is suited to the construction of knowledge resources. Examples include DBpedia [1] and YAGO2 [2].

However, information about the Wikipedia category structure has not been used well. One approach aimed to extract particular types of information using pattern-based rules [3, 4]. Another example is YAGO2 that uses leaf category information to estimate the class of the pages. This approach does not consider Wikipedia editorial policy; thus, only unsystematic aspects of the information about the Wikipedia category structure are utilized.

⋆⋆⋆ He is now working at KDDI corporation, but this work was conducted during his study at the Hokkaido University

[5] `http://www.wikipedia.org/` : All pages accessed on May 13, 2020.

We have been working on a project that aims to analyze Wikipedia category structures based on the definitions of Wikipedia categories in Wikipedia itself [5]. In this research, we classify Wikipedia categories using *set* (representing classes such as "Cities"), *topic* (representing instances such as "Japan"), and *set-and-topic* (a combination of *set* and *topic* such as "Cities in Japan").

In this paper, we propose a "Wikipedia category ontology" (WCO) based on our analysis of Wikipedia category types and our exhaustive analysis of Japanese Wikipedia categories. This ontology aims to act as a basis for interpreting the Wikipedia category structure when reorganizing (extracting) the Wikipedia category structure for a particular purpose. In addition, because one of the reorganization results represents class hierarchy information, WCO can also be used as a class-hierarchy component of the Wikipedia ontology. This ontology is available online in the form of linked open data (LOD) at `http://wcontology.org/`.

The main contributions of this paper can be summarized as: (1) analyzing the Japanese Wikipedia category structure to understand its characteristics, (2) providing a basic vocabulary for representing the Wikipedia category structure, and (3) providing LOD material that enables the extraction of useful information from the Japanese Wikipedia category structure.

2 WCO

2.1 Editorial Policy for Wikipedia Categories

Because Wikipedia editors edit the category structure based on the editorial policy described in Wikipedia itself, it is important to understand that policy. In Wikipedia, the category structure is organized as overlapping "trees" using subcategory relationships[6]. There are two main types of category. The *topic* category refers to an entity (e.g., "Japan") and the *set* category refers to a class (e.g., "Cities"). Sometimes, for convenience, the two types are combined to create a *set-and-topic* category (e.g., "Cities in Japan")[7]. Figure 1 shows an example of category names where the brown and green colors indicate names for set and topic categories, respectively. Names with both colors are set-and-topic categories.

Another important editing policy relates to the size of the categories. In Wikipedia, a large category will often be broken down ("diffused") into smaller, more-specific subcategories. For example, "Rivers of Europe" is broken down by country using subcategory "Rivers of Europe by country" and its subcategories such as "Rivers of Albania" and "Rivers of Austria." Most of the case, such big categories are divided into subcategories using constraints for selecting a part of pages that satsify such constraints. (e.g., instances of "country" (Topic category such as "Albania", "Austria") for "river", instances of "artist" (Topic category such as "The Beatles") for "song"). Those categories are typical examples of *set-and-topic* categories.

As a result, it is necessary to traverse link to find out appropriate broken down categories for making a list of pages categorized for such large category.

[6] `https://en.wikipedia.org/wiki/Help:Categories`

[7] `https://en.wikipedia.org/wiki/Wikipedia:Categorization`

This is different from collecting all progeny categories. For example, “Rivers of Austria” have subcategory such as “Danube” and “Drava” and those categories have “Populated places on the Danube” and “Bridges over the Drava” whose pages are not appropriate for “Rivers of Europe” category.

2.2 Analysis of the Japanese Wikipedia Category Structure

Most previous work [3, 4] aimed to extract category information using patterns without considering other issues. This lack of consideration leads to an inadequate understanding of the entirety of the Wikipedia structure; therefore, we conducted an exhaustive manual analysis of the categories in the Japanese Wikipedia as an extension of our previous work [5].

We created a database dump of the Japanese Wikipedia on October 20, 2017, which comprised 212,346 categories. Because some categories referred only to Wikipedia maintenance issues, we excluded these categories, which left 183,600 categories (and 451,074 parent–child category relationships between them) for our exhaustive analysis.

One of our coauthors then exhaustively checked and classified all of these Wikipedia categories, initially in terms of the three category types *set*, *topic*, and *set-and-topic*. However, because there were categories based on a combination of topics (e.g., “1990s in Japan”), we added this as a fourth type of category and classified the categories into the following four types (numbers in parentheses indicate the number of categories for that type).

Set category (10,748) indicates a class (usually in the plural).

Topic category (44,525) indicates a topic (usually sharing its name with a Wikipedia article on that topic).

Constrained set category (117,994) is a diffused version of a *set* category, with constraints.

Constrained topic category (10,333) is a diffused version of a *topic* category, with constraints.

The most numerous type is *constrained set* and a category of this type involves diffusion from an ancestor-set category. Therefore, it is necessary to provide a framework for analyzing such diffusion-related categories if the Wikipedia category structure is to be properly understood.

It is also necessary to classify subcategory relationships among the Wikipedia categories. One of the important issue for this classification is the role of transitivity in the relationships.

Fig. 1 shows the types and examples of category relationships. The red lines indicate transitive relationships. In this case, “Cities” is not interpreted as the ancestor category of “People from Mitaka, Tokyo.” “Geographically part of” and “Age” are special cases of a “Specified constraint” when used as constraints. “Narrower” and “Narrower transitive” are used for intransitive and transitive category relationships respectively that are difficult to categorize using other types.

3 WCO Resources for the Japanese Wikipedia

Based on our analysis of the categories in the Japanese Wikipedia corpus, it will be necessary to reorganize the Wikipedia category structure for those knowledge

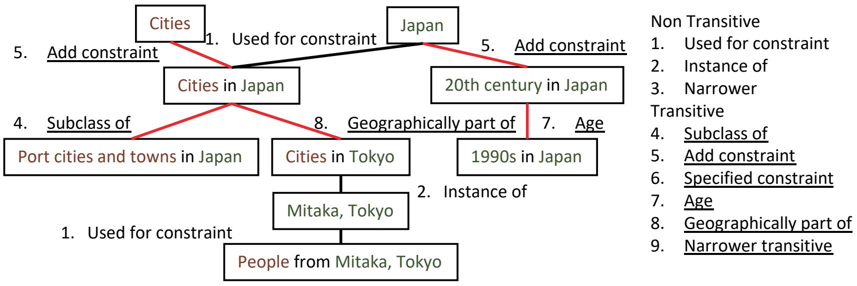

Fig. 1. Types of category relationships

engineers who would like to extract knowledge by making use of the Wikipedia structure. To address this problem, we propose WCO, which provides a reorganization of the Wikipedia category structure by redefining the types of categories and the relationships between them.

Fig. 2 gives definitions of the core vocabulary used in WCO using the resource description framework (RDF) for Linked Open Data(LOD)[6].

```
@prefix wcoc: <http://wcontology.org/core#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos:<http://www.w3.org/2004/02/skos/core#>
wcoc:setCategory rdfs:subClassOf wcoc:category .
wcoc:topicCategory rdfs:subClassOf wcoc:category .
wcoc:constrainedSetCategory rdfs:subClassOf wcoc:setCategory .
wcoc:constrainedTopicCategory rdfs:subClassOf wcoc:topicCategory .
wcoc:narrower rdfs:subPropertyOf skos:narrower .
wcoc:instanceOf rdfs:subPropertyOf wcoc:narrower .
wcoc:usedForConstraint rdfs:subPropertyOf wcoc:narrower .
wcoc:narrowerTransitive rdfs:subPropertyOf wcoc:narrower .
wcoc:narrowerTransitive rdfs:subPropertyOf skos:narrowerTransitive .
wcoc:subclassOf rdfs:subPropertyOf wcoc:narrowerTransitive .
wcoc:age rdfs:subPropertyOf wcoc:narrowerTransitive .
wcoc:geography rdfs:subPropertyOf wcoc:narrowerTransitive .
wcoc:addConstraint rdfs:subPropertyOf wcoc:narrowerTransitive .
wcoc:specifiedConstraint rdfs:subPropertyOf wcoc:narrowerTransitive .
```

Fig. 2. Definition of the core vocabulary used by WCO

Based on the WCO core vocabulary and the results of our analysis results, we construct resources for representing the Japanese Wikipedia category structure. These resources are connected to the English Wikipedia category structure via the language links in Wikipedia and DBpedia that use `owl:sameAs` (prefix `owl` is used for `<http://www.w3.org/2002/07/owl#>`). These resources are accessible in LOD form (http://wcontology.org/) with SPARQL endpoint using the Virtuoso Open Source Edition 7.2.6[8].

We can utilize these resources using the SPARQL endpoint `http://wcontology.org/sparql`. Several example queries are shown on the

[8] `https://github.com/openlink/virtuoso-opensource`

WCO home page, `http://wcontology.org/`. The example queries are given in both English and Japanese, where the original example queries were taken from the Japanese version and the English-version queries were constructed using an `owl:sameAs` link to an English-language Wikipedia category. There are fewer English-language queries than Japanese queries because there are categories without a language link to the English-language Wikipedia.

The following are examples of queries using WCO.

- Collection of Diffused Categories
  By selecting transitive relationships, we can select a set of categories that have been diffused from the target category.
- Collection of Subclasses (setCategory) of a Given Category
  By checking the transitive subcategories, we can extract the subclasses (setCategory) of a given category.

4 Conclusion

In this paper, we have proposed WCO, a framework that aims to act as a basis for interpreting the Wikipedia category structure by enabling a classification of its category types and the types of relationship between them. Resources and examples are available as LOD `http://wcontology.org/`.

For the future works, we plan to utilize this WCO resource for Japanese Wikipedia as a training data to construct ones for other languages.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 18H03338.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web **7** (2009) 154 – 165
2. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194** (2013) 28 – 61
3. Tamagawa, S., Sakurai, S., Tejima, T., Morita, T., Izumi, N., Yamaguchi, T.: Learning a large scale of ontology from japanese wikipedia. Journal of Japanese Society of Artificial Intelligence **25** (2010) 623–636 (in Japanese).
4. Heist, N., Paulheim, H.: Uncovering the semantics of wikipedia categories. In Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F., eds.: The Semantic Web – ISWC 2019, Cham, Springer International Publishing (2019) 219–236
5. Yoshioka, M.: Analysis of japanese wikipedia category for constructing wikipedia ontology and semantic similarity measure. In: Information Retrieval Technology 10th Asia Infomation Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014 Proceedings. Springer-Verlag GmbH (2014) 470–481 LNCS8870.
6. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems **5** (2009) 1–22