Study on Book Ontology in the Digitization Process of Books

Xiao-fei Li
Nanchang Normal University, Nanchang, China
Email:826907721@qq.com

Abstract. In the construction of digital library, it is difficult to find the content information of massive electronic books. Thus, the problem of many islands of information is presented, which hinders the efficient utilization of resources. This paper proposed an ontology framework of the electronic book domain to realize the semantic description of information resources. The KeyGraph algorithm was used to extract the key words of the book content abstract and chapter content of the electronic books, and the corresponding classes in the domain ontology were instantiated with the keyword content. Finally the semantic retrieval model has been proposed for extracting the query keywords. Content abstract and chapter content were obtained from book ontology. Compare query keywords with their similarity to get the final query result.

1. Introduction
With the increasing maturity and popularity of digital libraries, the digitization of library collection resources has also been carried out in some relatively well-funded libraries. According to the survey, libraries in 12 provinces and cities across the country were carrying out the digitization of their collection books. At present, the main objects of digitization are special collections, professional books, books with high rate of circulation and foreign language books. However, there are few libraries carrying out the digitization of all the books in their collection, which may be related to the insufficient digital funds. The service providers represented by Super Star group corporation of China has made great efforts in the digitization of book resources[1]. The Super Star digital library has realized the digitization of millions of books within 22 categories of Chinese library classification and the annual processing volume reaches 300,000, but the deficiency is that the Huiya ebook of their products is the image format book, only the Super Star Book World is the full text format[2]. Some libraries realize the digitization of print books in the form of entrusting the company's processing and scanning, which creates the resource barrier for the construction of digital library based on full-text retrieval in the future. The main goal of libraries carrying out the digitization of library collection is only to reduce the duplication of library collection and ease the crowding of library collection. They do not think about the retrieval and utilization of library resources in a deeper level, and emphasize the preservation and override utilization, so that the utilization rate of electronic books is not high. Compared with print publishing, digital publishing has the inherent advantages of timely updating, fast transmission and saving resources. However, these are only the basic advantages of digital publishing, and there is a lot of space for exploring digital resource retrieval and utilization. This paper proposed a book resource platform model based on domain ontology, which is used to describe the knowledge with ontology, establish the cognitive structure of knowledge points, and realize the convenient
retrieval and utilization of book resource content. This model can be applied in both the digital model of the collection and the business model of book publishing.

2. Semantic description of book resources
In face of the low utilization rate of literature resources of libraries, and the value of the library resources are not excavated, the reason lies in the traditional and extensive resource retrieval method and the single resource organization system. Because of the lack of in-depth disclosure of the inherent characteristics of resources, often due to the inconsistency between users and the concept expression in the resource library, many undiscoverable "resource islands" have been formed[3]. Without systematic organization and teasing out the knowledge resources, it is easy to get lost in information. Therefore, it is necessary to add a standardized semantic information to the resource concept and give a multidimensional logical relationship description between resources. In other words, ontology is adopted to describe resources, and then the relationship between domain ontologies is acquired by linguistic and statistical methods to make resources interconnected and shared[4]. The definition of ontology is an explicit formal specification of shared concepts. Ontology is represented by five-tuples, O=(C, P, V, R, Ru), C represents concept or classe, P represents attribute, V represents instance, R represents relationship, and Ru represents rule. Using ontology to describe resource needs to acquire the domain concept in resource content, and acquire the relationship between concepts, including hierarchical relation, part-whole relation, synonymy relation, etc., which are collectively referred to as ontology learning[5]. Ontology learning methods are based on linguistic method, statistical method and combined methods of the two. Domain ontology refers to the knowledge of a specific domain described by ontology. The conceptual hierarchy system, conceptual semantic relations, reasoning rules and constraints of the domain have characteristics different from those of other domains[6]. Domain ontology construction is generally achieved through machine learning and domain expert artificial construction[7].

3. Construction of book ontology
Ontology components include classes, attributes, relationships, and instances. You need to define class and class attributes, define class relationships, create instances of each class, and assign attribute values. We adopt Protégé as our ontology construction tool, and the established model can be exported to data files in other formats, which is convenient to realize visualization with the synthesis of other information systems[8].

3.1. Define class
This paper takes book as the research object, and the application goal is to realize the retrieval and utilization of book resources, among which the main resource network relations include knowledge - knowledge network, scholar - knowledge network, scholar - scholar network, and scholar – work unit network. Therefore, the basic classes in ontology can be defined as book (attribute: Chinese title, English title, secondary title, version, classification number), author (attribute: name, work unit, team author), chapter catalogue (keyword 1, keyword 2,...), book content abstract, press (attribute: name, publication date), citation, as shown in Figure 1. Instance is the most basic step in ontology construction. Only when the instance is clear, can the class and relationship accurately reflect the ontology[9]. In Figure 1, the oval represents the instance. Knowledge - knowledge network is a kind of hierarchy structure, which shows the complex relation between knowledge and standardizes the mode of knowledge association. As the ontology is constructed by using the domain concept, the ontology is a common and consistent understanding within the domain, so the domain ontology can be Shared. The relationships presented by the domain ontology can reveal the frontier knowledge of the subject, the context of the subject knowledge, the network relationship of the scholar, and the professional expertise of the scholar.
3.2. Define the relationship of the class
Defining inter-class relationships is the key to forming knowledge network diagrams. The main inter-class relationship in the literature ontology can be defined as: author, which refers to the relationship between book name and author. Publishing relationship is defined as published by, the relationship between books and publishers. Chapter is refers to the relationship between book and chapter directory. Chapter keyword relation is defined as chapter_have, it refers to the important keywords contained in the chapter content. Book abstract relation is defined as Abstract_is, it refers to the abstract content of a book. Bibliographic citation relationship is defined as citation_is which refers to bibliographic references. The basic relation dictionaries were shown in Table 1.

Table 1. Basic relation dictionary

| Relations                  | Annotation                        |
|----------------------------|-----------------------------------|
| author                     | Book author                       |
| is_published_by            | Books and publisher               |
| chapter_is                 | specific book chapters            |
| chapter_have               | Chapter content have keywords     |
| abstract_is                | Abstract of book contents          |
| citation_is                | The book’s citation               |

4. Instantiate the book ontology
Instantiation is adding concrete instances to a class. The Chinese title, English title, edition, classification code and other attributes of the book class, the attributes of the author class and the attributes of the publisher can directly import instances from the MARC data of the library collection. The chapter catalog, content abstract and citation need to be instantiated according to the specific book content, which is more complicated. Each chapter in the chapter catalog is a complete unit of knowledge. Each chapter has key words that represent its core idea. In this paper, keywords of each chapter are extracted. Keywords contain domain concepts, and chapter catalog classes are instantiated with chapter keywords as the basis for subsequent similarity matching. The content abstract part represents the basic outline of the book. In this paper, keywords containing domain concepts in the abstract are extracted, and the content abstract class is instantiated with the keyword sets, which is also used as the basis of subsequent similarity matching.

4.1. Keywords extraction
There are several methods to extract text keywords, such as density clustering method and text opportunity detection algorithm KeyGraph. This paper adopts the KeyGraph algorithm, which was first proposed by Japanese scholar Y.Oshawa in 1998. Based on the co-occurrence relationship of
words, the KeyGraph model is established to extract keywords. Compared with other keyword extraction algorithms, low frequency keywords can be found\cite{10,11}. The process of extracting keywords is shown in Figure 2. Firstly, the foundation $G_a, G_b$ are extracted. The roof are the keywords supported by the column.

**Figure 2.** The process of extracting keywords

4.1.1. **Document predisposition.** This article extracts the content summary of the book as a document, and extracts the content of each chapter as a document. After the abstracts of book contents are extracted as documents, the preprocessing of word segmentation, dividing the article into sentences and removal of stop words is carried out. The NLPIR word segmentation tool was used for Chinese word segmentation, with custom domain concept dictionary, in order to correctly identify the domain concept in the document\cite{12}. Set the word to $W \{w_1, w_2, ..., w_{N_W}\}$. The set of sentences is $S \{s_1, s_2, ..., s_{N_S}\}$. After the word segmentation, the inverted index of the word to the sentence is established, which is stored as $WS \{ (w_1, w_2, ..., w_{N_W}), (s_1, s_2, ..., s_{N_S}) \}$ in the form of matrix. $W$ and $s$ respectively represent the rows and columns of the matrix, the row and column values are the frequency of $w$ appearing in $s$. $N_W$ is the number of words, and $N_S$ is the number of sentences.

4.1.2. **Extract high-frequency words.** Words that occur more frequently than the threshold $H_f$ are classified as set $WH_F$. The correlation degree between words is below:

$$assoc(w_i, w_j) = \sum_{D} \min \left( |w_i|, |w_j| \right)$$

$|w_i|$ represents the number of occurrences in the sentence $s$ of document $D$, and takes the minimum co-occurrence frequency as the correlation value. Establishing the co-occurrence index matrix of high-frequency words, $WH_FCon = \{(whf_1, whf_2, ..., whf_{N_{HF}}), (wcon_1, wcon_2, ..., wcon_{N_{W}})\}$. $Whfi$ is a high-frequency word, $wcon_i$ is a co-occurrence word. The row and column values are correlation degrees $assoc(whf_i, wcon_j)$, and $Nh_f$ is the number of high-frequency words.

4.1.3. **Extract the foundation.** The high frequency word connected subgraph is established to extract foundation. Building graph $G$ with high frequency word as vertex. If the critical degree $ASSOC_{ConHF} = assoc(ConHF_j, ConHF_i)$ between the high-frequency words is greater than the threshold value, an edge is established between $ConHF_j, ConHF_i$. The maximum connected subgraph is found in graph $G$, which is divided into several connected subgraphs. Each connected subgraph is the group $g$ that represents the foundation. The co-occurrence word set of $g$ is denoted as $g_{-ConWs}(g)$, and the co-occurrence sentence set is denoted as $g_{-Sen}(g)$.

4.1.4. **Obtain the concept word key.** First, two auxiliary functions are defined:

$$based(w,g) = \begin{cases} \quad \sum_{w \in g_{-ConWs}(g)} |w|, & w \in g_{-ConWs}(g) \\ 0, & w \notin g_{-ConWs}(g) \end{cases}$$
If the document author takes into account all group $G(g \in G)$ that make up the document, the conditional probability that the concept $w$ is used is that:

$$
\text{key}(w) = \text{probability}(w|g) = 1 - \prod_{g' \in \text{neighbors}(g)} \left(1 - \frac{\text{base}(w, g)}{\text{neighbors}(g)}\right)
$$

(4)

$\text{neighbors}(g)$ denotes the proportion of the concept $w$ in the adjacent domain of $g$.

4.1.5. Extract the keywords of roof. The concept that $\text{key}(w)$ value is greater than the threshold value is used as a candidate keyword. The supporting force between the high key-value concept HKW and other high frequency concept words HFW is:

$$
\text{column}(hkw, hfw) = \sum_{g \in D} \min(|hkw|, |hfw|)
$$

(5)

This can be obtained by scanning the high-frequency word co-occurrence index WHFCOn. The column that has a high column value and connects the concept words to more group $g$ is added to $g$. For each high-key concept term, the sum of the values of column is calculated to obtain the score of candidate keywords, which is sorted from high to low according to the score, and the concept term above the threshold is used as the roof namely keyword.

5. Storage and query of domain ontology
Exporting an ontology as an RDF file can be stored as a file or in a database. Large RDF files are typically stored in distributed databases, such as HBase. Jena is used to convert RDF files into triples, and after uploading, the MapReduce module is responsible for combining them into corresponding key types and output them to HBase storage. The query of domain ontology adopts SPARQL query language based on HBase API, which converts the query statement of triad mode into HBase API for query, and returns the result of triad.

The syntactic analysis tool is used to word segmentation and label the query statement, and the KeyGraph algorithm is used to extract the key words. The key words are formed into query vectors, and the content abstract and chapter catalogue in book ontology are extracted to form vectors. Calculating the cosine similarity between two vectors using this formula:

$$
\cos \theta = \frac{a \cdot b}{\|a\| \|b\|}
$$

(6)

Comparing the similarity with the summary first, and if the threshold is exceeded then continue to compare the similarity with the chapter catalogue. The query results are sorted by similarity. When the user clicks the link to the query results, the specific chapter content will be opened. The semantic retrieval model is shown in Figure 3.
6. Conclusion
This paper proposed the domain ontology model of electronic books based on domain knowledge. The process of ontology construction was presented from the perspective of class, relation and instantiation of ontology. Instantiating the relevant class by extracting key words those represent the central idea of the book summary and the content of the chapter. The query results were obtained by comparing the similarity between the query keywords, the book abstract and chapter content.

Acknowledgments
This paper was supported by Science and Technology Project of Jiangxi provincial department of education. The project name is "Research on Construction and Visualization of Digital Library Domain Ontology", (Grant numbers: GJJ181089).

References
[1] Wan Yajuan. Current situation and development strategy of digitization of books in library. Library Theory and Practice, 2015, (10): 10-13.
[2] Super star digital library. http://about.chaoxing.com.
[3] Yang Xuanhui, Long fan. Study on library digital resource service model based on ontology learning. Research on Library Science, 2016, (9): 90-94.
[4] Ou Shiyan. Semantic Representation of Chinese Thesauri. Library and Information Service, 2015, (16): 110-118.
[5] Huang xuemei, Li Baiyang. The Construction of Specific Area of Web Archive Based on Domain Ontology. Research on Library Science, 2015, (9): 13-18.
[6] Li Zilong, Wang Daopin, Guan Zhongxin. Research on the Method of User Interest Model Construction Based on Domain Ontology. Information Science, 2015, (11): 69-73.
[7] Ding Guozhu, Yu Shenquan. Research on Construction of Discipline Domain Ontology base on Skeleton Ontology and Ontology Learning——Taking the Construction of Chinese Discipline Domain Ontology in Learning Cell as an Example. China Educational Technology, 2015, (3): 81-89, 124.
[8] QIAO Li-hong, DONG Wei. Research on semantic retrieval and decision technology based on ontology. Aviation Precision Manufacturing Technology, 2010, 46(2): 45-49.
[9] PENG Weiying, SONG Jihua. Research on Zizhi Tongjian Historical Ontology Construction and Application. Journal of Chinese Information Processing, 2010, 24(2): 33-38.
[10] Bai Ruijiang. Study on long sentence query extension based on KeyGraph keyword extraction. Information Studies: Theory & Application, 2014, 37(6): 123-132.
[11] Jin Xin. Identification of common and non-consensus tags based on text opportunity discovery. Northeastern University, 2011.
[12] Zhang Huapin. NLPIR Chinese word segmentation system. http://ictclas.nlpir.org.