Bridging Community Resource Gateways by Linking Community Taxonomies

Wonsook Lee, Mitsuharu Nagamori, and Shigeo Sugimoto

University of Tsukuba, Tsukuba 1-2, Ibaraki, Japan
{wonsook,nagamori,sugimoto}@slis.tsukuba.ac.jp

Abstract. Many communities provide Web resource directories to help users find useful resources in the community. A typical example is a resource directory in a homepage of a local government. Crosswalk of the directories of neighboring communities is a crucial function for users to collect useful resources from the communities. However, an appropriate scheme bridging the community directories is required. This paper proposes a few mapping schemes to connect community directories and compares them by applying them to the resource directories of three local governments - Tokyo and Hokkaido in Japan and Chungcheongnam-do in Korea. The mapping schemes use National Diet Library Subject Heading (NDLSH) and/or Nippon Decimal Classification (NDC) as a switching language. Evaluation of the proposed schemes shows their advantages and limitations.

Keywords: Community Taxonomy, Crosswalk of Web Directories, Switching Language, Subject Headings, Knowledge Organization.

1 Introduction

There are many Web sites which provide a directory of useful Web resources in a specific domain for a specific community. Each of the directories is useful for the community to find valuable resources in their domain. A nice example is a resource directory of a local government - prefectures and cities – which provides a categorized list of useful resources for their community members. A crosswalk of resource directories of neighboring communities is a crucial function to collect resources from the communities. Those community resource directories are usually not huge. Their organizing taxonomies are much smaller than those of comprehensive resource directories – some tens to hundreds of terms. It is not straightforward, however, to crosswalk the resource directories of neighboring communities because of the difference of their taxonomies. In this paper, we propose a few schemes for mapping between the community resource directories and compare them.

2 Related Works

Jens-Erik proposed to use Dewey Decimal Classification (DDC) as a switching language among several taxonomies [1]. There are a few collaborative projects to merge resource directories in specific domains, e.g. Renardus project and Resource Discovery Network [2][3]. These projects use comprehensive conventional classification schemes

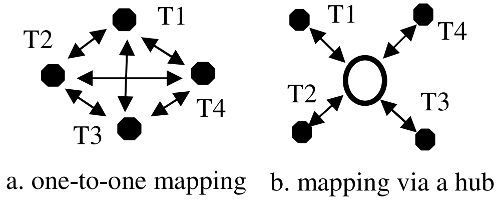
D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 482–486, 2007.

to merge the subject directories. Community resource directories have different features from comprehensive directories. In our previous work, we examined the characteristics of subject vocabularies of a community oriented directory of Web resources provided by Okayama Prefecture Library in Japan [4]. The library uses three taxonomies to classify resources – NDC, a taxonomy for children, and a taxonomy for the resources provided by the prefecture government. NDC is a comprehensive classification scheme and a major standard for Japanese libraries. On the other hand, both of the two local taxonomies have 300 to 400 categories. They each have a quite rich set of terms in specific subjects in accordance with their purpose.

3 Connecting Community Taxonomies

3.1 An Underlying Model

As shown in Fig. 1, there are two general schemes to connect taxonomies – one-to-one mapping and mapping via a hub taxonomy called *switching language*. It is obvious that mapping via a switching language is advantageous in terms of complexity. The mapping schemes proposed in the next section are based on this switching language model. The vocabularies used for the switching language in this paper are NDC and NDLSH. NDLSH is one of the authoritative subject headings in Japan and is maintained by the National Diet Library of Japan. These two authoritative vocabularies are primarily designed for Japanese resources but the sections used in this research are mostly not specific to Japanese contents.



a. one-to-one mapping b. mapping via a hub

Fig. 1. Mapping between Taxonomies

3.2 Community Taxonomy Mapping Models

The four mapping schemes (MS1-4) shown below are evaluated for community taxonomy mapping in this paper. Each MS is explained as a procedure to obtain a mapping function. The taxonomies have a hierarchical structure of category terms.

MS1: All resources are assigned one or more NDLSH terms in advance.

1. For each leaf category of taxonomy A and B, create a test set of resources by randomly choosing resources classified into the category.
2. Create a NDLSH term list for every category by gathering NDLSH terms assigned to the resources of the test set of the category.
3. For every leaf category pair of A and B, create an intersection of their NDLSH term sets. Neglect the pairs whose intersection is empty.
4. Calculate an inter-category connectivity value of all of the leaf category pairs using (1) and (2).
5. Choose category pairs which have inter-category connectivity values above a heuristic threshold value. The set of category pairs obtained is a mapping function between A and B.

A connectivity value CV_{cn} from a category term c to an NDLSH term n and an inter-category connectivity value ICV_{th} between category t and h connected via an NDLSH term n are defined as follows.

$$CVcn = Rcn / Rn \tag{1}$$

$$ICVth = \sum CVtn \times CVhn \tag{2}$$

where R_{cn} is the number of resources classified in category c and assigned term n, and R_n is the number of all resources classified in category c.

MS2: By definition, every NDLSH term is associated with one or more NDC terms. MS2 uses the NDC in addition to the NDLSH terms.

1. Same operations as Step 1 and 2 of MS1.
2. For each leaf category of A and B, create a frequency list of NDC terms by accumulating every NDC term that is assigned to the NDLSH terms of the resources included in the test set of the category.
3. For each leaf category of A and B, find the most frequently used NDC term(s) and call it (or them) the surrogate(s) of the category.
4. For every category of A, find one or more categories of B which have the same surrogate and make a category pair. The set of category pairs obtained is the mapping function from A to B.

MS3: In MS3, all resources are assigned one or more NDC terms (up to five).

1. Find the most frequently used NDC term(s) for each category of A and B. Use the most frequently assigned NDC term(s) as the surrogate of the category.
2. Find source- and target-category pairs which have the identical NDC term(s) as their surrogate. The set of pairs created is the mapping function from A to B.

MS4: MS4 uses a direct mapping from a category to NDLSH terms.

1. Assign one or more NDLSH terms to each category of A and B. If a category is a compound term, split the term into simple terms and assign an NDLSH term(s) to each of the simple terms; for example, split “School Education and Life Long Education” into “School Education” and “Life Long Education”.
2. Create a set of category pairs by coupling an A category and a B category that are assigned to at least one identical NDLSH term. The set of pairs created is the mapping function between A and B.

4 Experiments – Mapping Directories Between Local Governments

We applied all mapping schemes to resource directories of Tokyo, Hokkaido in Japan and Chungcheongnam-do in Korea. The directory of Tokyo has two layers and 56 leaf categories. The Hokkaido directory has two layers and 47 leaf categories. The Chungcheongnam-do directory has three layers and 118 leaf categories.

Table 1. Precision and Recall values of Community Taxonomy Mappings

	MS1		MS2		MS3		MS4	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
T to H	0.60	0.49	0.33	0.38	0.48	0.49	0.73	0.61
H to T	0.40	0.34	0.28	0.37	0.35	0.48	0.71	0.78
C to T	0.57	0.45	0.18	0.34	0.30	0.48	0.57	0.59
C to H	0.28	0.57	0.23	0.45	0.17	0.41	0.50	0.73

Precision =(number of correct mappings by MSx) / (number of all mappings by MSx)
Recall=(number of correct mappings by MSx) / (number of manual mappings)
T: Tokyo, H: Hokkaido, C: Chungcheongnam-do

In this experiment, we selected up to five resources from each category in MS1, MS2 and MS3. In the case that the number of resources under a category was smaller than five, we used all of the resources. We applied all three mapping schemes to the pairs of the directories. Assignments of NDLSH terms to the resources and the categories were manually done. We used 193, 218 and 149 resources of Tokyo, Hokkaido and Chungcheongnam-do, respectively. Table 1 summarizes the recall and precision values. This evaluation is performed by comparing the mapping functions created by the four mapping schemes with the answer set that were created by manual mapping of the categories between the taxonomies.

The recall and precision scores of MS4 are generally better than others. Among the four schemes, MS4 is the only scheme which directly maps a category term to another. These scores are calculated based on category-to-category mappings. Mapping schemes other than MS4 use subject terms of resources to create the mappings. Therefore, we consider that those schemes could be extended to mapping by sub-leaf-category which is implicitly formed by a group of resources or to mapping by single resource. These issues are left for our future work.

5 Conclusion

Community people know better and more deeply about community resources than Google. Crosswalking community resource directories is an essential function for end-users trying to find valuable resources. Taxonomies used in the community gateways are semi-controlled but not designed for interoperability with neighboring communities. Because of the nature of the human-created taxonomies, sharing the categories of community gateways is hard even when the taxonomies are not large. The mapping schemes and evaluation results discussed in this paper give us useful clues about how to create a crosswalk function for community gateways.

References

1. Mai, J.-E.: The Future of General Classification. Cataloging & Classification Quarterly 37 (1/2), 3–12 (2003)
2. Renardus, http://renardus.lub.lu.se/

3. Intute, http://www.intute.ac.uk/
4. Lee, W., Sugimoto, S.: Toward Core Subject Vocabularies for Community-oriented Subject Gateways. In: DC-2005: International Conference on Dublin Core and Metadata Applications, pp.15–24, Madrid (2005)