IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 23 July 2023, date of publication 2 August 2023, date of current version 11 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301134

RESEARCH ARTICLE

Normalized Storage Model Construction and Query Optimization of Book Multi-Source Heterogeneous Massive Data

DAILIN WANG[1,*], LINA LIU[2,*], AND YALI LIU[1]

[1]Library, Northeast Forestry University, Harbin, Heilongjiang 150040, China
[2]College of Computer and Control Engineering, Northeast Forestry University, Harbin, Heilongjiang 150040, China

Corresponding author: Lina Liu (lln@nefu.edu.cn)

This work was supported in part by the Research Project on Higher Education Teaching Reform in Undergraduate Universities in Heilongjiang Province under Grant SJGY20220137, and in part by the National Natural Science Foundation of China under Grant 71473034.

*Dailin Wang and Lina Liu are co-first authors.

ABSTRACT According to the characteristics of massive, multi-source, heterogeneous, and rapid growth of book literature data information from the perspective of the metaverse, in order to meet the requirements of efficient management and rapid retrieval such as standardized storage, effective extraction, and scientific library construction for unstructured massive and heterogeneous book in-formation, this study focuses on the normalization of multi-source heterogeneous massive book data, the construction of a warehouse model for book data in the metaverse perspective, and the query and optimization of book data. Systematic research and implementation were conducted to solve the problem of how to process, manage, and query multi-source heterogeneous massive book data in the metaverse, improving the utilization value and query efficiency of the data. This study utilized the semi-structured features of book text data to construct an extraction rule model for heterogeneous book data, and effectively extracted massive heterogeneous book information. Based on the HBase distributed storage structure and parallel computing technology, the storage scheme has been optimized and query efficiency has been improved to ensure efficient management and retrieval of massive heterogeneous book data. The experimental results show that compared with traditional methods, there are significant improvements in multiple aspects such as the accuracy and recall rate of book text data extraction, the management methods and query efficiency of book information.

INDEX TERMS Heterogeneous information, multi-source book data, extraction model, HBase distributed storage, query optimization.

I. INTRODUCTION

The construction of the metaverse is a complex process involving multiple fields, including massive data storage and query. The combination of Metaverse and massive data storage and query technology can not only solve the storage and management problems of digital assets, but also provide more functions and possibilities for the Metaverse platform, thus creating a more intelligent and efficient digital world. With the advent of the metaverse era, book literature information presents a large-scale, multivariate, and heterogeneous growth trend. The processing of this kind of unstructured and heterogeneous book data information, especially the standardized storage, effective extraction and scientific library construction of book information, has become the primary issue of big data processing and analysis of book information. Traditional document extraction, sorting, sorting and storage retrieval methods can no longer meet the needs of book data information management and fast retrieval.

Information extraction aims to cope with the challenges brought about by the explosive growth of information, to help

The associate editor coordinating the review of this manuscript and approving it for publication was Chong Leong Gan.

people use massive amounts of information more effectively, and to fully tap the value of information. Text information extraction technology is to automatically extract relevant or specific types of information from text, which is an important part of the field of natural language processing. Currently, the mainstream text information extraction model is the hidden Markov model based on the maximum entropy method. Zhang et al. [1] proposed to use the maximum entropy model and the rule-based method to classify electronic medical records by treatment type, and use the support vector machine (SVM) model to deduplicate the first course records and perform automatic differential analysis. Guo and Bao [2] used the six-tuple optimization Hidden Markov Model to determine the text key information to be extracted based on the solved maximum probability state sequence and design the unstructured text key information extraction model. Zhu and Qiu [3] used the self-encoding network to reduce the dimensionality of high-dimensional text in-formation, and clustered the text information on the basis of the similarity between words and text, and integrated the HMM (Hidden Markov Model) in machine learning Extraction tasks applied to different text information.

With the rapid development of Internet commerce in recent years, the research on the extraction of commodity information of web pages has become a hotspot, and many researches [4], [5], [6], [7], [8] mainly focus on the information extraction of semi-structured text on web pages, but there are really not many bibliographic information extraction methods for the characteristics of semi-structured book web pages. At present, the research work on web page information extraction mainly includes the following aspects:

(1) The method based on the DOM tree of the web page. Ban [9] took the forum webpage as the research object, proposed a depth-weighted DOM subtree similarity algorithm to extract comment information, and compared the extracted comment information with the standard value to improve the extraction accuracy.

(2) Visual feature based approach. Wu [10] used visual block center of gravity offsets to locate the data region, and used spectral clustering algorithm to find clusters of structurally similar nodes and localized the data with the text diversity. Wang [11] proposed WEMLVF, a visual feature based web page information extraction framework using supervised machine learning to extract information from forum websites and news review websites.

(3) Based on statistical methods. Zhen and Zhang [12] considered the local information and global information of word features, and used three methods of range, variation coefficient and deformation KL divergence to measure the importance of words from a global perspective, and proposed a text based on statistical range and variation coefficient. feature extraction method. Xie [13] proposed a semi-supervised Chinese key phrase extraction model, which uses a pre-trained language model to represent phrases and articles, and combines a number of statistical features to further improve the accuracy of phrase evaluation.

(4) Methods based on deep learning. Liu [14] proposed a core short text ex-traction model based on knowledge association, using MRF (Markov random field) to ex-tract the relationship and distribution contained in the event, and using the gradient method to extract a small amount of short text to maximize the relationship between events. Lai and Hong [15] proposed a deep learning network model based on rule constraints to solve the problem of performance degradation due to insufficient training samples in text information extraction. Shen et al. [16] proposed a fine-grained hierarchical generative confrontation network, designed stacked hierarchical modules, namely the spatial affine generation module and the cumulative combination module, and fully ''mined'' the se-mantic features of the text by using a multi-dimensional text feature extractor. Sarkhel et al. [17] et al. pseudo-tagged a large number of untagged web pages using some manually tagged pages,and then self-trained a transferable web extraction model on both manually tagged and pseudo-tagged samples to achieve efficient extraction of valid information from HTML documents.

The above research on web page information extraction has made great progress, but the traditional method relies on a specific data set, its robustness is poor, and the extraction efficiency is not high; the method based on visual features is easily affected by the structure of the web page and the interference of web page noise; Statistics-based methods have certain requirements on the quantity and quality of data, and need enough and high-quality sample data as input. In addition, supervised or semi-supervised extraction models need to have data annotations; although DOM tree-based methods have high precision, However, the entire web page needs to be parsed, and its extraction efficiency is low; although the deep learning method can extract text information well on the specified data set, its interpretability is poor and its generalization performance is poor. When it is applied to data with different structures Underperformed.

In view of the problems of the above web page information extraction algorithms, such as poor algorithm adaptability, great influence by web page structure, low extraction efficiency, unsatisfactory effect, and low accuracy rate of data collection recall, at the same time, considering multiple sources of data: from web pages, pdf documents, word document, ppt document; data heterogeneity: structured, semi-structured and unstructured data; massive data: PB, TB, GB-level data capacity, normalized storage requirements for massive multi-source heterogeneous data, this paper proposes a method based on heterogeneous book data extraction and normalized storage, the main research contents are as follows:

(1) According to the semi-structured characteristics of web page book information, analyze and model, and build a book data extraction rule model. Design an effective regular expression, set a specific threshold, calculate the similarity between each paragraph of text and the title of the book, and extract the most similar paragraphs as bibliographic information. Design a reasonable information

extraction algorithm, perform positioning and interception through keyword matching and identification, obtain field + content key-value pairs, use the backtracking algorithm to perform multiple backtracking, and improve the accuracy and recall rate of extracted data.

(2) Aiming at multi-source heterogeneous data of book information: web pages, pdf documents, word documents, ppt documents, a corresponding conversion method into text file.txt is proposed. Perform standardized storage management for multi-source heterogeneous data, and provide basic data for upper-layer query and other application ser-vices.

(3) For the massive data of book information, a data storage scheme based on HBase is proposed. Utilizing the characteristics of columnar storage in distributed non-relational databases, the storage cost of massive sparse book data is reduced, and the query efficiency is optimized with the help of parallel computing technology to meet the storage and query requirements of massive book text information.

In the metaverse era, the growth trend of book literature information is even more rapid. Using the heterogeneous book data extraction rule model based on semi-structured features proposed, which can effectively process multi-source and heterogeneous book in-formation and build a scientific library. In this process, the optimization scheme of distributed storage and parallel computing can ensure the efficient management and retrieval of massive book data. This processing method can not only improve the accuracy and recall rate of book text data extraction, but also improve the efficiency of book information management and query use, providing strong support for book information processing and analysis in the metaverse era.

II. MATERIALS AND MERHODS

A. BOOK INFORMATION EXTRACTION RULE MODEL

Text information extraction is a text information mining technology that targets semi-structured or unstructured natural language texts. It is used to automatically extract the information that users care about from text paragraphs and convert it into structured information. Book text is a typical semi-structured text, which is between fully structured text and unstructured text, and can be expressed as a collection of several information items.

1) BOOK INFORMATION REPRESENTATION

In this paper, each book information is first represented as a collection of information items. Each item represents an information content with independent semantic content and only one aspect. Each information item is represented by a noun or a combination of nouns, as shown in Figure 1.

In the design of the book data extraction rule model, firstly, it is clear that the set of target information items to be extracted is {title, ISBN, author, publisher, publishing time, category, page number, word number, content introduction, catalogue, book resources, book characteristics } 12 target information items, of which {title, ISBN, author, publisher, publishing time, category, page number, word number } 8 items are defined as basic in-formation items, {content introduction, catalogue, book resources, book characteristics } 4 items are defined as detailed information items. In fact, each book information will contain some of the target information items. The establishment of extraction rules makes full use of the semi-structured features of book information text, considering the influence of its flexible writing format on the extraction algorithm model. Through the analysis and summary of the structural characteristics of book information text, the structural characteristics of book information text include the strong and weak identifiers of book information, layout features, separators and local features of some information items.

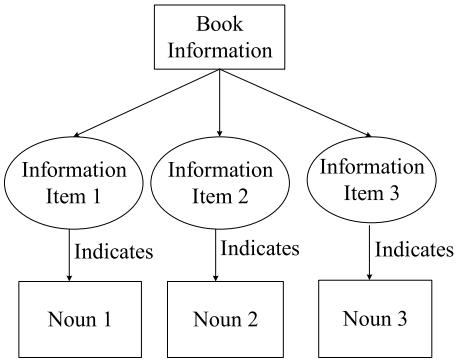

FIGURE 1. Book information item representation.

2) RULE MODEL DESIGN

Starting from the actual needs of the book data extraction scene, the extracted objects include documents in four formats: html, word, pdf, ppt. In order to unify the format and remove the text format and layout differences inside the document, the first step is to uniformly format various documents, use different conversion methods according to different document types, and convert them into.txt text format uniformly. During the conversion process, abnormal conditions such as encryption, damage, and empty documents in the document are handled differently, as shown in Figure 2.

Convert html format to txt document, use urllib.request+re to achieve. First use open to open the file, urllib.request reads the entire webpage, and the read webpage needs to be encoded to become a str type, and a regular expression is used for matching. When re.findall matches a string that meets the conditions, it returns a List, organize all items into a string by traversing the list, use write to write str and then close the file.

Word format is converted to txt document, and the third-party Python tool win32com is used for format conversion. docCom.Documents. Open opens the doc file, doc. Save aa saves it as a txt file, judges the status of the document according to the read content, and directly copies the encrypted document to the preset folder, and writes the unencrypted document into txt and stores it in the specified Folders awaiting extraction processing.

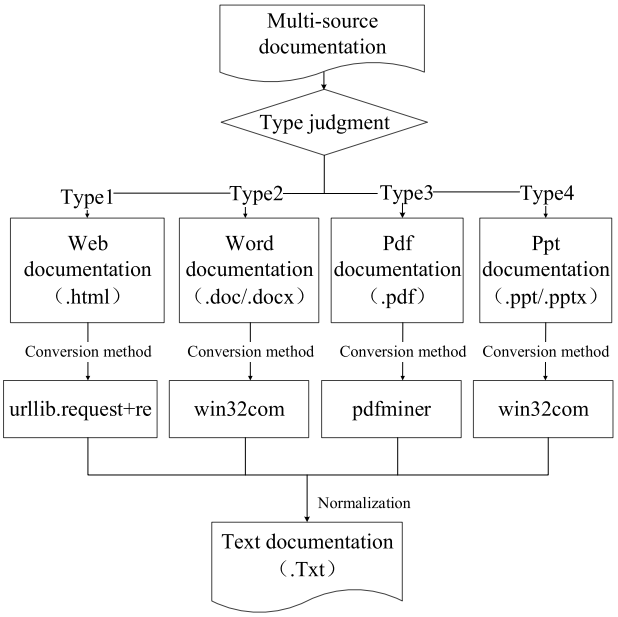

FIGURE 2. Flow chart of normalization preprocessing for multiple source documents.

Convert pdf format to txt document, using pdfminer tool. Import the corresponding packages: PDFResourceManager, PDFPageInterpreter, TextConverter, LAParams, import PDFPage, use TextConverter to convert pdf documents into documents.

To convert ppt format to txt document, the third-party Python tool win32com is also used for format conversion. win32com.client imports Dispatch, constants, pptCom.Presentations. Open opens the ppt file, reads the characters of the ppt file and writes them to the txt file in a loop.

In the preprocessing stage, in the process of converting the html format of web pages to txt files, the extraction of effective book information is involved. The information in the web pages is divided into two categories, one is valid book information, and the other is various information that modifies the information. Class web page tags, the basic frame-work of HTML web pages are:

```
<html>
  <head>
    <meta charset=“utf-8”>
    <title>Title of webpage</title>
  </head>
  <body>
    <h1> Main title of the page </h1>
    <p> This is the main content of the page </p>
  </body>
</html>
```

Among them, <body> is the main body of the html web page, which stores text, tables, pictures, hyperlinks, forms and other content in the web page. In <body>, use tags such as <table></table>, , <tr></tr>, <td></td> to define objects such as tables and rows. Use tags such as <h1></h1>, , , <u></u>, <font>,
 to format objects and highlight information display.

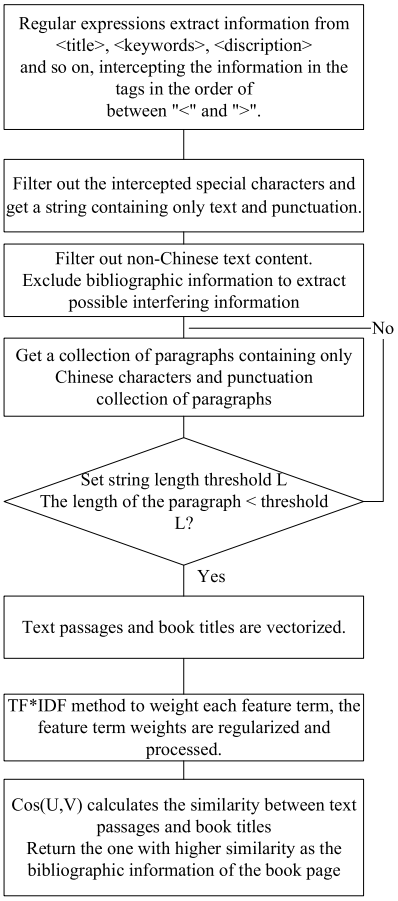

FIGURE 3. Flow chart of book page information extraction.

Through the analysis of the structure of the HTML web page, the format tags do not contain the main information of the web page. In the preprocessing stage, the format tags that have no practical meaning should be filtered out first, and the regular matching method should be used to sequentially traverse each paragraph to filter out the format tags to avoid Noise interference is caused when information is extracted. Object definition tags such as tables, images, etc. contain a small amount of web page information and can be reserved. In the main tag <body> of HTML, <title>, <keywords>, and <discription> mainly introduce bibliographic information such as the name of the book, the author of the book, and the publishing house. The content in these tags highly summarizes the main content of the webpage, and this part is the key point to be extracted. After the preprocessing of removing format tags, the book page information extraction process is shown in Figure 3.

$$w(t_j, d_i) = \frac{tf_i \times \log(\frac{|D|}{df_i} + 1)}{\sqrt{\sum_{i=1}^{n} (tf_i \times \log(\frac{|D|}{df_i} + 1))^2}} \quad (1)$$

Among them, represents the frequency of feature words appearing in the current text, |D| represents the total number of texts, and represents the frequency of feature words appearing in all texts [17]

$$\cos(U, V) = \frac{\sum_{i=1'}^{n} W_{ui} \times W_{vi}}{\sqrt{\sum_{i=1'}^{n} W_{ui}^2} \times \sqrt{\sum_{i=1'}^{n} W_{vi}^2}} \quad (2)$$

After the unified formatting process, for the convenience of data extraction, the second step is to split the.txt text segment twice. For the first segmentation, it is divided into two parts, the basic information block and the detailed information block, according to the characteristics of the book text. The basic information item of a book document is often at the beginning of the document, and the segmentation mark of rough segmentation is the keyword of the first detailed information item appearing in the text. Fine segmentation is to divide the text information into short text segments in units of target information items as much as possible, so as to reduce the impact of noise in the information extraction stage. The segmentation basis of the fine segmentation stage is the layout characteristics, separators, local characteristics of information items, etc. that are analyzed and summarized from a large number of book documents. The schematic diagram of the segmentation process is shown in Figure 4.

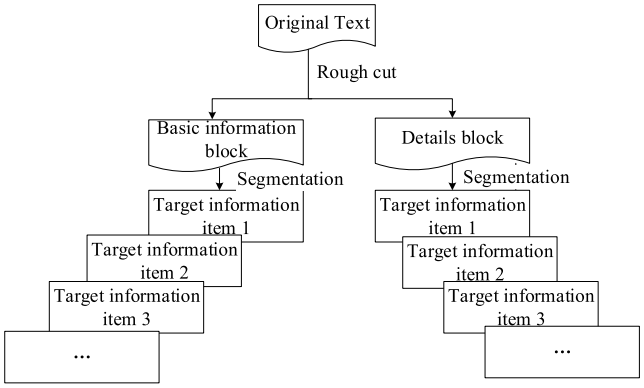

FIGURE 4. Segmentation of book text.

The third step is to extract the book information. The implementation of the data ex-traction algorithm is analyzed and designed on the basis of the segmentation results, as shown in Figure 4. Also according to the characteristics and types of information items, it is divided into strong and weak identification information for matching extraction. In actual book texts, the boundaries between different types of information are not very obvious, and there are overlaps between them. Therefore, the identification and location of information and the design of extraction algorithms must be layered, emphasizing the role of markers, and considering the weak markers. Positioning, algorithm design is shown in Figure 5.

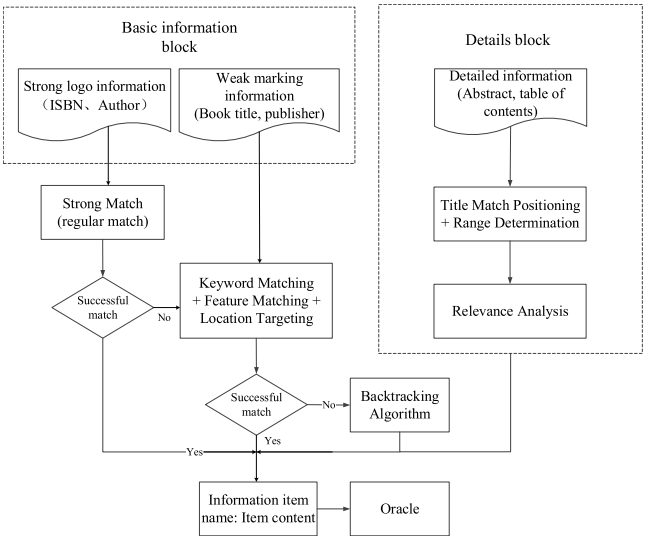

FIGURE 5. Flow chart of extraction algorithm.

In the whole extraction algorithm design, for the basic information block, the strong identification information is directly matched and extracted by regular strong matching method because of its obvious characteristic words or formats; the weak identification in-formation is carried out according to the method of keyword + feature matching + position positioning Match extraction. Detailed information blocks often appear in the form of sub-titles + natural text segments, so the title is used to initially locate the scope, and then the correlation analysis is used to make the final judgment on the relevant text segments.

The extraction of information items in the algorithm, through the matching and identification of keywords, combined with various characteristics of book data for positioning and interception, combined with algorithm extraction to obtain the content information corresponding to different required fields, and obtained key-value pairs of field + content, in order to avoid the segmentation process Errors and omissions in the matching process affect the extraction of book text information. The backtracking algorithm is used to per-form multiple backtracking to improve the accuracy and recall of the extracted data, and finally the extracted data is formatted and stored in the Oracle database.

B. MASSIVE BOOK DATA STORAGE AND QUERY

1) SYSTEM DATA STORAGE FRAMEWORK

From the analysis of data processing flow, the data processing and storage process of the system is divided into six steps: information acquisition, information extraction, data temporary storage (Oracle), data migration, data permanent storage (HBase), and data application.

(1) Information acquisition

Document acquisition is mainly to acquire the original document data processed by the system. As the original data of the system, it mainly exists in the form of documents, and

the sources of documents are web pages, word documents, pdf documents and ppt documents. Among them, the main source of massive book information is web pages, mainly by using Python to crawl book information from websites such as Dangdang, JD.com, and Amazon.

(2) Information extraction

The information extraction part is to use the automatic processing of computer calculation, use the text information extraction algorithm to match, identify and extract the information in the book documents, extract the book information, and form structured data for the next step of storage and query management.

(3) Data temporary storage (Oracle)

The temporary data storage object is mainly the structured data composed of the ex-traction results in the information extraction stage. The storage method is Oracle storage, which is convenient for the fast storage of the extraction results and facilitates the data migration operation.

(4) Data Migration

In order to deal with the problem of storing and querying massive book data, it is necessary to migrate data from the temporarily stored Oracle database to the distributed database HBase. The owner of the data migration needs to use the third-party tool Sqoop to periodically migrate the temporarily stored data in the Oracle database table. Prepare for effective storage and efficient query of massive data.

(5) Permanent data storage (HBase)

The last step of data storage is that the data needs to be stored in the distributed non-relational database HBase, and the data can be fully and effectively utilized with the help of the storage space and efficient retrieval and query capabilities of the distributed system.

(6) Data application

The data application here is mainly the application of data query. At this stage, the book information that has been extracted and stored in Hbase will be quickly and efficiently retrieved from the massive data through the user's query on the Web side.

In order to realize the distributed storage and access of massive book data, the com-bination of Oracle and HBase is adopted for the entire storage part in view of the sparse characteristics of the book extraction results. In order to speed up the extraction speed, the extraction results are temporarily stored in the Oracle database, and the data is finally migrated to HBase for persistent storage with the help of the distributed tool Sqoop, and the HBase database table is designed according to the characteristics of the book data to optimize the efficient retrieval efficiency. According to the demand analysis of system data storage and efficient query, the source file storage is directly stored in the server disk in the form of documents without excessive processing. In addition, the data storage processing of the entire system is designed into 4 modules: Oracle storage module, Sqoop Data mgration module, HBase storage module, Thrift-based HBase connection module. The system storage architecture design is shown in Figure 6.

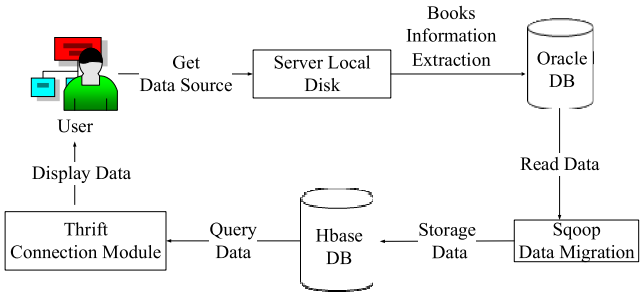

FIGURE 6. Architecture of system data storage.

2) BOOK DATA PRE-STORAGE

As the original data of the system, book documents can be obtained through various channels. The book information extraction algorithm model in this paper is used to extract information from book documents and obtain target item information. In order to facilitate the fast storage of the extraction results, temporary storage is first performed in Oracle to ensure the efficiency of the extraction process and the phased storage of data. According to the demand analysis, the data entities contained in the whole system include user entities, book entities, historical data entities, etc. The table structure of the system database is shown in Table 1-3.

TABLE 1. User.

NAME	TYPE	LENGTH	NULL	PRIMARYKEY	NOTE
ID	VARCHAR2	20	N	Y	MARKING
USERNAME	VARCHAR2	20	N	N	NAME
PASSWORD	VARCHAR2	10	Y	N	CODE
AUTHORITY	NUMBER	1	Y	N	PERMISSIONS
TELNUM	NUMBER	11	Y	N	PHONE

TABLE 2. Book.

NAME	TYPE	LENGTH	NULL	PRIMARY KEY	NOTE
ID	VARCHAR2	20	N	Y	MARKING
BOOK NAME	VARCHAR2	20	N	N	BOOK TITLE
AUTHOR	VARCHAR2	40	N	N	AUTHOR
PUBLISHER	VARCHAR2	40	N	N	PUBLISHER
DATE	DATE	DEFAULT	Y	N	PUBLICATION DATE
PRICE	NUMBER	4,1	Y	N	PRICE
CONTENT	CLOB	DEFAULT	Y	N	CONTENT INTRODUCTION

TABLE 3. Historical data.

NAME	TYPE	LENGTH	NULL	PRIMARY KEY	NOTE
ID	VARCHAR2	20	N	Y	MARKING
TIME	DATE	DEFAULT	Y	N	STATISTICS TIME
TYPE	CHAR	1	Y	N	FILE TYPE
TOTAL	VARCHAR2	40	N	N	TOTAL

3) BOOK DATA MIGRATION

In order to finally reduce the storage cost of massive book data and improve the query efficiency of data, the data needs to be stored persistently in HBase, so the data needs to be migrated from Oracle to HBase, and data migration from Oracle to HBase involves schema migration and data migration Two ways.

In the schema migration, firstly, the metadata of the database needs to be obtained according to the specified database, and the tags and metadata of each table are obtained through the original data. For key information, perform row key and column family de-sign as required, and create corresponding tables in HBase. In the relational database Oracle data table, there is foreign key information, and there is an association relationship between tables, while the HBase table is an independent table structure, and there is no foreign key and association relationship between tables, so in Oracle to When HBase per-forms schema migration, consider how to convert the relationship between tables in Oracle to a table in HBase database. Based on the above considerations, the schema migration design for data migration is shown in Figure 7.

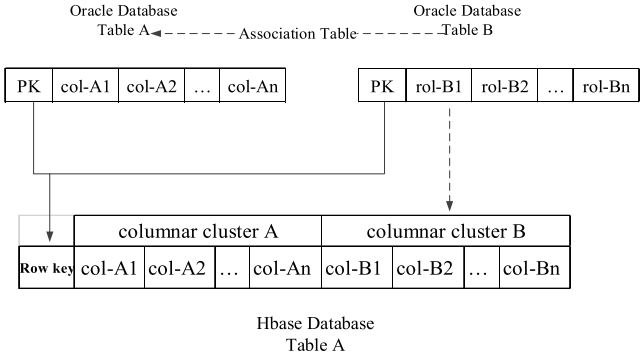

FIGURE 7. Schematic diagram of mode conversion.

The specific conversion rules are as follows:

The Oracle table name is used as the table name of the HBase data table, and at the same time, the table name is used as the first column family name of the HBase data table, and the table field corresponds to the column name of the column family;

The table name of its associated table is used as the column family name of the second column family, and its table fields are used as the column name of the second column family.

In this way, the relationship between tables can be recorded between the column families of HBase. In addition, in order to ensure fast and efficient query, the primary key or index column of the associated table is combined as the row key of HBase.

4) STORAGE MODEL AND QUERY OPTIMIZATION

The storage model design focuses on solving the storage structure of the data table and the optimization of query performance. The row key field in the HBase data table is similar to the primary key of a relational database table and is unique. Each row key corresponds to a data record. HBase distributed storage is sorted by the row key and divided into different Regions Thus realizing distributed data storage. At the same time, HBase provides three query methods: row key single-point query, row key range query, and full table scan.

To sum up, HBase uses a B+ tree model for storage, and the retrieval method based on the primary key is more efficient. Therefore, to design the primary key model reasonably, we must fully consider the retrieval efficiency and also consider the storage performance to avoid reading and writing hotspots. Therefore, in the storage design, split the data into different regions as much as possible to improve the parallel query capability. According to the above design ideas, the primary key design of the HBase table adopts the composite primary key model of the combination of the main query fields, and sets the identification prefix before the primary key to ensure the performance of distributed storage. The design strategy of the primary key model in this paper is shown in Table 4. In this design strategy, the primary key is composed of the identification prefix time and the basic information fields of the book.

TABLE 4. Composite primary key.

MARKING PREFIX	BOOK TITLE	ISBN	AUTHOR
2020-12	C PROGRAMMING LANGUAGE	9787302481447	HAOQIANG TAN
2020-11	MACHINE LEARNING	9787302423287	ZHIHUA ZHOU
2019-12	COMPUTER VISION: A MODERN APPROACH	9787121276170	DAVID A. FORSYTH
2019-08	LEARNING OPENCV 3	9781491937990	ADRIAN KAEHLER

III. EXPERIMENTAL TEST

In the experiment, 3,000 books including html, word, pdf, and ppt formats were used to test the information extraction performance of book information and the storage and query performance of massive book data, and to verify the efficiency of automated management of library information materials proposed in this paper. Among them, 2,500 books are used as basic data for feature collection, which is used to train and debug the model, and another 500 books are used

as test samples to evaluate the extraction performance of the model. For efficient query based on HBase, a certain amount of book data is stored in Oracle and HBase as basic data for performance comparison.

A. EXPERIMENTAL ENVIRONMENT

The test environment in this paper includes two parts, one is a local server, and the other is a distributed HBase storage system built using distributed clusters. HBase is implemented based on Hadoop, and the amount of experimental data is not very large. Therefore, the experimental environment is four nodes virtualized by a physical machine, including a master node and three slave nodes. The node configurations are all consistent, as shown in Table 5.

TABLE 5. Virtual node configuration.

PROJECTS	CONFIGURATION
CPU	INTEL(R) CORE (TM) I7-9700
MEMORY	16.00G
VIDEO CARDS	AMD RADEON R7 M260
HARD DISK	ST500LM021-1KJ152 (500GB)
MOTHERBOARD	20DCA01PCD(SDK0E50518 STD)
NIC	INTEL(R)ETHERNET CONNECTION(3)I218-V
DISPLAY	LEN:A640 RESOLUTION:1920X1080
SYSTEM	WINDOWS 10 64BIT

TABLE 6. Physical machine configuration.

ITEM	ACCURACY（P）/%	RECALLRATE（R）/%	COMPREHENSIVE INDEX（F）/%
BOOK TITLE	91.41	92.15	91.78
AUTHOR	93.52	91.02	92.27
ISBN	94.24	94.84	94.54
PUBLISHER	90.35	86.55	88.45
ELSE	86.32	85.25	85.76

The configuration of a single physical machine is shown in Table 6:

B. SYSTEM TEST ARCHITECTURE

Compared with the tag-based book search system architecture in Reference 20, combined with the core algorithm of this paper, the system architecture is designed and tested. The test architecture of the system is shown in Figure 8.

C. EXPERIMENTAL RESULT

In this system, according to the core functional requirements, the test experiment is divided into two parts: the book information extraction experiment and the massive book data query performance experiment.

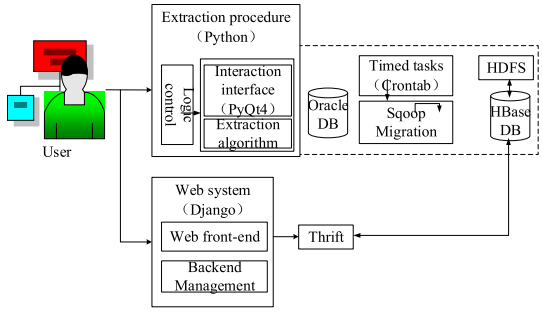

FIGURE 8. Architecture of test.

1) BOOK INFORMATION EXTRACTION EXPERIMENT

The book information extraction experiment is mainly an experiment to test the query performance of the model from the perspective of accuracy. The experimental steps are as follows:

① The experimental data is the 500 test samples of books mentioned above, including book documents in four formats: word, pdf, ppt, and txt. These book documents are tagged by counting the total number of information items in each book and tagging them as C3. In addition, the total number of encrypted or damaged books that cannot be processed normally is recorded as Abnormal_number.

② Use the automatic extraction program to extract the test books, and count the total number of information items C2 extracted from each book and the number of correctly ex-tracted information items C1.

③ According to the formula, calculate the accuracy rate P=C1/C2 and the recall rate R=C1/C3 for each non-abnormal book, and since there is more than one test book, the P and R indicators are respectively averaged and averaged Encrypted or damaged parts are not considered. The β value in the comprehensive index F is taken as 1 here, assuming that the recall rate and accuracy rate are of the same importance, calculate $F = (\beta^2+1)PR/(\beta^2P+R)$.

The experimental test data includes 500 books, including 10 encrypted or damaged book information, and 490 actual book information. The performance evaluation indicators: precision (P), recall (R) and comprehensive indicators (F) are used for statistics. The results of the information extraction experiments are shown in Table 7.

TABLE 7. Experimental results of information extraction.

PROJECT	CONFIGURATION
CPU	INTEL(R) CORE(TM) I7-9700
MEMORY	2G
HARD DISK	60G
SYSTEM	CENTOS6.5
JDK	1.7
HADOOP VERSION	2.5.2
DEVELOPMENT LANGUAGES	JAVA

The experimental results show that the extractions of authors and the ISBN are relatively better, with an accuracy rate of 93.52% and a recall rate of 91.02 for authors, and an

TABLE 8. Comparing the precision by each model.

Extracting Information Items	1	2	3	4	5	6	7	8	9	10
Xiangdong Li[15?]	82.25	83.25	82.57	81.36	81.14	79.35	78.57	78.14	76.12	75.09
Juan Lai[16?]	88.54	88.25	89.25	88.27	87.68	87.94	88.35	84.36	83.27	81.27
Xiaolong Zhu[17?]	92.58	93.67	93.54	91.24	90.87	88.59	87.18	86.54	84.12	83.18
OURS	94.54	94.27	92.57	93.49	92.24	90.68	91.57	91.02	90.81	89.36

TABLE 9. Comparing the recall by each model.

Extracting Information Items	1	2	3	4	5	6	7	8	9	10
Xiangdong Li[15?]	81.75	82.15	81.67	80.01	80.23	77.59	77.68	76.36	75.87	74.25
Juan Lai[16?]	86.57	86.16	87.58	86.79	86.48	85.85	85.01	82.82	81.69	79.65
Xiaolong Zhu[17?]	90.98	91.58	91.05	90.28	87.58	87.47	86.15	84.51	83.16	81.29
OURS	93.13	93.16	91.85	92.03	91.27	89.25	90.23	89.29	87.49	87.53

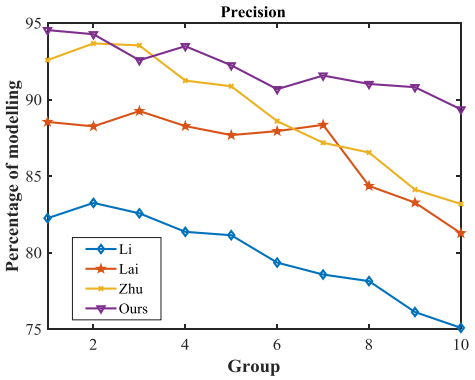

FIGURE 9. Precision trend for each model.

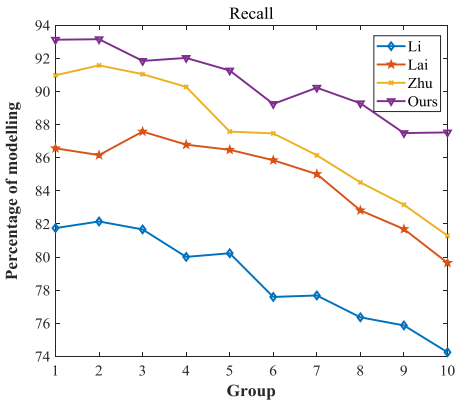

FIGURE 10. Recall trend for each model.

accuracy rate of 94.24% and a recall rate of 93.84 for ISBN, which is due to the fact that these two kinds of data belong to strong identification information, and the features are more easily recognized relative to other data. Other parts of the data are relatively slightly lower, but overall, the combined accuracy and recall rate can reach more than 85%, and the combined indicator F also reached 85.78%. Tables 8,9 and figure 9,10 show the comparison of Precision and Recall of various algorithms when extracting different numbers of information items. Li Xiangdong et al. used the traditional method, and its Precision and Recall are relatively low. Lai Juan and Zhu Xiaolong are both deep learning network models with much improved model accuracy, the accuracy of Zhu Xiaolong's model is closer to that of this paper's model, and by comparing with the above three models, this paper's model has higher Precision and Recall.

The response time of the book information query request is mainly to test the query performance of the model from the perspective of query speed. Through the query access test, the average query response time based on HBase is within 0.11s, and the average response time of Oracle query is within 0.09s. Almost, can meet the general query needs. However, Hbase has absolute advantages in query speed and storage performance when querying massive amounts of data.

2) DATA QUERY PERFORMANCE EXPERIMENT

The response time of the book information query request is mainly from the perspective of query speed to test the query performance of the model, through the query access test, based on the HBase query response time within an average of 0.11s, Oracle query response time within an average of 0.09s, both are similar to meet the general query requirements.

The data storage performance experiment is to compare the disk space occupied by the traditional relational database Oracle and the distributed non-relational database HBase with the same amount of data, and evaluate the storage advantages of distributed HBase. In the actual test environment, the same data is stored in the Oracle database and the HBase distributed database respectively, and then the

space occupied by the Oracle data table and the HBase database table is statistically compared and compared. The experimental results are shown in Figure 11.

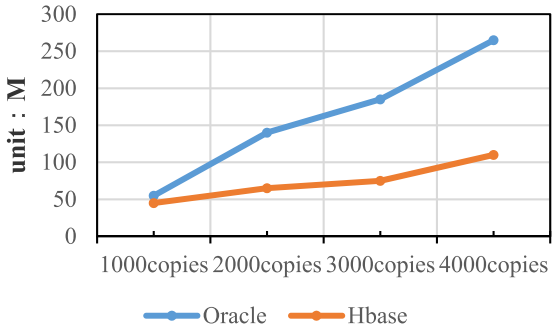

FIGURE 11. Comparison of storage space between Oracle and HBase.

It can be seen from the figure that the same amount of book data is extracted and stored in Oracle and Hbase databases respectively. The storage space occupied is significantly higher than that of Hbase, and with the increase of the number of books, the space occupied by HBase storage grows steadily and slowly, while the space occupied by Oracle grows rapidly, and the query speed and query accuracy brought about by the rapid growth of storage space And the system performance degradation such as stability is also more obvious. It can be seen from this that the performance of Hbase database is better than that of Oracle database for storage and query of sparse and massive book data. In the actual database system application, Oracle and Hbase are combined. Oracle first receives a small amount of real-time data for temporary storage, and then continuously migrates the temporary data to the Hbase database for permanent massive storage and query. The combination of the two can give full play to their respective advantages, to improve the overall performance of the system, to give full play to the economical and practicality of the system and to meet the actual application requirements in practical applications.

IV. DISCUSSION

In the book information extraction experiments, the information extraction results in Table 7 show that ''author'' and ''ISBN'' extraction results are significantly better than other items, thanks to the strong identification information of these two data, features are easier to identify than other data. The comparison of Precision and Recall of the four algorithms in Tables 8 and 9 reveals that the overall trend is that as the number of information extraction increases, the Precision and Recall of each model decrease, but Zhu Xiaolong's model has a larger decrease, while this paper's model decreases relatively gently, which shows that this paper's model has a better stability. Li Xiangdong et al. extracted bibliographic information from book web pages by using traditional techniques such as generalized rules, co-occurring words and page analysis, and had relatively low extraction accuracy and recall. The deep learning network model constructed by Lai Juan et al. does not apply dimensionality reduction to high-dimensional text information, and the complexity of information extraction is high, and the accuracy of the algorithm is not ideal. The algorithm proposed by Zhu Xiaolong is also a deep learning network model, although the dimensionality reduction process is implemented on the information before extracting the semi-structured text information to reduce the complexity of information extraction, but its dependence on the data and information structure is large, which has certain limitations, and the accuracy of the test on the dataset of this paper is not as good as the original text. The model in this paper uses different methods to match and extract book information according to the strong and weak identification information, and uses the backtracking algorithm to backtrack many times, which can effectively improve the extraction accuracy of the model, and the method in this paper is better by comparing with the above three models.

The test in terms of information extraction efficiency, the model speed of this paper is an average of 615 documents per second extraction and processing, which meets the expectations and satisfies the actual demand. In the experimental test of data query and storage performance, the average query response time of Hbase and Oracle is 0.11s and 0.09s respectively, which can meet the general query requirements. In terms of storage performance, when the data volume is large, the storage space occupied by Oracle for storing the same amount of book data is significantly higher than that of Hbase, and the query speed and storage performance of Hbase are more absolute advantages when querying massive data. Oracle and Hbase will be combined in the actual database system application, Oracle for temporary storage, Hbase for permanent mass storage, the combination of the two can fully utilize their respective advantages, improve the overall performance of the system, and give better play to the economic practicality of the system.

V. CONCLUSION

Based on the rapid growth of book and document information in the Metaverse era, which presents large-scale, multivariate, and heterogeneous trends, this paper studies the extraction rule analysis method and storage management mechanism of book information. Based on the semi-structured characteristics of book documents, a heterogeneous. The book data extraction rule model is constructed to effectively extract massive heterogeneous book information. Based on HBase distributed storage structure and parallel computing technology, optimize the storage scheme of book text data and improve query efficiency to ensure efficient management and real-time retrieval of massive heterogeneous book data. In the experimental results, the accuracy rate and recall rate can reach more than 85%, and the highest comprehensive index F reaches 94.54%, which is better by comparing with the existing three models. The comprehensive experimental results show that the system performance meets the practical application requirements compared with the traditional method in terms of book text data extraction accuracy, recall

rate, book information management and query efficiency and so on. The system architecture approach designed in this paper has better performance of book data management, with economy of mass data storage and high efficiency of query management, which provides strong support for book information processing and analysis in the era of meta-universe. Due to the diversity of book data in the meta-universe, this study mainly focuses on the processing of several types of textual book information, while the system cannot meet the requirements for the processing of data in formats such as pictures, audio and video. In addition, with the emergence of a new generation of AI technology applications such as ChatGPT, in future work, in-depth research should be conducted to introduce highly intelligent text processing techniques into the model, and continuously improve the robustness and intelligence of the system, so that it can meet the processing requirements of the intricate and diversified book data in the meta-universe.

ACKNOWLEDGMENT

The authors are grateful to the Northeastern Forestry University Library for providing important library data support, the Northeastern Forestry University Artificial Intelligence Laboratory for providing experimental techniques and instrumentation support, and Prof. Wang Mingyang for the oracle of the manuscript and for her project-related support.

REFERENCES

[1] K. Zhang et al., "The study of Chinese obstetric electronic medical records based on natural language processing," *J. Zhengzhou Univ., Natural Sci. Ed.*, vol. 49, no. 4, pp. 40–45, 2017, doi: 10.13705/j.issn.1671-6841.2017005.

[2] W. Guo and X. Bao, "Key information extraction model of unstructured text in knowledge database," *Comput. Simul.*, vol. 38, no. 9, pp. 357–360 and 394, 2021.

[3] X. Zhu and L. Qiu, "Simulation of semi?structured text information extraction using simulation of machine learning," *Comput. Simul.*, vol. 40, no. 2, pp. 540–544, 2023.

[4] H. Zhao and F. Wang, "Information extraction and integration of large-scale heterogeneous socio-economic statistical statements," *J. China Soc. Sci. Tech. Inf.*, vol. 39, no. 9, pp. 938–948, 2020.

[5] B. Yu, J. Du, and Y. Shao, "Web page content extraction based on multi-feature fusion," 2022, *arXiv:2203.12591*.

[6] V. Nundloll, R. Smail, C. Stevens, and G. Blair, "Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science," *Heliyon*, vol. 8, no. 10, Oct. 2022, Art. no. e10710.

[7] F. Ciravegna, A. Dingli, and D. Petrelli, "Active document enrichment using adaptive information extraction from text," in *Proc. 1st Int. Semantic Web Conf. (ISWC)*, Sardinia, Italy, Jun. 2002.

[8] X. Li, Y. Hou, and L. Huang, "Study of book pages automatic identification and bibliographic information extraction," *New Technol. Library Inf. Service*, no. 4, pp. 71–77, 2014.

[9] X. Ban, "Research on web information extraction and sentiment classification based on forum," Tianjin Univ., Tianjin, China, 2019.

[10] K. Wu, "Research on Web data extraction technology based on template and visual features," Chongqing Jiaotong Univ., Chongqing, China, 2018.

[11] X. Wang et al., "Research on web page information extraction based on visual features," *J. Chin. Inf. Process.*, vol. 33, no. 5, pp. 103–112, 2019.

[12] Z. Zhen and J. Zhang, "Research on feature extraction based on statistical range and coefficient of variation," *Statist. Decis.*, vol. 38, no. 23, pp. 43–47, 2022, doi: 10.13546/j.cnki.tjyjc.2022.23.008.

[13] H. Xie et al., "A semi-supervised method for Chinese key phrase extraction based on statistical features and graph model," *J. Chin. Inf. Process.*, vol. 36, no. 4, pp. 57–65, 2022.

[14] W. Liu, "Web short text oriented knowledge association model and semantic coherence computation method," Shanghai Univ., Shanghai, China, 2016.

[15] J. Lai and Y. Hong, "Deep learning network based on rule constraints for text information extraction," *Comput. Eng. Des.*, vol. 42, no. 12, pp. 3548–3554, 2021, doi: 10.16208/j.issn1000-7024.2021.12.033.

[16] H. Shen et al., "Exploration and exploitation: A fine-grained hierarchical network for text-to-image synthesis," *China Sci.Paper*, vol. 18, no. 3, pp. 238–244, 2023.

[17] R. Sarkhel, B. Huang, C. Lockard, and P. Shiralkar, "Label-efficient self-training for attribute extraction from semi-structured web documents," in *Proc. ACM Conf.*, 2022, pp. 1–12.

DAILIN WANG received the master's degree, in 2005. She is currently a Librarian with Northeast Forestry University. She has been engaged in the research of digital sharing platform service of library intelligence and document management, cooperated with the discipline team of related colleges, participated in one project of data science direction at the departmental level and one project of SCAL in digital library construction, published nearly ten articles and intellectual property rights (invention patents) in related fields, and presided over and participated in five related topics. She is a Provincial Member of the Heilongjiang Provincial Library Association. She received one provincial award and two other academic awards.

LINA LIU is currently pursuing the Ph.D. degree with the College of Computer and Control, Northeast Forestry University. Her current research interests include big data processing technology and machine vision.

YALI LIU received the Ph.D. degree in agronomy. She is currently an Associate Research Librarian. Her main research interests include library research work, intellectual property information service work, and institutional repository construction.