Cluster Comput (2015) 18:103–112
DOI 10.1007/s10586-013-0339-7

Mapping discovery modeling and its empirical research for the scientific and technological knowledge concept in unified concept space

Lijun Zhu · Chen Shi · Jianfeng Guo

Received: 28 August 2013 / Revised: 1 December 2013 / Accepted: 2 December 2013 / Published online: 1 January 2014

Abstract For heterogeneous scientific and technical knowledge organization systems (HSTKOSs), computer-aided concept mapping discovery becomes very difficult among HSTKOSs. First, this paper puts forward and establishes a common data model (CDM) oriented to the HSTKOS used for the standardized description of scientific and technological knowledge concepts. Then, in the mapping discovery algorithm, the algorithms to discover the relations of inheritance, "is a characteristic of", "is a part of", relevance and other partial ordering relations between heterogeneous concepts are put forward and designed through mapping transfer. Finally, the empirical results show that in public concept space, the mapping discovery algorithm, put forward and designed by this paper, is feasible and can have certain practical significance.

Keywords Knowledge organization system · Concept space · Mapping discovery · Mapping transfer

L. Zhu · C. Shi
Institute of Scientific and Technical Information of China,
Beijing, China

L. Zhu
e-mail: zhulj@istic.ac.cn

C. Shi
e-mail: shichen2012@istic.ac.cn

J. Guo (✉)
Center for Energy & Environmental Policy Research, Institute of
Policy and Management, Chinese Academy of Sciences, Beijing,
China
e-mail: Guojf@casipm.ac.cn

1 Introduction

With the explosive growth of information, large numbers of classification systems based on different criteria and thesauri of each disciplinary field have occurred, representing different knowledge organization systems. Every country attaches great importance to the classification of standard literature and nearly all advanced industrial countries have their own classification systems. In 1983, China compiled the Chinese Classification for Standards (CCS) in accordance with Chinese reality (conditions). Among all the classifications specifications, International Patent Classification (IPC), International Classification for Standards (ICS), CCS, Universal Decimal Classification (UDC), Dewey Decimal Classification (DDC), and Chinese Library Classification (CLC) are widely used. Thesauri, different from classifications, are mainly used in China. Types of thesauri include the

- Chinese Classified Thesaurus, which is the first Chinese large-scale comprehensive thesaurus with the integration of classification and subject;
- Chinese Thesaurus, which is the supporting project of the Chinese Characters Information Processing System (also called Project 748);
- Social Science Thesaurus, which is used for storing and retrieving intelligence on social science and other general thesauri;
- Aerospace Scientific and Technical Thesaurus, referred to as the Aerospace Thesaurus, which is the first Chinese industrial thesaurus;
- Chinese Thesaurus of Petrochemical Industry, referred to as the Petrochemical Thesaurus, which is used for indexing, storing, and retrieving scientific and technological intelligence on the petroleum industry; and

- Thesaurus for Environmental Sciences, which is used for retrieving documents and materials on environmental science and other special thesauri of different academic fields.

The current situation of scientific and technological knowledge organization systems (STKOSs) is that in content, the scientific and technological knowledge organization system has multi-discipline and multi-field sources and that in structure, there are various heterogeneous scientific and technological knowledge organization systems, such as thesauri and classifications. Therefore, from the view of knowledge concept mapping, all STKOSs must be accurately correlated. Building an accurately correlated and well-organized knowledge organization system for multi-discipline and multi-category STKOSs has been a research focus in recent years. Among them, computer-aided scientific and technological knowledge concept mapping is a fundamental work practice and an important research direction [1–3].

Many experts focus on the study of integration between heterogeneous knowledge organization systems. Wiederhold proposed the concept of a mediator that uses knowledge from multiple sources for a higher layer of applications [4]. In 1994, Sibel et al. put forward a uniform declarative and operational framework amalgamating multiple knowledge bases and data structures regarding the mediator [5]. Furthermore, a series of approaches and algorithms have been proposed to address heterogeneous knowledge and information; these are outlined as follows. The multi-agent learning algorithm (MALA) [6] and adaptive resource-provisioning scheme [7] sought to minimize the total project duration. Federated repository and automated mechanisms aimed to discover resources and implement reuse [8]. Spectral clustering algorithms addressed issues of ambiguity and redundancy of metadata [9]. Systematic approaches analyzed both structural and un-structural content and their inter-relationships [10]. Application profile (AP) dealt with the problem of heterogeneity and a root application ontology (AAO) based on AP sought to extend the domain knowledge [11]. The concept based approach processed heterogeneous data in pub/sub systems [12]. MR-Radix aimed to mine multi-relational data [13]. Model-based AEC/FM systems addressed the integration of heterogeneous data representations [14]. Heterogeneous knowledge was addressed at the global level [15, 16] and researchers studied the construction and problems of multi-lingual knowledge organization systems. Hudon studied the development of a multi-lingual thesaurus based on the premise that in a multi-lingual thesaurus, all languages are equal [17]. In addition, trends of multi-lingual services and products involving multiple knowledge organization systems will continue [18].

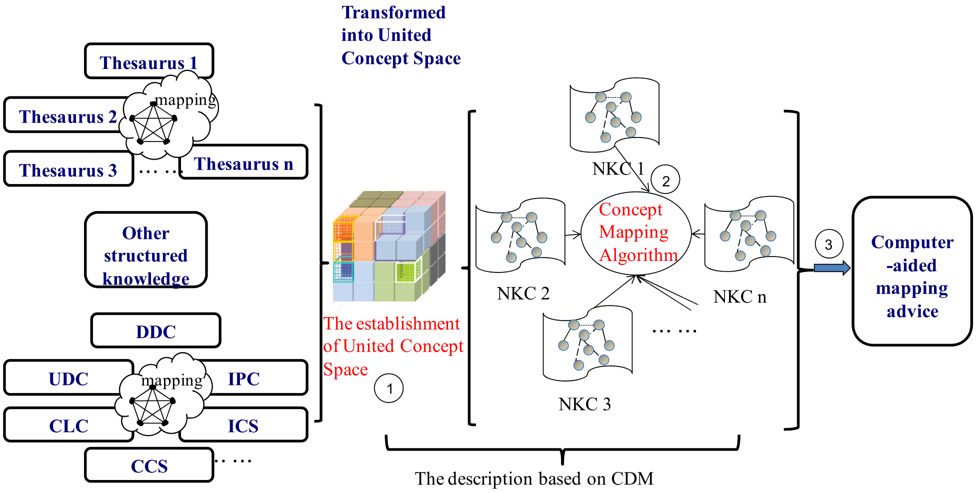
Research on the mapping between knowledge organization systems is also popular. Whitehead investigated the Art and Architecture Thesaurus (AAT) model to implement mapping from Library of Congress Subject Headings (LCSH) to thesauri [19]. The German Federal Ministry for Education and Research funded a major terminology mapping initiative to organize, create, and manage “cross-concordances” between controlled vocabularies (thesauri, classification systems, subject heading lists) [20]. Margaret attempted to access the potential for automatic vocabulary switching from a thesaurus [21]. Doerr concluded possible semantic differences that may hinder the unambiguous mapping and transition from one thesaurus to another [22]. Brahami et al. and Nagpal et al. sought to improve the knowledge mapping process based on the use of different data sources via the data mining technique [23, 24]. Similarity computations, improved instance-based mapping techniques [25], and ontology-mapping frameworks with hybrid architecture [26] were introduced to find mapping between compatible ontologies.

Above all, the unification and interoperability of heterogeneous knowledge organization systems from multiple fields, structures, and languages cannot be avoided [16]. Specifically, mappings for similarity relations between knowledge organization systems have been widely studied, using many techniques and approaches. However, in previous research, few address mapping transfer algorithms of partial ordering relations, which have largely limited the efficiency of semantic relation discovery between knowledge concepts [27–29].

To express the degree of closeness of different objects described by concepts, Gerard Salton [30] put forward the concept of “Vector Concept Space,” based on the algebraic model of information retrieval systems, in his book Mathematics in Library and Information. Concepts and their relations are spatial and multi-dimensional and every complex concept can be assembled by combining easy concepts. The essence of concept space is the algebraic expression of knowledge organization systems, including concept, relation, attribute, classification, and equivalent in implication.

In early mapping discovery research, the expressions of knowledge concepts by different scientific and technological knowledge organization systems are not consistent, making concept mapping in heterogeneous scientific and technical knowledge organization systems (HSTKOS) difficult. Conversely, any structured knowledge can be described with consistent and standard descriptions using a united knowledge organization system (or concept space). Therefore, if HSTKOS is first projected into a neutral concept space and then research on the method of concept mapping is completed, the computer-aided mapping of heterogeneous knowledge concept will be more effective. Considering the mapping discovery algorithm itself, structured deduction,

Fig. 1 Research technology roadmap


Thesaurus 1
Thesaurus 2
Thesaurus 3
Thesaurus n
mapping
Other structured knowledge
DDC
UDC
IPC
CLC
ICS
CCS
mapping
Transformed into United Concept Space
The establishment of United Concept Space
1
NKC 1
2
Concept Mapping Algorithm
NKC 2
NKC n
NKC 3
3
Computer -aided mapping advice
The description based on CDM

which is in essence a mapping transfer method, is easy to use as the mapping deduction algorithm due to its similarity relation.

In this paper, first, a unified concept space model oriented to HSTKOS, namely the Common Data Model (CDM), was built and used as a standardized description of scientific and technological knowledge. Then, heterogeneous knowledge concepts in CDM were projected into relevant the Neutral Knowledge Concept (NKC), integrating the current semantic analysis technologies, to discuss computer-aided concept mapping algorithms among different knowledge organization systems. Finally, efficiency of the unified concept space model and mapping discovery algorithm was verified using practical calculations. The research objects HSTKOS includes thesauri and common classifications, such as IPC, UDC, DDC, CLC, CCS, ICS, etc. The technology roadmap is shown in Fig. 1.

The contribution of this article includes two parts. First, based on the thesauri and various heterogeneous scientific and technological knowledge classifications, the CDM is established, making the standardized description of HSTKOS possible. Second, based on the semantic computation and structured deduction, a set of discovery algorithms for knowledge concept semantic mapping, which are oriented to the specified semantic relation computation between concepts, are put forward and verified.

2 HSTKOS oriented CDM

2.1 Structure analysis of common classifications and thesauri

There are differences between different classification criteria and thesauri in architecture; however, there are two generalities between them. First, IPC, ICS, CCS, UDC, DDC, and CLC have their own hierarchy structures. Thesauri do not have direct codes hierarchies; the relations in the thesauri such as ‘U’, ‘UF’, ‘BT’, ‘NT’, and ‘CT’ can also organize all concepts in the thesauri into a hierarchy system. Second, the classifications and thesauri, such as IPC and CLP, are all mainly classified according to the profession and discipline. A concrete analysis of similarities and differences of research objects in several classifications and thesauri is completed next, mainly by three aspects: (1) the names of the categories or words contained in the knowledge concepts; (2) the attributes constitution contained by the categories or words in the knowledge concepts; and (3) the relationship between the category or words of each level and the subordinate category or words.

Through comparative analysis, the definition of relations between all concepts can be summarized into five types: equivalence, inheritance, relevance, “is a characteristic of”, and “is a part of”. The inheritance relation can be subdivided into three types of relations: “is a kind of”, “apply to”, and “have a characteristic of”. The relation of “is a characteristic of” can be further subdivided into two types of relations: “is an attribute of” and “is a manifestation of”. The CDM we defined is not the unique one for these HSTKOSs involved in the research. But any qualified CDMs must be able to cover the all semantic information containing in the HSTKOSs, and contains the few types of nodes and relations the better for converting HSTKOS into NKC.

2.2 Establishment of CDM

2.2.1 Basic composition of CDM

Based on the analysis of all aspects of all classifications and thesauri, the basic constitution of CDM for the HSTKOS is summarized as follows:

– The nodes in CDM—The sections of IPC are one kind of node, defined as a section node, denoted by A. The classes

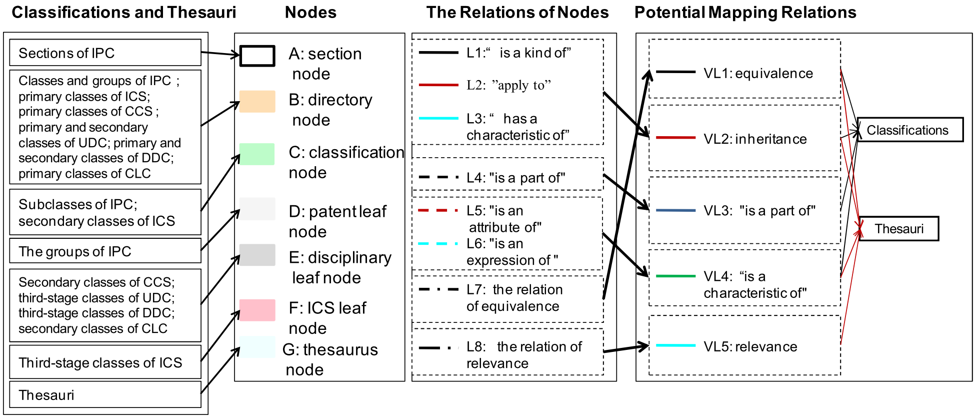

Classifications and Thesauri
Nodes
The Relations of Nodes
Potential Mapping Relations
Sections of IPC
Classes and groups of IPC ; primary classes of ICS; primary classes of CCS ; primary and secondary classes of UDC; primary and secondary classes of DDC; primary classes of CLC
Subclasses of IPC; secondary classes of ICS
The groups of IPC
Secondary classes of CCS; third-stage classes of UDC; third-stage classes of DDC; secondary classes of CLC
Third-stage classes of ICS
Thesauri
A: section node
B: directory node
C: classification node
D: patent leaf node
E: disciplinary leaf node
F: ICS leaf node
G: thesaurus node
L1:" is a kind of"
L2: "apply to"
L3: " has a characteristic of"
L4: "is a part of"
L5: "is an attribute of"
L6: "is an expression of "
L7: the relation of equivalence
L8: the relation of relevance
VL1: equivalence
VL2: inheritance
VL3: "is a part of"
VL4: "is a characteristic of"
VL5: relevance
Classifications
Thesauri

Fig. 2 The nodes and their relations in the architecture

and groups of IPC, the primary classes of ICS, CCS, and CLC, and the primary and secondary classes of UDC and DDC can be classified as one kind of node, defined as the directory node, denoted by B. The subclasses of IPC and the secondary classes of ICS are one kind of node, defined as a classification node, denoted by C. The subgroups of IPC are another kind of node, defined as a patent leaf node, denoted by D. The secondary classes of CCS and CLC and the third-stage classes of UDC and DDC can be classified as one kind of node, defined as a disciplinary leaf node, denoted by E. The third-stage classes of ICS are one kind of node, defined as an ICS leaf node, denoted by F. Subject headings can be seen as one kind of node alone, defined as a thesaurus node, denoted by G. In total, there are seven kinds of nodes.

- The relations in CDM—For the relations in CDM, "is a kind of", is denoted by L1; "apply to", a sort of L1, is denoted by L2; "has a characteristic of", also a sort of L1, is denoted by L3; "is a part of", denoted by L4; "is an attribute of", denoted by L5; "is an expression of ", denoted by L6; the equivalence relation, denoted by L7; and the relevance relation, denoted by L8. In total, there are eight kinds of relations.

CDM is defined, oriented to the thesaurus and the six classification systems, in which the nodes and relations are set up as displayed in Fig. 2. The sources of seven kinds of nodes in CDM are shown on the left of Fig. 2, and eight kinds of basic semantic relations and their derived potential mapping relations in CDM are shown on the right.

2.2.2 *Concept nodes and their basic semantic relations' constraints in CDM*

In CDM, the seven kinds of concept nodes' coverage are different, so they have different attributes. Attribute 1: name. All kinds of concepts have names or names listed in both Chinese and English. Attribute 2: source. With every node subordinate to its own classification or thesaurus, the sources indicate which classification or thesaurus they come from. Attribute 3: number. This is used for the codes existing in IPC, ICS, CCS, DDC, UDC, and CLC. Attribute 4: description. This attribute is aimed at those grade classifications with annotation and nodes with descriptive names. Attribute 5: characteristic. The structure of attributes of the concept nodes in CDM are shown in Fig. 3. Each concept's constitution of attributes is not only one of the basic components to describe the HSTKOS's CDM, but also basic data supporting the concept mapping discovery algorithm.

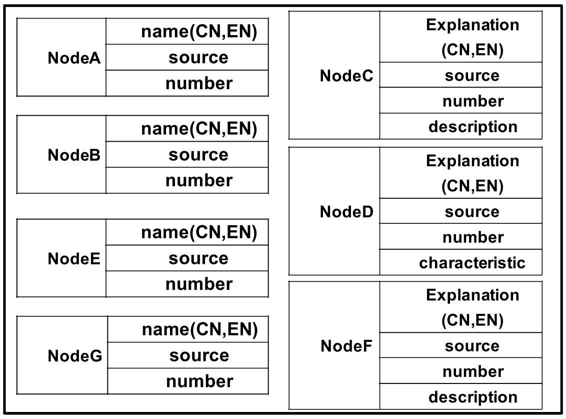

NodeA: name(CN,EN); source; number
NodeB: name(CN,EN); source; number
NodeE: name(CN,EN); source; number
NodeG: name(CN,EN); source; number
NodeC: Explanation (CN,EN); source; number; description
NodeD: Explanation (CN,EN); source; number; characteristic
NodeF: Explanation (CN,EN); source; number; description

Fig. 3 The attributes constitution of the concept nodes in CDM

In Fig. 2, the letters A to G are seven kinds of concept nodes in CDM. Between same kind of nodes, all of the basic semantic relations identified by L1 to L8 are not permitted to exist. The constraints between concept nodes and basic semantic relations are shown in Fig. 4.

2.2.3 *Basic semantic relations and potential mapping relations in CDM*

In Fig. 2, within eight basic relations, in fact, some basic relations that have the same semantic implications generally would be applied with 3. Concept mapping discovery algorithm in same structure, so just five kinds of potential

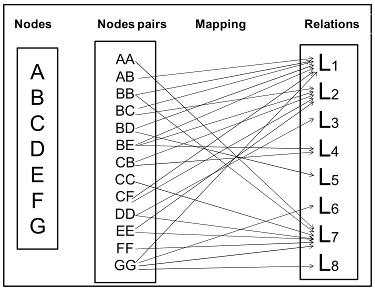


Fig. 4 The relations mapping model between nodes

mapping relations are given based on computer-aided advice, denoted by VL1, VL2, VL3, VL4 and VL5. The relevance relation is regarded as a type alone in potential mapping relations, denoted by VL5. Thus, it can be seen that suing concept mapping within various kinds of classifications, only the computer-aided advice of VL1, VL2, VL3, and VL4 can be output. When the concept mapping is within thesauri, only the computer-aided advice of VL1, VL2, VL4, and VL5 can be output. When the concept mapping is between the classification and the thesaurus, the five kinds of potential mapping relations all can be output.

Combined with the preceding analysis, the HSTKOS oriented CDM is described by the Ecore model, based on the Graphical Modeling Framework (GMF) defined by IBM. The HSTKOS oriented CDM described by the Ecore model is divided into two parts, according to basic semantic relations and potential mapping relations between the concept nodes. The concrete Ecore model is shown in Fig. 5: Fig. 5a defines the connection between seven kinds of concept nodes and eight kinds of basic semantic relations; Fig. 5b defines the connection between seven kinds of concept nodes and five kinds of potential mapping relations.

Based on the GMF, the Ecore model of HSTKOS's CDM is built. Visualized unified concept space construction tools can be derived from the Ecore model. The Ecore model is a standard set of modeling codes, easy to the interaction and transformation with the standard ontology description language, beneficial to HSTKOS's CDM's share and extension. The standardized and structured knowledge description is an important basis for the algorithm of potential mapping discovery.

3 Concept mapping discovery algorithm based on CDM

3.1 The architecture of concept mapping discovery algorithm

Based on the public data model, this article designs a set of algorithms for the scientific and technological concept mapping discovery, which can determine the specified semantic relations between heterogeneous scientific and technological knowledge concepts. Its architecture is shown in Fig. 6. At first, using the point-to-point similarity computation (including grammaticality judgment, "looking-up dictionary" and the edit distance computation) [31] among concepts, find the potential basic mappings, including the relations of equivalence and "is a characteristic of". The relation "is a characteristic of" (e.g., a safety coefficient "is a characteristic of" safety), generally has striking features on the grammatical level, which can be viewed from grammaticality judgment. The "looking-up dictionary" and the edit distance computation are used to discover and compute the similarity relations between concepts. Thereafter, based on the potential basic mapping, using the algorithm of structured deduction, infer the possible potential concept mapping advice of the relations of equivalence, inheritance, "is a characteristic of", "is a part of", and relevance.

Using mature and traditional computing methods, the discovery algorithm of basic relations mapping will not be described in detail in the article. The discoveries of the four mapping relations of inheritance, "is a characteristic of", "is a part of" and relevance are largely triggered by the discovery of the equivalence relation. In the following sections, the algorithm using structured deduction, based on potential basic mapping, will be emphasized.

3.2 Mapping discovery algorithm based on structured deduction

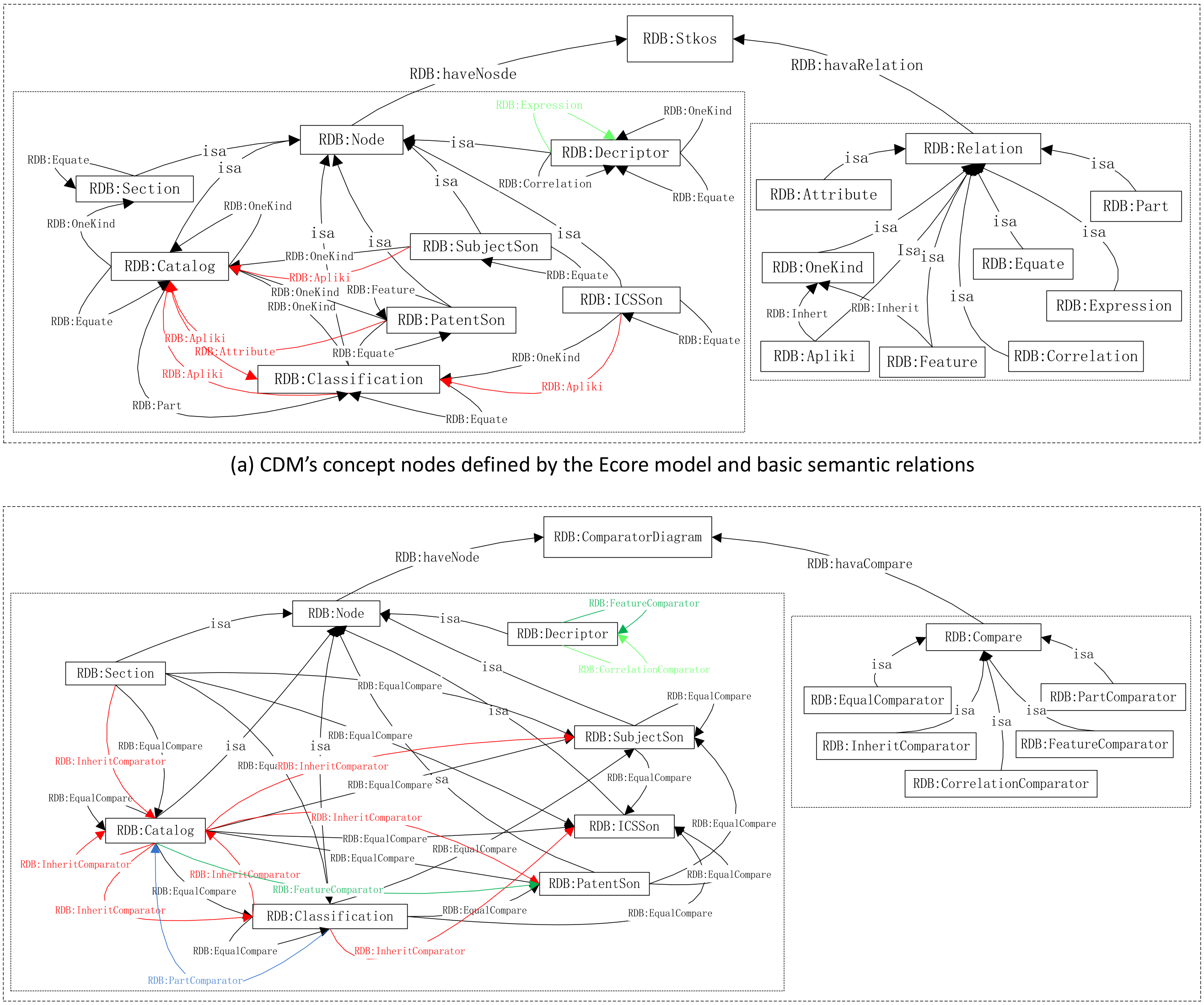
In contrast to similarity relations, relations of partial order should consider the attenuation effect in mapping transfer. This research includes four kinds of partial order relations: inheritance, "is a characteristic of", "is a part of", and relevance. In data structure, all kinds of classifications and thesauri are tree structures. The transfer toward the tree root is called upward transfer and the transfer away from the tree root is called downward transfer.

3.2.1 Situation of upward transfer

In the tree data structure composed of knowledge concepts, the concept nodes show an increase of attributes from top to bottom; that is to say, knowledge concepts show shrinkage of coverage layer by layer. Conversely, if the attributes decrease, the coverage will expand layer by layer. Therefore, the axiom can be summarized as:

Axiom 1 In the tree knowledge organization system, mappings of similarity relations or of other partial ordering relations show no decay in the upward transfer. Namely, the transfer decay coefficient is 1 on each step.

However, a special case could occur in the upward transfer. Inconsistent results can occur transferring concepts from

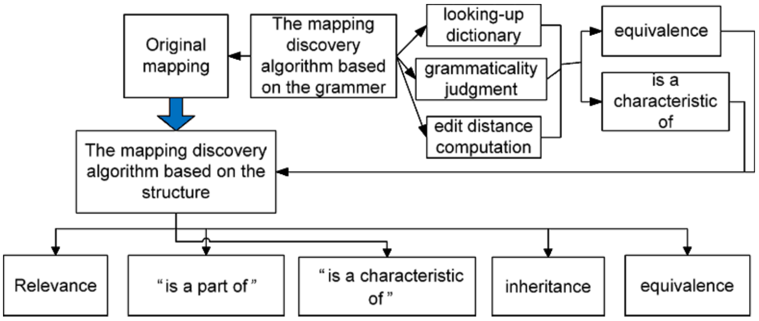


(a) CDM’s concept nodes defined by the Ecore model and basic semantic relations

(b) CDM’s concept nodes defined by the Ecore model and potential mapping relations

Fig. 5 CDM defined by Ecore model

Fig. 6 The architecture of concept mapping discovery algorithm



brother nodes to their same father node, when there is a probability that multiple brother nodes are equivalent to a heterogeneous concept, because the source node of this kind of transfer is not unique. In this case, to make it easier, the maximum of the computed value could be considered.

3.2.2 *Situation of downward transfer*

In the tree knowledge organization system, semantic relations of a partial order show a decrease in the downward transfer; however, different partial ordering relations have

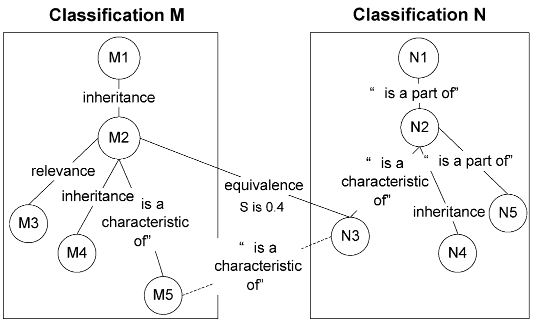


Fig. 7 The example of mapping discovery algorithm based on structured deduction

the same algorithm structures and principals. Thus, this article takes the potential probability computation of "is a characteristic of" relation, for example, which can be seen in Formula (1)

$$S_{abF} = \sum_{r=1}^{R} \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_C^{L_1} \times \alpha_I^{L_2} \alpha_P^{L_3} \alpha_R^{L_4} f_r \tag{1}$$

$$R = 1, 2, \ldots, A_{m+1} + B_{m+1} - A_{m+1} 1 B_{m+1}$$

In Formula (1), S_{abF} is the probability computed after the transfer of the "is a characteristic of" relation. And $\alpha_C, \alpha_I, \alpha_P$, and α_R are transfer decay coefficients between some two concept nodes that connected directly with the four relations of "is a characteristic of", inheritance, "is a part of", and relevance, respectively. L_1, L_2, L_3, and L_4 are the number of times that the relations of "is a characteristic of", inheritance, "is a part of", and relevance appear. f_r is the similarity of the rth attribute between the two concepts. A_{m+1} and B_{n+1} are the number of attributes belonging to concept A and concept B. $A_{m+1} \cap B_{n+1}$ is the number of the attributes of same type in concept A and concept B.

If only attributes of English names are compared in concept nodes, Formula (1) can be simplified as the follows, into Formula (2).

$$S_{abF} = \alpha_C^{L_1} \alpha_I^{L_2} \alpha_P^{L_3} \alpha_R^{L_4} f_{EN} \tag{2}$$

Generally, in the same STKOSs, the probability that some semantic relation appears is evenly distributed, namely, independent of the sample size. Assuming the decay coefficients $\alpha_C, \alpha_I, \alpha_R$, and α_R should be relevant to the average value $\bar{P}$ of $\bar{P}_A$ and $\bar{P}_B$, which are the proportions that the corresponding semantic relations account for all relations, respectively, in two classifications (also a classification and a thesaurus or two thesauri), and relevant to the probability of equivalence S between the transfer's original two nodes. Moreover, the decay coefficients $\alpha_C, \alpha_I, \alpha_P$, and α_R must have some constraints: First, $\alpha_C, \alpha_I, \alpha_P$, and α_R must be monotone functions. Then, if S equals 0, $\alpha_C, \alpha_I, \alpha_P$, and α_R must be 0 and if S equals 1, $\alpha_C, \alpha_I, \alpha_P$, and α_R must be 1. Finally, the range of S is from 0 to 1. Therefore, the calculation functions of $\alpha_C, \alpha_I, \alpha_P$ and α_R are defined as $a = \frac{(\bar{P}+1)S}{\bar{P}+S}$, with $\alpha_C, \alpha_I, \alpha_P$, and α_R corresponding to $\bar{P}_C, \bar{P}_I, \bar{P}_P$, and $\bar{P}_R$, respectively. The final formula, Formula (3), can be obtained by combining $\alpha_C, \alpha_I, \alpha_P$, and α_R.

$$\begin{aligned} S_{abF} &= \left[\frac{(\bar{P}+1)S}{\bar{P}+S}\right]^{L_1} \left[\frac{(\bar{P}_I+1)S}{\bar{P}_I+S}\right]^{L_2} \\ &\quad \times \left[\frac{(\bar{P}_P+1)S}{\bar{P}_P+S}\right]^{L_3} \left[\frac{(\bar{P}_C+1)S}{\bar{P}_C+S}\right]^{L_4} f_{EN} \\ S_{abF} &= \left[\frac{(\bar{P}_C+1)S}{\bar{P}_C+S}\right]^{L_1} \left[\frac{(\bar{P}_I+1)S}{\bar{P}_I+S}\right]^{L_2} \\ &\quad \times \left[\frac{(\bar{P}_P+1)S}{\bar{P}_P+S}\right]^{L_3} \left[\frac{(\bar{P}_C+1)S}{\bar{P}_C+S}\right]^{L_4} f_{EN} \end{aligned} \tag{3}$$

The algorithm example is shown in Fig. 7.

In the example, there are two knowledge organization systems, a thesaurus M and a classification N. While the probability that concept M2 in thesaurus M is equivalent to the concept N3 in classification N is already known, as the relation between M2 and M5 is the relation of "is a characteristic of", the probability that the relation between M5 and N3 is "is a characteristic of" can be calculated by transfer. In Fig. 7, the numbers of M's relations and N's relations are both four, among which the number of the "is a characteristic of" relation is 1. So P_M and P_N are both 0.25, with $\bar{P}$ computed at 0.25. In Formula (3), the f_{EN} is S, namely 0.4, and L_1 is 1. Introducing the values obtained above to Formula (3), then the probability S_{abF} that the relation between M5 and N3 is "is a characteristic of" can be computed at 0.308.

4 Empirical results

- Experimental data:

The HSTKOS includes Petroleum Thesaurus (PT), IPC1 and UDC2. The data is chosen from the concepts of the petroleum-related area. The number of concepts used in the experiment extracted from PT is 1,500. In addition, 900 concepts possibly related to petroleum are extracted from UPC and 700 from UDC.

- Experimental content:

Based on the chosen experimental data, the data from PT, IPC, and UDC are described by CDM. In CDM, every two concepts are computed by the concept mapping discovery algorithm put forward in this article.

- Experimental output:

Table 1 The results of the mapping relations found and verified artificially

Sources		Total	E-R	I-R	R-R	IACO-R
PT&IPC	P	231	161	42	28	0
	V	\	11	17	26	\
PT&UDC	P	238	105	112	21	0
	V	\	7	37	19	\
IPC&UDC	P	161	29	97	28	7
	V	\	7	16	23	2

P is the number of potential mapping relations found and V is the number of mapping relations that are verified artificially. PT&IPC is the number of the mapping relations found between the Petroleum Thesaurus and IPC. PT&UDC is the number of the mapping relations found between the Petroleum Thesaurus and UDC. IPC&UDC is the number of the mapping relations found between IPC and UDC. E-R is equivalence relation. I-R is inheritance relation. R-R is relevance relation. IACO-R is "is a characteristic of" relation

For the HSTKOS mentioned in Sect. 3, the potential concept mapping pairs of five kinds of relations of equivalence, inheritance, "is a characteristic of", "is a part of" and relevance, and their probabilities are listed. Please see Table 1 for detail.

- Experimental results:
  - A total of 231 potential mapping relations are found between PT and IPC and 54 of them are verified artificially. The precision is 23.38 %. The number of the potential equivalence relations is 161 and the number verified artificially is 11. The number of the potential inheritance relations is 42 and the number verified artificially is 17. The number of the potential relevance relations is 28 and the number verified artificially is 26.
  - A total of 238 potential mapping relations are found between PT and UDC and 63 of them are verified artificially. The precision is 26.47 %. The number of the potential equivalence relations is 105 and the number verified artificially is 7. The number of the potential inheritance relations is 112 and the number verified artificially is 37. The number of the potential relevance relations is 21 and the number verified artificially is 19.
  - A total of 161 potential mapping relations are found between IPC and UDC and 48 of them are verified artificially. The precision is 29.81 %. The number of the potential equivalence relations is 161 and the number verified artificially is 29. The number of the potential inheritance relations is 97 and the number verified artificially is 16. The number of the potential relevance relations is 28 and the number verified artificially is 23. The number of the potential "is a characteristic of" relations is 7 and the number verified artificially is 2.
  - In various HSTKOSs, it is rather difficult to simultaneously discover potential semantic mappings by computer. Overall, the precision of the algorithm is not high, but the algorithm already has preliminary practical value.

5 Conclusion, deficiency, and prospect

- Conclusion:

The HSTKOS CDM, which is put forward and established in this article, projects HSTKOS in the same concept space without losing information, making computer-aided discovery possible between various HSTKOSs.

In the mapping discovery algorithm, the algorithm to discover the relations of inheritance, "is a characteristic of", "is a part of" relevance and other partial ordering semantic relations are put forward and designed through mapping transfer, providing a new research thought for the computer-aided discovery of the defined semantic relations. The experimental results indicate that the algorithm has practical value.

- Deficiency and prospect:

With the algorithm experiment completed between the IPC and UDC thesauri the potential semantic mapping of the "is a part of" relation is not found in the experimental results, and the percent of the effective mapping is not high, especially for the relations of inheritance, "is a characteristic of" and "is a part of". The circumstance shows that the semantic relation of "is a part of" defined in CDM does not exist in the IPC and UDC thesauri, which is related to the structures of the three knowledge organization systems. The accuracy of the projection from HSTKOS to CDM should be improved. To be more reasonable and efficient in future studies, the design of CDM must be further analyzed and optimized.

In the process of mapping discovery based on structured deduction, the computation of the transfer decay coefficient, oriented to different specified semantic relations, will be a key task in further study.

Acknowledgements Support from the National Natural Science Foundation of China under Grant No 71271200 and National Twelfth "Five-Year Plan" for Science and Technology Support Program under Grant Nos 2011BAH10B04, 2011BAH10B02 are greatly acknowledged.

References

1. Adali, S., Emery, R.: A uniform framework for integrating knowledge in heterogeneous knowledge systems. In: IEEE Proceedings of the Eleventh International Conference on Data Engineering, pp. 513–520 (1995)
2. Teraoka, T.: Organization and exploration of heterogeneous personal data collected in daily life. Hum.-Cent. Comput. Inf. Sci. **2**(1), 1–15 (2012)
3. Gruber, T.R.: A translation approach to portable ontology specifications. Technical Report of Knowledge System Laboratory, 513–520 (1993)
4. Wiederhold, G.: Mediators in the architecture of future information systems. Computer **25**(3), 38–49 (1992)
5. Adali, S., Subrahmanian, V.S.: Amalgamating knowledge bases II: distributed mediators. Int. J. ntell. Coop. Inf. Syst. **3**(04), 349–383 (1994)
6. Omid, M., Mohammad, R., Akbarzadeh, T.: A novel learning algorithm based on a multi-agent structure for solving multi-mode resource-constrained project scheduling problem. J. Converg. **4**(1), 47–52 (2013)
7. Kim, B., Youn, C., Park, Y., Lee, Y., Choi, W.: An adaptive workflow scheduling scheme based on an estimated data processing rate for next generation sequencing in cloud computing. J. Inf. Process. Syst. **8**(4), 555–566 (2012)
8. Yen, N.Y., Kuo, S.Y.: An integrated approach for Internet resources mining and searching. J. Converg. **3**, 37–44 (2012)
9. Pan, R., Xu, G., Fu, B., Dolog, P., Wang, Z., Leginus, M.: Improving recommendations by the clustering of tag neighbours. J. Converg. **3**(1) (2012)
10. Hsueh, H.Y., Chen, C.N., Huang, K.F.: Generating metadata from web documents: a systematic approach. Hum.-Cent. Comput. Inf. Sci. **3**(1), 1–17 (2013)
11. Liang, A.C., Salokhe, G., Sini, M., Keizer, J.: Towards an infrastructure for semantic applications: methodologies for semantic integration of heterogeneous resources. Cat. Classif. Q. **43**(3–4), 161–189 (2007)
12. Cilia, M., Antollini, M., Bornhovd, C., Buchmann, A.: Dealing with heterogeneous data in pub/sub systems. The concept-based approach, pp. 26–31 (2004)
13. Valencio, C.R., Oyama, F.T., Neto, P.S., Colombini, A.C., Cansian, A.M., de Souza, R.C.G., Correa, P.L.P.: MR-Radix: a multi-relational data mining algorithm. Hum.-Cent. Comput. Inf. Sci. **2**(1), 1–17 (2012)
14. Kosovac, B., Froese, T., Vanier, D.: Integrating heterogeneous data representations in model-based AEC/FM systems. Constr. Inf. Technol. **2**, 556–567 (2000)
15. Castano, S., De Antonellis, V.: Global viewing of heterogeneous data sources. IEEE Trans. Knowl. Data Eng. **13**(2), 277–297 (2001)
16. Reddy, M.P., Prasad, B.E., Reddy, P.G., Gupta, A.: A methodology for integration of heterogeneous databases. IEEE Trans. Knowl. Data Eng. **6**(6), 920–933 (1994)
17. Hudon, M.: Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. Inf. Serv. Use **17**(2–3), 111–123 (1997)
18. Leizeng, M., Maichan, L.: Trends and issues in establishing interoperability among knowledge organization systems. J. Am. Soc. Inf. Sci. Technol. **55**(5), 377–395 (2004)
19. Whitehead, C.: Mapping LCSH into thesauri: the AAT model. In: Petersen, T., Molholt, P. (eds.) Beyond the Book: Extending MARC for Subject Access, p. 81. G.H. Hall, Boston (1990)
20. Mayr, P., Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. arXiv:0806.3765 (2008)
21. Chaplan, M.A.: Mapping “Laborline thesaurus” terms to Library of Congress subject headings: implications for vocabulary switching. Libr. Q. **65**, 39–61 (1995)
22. Doerr, M.: Semantic problems of thesaurus mapping. J. Digit. Inf. **1**(8) (2006)
23. Brahami, M., Atmani, B., Matta, N.: Dynamic knowledge mapping guided by data mining: application on healthcare. J. Inf. Process. Syst. **9**(1), 1–30 (2013)
24. Nagpal, G., Uddin, M., Kaur, A.: A comparative study of estimation by analogy using data mining techniques. J. Inf. Process. Syst. **8**(4), 621–652 (2012)
25. Wang, S., Englebienne, G., Schlobach, S.: Learning concept mappings from instance similarity. In: The Semantic Web-ISWC 2008, pp. 339–355. Springer, Heidelberg (2008)
26. Liping, Z., Guangyao, L., Yongquan, L., Jing, S.: Design of ontology mapping framework and improvement of similarity computation. J. Syst. Eng. Electron. **18**(3), 641–645 (2007)
27. Farquhar, A., Fikes, R., Rice, J.: The Ontolingua server: a tool for collaborative ontology construction. Int. J. Hum.-Comput. Stud. **46**(6), 707–727 (1997)
28. Czarnecki, K., Antkiewicz, M.: Mapping features to models: a template approach based on superimposed variants. In: Generative Programming and Component Engineering, pp. 422–437. Springer, Heidelberg (2005)
29. Gruber, T.: Collective knowledge systems: where the social web meets the semantic web. J. Web Semant. **6**(1), 4–13 (2008)
30. Salton, G.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
31. Bringer, J., Chabanne, H.: Embedding edit distance to enable private keyword search. Hum.-Cent. Comput. Inf. Sci. **2**(1), 1–12 (2012)

Lijun Zhu Professor at Institute of Scientific & Technical Information of China, Deputy Director of Knowledge Organization System Association. His research interesting includes Knowledge Engineering, Large Scale Semantic Computing.

Chen Shi received the B.I.T from Nanjing University, China. She is a master candidate at Information Technology Supporting Center, Institute of Scientific and Technical Information of China. Her research interests include multidimensional clustering, ontology, and competitive intelligence.

Jianfeng Guo received a PHD Degree from the Zhejiang University, China. He has been working in Tsinghua University as a postdoctor for two years and in NEC Labs China as a Senior visiting scholar for seven months. He is an associate professor of Institute of Policy and Management, Chinese Academy of Sciences now. His research interests mainly focusing on semantic web, MIS, DSS, KM et al.