Federated Searching System for Humanities Databases Using Automatic Metadata Mapping

Fuminori Kimura
College of Information Science and Engineering, Ritsumeikan University, Japan
fkimura@is.ritsumei.ac.jp

Takushi Toba
Production Bureau, The Yomiuri Shimbun, Japan
toba1621@yomiuri.com

Taro Tezuka
College of Information Science and Engineering, Ritsumeikan University, Japan
tezuka@media.ritsumei.ac.jp

Akira Maeda
College of Information Science and Engineering, Ritsumeikan University, Japan
amaeda@media.ritsumei.ac.jp

Keywords: federated searching system, metadata mapping.

Abstract

Recently, many collections and resources in libraries, museums and research institutes are digitized and opened to the public. As there are many humanities databases and each database has its own user-interface and metadata schema, it is not easy for users to find desired information. Users must input the same query in different ways for each database in order to access all related databases (Chang et al., 1999).

Our goal is to construct a federated searching system for humanities databases. A federated searching system refers to the retrieval system that a user can access multiple humanities databases with only one query input. In order to realize a federated searching system for existing heterogeneous databases, a method for metadata mapping of attribute names is needed to cope with the differences in schema for each database. Therefore, we propose a method of automatic metadata mapping using metadata elements that are revised from Dublin Core Metadata Element Set (DCMI, 2008).

Our proposed method consists of two preprocessing phases and four mapping phases. Figure 1 shows the flow of metadata mapping of our proposed method.

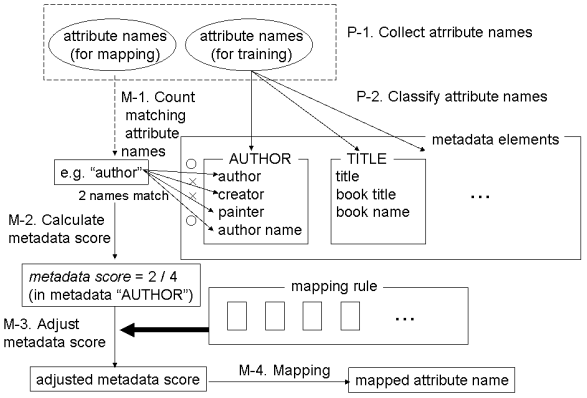

FIG. 1. Flow of metadata mapping.

The procedure of preprocessing is as follows:

P-1. Collect attribute names from humanities databases for training and mapping.

P-2. Classify attribute names for training into appropriate metadata elements manually.

The procedure of mapping phase is as follows:

https://doi.org/10.23106/dcmi.952109699

M-1. Count the number of partial string matches between the attribute name for mapping and each metadata element.

M-2. Calculate the *metadata score* of each metadata element. Metadata score is calculated by the number of partial string matches per total number of attribute names in the metadata element.

M-3. Adjust the metadata score for each metadata if the target attribute name matches one or more of the mapping rules consisting of partial string match with keywords relevant to metadata elements. (e.g. if the attribute name includes "year", increase the metadata score for "TEMPORAL")

M-4. Map the target attribute name into a metadata element that has the highest metadata score. If the attribute name is given the metadata score value 0 for all metadata set, the attribute name is classified into "OTHER" metadata.

In our method, we use a revised DCMES (Dublin Core Metadata Element Set) instead of DCMES. In the revised DCMES, some metadata elements of DCMES are unified and some are divided. For example, DCMES "COVERAGE" is divided into "TEMPORAL" and "SPATIAL" in revised DCMES. Revised DCMES consists of 8 metadata elements ("TITLE", "SUBJECT", "AUTHOR", "PUBLISHER", "IDENTIFIER", "TEMPORAL", "SPATIAL", "OTHER").

We collected 334 attribute names in Japanese from 50 humanities databases, and conducted metadata mapping. We conducted experiments for three cases, using standard DCMES without the mapping rules, using standard DCMES with the mapping rules, and revised DCMES with the mapping rules. Judgments for the mapping results were conducted manually.

Table 1 shows the results of these three experiments. In using standard DCMES with the mapping rules, the precision is improved by 15.9% over using standard metadata without the mapping rules. This result shows that mapping rules contribute to the improvement of the metadata mapping. In the case of using revised DCMES with the mapping rules, the precision is improved by 14.9% over using the standard DCMES with mapping rules. This result shows that the revised DCMES contributes considerably to the improvement of the metadata mapping.

TABLE 1: Precision of metadata mapping.

Metadata	Rule	Average precision (%)
Standard Dublin Core Metadata Element Set	No rules	73.8
Standard Dublin Core Metadata Element Set	Mapping rules	79.0
Revised Dublin Core Metadata Element Set (8 elements)	Mapping rules	94.9

We proposed an automatic metadata mapping method for the federated searching system for humanities databases. We used eight metadata elements that are revised from DCMES. In this paper, we collected attribute names from various humanities databases in Japanese, and conducted the metadata mapping using a metadata score and mapping rules. The result of mapping experiments shows that our proposed method achieved sufficient mapping precision. In our future work, we need to cope with the ambiguity of notation.

References

Chang Chen-Chuan K., Hector Garcia-Molina (1999). Mind your vocabulary: query mapping across heterogeneous information sources, Proceedings of the 1999 ACM SIGMOD international conference on Management of data, p.335-346.

DCMI. (2008). Dublin Core Metadata Element Set, version 1.1: Reference description. Retrieved January 14, 2008, from http://www.dublincore.org/documents/dces/.

https://doi.org/10.23106/dcmi.952109699