Library Resources Semantization based on Resource Ontology

Fan Yu
Institute of Quality Development Strategy, Wuhan University, Wuhan, China
Junping Qiu and Wen Lou
School of Information Management, Wuhan University, Wuhan, China

Abstract

Purpose – The paper aims to solve the disadvantages of content-based domain Ontology (CDO) and metadata-based domain Ontology (MDO), and improve organization and discovery efficiency of library resources by Resources Ontology (RO).

Design/methodology/approach – The paper constructed a RO model. Methods of Informetrics are utilized to reveal semantic relationships among library resources. Methods of Ontology, Ontology-Relational database Mapping (O-R Mapping), and relational database modelling are utilized to construct RO. Take author co-occurrence for example, the paper demonstrated the capability of RO model.

Findings – RO not only revealed the deep-level semantic relationships of metadata of library resources, but also realized totally computer automatic processing. RO improved the efficiency of knowledge organization and discovery.

Research limitations/implications – Semantic relationships revealed by RO are limited to simple metadata, which makes it difficult to reveal fine-grained semantic relationships. Ongoing research focuses on the revelation of semantic relationships based on title and abstract.

Practical implications – The paper includes implications for utilizing methods of Informetrics to construct Ontology.

Originality/value – This paper proposed a standardized process of Ontology construction in library resources. It may be of potential interest for anyone who needs to effectively organize library resources.

Keywords: semantization, resource ontology, Informetrics, knowledge organization, library resources

1. Introduction

In 2000, Tim Berners-Lee proposed the seven-layer architectures of Semantic Web model (Berners-Lee, Hendler, & Lassila, 2001). Ontology, which is in its middle layer, is the most important layer. URI and XML, which are in the bottom layer, are basic language elements of Ontology. Reason and Proof, which are in the top layer, need support of Ontology. Researches of Ontology focus mainly on two aspects: researching basic theory; constructing Ontology to do knowledge semantization. Biological science, medicine science (Dongmei, & Zhi, 2007), computer science, military science, agriculture science (Yewang, Haibo, & Jinshan, 2012) and geography science (Fan, Lin, & Hong, 2011) gradually pay attention to Ontology. Domain Ontology (DO) in these research areas has been constructed. Knowledge organization of library resources based on DO has become research hotspot of Information Science.

DO of library resources is divided into two categories according to different data sources: CDO (Hui, Chuanming, & Ying, et al., 2006) and MDO (Chengchun, & Zhu, 2011). CDO extracts basic

elements sentence by sentence from the content of library resources, utilizes OWL to describe basic elements, and stores them into Ontology. MDO constructs Ontology based on metadata of library resources. Classes, properties, and individuals of MDO originated from metadata.

The procedures of CDO is more complex than that of MDO. CDO can reveal deep-level, abundant semantic relationships. However, experts' experience will change constantly along with aging. It makes the construction of CDO become more uncertain. Meanwhile, it is difficult for us to describe the whole knowledge structure because of the limitations of experts' experience. MDO is more executable than CDO. MDO can be operated automatically by computer. However, MDO can only reveal superficial semantic relationships. MDO is difficult to improve the efficiency of knowledge organization. In addition, as a visualization tool of semantization, protégé is utilized in the constructions of CDO and MDO. Protégé is easy to operate, yet it has one fatal flaw – the data can only be entered manually. For large-scale data, it is difficult to achieve semantizaiton if data can not be imported in batches.

Metadata of library resources are throughout data objects of Informetrics. Methods of Informetrics, such as: co-occurrence analysis, coupling analysis, and co-citation analysis, can reveal the relationships of metadata. These relationships are semantic relationships (Junping, & Fan, 2012). Resource Ontology (RO) is constructed in this paper which is based on method of Informetrics. An RO is an explicit specification of a conceptualization of library resource based on methods of Informetrics. Compared to the CDO, methods of Informetrics can be operated automatically by computer. Compared to the MDO, methods of Informetrics can reveal more complex semantic relationships. Methods of Informetrics absorb the advantages of CDO and MDO.

2. Literature Review

2.1 Semantization and Knowledge Organization

Knowledge organization is a process of knowledge serialization (Ning, 2012; Qiang, & Yongfu, 2006). Semantization is a kind of knowledge organization. The knowledge organization whose core idea is the revelation of semantic relationships to serialize knowledge is called semantization. Any method which can reveal semantic relationships among knowledge is regarded as semantization. Semantization is no correlation with languages and tools. Therefore, utilizing Ontology languages and tools to organize knowledge is not the only method to do semantization. RO choose relational database to organize knowledge.

2.2 CDO

Semantic annotation is prerequisite for the construction of CDO. Semantic annotation tools are developed by research institutions. Liang and Shumei (2004) compared and analyzed seven semantic annotation tools: SMORE, MnM, OntoMat Annotiser, AeroDAML, Annotea, COHSE, SHOE Knowledge Annotator. Different data sources should be annotated by different semantic annotation methods. Semantic annotation methods include: vector space model (Nianyun, & Chen, 2007), web page sense of vision information (Yulian, Shuai, & Xinglin, 2009), calculating the correlation between the ontology entity and the document and the co-appearance of the label-words frequents and the sematic context in local window (Yewang , Wen , & Xin, et. al., 2009), bootstrapping (Jun, Qi, & Yi, 2010),

The most difficult procedure of CDO is to abstract significant and correct words from sources.

Sources are text files. Words include subjects, predicates, and objects. Jun (2004) introduced the algorithm of automatically extracting ontological concepts by means of agency of liner conceptual graph, and proved its validity and complexity. Hui and Chuanming (2005) used the algorithm of statistics-based term extraction to construct automatic-learned ontology. Dufeng and Baisong (2010) proposed that C-value method and TF-IDF algorithm can be used to filter words and achieve automatic domain-specific term extraction.

The construction of CDO involves two procedures: (1) Constructing the structure with experts' experience. Different research areas have different knowledge conceptions. It is unrealistic to rely solely on Ontology builders to complete classification of conceptions and construct the structure of CDO. Therefore, in the process of construction, experts' experience is utilized to classify conceptions. The structure of CDO includes classes, and properties. After experts determine classification criteria of conceptions, Ontology builders map classification criteria to classes, and properties. (2) Extracting individuals with computer system. Individuals consist of subjects, predicates, and objects in each sentence of library resources. Automation segmentation system divides library resources into sentences, and then knowledge extraction system extracts subjects, predicates and objects from sentences. Finally, Ontology builders map each group of subject, predicate, and object to Triple (Triple is called in Ontology language), and store them into CDO.

The construction of CDO have a premise – the definitions of all conceptions that constitute the structure of CDO must be unique (Wenhua, 2005). As mentioned above, the conceptions, which will be influenced by experts' experience, are not unique. It gives a challenge to application, reusing and integration of CDO.

As mentioned above, automation segmentation system divides library resources into sentences, and then knowledge extraction system extracts subjects, predicates and objects from sentences. Seeing from its appearance, machine-based content extraction saves time and effort. It is a computer automatic processing system. In fact, there are many problems:

(1) Sentences include simple sentences and complex ones. Some simple sentences lack subjects, predicates, or objects, such as exclamatory sentences. Some complex sentences include attributive clauses or adverbial clauses. Machine-based content extraction can not extract these sentences effectively.

(2) Predicates include a wide range of words. Polysemy and synonymy really exist. How to extract these predicates is a major problem.

(3) There are still a lot of pronouns in the sentences, such as: personal pronouns, possessive pronouns, demonstrative pronouns. In order to determine their meaning, we should understand their context firstly. How to understand the meaning of context automatically is one of the problems urgently waiting to be solved.

2.3 MDO

Ontology is an advanced semantic web technology with prospect of application. The complexity of CDO hinders the development of Ontology (Jing, & Ping, 2004). The approach to build domain ontology from the base of thesaurus has advantages (Aimin, Zhen, & Jing, 2005). Chinese Classified Thesaurus (Yun, Dongyi, & Wende, 2007), SKOS (Chunyan, Shuping, & Yucheng, 2007), WordNet (Zili , & Yanna, 2011), HowNet (Wenjuan, & Feng, 2011) are well-structured vocabulary. They can be used to construct MDO. The key issue for MDO is that how to express property and relations among classes, properties, and individuals (Hua, 2010).

Appropriate methods should be used according to different characteristics of thesauri, such as scale, standardization and rigidness of semantic relationships (Junzhi, Rongjuan, 2009). How to map thesaurus to the hierarchy of ontology is the key procedure to gather conceptions and relationships (Junzhi, 2007).

The construction of MDO involves two procedures: (1) Constructing the structure based on thesauri. The conceptions in whole subjects are classified and coded by Classified Chinese Thesaurus (CCT) (Huanhuan, 2011). CCT determines what conceptions each research area includes. Conceptions are basic elements to construct the structure. At the same time, dictionaries in research area, which are more microscopic than CCT, can be its good expansions. (2) Fulfilling individuals with online databases. Individuals consist of metadata of library resources. There are many online databases, such as: CNKI, WOS, and CSSCI. They download metadata according to the CCT. Ontology builders map each metadata to Triple, and store them into MDO.

The definitions and hierarchies of all conceptions in Thesaurus are very clear. Compared to CDO, the construction of MDO is simpler. Yet MDO can only reveal the superficial semantic relationships among metadata. The semantic revelation of MDO is no more than that of two-dimensional tables. In addition, MDO can only reveal hierarchical relationships, it cannot reveal non-hierarchical relationships. Non-hierarchical relationships can perform more abundant semantics than hierarchical relationships. The degree of MDO should be deepened.

2.4 Methods of Informetrics

Co-occurrence analysis is a quantitative analysis method to analyze co-occurred information in various information carriers (Yuefen, Shuang, 2006, 2007). Co-occurrence analysis can reveal semantic relationships in information carriers. DO based on co-occurrence analysis can shorten construction period.

Methods of Informetrics which can reveal semantic relationships have the following characteristics: data sources of method include at least two or more literatures, among which semantic relationships can be revealed; analysis object of method is metadata of library resources, which is convenient to be processed; relationships revealed from literatures have difference of strong and weakness, which can be used to calculate semantic similarity. In this section, we will introduce the definitions of methods of Informetrics which have the above characteristics. These methods are properties of RO.

Co-occurrence: two different metadata characteristic values appear in one paper. Co-occurrence analysis methods that have been used in Informetrics include: keyword co-occurrence (Shenqin, Jilong, & Lei, 2011), author co-occurrence (Chunlin, & Yuguang, 2010), institution co-occurrence (Guohe, & Jingxue, 2011), and subject co-occurrence (Chunjuan, Haishan, & Baode, 2010).

Co-word: two different keywords appear in one paper. Co-word analysis method that has been utilized in Informetrics includes: keyword co-word (Yi, & Chuanjun, 2011).

Cooperation: two different authors or institutions appear in one paper. Cooperation analysis methods that have been utilized in Informetrics include: author cooperation (Yun, Wenyuan, & Yunlin, et al., 2009) and institution cooperation (Junping, & Hui, 2011).

Coupling: it includes bibliographic coupling and citation coupling (citation coupling is also called coupling). Bibliographic coupling refers to one metadata characteristic value appears in two or more papers. Citation coupling refers to one metadata characteristic value appears in two or more references. Coupling analysis methods involves at least two papers, which are different from

co-occurrence, co-word, and cooperation. Therefore, different metadata can be combined together to generate new coupling. For example, we can analyze keyword coupling first, and then analyze author coupling based on keyword coupling. By convention, combined coupling analysis method is called author keyword coupling. Bibliographic coupling analysis method that has been utilized in Informetrics includes: author keyword coupling (Junping, & Feifei, 2010). Citation coupling analysis method that has been utilized in Informetrics includes: citation coupling (Ming, & Guojun, 2011).

Co-citation: two or more papers are cited by one paper at the same time. Co-citation analysis methods that have utilized in Informetrics include: author co-citation (RuiMin, & Chaoqun, 2011), journal co-citation (Junping, & Weihua, 2008), and subject co-citation (Wenta, Haojie, & Gong, et al., 2008).

2.5 Mapping from Ontology to Relational Database

DO described by OWL has shown the unique advantage in knowledge organization and the readability of human and machine. However, there is a problem we can not ignore: DO has a lower efficiency of search when data are large. This is because there are not market-oriented query language and database technology to support. The scale of Ontology is more and more big. It becomes more and more difficult to query data. In contrast, relational databases have obvious advantages in data query. SQL Server, Oracle, and MySQL are excellent relational databases. They can satisfy easily users' demands in data storage and query. Structured Query Language (SQL) is very efficient and suitable for almost any relational databases. Relational database can play fully advantages of Ontology, which will not be limited by OWL (Yanzhang, & Liang, 2011). Data organization mode of OWL is similar to that of relational database. Conceptions between OWL and relational database have mapping relationships. For instance, classes of OWL are similar to tables of relational database. Data types of OWL are similar to that of relational database. Subclasses of OWL are similar to primary keys and foreign keys (Man, Yan, & Yiyu, 2005). OWL defines some class tags, such as: domain, range, and property tags, such as: FunctionalProperty, TransitiveProperty. Class and property tags have not corresponding objects in relational database. The usual practice is that each tag is stored into each table, and tables are connected by primary keys and foreign keys (Jun, Bo, 2010; ZhuoMing, & Yongjing, 2006; Suihua, Xiaodan, & Yue, 2011).

3. Research Questions

On the basis of revealing the problems existing in CDO and MDO, this paper constructs RO. Methods of Informetrics are introduced to integrate advantages of CDO and MDO. Based on the assumption that RO organizes knowledge better than CDO and MDO does, the study addresses three questions:

How does RO reveal semantic relationships of library resources?

Does RO solve problems in CDO and MDO?

Does RO improve the efficiency of knowledge organization?

4. RO

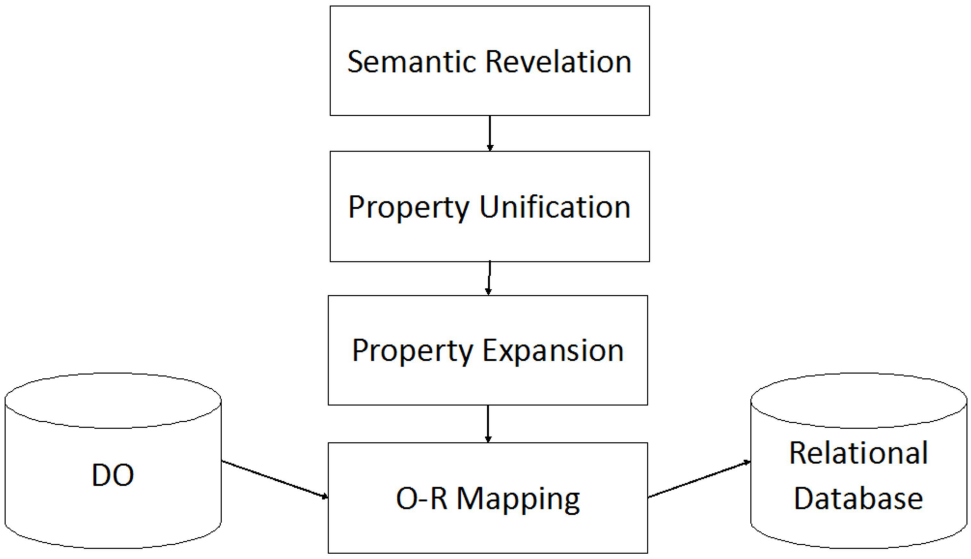


Fig.1 describes procedures of RO construction. The revelation of semantic relationships is the target of DO. RO utilizes methods of Informetrics to reveal semantic relationships. In the section of Semantic Revelation, the detail procedures of semantic revelation based on methods of Informetrics are described. DO properties are named after the names of Informetris methods. The names of some Informetrics methods are quite similar. They will by unified in the section of Property Unification. In the section of Property Expansion, DO properties are expanded based on present Informetrics methods. In the section of O-R Mapping, the structure of DO is constructed. Two O-R mapping methods are combined together to convert DO into relational database.

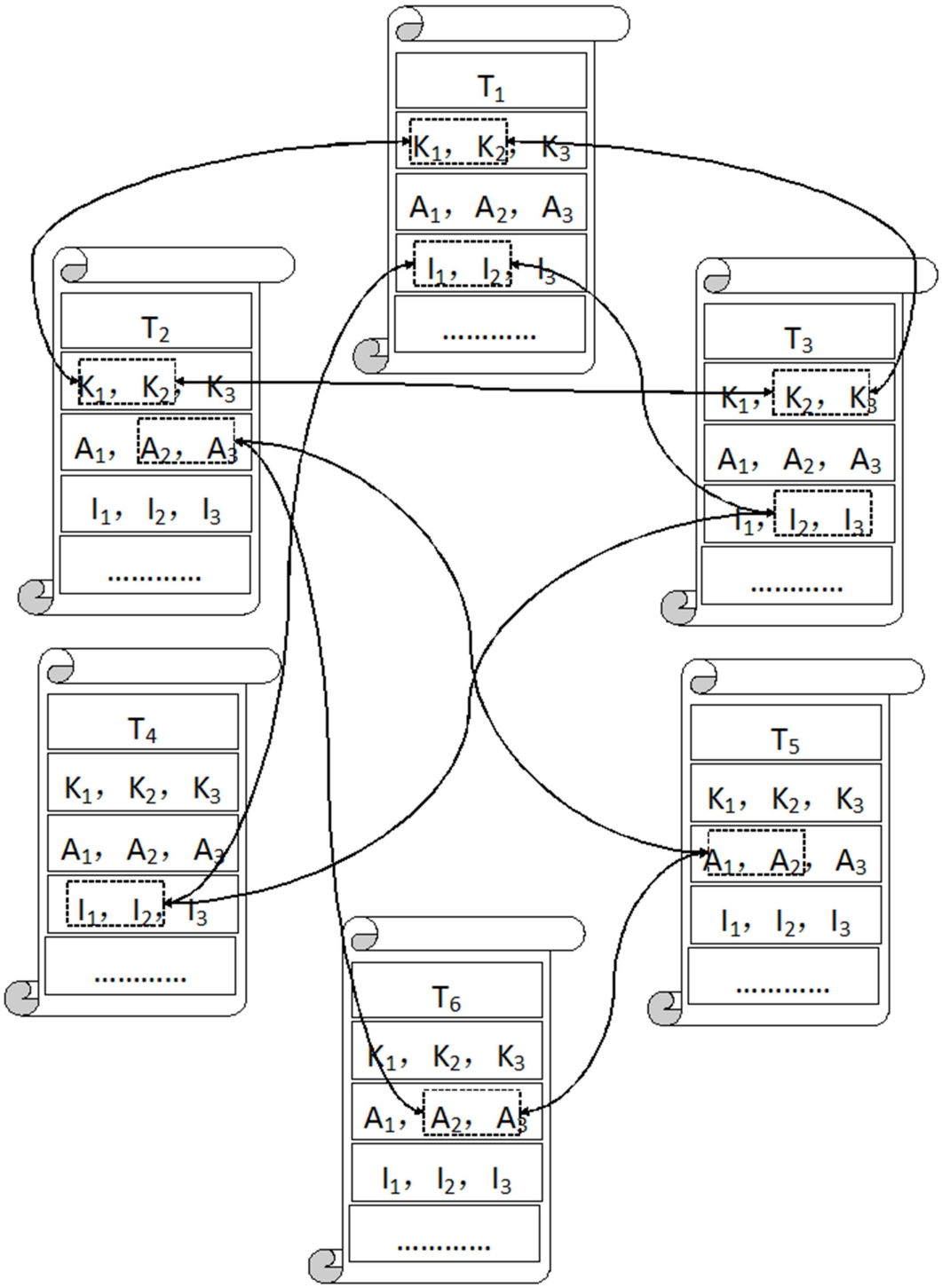
4.1 Semantic Revelation

The construction idea of RO is: corpus is constructed based on metadata of library resources, methods of Informetrics are utilized to do statistical analysis based on corpus, and the structure, semantic relationships and individuals of RO are stored into relational database.

Relation table whose fields are constructed by metadata has superficial semantic relationships. Semantization of library resources based on metadata is shown in Fig. 2.

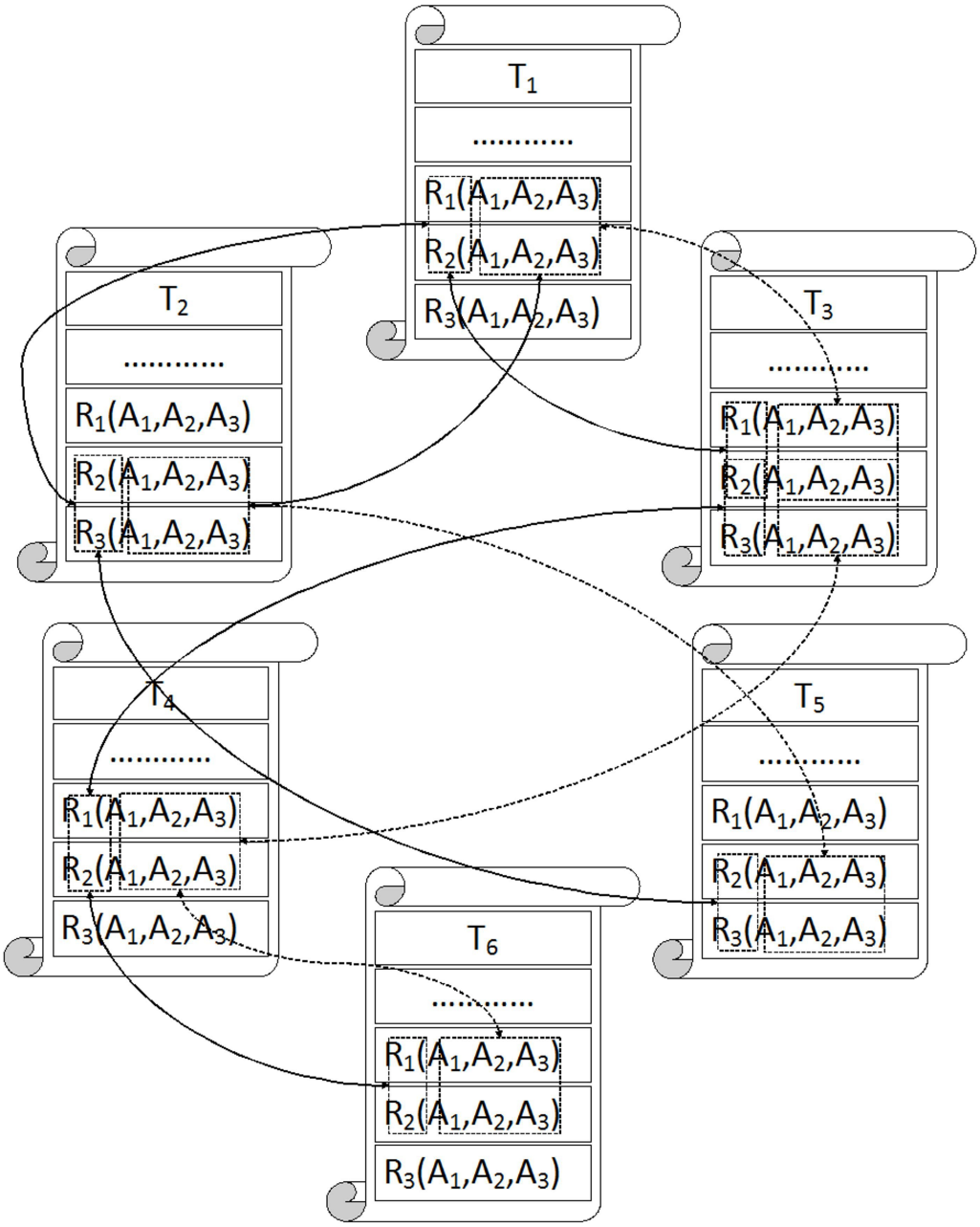
Fig.2 is similar to a relation table. Data in the first line are equivalent to table fields; data from the second to the last line are equivalent to table records. The values of fields in vertical dashed boxes have same characteristics. For example, author2 and author3, keyword1 and keyword2 are same characteristics. Fields in different records are integrated together with the same characteristics. Vertical semantic relationships are generated. Records in horizontal dashed boxes combine all fields together. Horizontal relationships are generated. Vertical and horizontal relationships make data in the whole table to be connected intimately. However, these relationships appear when data are imported into table. They are too simple. Relationships mining is needed.

Methods of Informetrics explore the relationships among fields based on relation tables, and reveal deep semantic relationships.



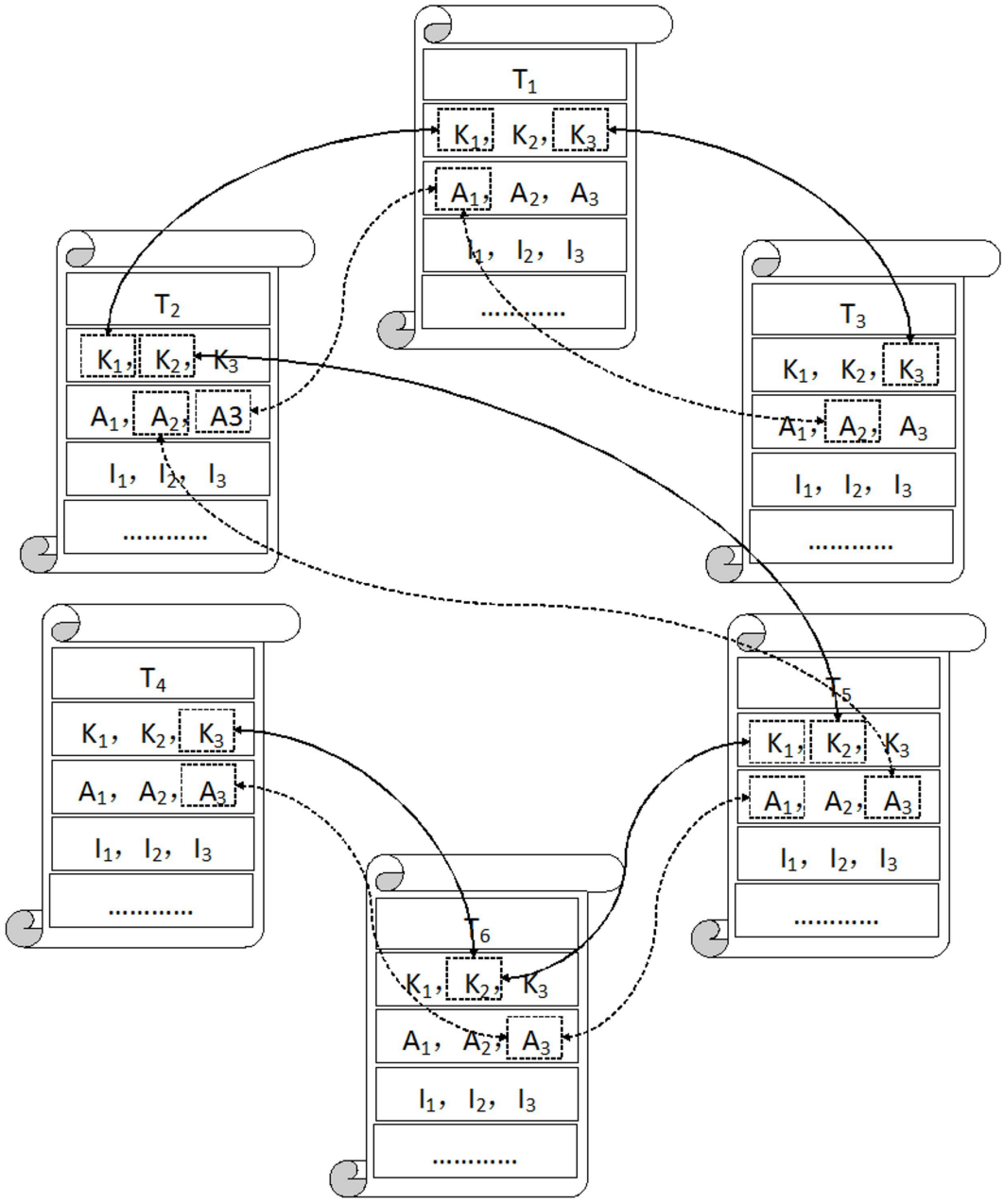
As shown in Fig. 3, $T_i(i \quad 1,2,...,n)$ refers to paper whose title is T_i . $K_i(i \quad 1,2,...,n)$, $A_i(i \quad 1,2,...,n)$, and $I_i(i \quad 1,2,...,n)$ refer to the ith keyword, author,

and institution. The ellipsis refers to the other metadata. Data in dashed boxes refer to keyword co-occurrence, author co-occurrence, and institution co-occurrence. Two papers whose keywords, authors, and institutions are connected by two-way arrows represent that there are co-occurrence relationships between them. Take the keywords co-occurrence for example, $Keywords\_\tilde{Co}\ occurrence$ refers to keywords co-occurrence, K_1 and K_2 refers to two different keywords. Then, $Keywords\_\tilde{Co}\ occurrence(K1, K2)$ refers to that K_1 and K_2 have semantic relationship of keyword co-occurrence. In Fig. 3, (K_1, K_2) in T_1, (K_1, K_2) in T_2, and $T_3(K_2, K_3)$ in T_3 have the same relationships of keyword co-occurrence. In order to distinguish the strength of keyword co-occurrence, the conception of co-occurrence strength is introduced. $Keywords\_\tilde{Co}\ occurrence(K1, K2, n)$ refers to that K_1 and K_2 have semantic relationship of keyword co-occurrence, and the co-occurrence strength is n. Irrelevant papers are connected by Keyword co-occurrence. Keyword co-occurrence analysis method reveals implicit relationships of library resources based on keywords.



In Fig. 4, each paper omits some basic metadata, such as: keywords, authors. It lists references in detail. $R_i(A_a, A_b, A_c)$ refers to the ith reference in paper T_i. T_i has three authors: A_a, A_b, and A_c. R_i in dashed boxes refers to references co-citation. A_a in dashed boxes refers to authors co-citation. Two-way arrows have two kinds of lines: solid lines and dashed lines. Two papers whose references are connected by two-way arrows with solid lines represent that there are reference co-citation between them. Two papers whose authors in references are connected by two-way arrows with dashed lines represent that there are author co-citation between them. Author co-citation reveals indirect semantic relationships, which is different from keyword co-occurrence. We should reveal references co-citation first, and then reveal authors co-citation. $Authors\_\tilde{Co}\ citation(A1, A2, n)$ refers to that A_1 and A_2 have semantic relationship of

authors co-citation, and the strength is n. In Fig. 4, T_1, T_2, T_5 and T_4, T_6 have relationships of author co-citation, respectively. Two groups of papers have not relationships. Yet T_1, T_3 and T_3, T_4 have relationships of author co-citation, respectively. Two groups of papers are connected by T_3. Author co-citation analysis method reveals implicit relationships of library resources based on authors in references.

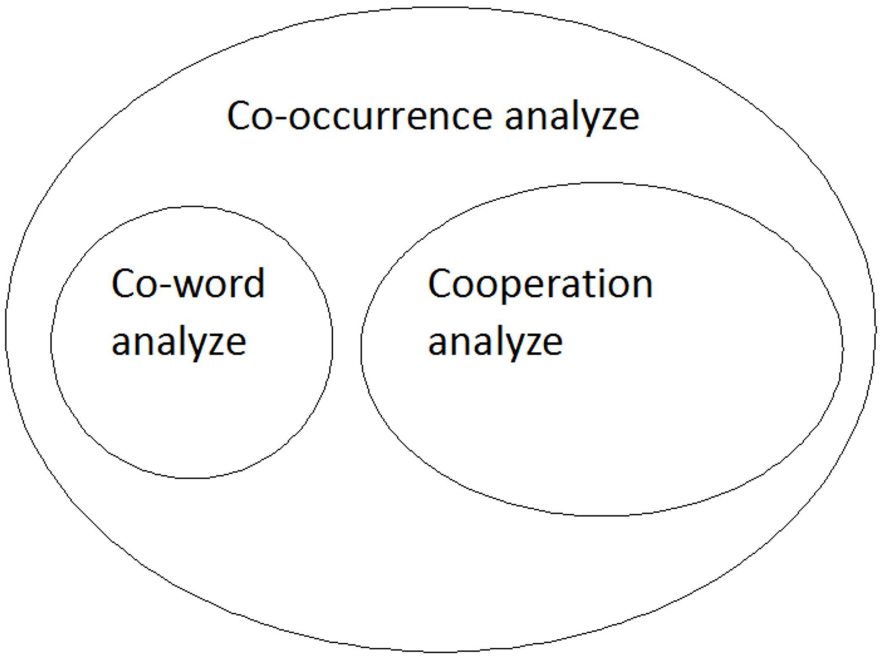


In Fig. 5, two papers whose keywords are connected by two-way arrows with solid lines represent that there are keyword coupling between them. Two papers whose authors are connected by two-way arrows with dashed lines represent that there are author keyword coupling between them. Author keyword coupling reveals indirect semantic relationship, which is the same as author co-citation. We should reveal keyword coupling first, and then reveal author keyword coupling.

$Authors\_Keywords\_Coupling(A_1, A_2, n)$ refers to that A_1 and A_2 have semantic relationship of author keyword coupling, and the strength is n. In Fig. 5, T_1, T_3 and T_4, T_6 have relationships of author keyword coupling. They are connected by T_2, T_5. Author keyword coupling analysis method reveals implicit relationships of library resources based on authors and keywords.

Property Unification

Definitions of some Informetrics methods listed above are quite similar. In order to simplify the construction of RO, they are unified.



As shown in Fig. 6, co-word only analyzes keywords. Cooperation analyzes authors and institutions, which is more than co-word. Co-occurrence analyzes subjects besides keywords, authors, and institutions. Because the analysis procedures of co-word and cooperation are the same with co-occurrence, we unified use the definition of co-occurrence.

4.2 Property Expansion

Methods of Informetrics focus mainly on five metadata: keywords, authors, institutions, journals, and subjects. The present Informetrics methods are expanded based on five metadata to enrich semantic relationships of RO.

Co-occurrence: Besides methods of keywords, authors, institutions, and subjects co-occurrence, expanded method includes: journals co-occurrence.

Coupling: Besides methods of bibliographic coupling mentioned above, expanded methods include: keyword coupling, author coupling, institution coupling, journal coupling, and subject coupling. Combined coupling analysis methods include: keyword keyword coupling, keyword author coupling, keyword institution coupling, keyword journal coupling, keyword subject

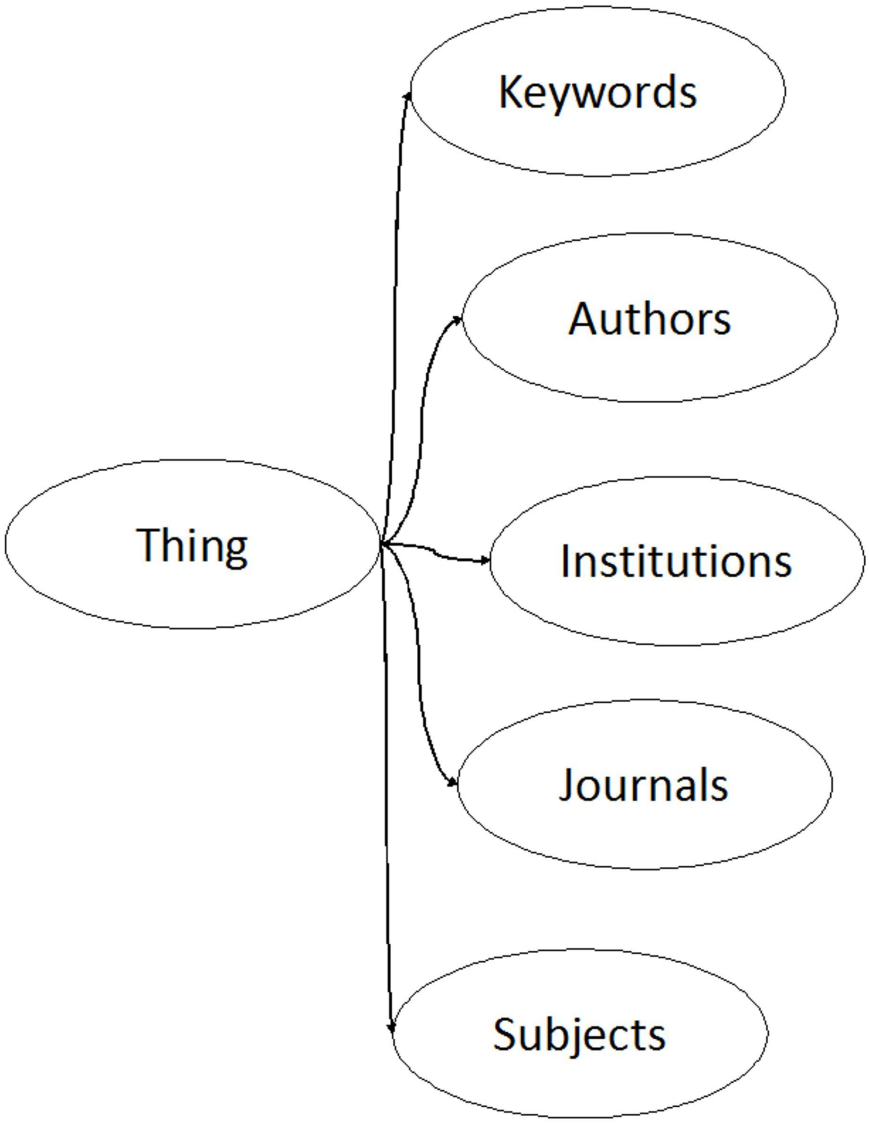
coupling, author author coupling, author institution coupling, author journal coupling, author subject coupling, institution keyword coupling, institution author coupling, institution institution coupling, institution journal coupling, institution subject coupling, journal keyword coupling, journal author coupling, journal institution coupling, journal journal coupling, journal subject coupling, subject keyword coupling, subject author coupling, subject institution coupling, subject journal coupling, and subject subject coupling.

Expanded methods of citation coupling include: keyword citation coupling, author citation coupling, institution citation coupling, and subject citation coupling.

Co-citation: Expanded methods of co-citation include: keyword co-citation, and institution co-citation.

4.3 DO

DO consists of three parts: classes, properties and individuals. Methods of Informetrics analyze five metadata of library resources. DO of library resources are as shown in Fig. 7:



In Fig. 7, RO classes have two layers. The classes of keywords, authors, institutions, journals, and subjects are subclasses of Thing. The structure is very clear. Individuals are abstracted from

five metadata.

Table 1 Partial properties of RO

Property Name	Domain	Range
Keywords_Co-occurrence	Keywords	Keywords
Authors_Co-occurrence	Authors	Authors
Institutions_Co-occurrence	Institutions	Institutions
Journals_Co-occurrence	Journals	Journals
Subjects_Co-occurrence	Subjects	Subjects
Keywords_Coupling	Keywords	Keywords
……	……	……
Authors_Keywords_Coupling	Authors	Authors
Authors_Authors_Coupling	Authors	Authors
Authors_Institutions_Coupling	Authors	Authors
Authors_Journals_Coupling	Authors	Authors
Authors_Subjects_Coupling	Authors	Authors
……	……	……
Keywords_Co-citation	Keywords	Keywords
Authors_Co-citation	Authors	Authors
Institutions_Co-citation	Institutions	Institutions
Journlas_Co-citation	Journlas	Journlas
Subjects_Co-citation	Subjects	Subjects
……	……	……

In table 1, property name consists of two parts: metadata name and Informetrics methods. They are connected by "_". The domain and range of each property is also defined.

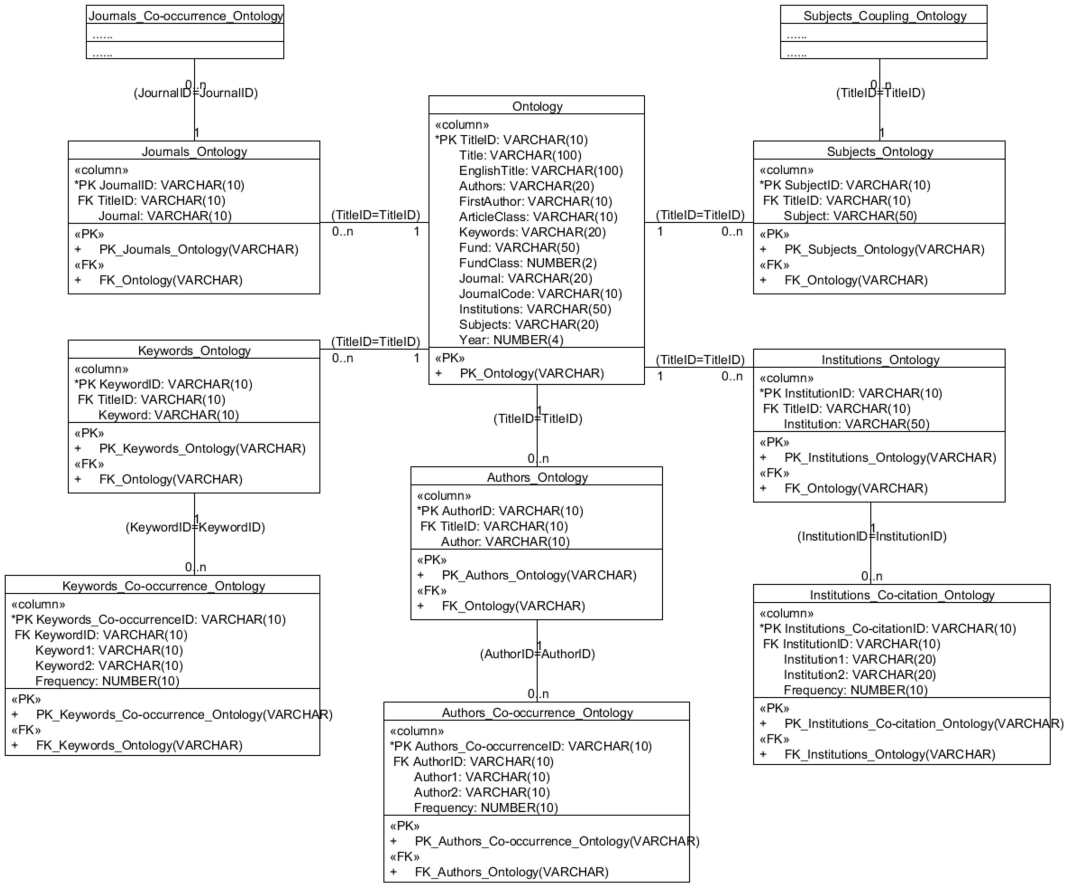
4.4 O-R Mapping

The data of Ontology can only be entered manually. Meanwhile, it is difficult to search data effectively in large-scale data. How to utilize relational database to store DO has become research hotspot. The storage modes of O-R mapping include: horizontal mode (Agrawal, & Somani, 2001), vertical mode (Yahong, & Zhuoming, 2002), and decomposition model (Alexaki, Christophides, & Karvounarakis, et al., 2001). Horizontal mode only creates a table. Fields represent properties. Records store Individuals and semantic relationships. The structure of horizontal mode is quite simple. However, data based on horizontal mode are difficult to understand and search. Vertical mode also creates a table. There are three fields in table: subject, predicate, object, whose structure is similar to DO triples. Each record corresponds to one triple. Partial information is omitted in vertical mode. Decomposition model has two decomposing methods: decompose with classes as unit, and decompose with properties as unit. In two decomposing methods, each class or property create a table. Relationships among classes and properties are connected by primary keys and foreign keys. In order to preserve complete information, two decomposing methods are combined together. Each class and property creates a table.

4.5 Relational Database

According to the rules of O-R mapping mentioned above, each class table is generated to store

one class. Each property and individual in each metadata table is abstracted, and property-individual tables are generated to store individuals.



In Fig. 8, storage mode of research areas – Ontology is constructed. Table named “Ontology” in the middle of the graph is metadata table. It includes basic fields of library resources, such as: title, author, keyword, institution, and fund, etc. Five tables around “Ontology” are Journals_Ontology, Keywords_Ontology, Authors_Ontology, Institutions_Ontology, and Subjects_Ontology. They represent journal class, keyword class, author class, institution class, and subject class. Five class tables are connected with “Ontology” through the fields of TitleID. The relationships between five class tables and “Ontology” are n:1, which means many records in five class tables correspond to one record in “Ontology”. Property-individual tables are connected with five class tables. The names of Keywords_Co-occurrence_Ontology, Authors_Coupling_Ontology, and Institutions_Co-citation_Ontology refer to semantic relationships of Keywords_Co-occurrence, Authors_Coupling, and Institutions_Co-citation, respectively. Three property-individuals tables store individuals, such as Keyword1 and Keyword2, and strength, such as Frequency, into Keywords_Co-occurrence_Ontology. The relationships between property-individual tables and class tables are also n:1. They are connected by KeywordID, AuthorID, and InstitutionID, respectively. The hierarchy of this storage mode is very simple. Metadata tables are connected with property-individual tables by class tables. Three categories of tables can be expanded easily.

5. Empirical Analysis

In this section, the author co-occurrence analysis is utilized to extract semantic relationships from author metadata.

5.1 Data Sources

Data of authors co-occurrence analysis are from Web of Science. Search strategy is: inputting "TI=(Ontology)" in text box of advanced search, selecting "English" as paper language, selecting "Article" as paper type, selecting "2001-2011" as publish time. 3406 records are founded in Web of Science.*5.2 Methods*

Price's law is used to calculate productive authors. It can be used to filter core authors vaguely. The results calculated by Price's law are approximate numbers. These approximate numbers should be taken into metadata to test their fitness. Test includes two aspect: integrity and simplicity. On the one hand, we should ensure that all core authors can be filtered out. On the other hand, we should ensure that authors who published a few papers can be excluded. Through comparative analysis, the most appropriate number can be filtered out.

Matrix analysis is the most common method of Informetrics. Through matrix construction, the relationships and hierarchies of co-occurrence can be showed clearly.

Social network analysis is a method that is used to analyze various relationships and properties in social network. It has been confirmed that social network analysis can be used to analysis authors co-occurrence relationships (Otte, Rousseau, 2002).

5.3 Data Preprocess

Excel is used for simple data process, such as sort, filter, breakdown. VBA (Visual Basic for Applications) is used for complex data process, such as frequency statistics, generating co-occurrence matrix.

Data downloaded from Web of Science are text files. Text files can be imported into Excel. Metadata include many fields, such as: AU, TI, SO. Authors co-occurrence analysis needs the field of AU. Each paper has several authors. Authors in each AU field are separated by semicolons. We use breakdown of Excel to divide authors into single, each author occupies a cell. All authors store into Sheet1. We use VBA TO write a program to adjust authors in columns into one column. We write word frequency statistics program to count the number of each author's published papers. The results of statistics store into Sheet2.

The number of published papers varies from person to person. In the same period, some authors published a lot of papers, yet some only published a few papers. It is difficult to establish the co-occurrence relationships among authors who published a few papers. On the contrary, it is easy to establish the co-occurrence relationships among authors who published many papers. Therefore, it is necessary to select out core authors before analyzing authors co-occurrence.

Price's law is used to filter out the core authors to reduce data redundancy and improve the efficiency of data processing. Two formulas of Price's law are as follows: $\sum_{m}^{i} n(x) \quad \sqrt{N}$ and $N \quad 0.749(\ _{\max})^{1/2}$. The first formula refers to that the number of productive authors $\sum_{m}^{i} n(x)$ is equal to the square root of the number of all authors $\sqrt{N}$. The second formula counts the

number of eminent authors whose published papers are more than N . $_{\max}$ refers to the number of author(s) who published most papers. The calculation results of two formulas are 8 and 4, respectively. According to the method in 5.2 Methods, we analyze the authors who published 4 to 8 papers, and choose authors who published more than 6 papers as core authors. There are 122 core authors.

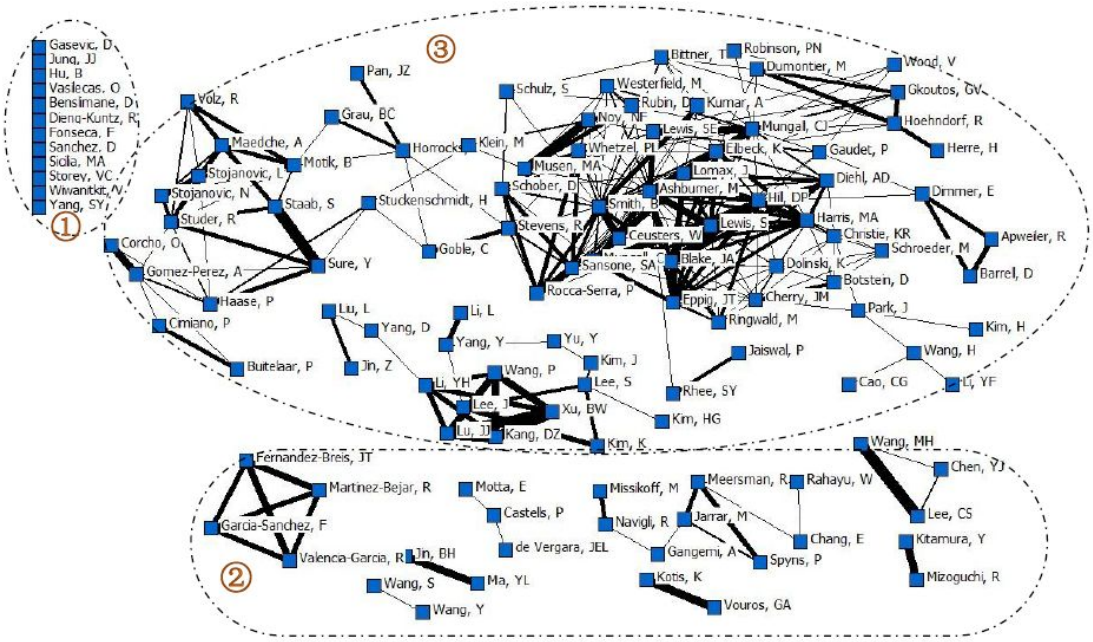
After determining core authors, we write a program based on three-loop algorithm to generate author co-occurrence matrix based on Sheet1 and Sheet2.

	Smith, B	Staab, S	Gomez-Per	Sure, Y	Mungall,	Stevens,	Blake, J	Ashburner	Horrocks,	Meersman,	Mizoguchi	Musen, M	Kim, J	Lee, CS	Maedche,	Apweiler,	Noy, NF	Park, J	Studer, R	Spyns, P
Smith, B	31	0	0	0	2	1	4	2	0	0	0	1	0	0	0	0	1	0	0	0
Staab, S	0	23	0	9	0	0	0	0	0	0	0	0	0	0	4	0	0	0	3	0
Gomez-Per	0	0	20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Sure, Y	0	9	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
Mungall,	2	0	0	0	17	0	2	6	0	0	0	1	0	0	0	0	1	0	0	0
Stevens,	1	0	0	0	0	18	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Blake, JA	4	0	0	0	2	0	15	4	0	0	0	0	0	0	0	0	0	0	0	0
Ashburner	2	0	0	0	6	0	4	14	0	0	0	1	0	0	0	0	1	0	0	0
Horrocks,	0	0	0	0	0	1	0	0	16	0	0	0	0	0	0	0	0	0	0	0
Meersman,	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	5
Mizoguchi	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0
Musen, MA	1	0	0	0	1	0	0	1	0	0	0	14	0	0	0	0	8	0	0	0
Kim, J	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0
Lee, CS	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0
Maedche,	0	4	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	3	0
Apweiler,	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0
Noy, NF	1	0	0	0	1	0	0	1	0	0	0	8	0	0	0	0	13	0	0	0
Park, J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0
Studer, R	0	3	1	4	0	0	0	0	0	0	0	0	0	0	3	0	0	0	13	0
Spyns, P	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	12

In Fig. 9, partial author co-occurrence matrix is listed. It is generated by macros of Excel. The first row and column are author's name. Nonzero values in matrix refer to that the authors on corresponding rows and columns have semantic relationships of author co-occurrence. The values in matrix are the strength of author co-occurrence.

Finally, three tables are created. Metadata table named "Ontology" stores metadata of library resources. Data structure is <TitleID, Title, EnglishTitle, FirstAuthor, ArticleClass, Keywords, Fund, FundClass, Journal, JournalCode, Institutions, Subjects, Year>. Class table named "Authors_Ontology" stores authors of metadata. Data structure is <AuthorID, TitleID, Author>. The primary keys of TitleID in "Ontology" are connected with the foreign keys of TitleID in "Authors_Ontology". Individual table named "Authors_Co-occurrence_Ontology" stores individuals. The table name "Authors_Co-occurrence_Ontology" refers to author co-occurrence analysis of Informetrics is utilized to abstract semantic relationships. Records in "Authors_Co-occurrence_Ontology" have semantic relationships of author co-occurrence. Data structure is <Authors_Co-occurrenceID, AuthorID, Author1, Author2, Frequency>. In each record, Author1 and Author2 have semantic relationships of author co-occurrence. Frequency refers to strength of semantic association. The primary keys of AuthorID in "Authors_Ontology" are connected with the foreign keys of AuthorID in "Authors_Co-occurrence_Ontology".

6. Results



We import author co-occurrence matrix into Ucinet, a graph of author co-occurrence is generated. We choose component analysis and n-clique analysis of Ucinet, semantic relationships of author co-occurrence are divided into three categories. The first category is isolated nodes. Most of them published papers individually. They have not any relationship with other authors. Isolated nodes contribute little to revelation of semantic relationships of library resources. The second category is simple cluster cliques. The number of nodes are 2 to 8. There are 8 simple cluster cliques. Authors in each cluster are no more than 5 generally. The strength of co-occurrence is low. The authors who cooperated with them are in a relatively narrow range. They reveal partial semantic relationships. The third category is complex cluster cliques. The number of nodes are more than 8. There are 2 complex cluster cliques. Each cluster includes many authors. They have strong co-occurrence relationships. Cooperation relationships among them are wide and close. They contribute greatly to revelation of semantic relationships of library resources.

We use the following formula to calculate the number of library resources discovered based on RO and metadata. A_i refers to author, $Frequency$ refers to the frequencies of authors co-occurrence. The occurrences of authors corresponds to the number of published papers. The number of authors is equal to the number of library resources.

$$N_{RO} \sum_{\substack{i=1 \\ j=1}}^{n} (A_i \quad \tilde{A_j} \quad Frequency)$$

$$N_M \sum_{i=1}^{n} A_i$$

The results are shown in Table 2. We choose four groups statistics in order to analyze conveniently. The frequencies of authors co-occurrence are more than 0, 1, 3, and 5, respectively. In Table 2, the values of RO are significantly more than that of Metadata.

Table 2 The number of library resources discovered by RO and Metadata (Partially)

Frequencies > 0		Frequencies > 1		Frequencies > 3		Frequencies > 5	
RO	Metadata	RO	Metadata	RO	Metadata	RO	Metadata
185	17	68	17	34	17	28	17
67	23	67	23	43	23	33	23
46	14	28	14	23	14	19	14
61	15	56	15	39	15	18	15
202	31	118	31	47	31	34	31
133	14	55	14	31	14	16	14
16	9	16	9	16	9	11	9
12	8	12	8	12	8	10	8
12	10	12	10	12	10	12	10
13	10	13	10	13	10	11	10
12	9	12	9	12	9	10	9
12	9	12	9	12	9	10	9
21	14	21	14	15	14	15	14
40	19	33	19	33	19	19	19
22	10	22	10	22	10	10	10
51	14	51	14	24	14	14	14
43	16	26	16	26	16	16	16
102	10	43	10	18	10	10	10
15	7	15	7	14	7	7	7
51	13	43	13	19	13	13	13
13	7	13	7	13	7	7	7
13	7	13	7	13	7	7	7
25	12	17	12	17	12	12	12
15	10	15	10	15	10	10	10
59	18	32	18	22	18	18	18
14	10	14	10	14	10	10	10
46	9	31	9	12	9	9	9
38	8	29	8	11	8	8	8
31	8	11	8	11	8	8	8
11	8	11	8	11	8	8	8

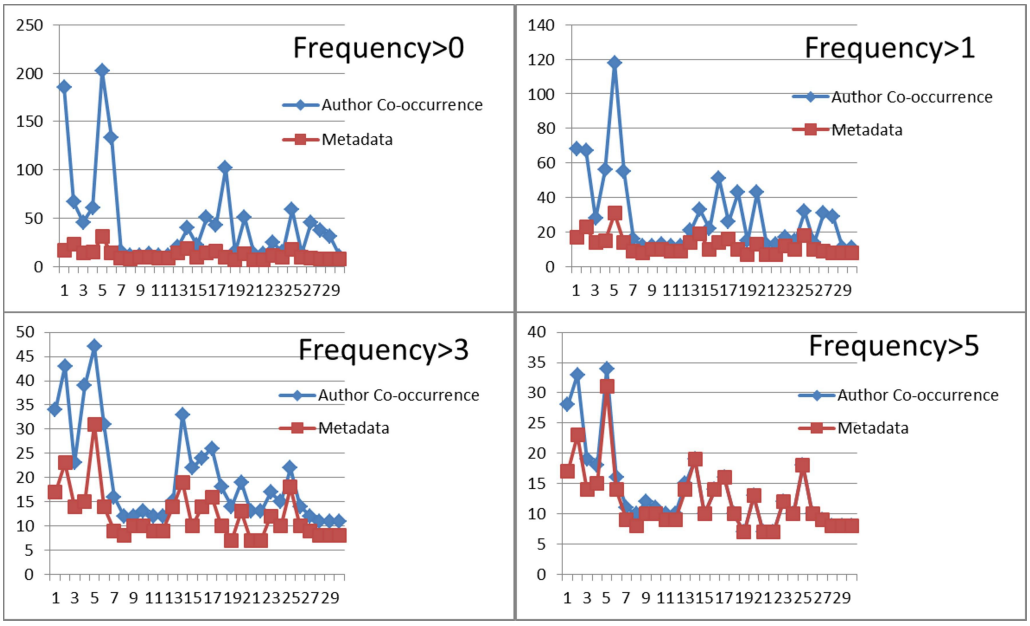


Fig. 11 is generated based on statistics in Table 2. Lateral and longitudinal axis of each graph refers to the number of authors and library resources, respectively. Each graph includes two lines. Author co-occurrence analysis excavates semantic relationships among different authors. Papers published by different authors are connected by semantic relationships of author co-occurrence. The blue line represents the corresponding relationships between papers and authors based on semantic relationships of author co-occurrence. The red line represents the relationships based on metadata. As shown in Fig. 11, the values of blue line are significantly more than that of red line. RO contributes greatly to knowledge organization and discovery of library resources. The differences between two lines decline gradually with the increasing frequency. The lower the frequency is, the greater the differences between two lines are, the lower the relativity among papers is, and vice versa.

7. Conclusions and Further Research

From the above empirical analysis, we can see that methods of Informetrics can reveal semantic relationships of library resources. Classes, properties, and individuals can be stored in relational database easily. If all methods of Informetrics are utilized to do semantic revelation, then complexity of semantic relationships among metadata will be presented. Metadata form a complex semantic network, which contributes greatly to enhance the semantic relationships among library resources. RO not only reveals the deep-level semantic relationships of metadata of library resources, but also realizes totally computer automatic processing. RO can better solve the disadvantages of CDO and MDO.

RO can be used for knowledge organization, knowledge mining. Information organization based on relational two-dimensional table cannot connect data tightly. Knowledge organization mode of RO is a net. It is like the Internet. Each node can connect others through many different paths. Meanwhile, if a single node is broken, the whole net can operate normally. RO is a well-structured organization form. Each node has many neighboring nodes. Each node can expand to the surrounding nodes unlimitedly. New knowledge will be found easily through RO.

Automatic construction of ontology is always a hot topic because of its complex procedures. This paper introduces methods of Informetrics to construct ontology automatically from a new

perspective. However, RO has several limitations: (1) RO is based on metadata. The level of semantic revelation is relatively shallow. (2) Whether the frequencies of co-occurrence and similarity of library resources have positive correlation or not needs to be proved. (3) RO can only store data with only one language, and process data with only one database.

We will focus on the following problems in the future:

(1) Automatic segmentation technology will be introduced to deal with title, abstract. We will store them into RO to improve the level of semantic revelation.

(2) Dice index, cosine index will be introduced to prove the correlation between the frequencies of co-occurrence and similarity of library resources.

(3) Cross-language issue. Realization the language seamless integration of English and Chinese can improve the application scope of RO. Subject classification problem. Different databases, such as: CNKI, CSSCI, WOS, use different subject classification systems. It will hinder the integration of different databases. Different subject classification systems should be unified to realize cross-database data organization.

References

Agrawal R, & Somani A. (2001). Storage and Querying of E-commerce Data, *Proceedings of the 27th VLDB Conference. Roma, Italy*,p.1.10.

Aimin T., Zhen Z., & Jing F. (2005), "New Technology of Library and Information Service", Vol. No. 4, pp. 1-5.

Alexaki S, Christophides V, & Karvounarakis G, et al. (2001). On Storing Voluminous RDF Descriptions: The Case of Web Portal Catalogs, *Proceedings of the 4th International Workshop on the Web and Databases. Santa Barbara, California, USA*,p.1.6.

Berners-Lee T., Hendler J., & Lassila O. (2001), "The semantic web", *Scientific American*, Vol. 284 No. 5, pp. 34-43.

Chengchun D., & Zhu F. (2011), "Research on the Construction of Semi-automation of Domain Ontology Based on aerospace thesaurus", *Journal of Information Theory and Practice*, Vol. 34 No. 11, pp. 113-116.

Chunjuan L., Haishan L., & Baode J. (2010), "Core Journal & Category Distribution of International Patent Studies", *Information Science*, Vol. 28 No. 11, pp. 1689-1692.

Chunlin J., & Yuguang C. (2010), "Transform CSSCI Data to Bibexcel Data to Actualize Co-occurrence Matrix and A Case Study", *Library Journal*, Vol. 29 No. 4, pp. 58-63.

Chunyan L., Shuping C., & Yucheng W. (2007), "The Transformation from Thesaurus to Ontology Based on SKOS", New Technology of Library and Information Service, Vol. No. 3, pp. 32-35.

Dongmei M., & Zhi F. (2007), "Building and Reasoning of Digital Library's Domain Ontology", *Library and Information Service*, Vol. 51 No. 8, pp. 26-30.

Dufeng Z., & Baisong L. (2010). "Automatic Domain-specific Term Extraction in Administrative-domain Ontology", *New Technology of Library and Information Service*, Vol. No. 4, pp. 59-65.

Fan P., Lin L., & Hong W. (2011), "Application of concept lattice in the hierarchies construction of fundamental geographic information ontology", *Science of Surveying and Mapping*, Vol. 36 No. 6, pp. 235-237.

Guohe F., & Jingxue W. (2011), "Visualization Analysis on Domestic Research of Institutional Repositories", *Library and Information Service*, Vol. 55 No. 22, pp. 95-100.

Hua B. (2010). “Some Important Issues for the Ontological Semantic Description of Thesarus in Chinese Language”, Library Journal, Vol. 29 No. 11, pp. 21-25.

Huanhuan C. (2011), “Research on the Construction of Domain Ontology in Library and Information Science”,*Research on Library Science*, Vol. 44 No.21, pp. 11-16.

Hui D., & Chuanming Y. (2005), “Analysis of Automatic Chinese Ontology Learning and Its Evaluation Algorithm”, *Information Studies: Theory & Application,* Vol. 28 NO. 4, pp. 415-418.

Hui D., Chuanming Y., & Ying J., et al.(2006), “Research on the Ontology-based Retrieval Model of Digital Library (Ⅱ)”, *Journal of the China Society for Scientific and Technical Information*, Vol. 25 No. 4, pp. 451-461.

Jing L., & Ping Q. (2004), “The Journal of The Library Science In China”, Vol. 30 No. 149, pp. 36-39.

Jun L., & Bo C. (2010), “Research on Method About Relational Database Storing OWL Ontology”, *Computer Engineering*, Vol. 36 No. 21, pp.71-73.

Jun L., Qi G., & Yi W. (2010), “Ontology Annotation Method Based on Bootstrapping”, *Computer Engineering*, Vol. 36 No. 23, pp.85-87.

Jun M. (2004), “An Algorithm of Automatically Extracting Ontological Concepts from Linear Conceptual Graph”, *Computer Engineering and Applications*, Vol. NO. 23, pp. 161-164.

Junping Q., & Fan Y. (2012), “Theoretical Research on Semantization of Library Resources Based on Informetrics Analysis”, paper presented at Journal of Library Science in China.

Junping Q., & Feifei W. (2010), “Analysis on Author Cooperation Relationship of Competitive Intelligence Research in China based on SNA”, *Library Tribune*, Vol. 30 No. 6, pp. 35-40.

Junping Q., & Hui Q. (2011), “Study on the Knowledge Diffusion in the Network of Chinese Research Institutions”. *Knowledge of Library and Information Science*, Vol. 144 No. 6, pp. 5-11.

Junping Q., & Weihua Z. (2008), “A Metric Demonstration of Journal Co-citation”, *Information Science*, Vol. 26 No. 10, pp. 1447-1450.

Junzhi J. (2007), “On the Conversion of Classified Chinese Thesaurus to an Ontology”. Journal of Library Science In China, Vol. 33 No. 170, pp. 41-44.

Junzhi J., & Rongjuan W. (2009), “Analysis on the Methods of Converting Thesauri intoOntology”, Information Science, Vol. 27 No. 9, pp. 1363-1366.

Man L., Yan W., & Yiyu Z., et. al. (2005),” A study on storage schema of large ontology based on relational database”, *Journal of Huazhong University of Science and Technology*, Vol. 33 No.s1, pp. 217-220.

Ming X., & Guojun L. (2011), “Research on Visualization of Competitive Intelligence in China based on Coupling Analysis and Keyword Analysis”, *Journal of Information Theory and Practice*, Vol. 34 No. 1, pp. 100-102.

Nianyun S., & Chen Y. (2007), “Towards domain ontology-based semantic annotation research”, *Computer Engineering and Design*, Vol. 28 No. 24, pp. 5985-5987.

Ning Z. (2012), “The Research of Knowledge Serialization and Control Based on Linked Data”,*Research on Library Science*, Vol. No. 7, pp. 48-51.

Otte E., & Rosseau R. (2002). “Social network analysis: a powerful strategy, also for the information sciences”, Journal of Information Science, Vol. 28 No. 6, pp. 443-455.

Qiang S., & Yongfu J. (2006), “Knowledge Serialization of Library”, *Library Theory and Practice*, Vol. No. 4, pp. 35-36.

RuiMin M., & Chaoqun N. (2011), "A Study on Intellectual Structure and the Evolution of Library and Information Science in China Based on Author Co-citation Analysis", *Journal of Library Science in China*, Vol. 37 No. 196, pp. 17-26.

Shenqin Y., Jilong Z., & Lei R. (2011), "Research Hotspots Analysis of Digital Library Based on Keywords Co-occurrence Analysis and Social Network Analysis", *Journal of Academic Libraries*, Vol. No. 4, pp. 25-30.

Suihua W., Xiaodan Z., & Yue W. (2011), "An Approach to Building an OWL Ontology Relational Database", *Computer Engineering & Science*, Vol. 33 No. 12, pp. 143-147.

Studer R., Benjamins V R., & Fensel D. (1988), "Knowledge Engineering, Principles and Methods", *Data and knowledge Engineering*, Vol. 25 No. 122, pp. 161-197.

Wenhua D. (2005), "Research on Construction of Ontology and Application for Digital Library", Dissertation in Wuhan University, pp. 9-10.

Wenjuan J., & Feng H. (2011), "Research of Chinese Ontology Learning Based on HowNet", Computer Technology and Development, Vol. 21 No. 6, pp. 77-84.

Wenta Z., Haojie S., & Gong W., et al. (2008), "Research on Sino-Russia Cooperation from Co-authors of SCI", *Forum on Science and Technology in China*, Vol. No. 2, pp. 139-144.

Yanzhang Q., & Liang G. (2011), "Storage and Querying of Large Scale Web Ontology in Relational Database", *Microelectronics & Computer*, Vol. 28 No. 12, pp. 137-140.

Yahong X., & Zhuoming X. (2002), "How to Store RDF in a Relational Database", *Computer and Modernization*, Vol. 16 No. 11, pp. 70-72.

Yewang C., Haibo L., & Jinshan Y. (2012), "A Semantic Retrieval Model Based on Agricultural Field Ontology", *Journal of Huaqiao University (Natural Science)*, Vol. 33 No. 1, pp. 27-32.

Yewang C., Wen L., & Xin P. et. al. (2009), "Improved semantic annotation method for documents based on ontology", *Journal of Southeast University*, Vol. 39 No. 6, pp. 1109-1113.

Yi Y., & Chuanjun S. (2011), "Research Topic and Development Tendency Analysis Based on Information Lifecycle in China", *Journal of Information Theory and Practice*, Vol. 34 No. 10, pp. 17-21.

Yuefen W., Shuang S., & Lu M. (2006), "Application Study of Co-occurrence Analysis in Knowledge Service", *New Technology of Library and Information Service*, Vol. 33 No. 4, pp. 29-34.

Yuefen W., Shuang S., & Minghui X. (2007), "Research on Method of Text-knowledge Mining Based on Co-occurrence Analysis", *Library and Information Service*, Vol. 51 No. 4, pp. 66-70.

Yuefen W., Shuang S., & Ning N., et al. (2007), "Application of Co-Occurrence Analysis in Text knowledge Mining", *Journal of Library Science In China*, Vol. 33 No. 2, pp. 59-64.

Yulian Z., Shuai L., & Xinglin Z. (2009), "Research on Ontology-based Automatic Annotation for Deep Web", New Technology of Library and Information Service, Vol. NO. 9, pp. 45-50.

Yun F., Wenyuan N., & Yunlin W., et al. (2009), "Analysis on the author cooperation network in the field of science", *Science Research Management*, Vol. 30 No. 3, pp. 41-46.Yun X., Dongyi Y., & WenDe Z. (2007), "Journal of Information", Vol. No. 3, pp. 15-18.

Zhuoming X., & Yongjing H. (2006), "Conversion from OWL ontology to relational database schema", *Journal of Hohai University (Natural Sciences)*, Vol. 34 No. 4, pp. 95-99.

Zili Z., & Yanna W. (2011), "Information Content Security Ontology Construction Based on WordNet", Vol. 37 No. 20, pp. 136-138.