Research on Interoperability of Metadata in Classification Schemes —construction of automatic mapping system between CLC and DDC

Jianbo Dai Hanqing Hou Ling Cao

Dept. of Libr. & Inform. Sci., Nanjing Agri. Univ., Nanjing, China 210095

Mail: hqhou@njau.edu.cn

Abstract: The interoperability of classification schemes is an urgent need for information organization & retrieval. Now, it is often to recognize and build up the mapping relationship between classes by hand in EU project Renardus and other projects. Each class of CLC or DDC can be looked as a concept, and the integrity of concept can be divided into several concept elements expressed by words. We can computer the word similarity of concept elements to reflect the similarity of the integrity of class concept on the hypothesis that the integrities are similar if all the parts of them are similar. In the automatic mapping system between CLC4 and DDC21, the concept of class can be divided into concept elements expressed by class headings, note terms, lower-class terms and upper-class terms, then rank these matched concept elements of the classes between CLC and DDC according to the principle of maximum similarity. We formulated detailed mapping regulation according to three parameters: word similarity value of concept elements, margin of concept elements and kind of equivalent concept elements of class. The article introduced the structure, functions and use of the automatic mapping system.

Keywords: Library classification; *Chinese Library Classification; Dewey Decimal Classification;* Interoperability; Automatic mapping system; Word similarity.

1 Introduction

Library classifications either applied in the traditional organization of information resources or the web information resources have never been unified, in fact they cannot be unified since there are various types in the information organization field.

Diversity of library classification has been the barrier in the cross-database or cross-domain retrieval. The metadata interoperability between different classifications has been a serious problem waiting for being resolved quickly.

It is necessary to realize the metadata interoperability in library classification for resources share. To construct the mapping systems among several main classifications schemes in the world, thus some barriers exiting in such as browsing, indexing or retrieval can be avoided.

If the metadata interoperability in classification schemes can be realized, users browse and search the information organized by another type of library classification and browse portal web organized by different classifications through a unified semantic structure, a library classification. Users can freely browse the content of one concept existing in different systems, and only one query can get the result of all databases in the web.

2 The Principle of Automatic Mapping

The consistence of structure of class schemes and the similarity of concept expression of classes between CLC and DDC are the theory basis to realize the interoperability of metadata in classification schemes.

The two classifications are both general classification schemes. They are description and categorization of knowledge world, even both on the basis of conception logic and knowledge classification.

However, subject classification is according to the certain principles, such as impersonality, development and practicability. The subject domain and knowledge gross of these two classification schemes are almost similar. Under the circumstance of the

https://doi.org/10.23106/dcmi.952107783

similar principle of class division, the conception expressions of classes consequentially have similarity more or less.

Classes are the basic unit to express concept of subject content in documents. The essence of class of CLC and DDC is the mark of subject concept in numerical form, and only kinds and methods of the mark vary probably. Certainly, the system of classification differs. There are a lot of differences of class division and location between CLC and DDC. We can calculate the overlap degree of the concept of class, namely semantic similarity, then decide the mapping relationship of class.

The classes of both classifications are pre-coordination, so we can analyze the class concept into concept elements by semantic factor, then calculate the semantic similarity between the concept elements by certain weighting strategy. In sum, the basic idea of calculation is as follows: the class heading expresses a whole concept, which contain several concept elements. The whole similarity is on the basis of element similarity, namely a complex concept can be analyzed into several concept elements. We can get the whole similarity through the calculation of similarity of concept elements

The basic principle is the following: after semantic factor of classes, the semantic sum of these several concept elements should be identical or near identical to the pre-coordinated class, namely the semantic sum of concept elements should be equal to the whole class concept.

What terms can be used to express class concept? Usually, class headings in CLC and DDC are word or phrase, which can express the connotation and extension of conception. Class terms is standard, concise and it can exactly reflect its content. If cannot, it defined by class notes additionally, every class is restricted by upper-class and lower-class concept.

Words, which are directly extracted from class headings and headings of its upper-class and lower-class, can also express the concept of classes. Words mapping to a class consist of class headings, note terms, upper-class terms and lower-class terms. These words are called as index term_index entries of library classification are derived from these words actually_

The publication and application of kinds of *Classified Chinese Thesaurus* have proved the fact that descriptor from the thesauri can express the concept of classes.

Class concept is expressed by several words usually. If class heading C1 has *n* words and C2 has *m* words, there will be *n_m* kinds of mapping relationships in calculation of word similarity between the two classes.

But only the comparison between the parts, which play the same role, is available. For example, when we compare appearance of two men, we always compare the similarity between the same part such as their face, figure, eye and nose. And we cannot compare between eyes and nose.

Consequently, when compare the similarity of the two concepts, first, judge which parts are the most similar. It adopts the principle of max similarity value, namely, the part of max similarity value plays the same role in the two concepts.

After confirm the similarity value of every element, get the whole similarity by weighting similarity value of every elements. That is similarity of classes. Similarity of concept elements is actually the similarity value of words expressed the class concept. The method is the calculation of semantic

https://doi.org/10.23106/dcmi.952107783

similarity between the words.

3 Construction of Automatic Mapping System

Mapping of Classes can be one-to-one, one-to-multiple or multiple-to-multiple. The mapping system can process not only similarity calculation between two classes, but also the similarity among a lot of classes, thus can construct mapping relationship of one-to-one or multiple-to- multiple. The mapping relationship can be built dynamically, and also can be built to edit a concordance between CLC and DDC created by the mapping system in batch processing.

3.1 Regulation of Automatic Mapping

The mapping system refers the definition of mapping relationships classes in EU project Renardus. The definition is the following: assume A presents a class of CLC, and B presents a class of DDC, five types of mapping relationship can be defined according to the overlap degree of concept expression of class A and B: _fully equivalent _Narrower equivalent _Broader equivalent _Major overlap _ Minor overlap.

The type of mapping relationship is decided by three parameters:_ similarity value of two classes _margin of concept elements of two classes _kind of equivalent concept elements, that is, two classes waiting for mapping perhaps have equivalent class headings, notes terms, upper-class terms or lower-class terms.

The margin can affect the mapping, because if the expression conception of two classes have different connotation and extension, after semantic factor, the concept of class will get several concept elements, then the discrepancy of two classes will be reflected in the number of concept elements. For example, class C_1 has n subject terms, after semantic factor, and then C_1 will have n concept elements. C_2 has m subject terms, so C_2 will have m concept elements. If m is not equal to n, $|n\text{-}m|$ is the margin of these two class headings. As a result, there will have $|n\text{-}m|$ concept elements which cannot be matched, the certainty of mapping relationship will be interfered by these unmatched concept elements.

The system can calculate the mapping relationship of classes according to the values of those three parameters.

3.2 The Choice of Mapping Methods

The similarity of classes is calculated by the following three methods:_by index term;_by class term;_by subject term(descriptors). Index term refers to these words extracted from class headings, notes, its upper-class and lower-class. Class terms refer to some keywords directly extracted from class headings. The reason taking the method of calculation by class term is that the balance discrepancy of class notes and its lower-class between CLC and DDC. Subject terms of CLC comes from"*Chinese Thesaurus*" and DDC are from LCSH in Windows for Dewey. The system uses these words to calculate and decide the mapping relationship. At last, we found that the best method is to take index term as calculation resources among these three methods from a great deal of tests.

3.3 Structure and Functions of Mapping System

The automatic mapping system includes five modules:

3.3.1 Data processing module

This module is to build a concordance, which contains the relationship of class and index terms, class and class terms, class and descriptors. We select finance class of CLC4 and DDC21 as test data. Before building the concordance, some

https://doi.org/10.23106/dcmi.952107783

preprocessing must be done. Firstly, perfect the class headings and complete their meaning. Secondly, translate the class headings of DDC21 into Chinese. Thirdly, some necessary tags must be added to these classes in order to automatically recognition when data processing.

3.3.2 Automatic mapping module

This module is to decide mapping relationship. Select one of three calculation methods, index term, class terms and subject terms are taken respectively as calculation resources. The calculation procedure is followed by the mapping regulations. First, semantic factor of classes, then build a matrix of concept elements to do similarity calculation according to some weighting strategies. At last, check the calculation values and conclude the mapping relationship.

3.3.3 Batch processing module

In this module, mapping scope should be certain in order to decrease some useless computation. For example, we select finance class as test data, so religious class of DDC is no necessary to attend similarity computation. Batch processing is to calculate between some related classes in fact. The function of this module is to map multiple-to-multiple. In general, the system always transfers multiple-to-multiple to one-to-multiple and one-to-multiple is transferred into some one-to-one. Thus a multiple-to-multiple is transferred into several one-to-one. The processing method will be the same as automatic mapping module.

3.3.4 The module of geography subdivision processing

3.3.5 System maintenance module

4 Use of Automatic Mapping System

4.1 Processing of Class Data

The function of Processing of class data is to construct the concordances of classes. For example, the class term "332.404 forms and units of money _DDC class_"can be processed into follow forms:

_class terms: forms of money units of money

_lower-class terms: gold coins silver coins Token coin paper money decimalization of currency

_upper-class term: money

_subject terms: forms of money units of money gold coins Token coin

_ note terms: _vacant_

4.2 Building of Mapping Relationship

4.2.1 Mapping between two classes

Choose the classes of CLC and DDC in the tree box of figure 4-3. Index terms mapping to the classes can be automatically read into semantic calculation box. Then add tags to class terms, notes terms, lower-class terms and upper-class terms respectively. So when calculating equal words, it is easy to recognize the type of words. At the same time we use a semantic dictionary to assign semantic code to calculate word similarity.

In the result display area of Figure 4-2, the upper part displays results of semantic factor of F820.2 and 332.404 class. The lower part displays the matrix of similarity value, after calculation semantic similarity of a pair of words mapping to these two classes. Construction of the words matrix aims to match the most similar concept elements.Assume the matching of concept elements according to the max value, that is, the greater similarity value, the former ranking concept elements.

In Figure4-5, After automatic calculation of "F820.2 forms of money_CLC class_" and "332.404 forms and units of money_DDC class_", the result displays in the range of "fully equality" threshold

https://doi.org/10.23106/dcmi.952107783

value. That is, concept of note terms is equal to class terms, similarity value of classes is more than 65% and the margin of classes is less than 4.

4.2.2 Automatic mapping of batching processing

Batching mapping processing can realize the mapping of one-to-multiple and multiple-to-multiple. By the means of one-to-multiple, the mapping relationship between any class in one library classification and related class in another can be worked out. A concordance can be constructed by multiple-to-multiple mapping, however this kind of calculation is a dynamic computation in a certain scope. The interface of batching mapping processing can be seen in the following figure 4-6.

In the figure 4-6, select 6 class headings of CLC and 11 class headings of DDC, so the batch processing will calculate the similarity value of 66 pairs of class headings. The concrete calculation method is as follows: select a class heading of CLC, then compute the similarity value of this class heading to each class of DDC's, at last, select the class of DDC with the max similarity value as the mapping class to that CLC class. The sorted result of batch processing is also revealed in the figure 4-6.

Figures:

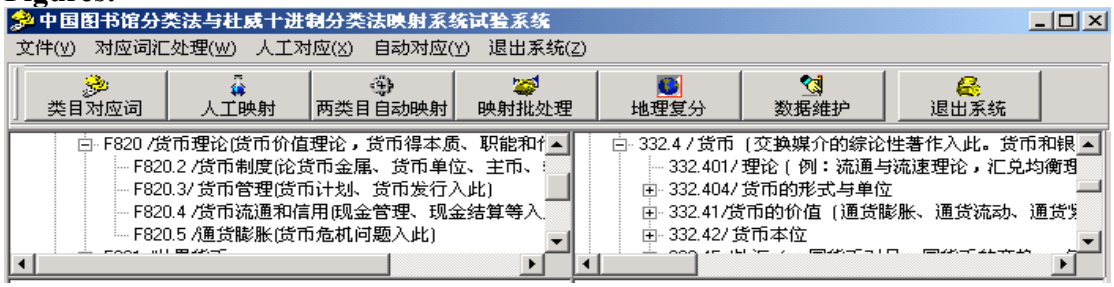

Figure4-1. The basic interface of automatic mapping system

类目	词一	词二	词三	词四	词五	词六	词七	词八
F820.2 /货币	货币制度	(货币金属	(货币单位	(主币)	(辅币)	(货币改革	(复本位制	《货币理论
332.404/ 货币	货币形式	货币单位	【金币】	【银币】	【辅币】	【纸币】	【通货十进	《货币》

Figure 4-2. Index terms mapping to class headings

https://doi.org/10.23106/dcmi.952107783

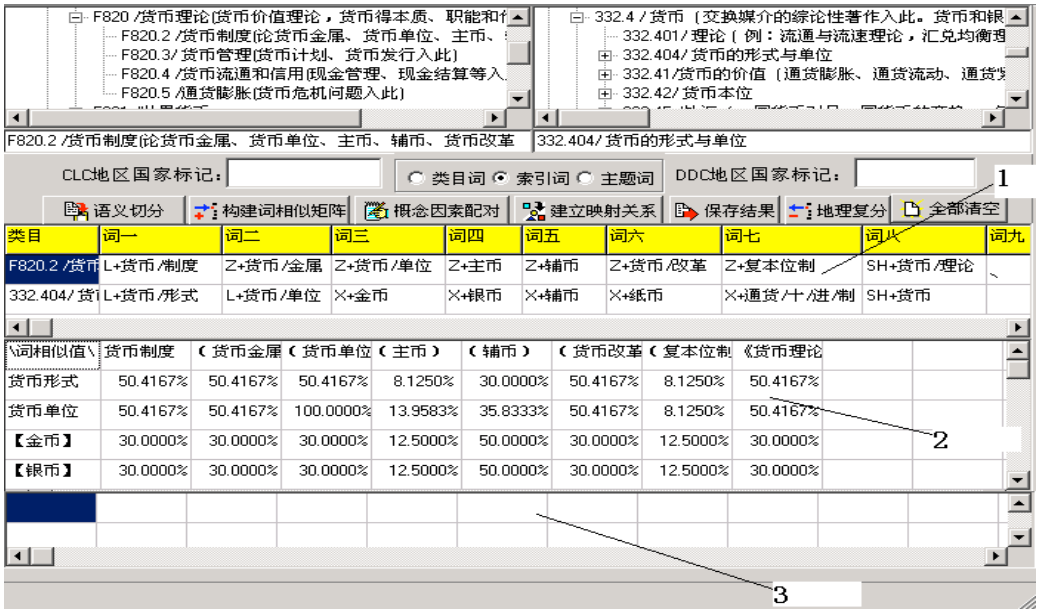

Figure 4-3. Automatic mapping between two classes

F820.2 /货币	《货币单位	《辅币》	货币制度	《货币金属	《货币理论	《货币改革	《主币》	《复本位制
332.404/ 货	货币单位	【辅币】	《货币》	【通货十进	货币形式	【金币】	【银币】	【纸币】
相似值	100%	100%	59.1%	58%	50.4%	30%	12.5%	12.5%

Figure 4-4. Concept elements matching results

CLC类目	映射关系	DDC类目	类目相似度	相等词词类型	词差额	地区
F820.2 /货币制度	相等	332.404/ 货币的形式与单位	68.7139%	2__注释词与类名词+注释词与下位		

Figure 4-5.Construction mapping relationship of class headings

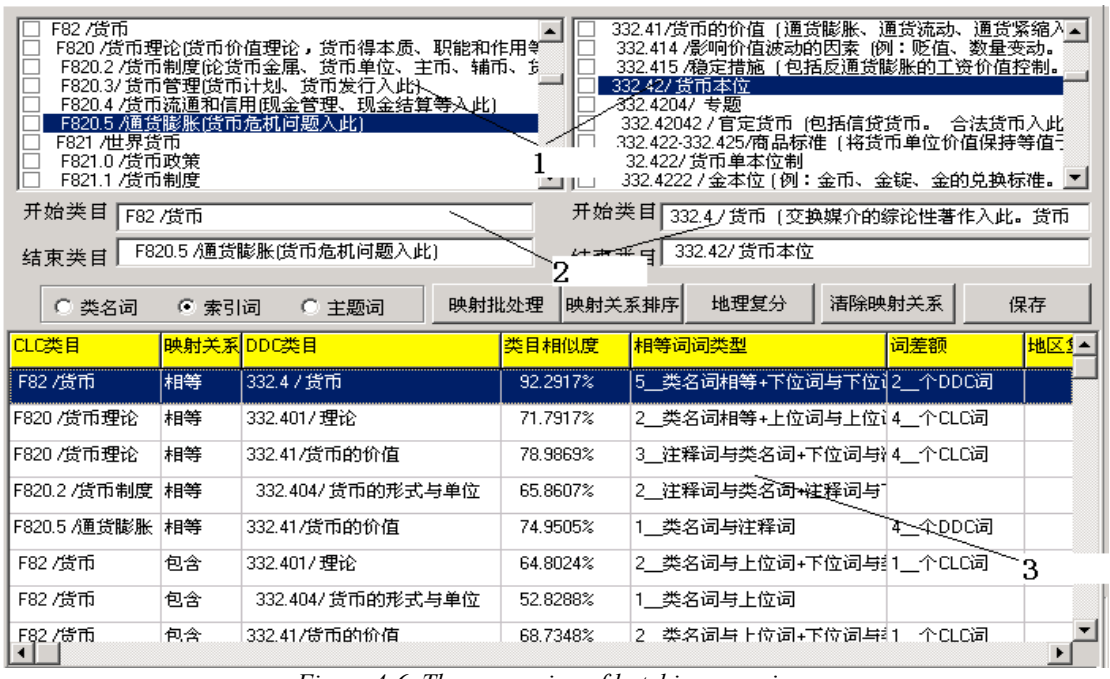

Figure 4-6. The processing of batching mapping

https://doi.org/10.23106/dcmi.952107783

References:

1. Heike Neuroth; Traugott Koch. Metadata mapping and application profiles. Approaches to providing the cross-searching of heterogeneous resources in the EU project Renardus. DC-2001, October 24-26,2001, NII Tokyo, Japan.
2. Lois Mai Chan; Marcia Lei Zeng. Ensuring interoperability among subject vocabularies and knowledge organization schemes:
   a methodological analysis, 68th IFLA Council and General Conference, 2002
3. Ali Shiri. Mapping UNESCO and MeSH mapping UNESCO, LSCH and DDC: Health section center for digital library research HILT project. 2002.11
4. Chengzhi, Zhang_2002_.Research on intelligent search engine of Chinese economic information based on knowledge database. Supervised by Hanqing Hou. Master Dissertation of Nanjing Agricultural University, 2001,6
5. F. W. Lancaster. Vocabulary Control for Information Retrieval. Arlington, Virgiria, Information Resources Pr., 1986.

https://doi.org/10.23106/dcmi.952107783