Review and prospect of research on ancient book information processing in China

Zhongbao Liu*, Zhenzhen Qin and Wenjuan Zhao
Institute of Language Intelligence,
Beijing Language and Culture University,
15 Xueyuan Road, Haidian District, Beijing 100083, CHINA
e-mail: *zbliu@blcu.edu.cn (corresponding author);
2365608907@qq.com; zhaowenjuan1118@163.com

ABSTRACT

The advent of the era of big data appears to be an unprecedented opportunity to the development of ancient book information processing in China. A comprehensive review of the research on ancient book information processing can help researchers understand the research progress and grasp the future research trend. Based on the life cycle of ancient book information processing, this paper reviews and summarizes the progress of digital resource construction, data mining, system construction and information service, and prospects the future research trend. Although progress has been made to some extent, the ancient book information processing is still in its infancy. It is expected that more researchers will pay attention to and engage in this field.

Keywords: Ancient book information processing; Ancient books; Digitization of books; Data mining; Chinese ancient literature.

INTRODUCTION

The ancient book is an important carrier of Chinese excellent traditional culture and a cultural treasure with Chinese characteristics. With the advent of the big data era, the number of ancient books continues to grow. Making use of the digital technology to mine valuable information or knowledge from ancient books determines the continuation and promotion of Chinese culture. Under this context, ancient book information processing is proposed. It refers to the processing of the pronunciation, form and semantic of ancient books with the help of digital technology, and based on which, the implicit information or knowledge can be obtained by data mining and knowledge discovery. The ancient book information processing involves a wide range of subject disciplines, such as history, sociology, linguistics, philology, and informatics. It covers ancient book digitalization, sentence breaking, automatic punctuation, word segmentation, part-of-speech (POS) tagging, semantic understanding, knowledge organization, and information services. With the development of information technologies, especially the improvement of natural

language processing technologies, the ancient book information processing has appeared an unprecedented opportunity for development, which has attracted more and more researchers' interest. It is necessary to review and analyze previous studies done on ancient book information processing to help researchers understand its current status and problems existed in the earlier studies, as well as grasp the future research trends. It can also provide an essential support of theories and methods for researchers.

The academic value of this paper includes three aspects besides its important theoretical value and practical significance to ancient book information processing. First, it is beneficial for much more researchers to comprehensively understand the advances of ancient book information processing in China, and it is also convenient for them to research and use the ancient book information resources. Second, it is easier for other countries to understand China with a long history and profound culture by reviewing the achievements of ancient book information processing. Third, it contributes to promote interdisciplinary research and international cooperation in ancient book information processing.

MATERIALS AND METHOD

The research data and the research methodologies used in this paper are as follows. The keywords such as "*ancient Chinese book information, ancient Chinese classics information, ancient Chinese book information processing, ancient Chinese classics information processing"* and some typical ancient Chinese books in Chinese and English are respectively used for retrieval on several databases The databases used are China National Knowledge Infrastructure (CNKI), IEEE Xplore, ScienceDirect and Web of Science. The retrieval results are arranged based on publication dates. Bibliometric analysis is applied to investigate the literatures with close correlation with research topic, high citation and download time. The related literatures are compared and analyzed with the help of comparative analysis method, and it identifies the contributions and limitations of existing studies, which provides a foundation for review and prospect of ancient book information processing in China.

The limitation in existing studies is that most of them are periodical summarization aiming at the digitization of ancient books, and these achievements have been made earlier in time. Meanwhile, there exists few studies on the whole process of ancient book information processing. Therefore, it is necessary to include the new achievements in recent years into the studies on ancient book information processing for a more comprehensive and detailed summarization and analysis. This paper aims to review and prospect the studies on the ancient book information processing, and it focuses on its whole process based on the information life cycle theory.

The research framework is constructed based on the life cycle of ancient book information processing, as shown in Figure 1. The research framework includes four parts, namely digital resource construction, data mining, system construction and information service.

The digital resource construction, as the basic data layer, provides necessary data support for subsequent researches. The data mining layer utilizes information technology to organize and analyze the digital resources of ancient books to form valuable information or knowledge. The system construction layer constructs an ancient book information system by using ancient book information or knowledge according to specific topics or practical needs. The ancient book information service layer provides users with information services such as collation, indexing, translation, retrieval and compilation.

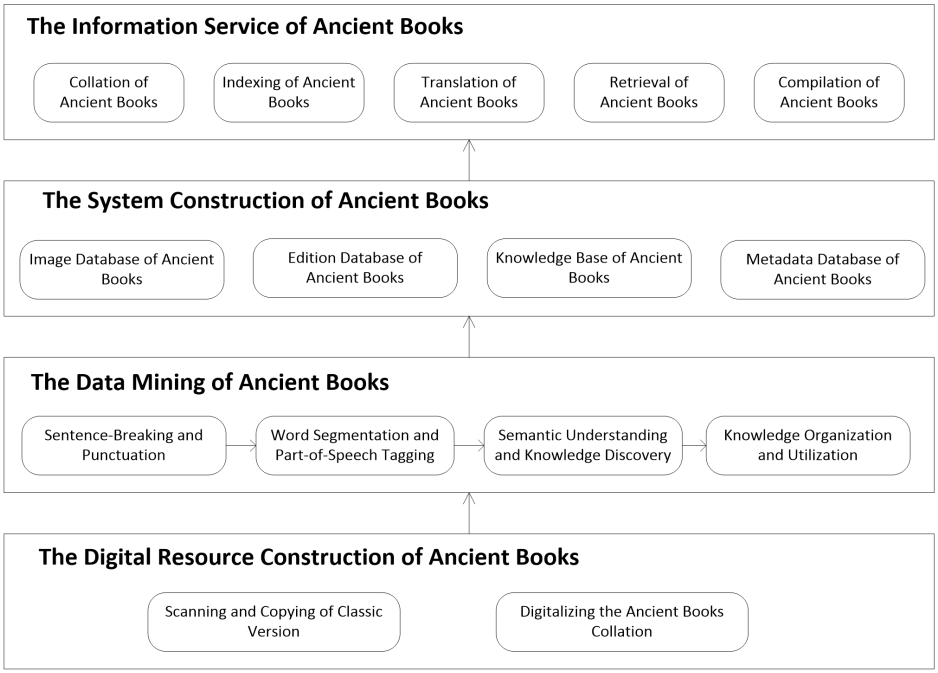

Figure 1: The Ancient Book Information Processing Research Framework

FINDINGS ON ANCIENT BOOK INFORMATION PROCESSING RESEARCH

Digital Resource Construction of Ancient Books

Digital resource construction of ancient books aims to transform ancient books such as manuscript, movable type, overprint, engraving, pictures, ancient paintings and type printing into digital symbols by information technology. There are two main sources of ancient book digital resources.

One is from scanning and copying of classic version. Many libraries and institutions have carried out several projects of digital resource construction, and lots of ancient books have been scanned into picture format. The United States pioneered the digital resource construction of ancient books, through the construction of a series of ancient book databases by implementing the “Digitization Plan of Ancient Rubbings”, “Chinese Classic Versions Bibliographic Database” and other projects (Yu and Guan 2017). Britain transformed *Ancient Chinese Maps* and *Ji Jiu Pian* into digital resources supported by grant from “Digital Library Project” (Xue 1995). In Japan, ancient book databases such as “Classical Books Collected in Comprehensive Library” and “National Bibliographic

Database of Chinese Books" were established (Mao 2006). In China, Taiwan was the first to practise the digital resource construction of ancient books. The Central Library of Taiwan established the "Joint Catalogue of Rare Books and Ancient Books in Taiwan". In recent years, the National Library of China has developed "The Reconstruction of Chinese Rare Book Database". In addition, Peking University Library has developed "Ancient Literature Resources of Peking University Digital Library" named *Mi Ji Ling Liang*. China Academic Library and Information System has built "Ancient Literature Resource Database in Colleges" named *Xue Yuan Ji Gu*.

The other is to digitalize the ancient book collations, which is convenient for reading and diverse retrieval. The "Chinese Classics Database" is most representative. It collected ancient books organized by Zhonghua Book Company, including "*Twenty-four Histories*", "*Elementary Series of Classical Literature*", and "*Zi Zhi Tong Jian*". The National Library of China has developed "Analysis System of Tang Poetry" and "Analysis System of Song Poetry". Beijing Erudition Digital Technology Research Center has developed the full-text retrieval system of "Chinese Classic Ancient Books Database", and Beijing UniHan Digital Technology Co. Ltd. has developed a series of full-text retrieval systems, including "*Four Series*", "*Kangxi Dictionary*", "*Shi Tong*", and "*Records of Imperial Laws and Regulations in Ming Dynasty*". Gao (2021) studied a fast sharing method of digital ancient book resources under cloud computing environment, so as to improve the sharing efficiency and effect of digital ancient book resources.

Sentence Breaking and Punctuation of Ancient Books

Different from modern books, sentences in ancient books do not have punctuation for segmentation. Therefore, the first process is sentence breaking and punctuation; the former is to split the successive sentences, and the latter is to add punctuation marks to the sentences in ancient books. According to the techniques used, the methods of sentence breaking and punctuation can be classified into three categories.

The first category is the rule-based method, which obtains the features of sentence breaking by analyzing the ancient books with broken sentences, based on which, the rules of sentence breaking are summarized to punctuation. Chen et al. (2007) proposed an algorithm of punctuating the sentences in archaic Chinese language based on context N-gram model. Huang and Hou (2008) probed into certain patterns on sentence breaking and punctuation model for ancient books on agriculture. The rule-based method performs efficiently and its result is accurate in some cases, but there exists two deficiencies in this method, i.e. the rule base of sentence breaking is laborious; and it is only effective for specified ancient books, but not applicable to others.

The second category is the statistical learning-based method. The machine learning model is introduced to sentence breaking and punctuation, which is considered as the problem of classification. Wang and colleagues designed a 6-tag set and proposed a method based on cascaded Conditional Random Field (CRF) (Wang, Xiong and Wang 2009). Zhang, Xia and Yu (2009) carried out a research on sentence breaking and

punctuation of ancient books with the CRF model based on mutual information and difference of t-test. In general, the statistical learning-based method performs much better than the rule-based method, but it is limited by two problems. One is the feature template should be given in advance for a specific ancient book. The other is the applicability of feature template is restricted, and it cannot be applied to other ancient books.

The third category is the deep learning-based method. Deep learning model, with good abilities of feature extraction, performs well in sentence breaking and punctuation. Wang et al. (2016) proposed a sentence segmentation method for ancient Chinese texts based on neural network language models. Wang, Shi and Su (2017) proposed a sentence segmentation method for ancient Chinese texts based on Recurrent Neural Network (RNN). Han et al. (2019) applied a joint model based on Bi-directional Long-Short Term Memory (BiLSTM) and CRF with character features in order to break sentences in ancient books. Yu, Wei and Zhang (2019) researched on the position and context of the unlabelled characters in ancient books by Bidirectional Encoder Representations from Transformers (BERT) model.

In general, the efficiencies of sentence breaking and punctuation have greatly improved in recent years, especially in automation, flexibility and applicability. We can draw the following conclusions based on the above analysis.

(a) Firstly, the precision of punctuation is obviously lower than that of sentence breaking in ancient books. The main reason is that the diversity of punctuation marks makes it difficult to punctuate the ancient books.

(b) Secondly, the deep learning-based method is superior to the rule-based method and the statistical learning-based method in the perspective of performance. The main reason is that the deep learning model, with good abilities of feature extraction, constructs a prediction model on the training corpus, and it is utilized to sentence breaking and punctuation on the test corpus.

(c) Finally, when the training and the test corpus are from the same ancient book, the performance of deep learning-based method is excellent. However, when both of them are from different ancient books, the performance of such method is greatly descended. The reason of its poor performance is the features extracted from the training corpus have little correlation with the features in the test corpus, it seriously affects the efficiencies of sentence breaking and punctuation. Therefore, how to apply the deep learning-based method in practice deserves further research.

Ancient Chinese Word Segmentation and Part-of-Speech Tagging

Similar with modern books, a word in an ancient book is treated as the basic semantic unit. It is necessary to utilize word segmentation technologies to separate words in order to easily understand its semantics. Each word in the ancient book has specific POS, such as nouns, verbs and adjectives, and POS tagging is to label words with part of speech. The method of ancient Chinese word segmentation and POS tagging can be classified into three categories.

The first category is the dictionary based method, which applies matching algorithms based on manual dictionary. Huang and Hou (2011) introduced a series of methods to word segmentation for ancient books on agriculture, including segmentation mark, word segmentation dictionary and N-gram. Xu and Chen (2012) applied a maximum matching algorithm to word segmentation on the corpus of *Zuo Zhuan.* Wang et al. (2018) combined the *Sinological Index Series* with the domain knowledge of *Mao Shi Index*, and studied the automatic word segmentation of *The Book of Songs* using the CRF model. The dictionary based method is convenient and feasible, but its performance is limited to some deficiencies. One is the manual dictionary is time-consuming and laborious. The other is the problem of non-login words due to dictionary limitation, which seriously affects its performance.

The second category is the statistical learning-based method. It is independent of word dictionary, and natually solves the problem of over-reliance on word dictionary encountered by dictionary based method. Jiang et al. (2010) proposed a new fast segmentation method for classic Chinese texts based on the tree pruning process. Liang and Chen (2013) proposed a word segmentation method based on the CRF model applied in *Mencius*. Qian et al. (2014) studied the ancient Chinese word segmentation and tagging technology by using Hidden Markov Model (HMM). Wang, Huang and He (2017) constructed a model of the POS tagging for the Pre-Qin literature based on the CRF model and combined feature template. Fu et al. (2019) developed a word segmentation system of ancient books by constructing a Thesaurus of Chinese medicine terminology and a special POS tagging method based on HMM and Java language. The statistical learning-based method takes advantage of the distribution of words in ancient books for word segmentation and POS tagging. It breaks through the restriction of dictionary, but requires a large-scale corpus to train the statistical model. Its performance is closely associated with the quality of training corpus.

The third category is the deep learning based method. The excellent performance of deep learning model on feature extraction has a good effect on the ancient Chinese word segmentation and POS tagging. Zhang, Chen and Xu (2013) used deep layers of neural networks to discover relevant features to Chinese word segmentation and POS tagging. Yao and Huang (2016) used the BiLSTM model for Chinese word segmentation. Yu et al. (2020) proposed the Multi-Stage Iterative Training (MSIT) for unsupervised word segmentation by combining non-parametric Bayesian models with BERT. The efficiencies of deep learning model depends on the scale and quality of training corpus, however, only a few corpora can be directly used until now. This contradiction restricts the popularization and application of deep learning based method.

The following conclusions are made based on the analysis above:
(a) Firstly, the precision of ancient Chinese word segmentation and POS tagging is obviously higher than that of sentence breaking and punctuation, but it is lower than that of modern books in most cases. The main reason for such appearance is that the expressions between ancient books and modern books are quite different, and

therefore, whether the method applicable to modern books is suitable to ancient books deserves to further research.

(b) Secondly, the effectiveness of current studies are verified on the small-scale corpus, however, the scale of existing ancient books is large. How to conduct the current studies on the larger scale corpus is quite important.

Semantic Understanding and Knowledge Discovery of Ancient Books

In recent years, semantic understanding and knowledge discovery of ancient books mainly focus on named entity recognition, semantic disambiguation and alignment, and knowledge mining. Named entity recognition is used to identify entities with specific meanings in ancient books. Word sense disambiguation and alignment are conducted to discriminate the semantics of specific words according to the semantic context of ancient books. Knowledge mining tries to discover comprehensible knowledge from massive and heterogeneous ancient books.

In the research of named entity recognition, Wang and Tsai (2013) implemented feature selection of bigram Chinese characters, phonemic representations, appellations, pronunciations and function words in *Chinese Buddhist scriptures* with the CRF model. MARKUS was a well-known online platform for automatically tagging a range of historical named entities in Chinese and Korean texts (Ho and De Weerdt 2014). Liu et al. (2015) proposed a constrained N-grams model for name entity recognition after mapping entity types to the longest strings in the China Biographical Database. Huang, Wang and He (2015) analyzed the internal and external characteristics of ancient Chinese places in *Zuo Zhuan*, and constructed the feature template. Gao, Jin and Zhang (2019) applied the joint model BiLSTM-CRF to identify the entities of Chinese medical classics. The recognition effect of named entities in ancient books is much lower than that in modern books. The reason of such appearance is investigated that the semantic understanding abilities of ancient books requires to be further improved. The difficulties of semantic understanding can be concluded as follows: (a) firstly, there are plenty of interchangeable words, rare words and polysemy in ancient books; (b) secondly, the sentence patterns of ancient books are complex, mostly long sentences, and many of rhetorical devices are utilized; and (c) thirdly, the representation of some entities lacks uniform standards.

In the research of semantic disambiguation and alignment, Yu et al. (2009) applied the CRF model with six different templates by choosing contextual words and adding linguistic features. Chang et al. (2013) proposed an improved unsupervised disambiguation method of ancient Chinese based on the vector space model. In order to solve the problem of term alignment in ancient books, Che and Zheng (2016) proposed a Maximum Entropy (ME) model based on sub-words. The above studies adopt the supervised learning method to semantic disambiguation and alignment, which effectively solves the sparse labeling problem, but requires to build a large-scale corpus with annotation.

In the research of knowledge mining, Ma, Chen and Qu (2013) proposed to organize the knowledge in ancient books and commentary literatures with structured knowledge representation method. Zhu and Bao (2015) applied Geographic Information System (GIS) in the development and utilization of Chinese ancient local chronicles on the *Local Chronicles of Guangdong*. Li et al. (2017) created a database annotating each person and in addition, each place was tagged the present name, and the geographic information in Baidu Map. Wang et al. (2018) studied on the classification of the questions related to ancient Chinese by introducing the models of Support Vector Machine (SVM), CRF, and a deep learning model. Zhou, Hong and Gao (2019) designed a Tang poetry ontology model driven by domain knowledge service. Yu and Wei (2019) took advantage of the LSTM model to analyze and process the character sequence in ancient Chinese, so as to solve the issues on the dating of ancient Chinese texts. Liu, Dang and Zhang (2020) proposed a historical event extraction method based on *Historical Records* based on BERT and LSTM-CRF. Cheng and Liu (2021) discussed the network training of deep convolution neural network and compared the algorithm with the traditional machine learning algorithms in order to effectively extract the ancient Chinese characters. The above studies have been effectively explored in knowledge mining and utilization of ancient books. However, there exists some problems in the these studies, such as small-scale manual annotation, poor performance of the model and low-quality knowledge. In addition, the process of knowledge mining is completely data-driven, and there is no domain knowledge to guide this process, which affects the knowledge mining efficiencies.

Knowledge Organization and Utilization of Ancient Books

The knowledge organization of ancient books is a series of methods to collect, process and regulate the ancient books knowledge. Liu (2004) proposed a knowledge representation method based on meta knowledge according to ontology and epistemology. Wang (2014) studied knowledge organization of handed down Chinese Dictionary in Japan, South Korea and China, and noted that the standardized knowledge classification principle was a prerequisite for East Asian countries to regulate the construction of database of handed dictionaries. Chang, Lu and Zhai (2019) introduced linked data to knowledge organization of ancient books. Summarizing the above analysis, it can be concluded that current studies on knowledge organization and utilization of ancient books mainly concentrated on theoretical introduction, technological exploration and application research, specially solving the problems of the knowledge representation and related technologies. However, due to the limitation of cognition and technology, current studies were far from expectation and the unified research system integrating theory with practice has not yet been formed.

Ancient Book System Construction

The construction of ancient book system includes image database, edition database, knowledge base and metadata database. There are two ways to construct ancient book system. One is man-made construction, and the other is by means of information technology. As the former is time-consuming and inefficient, more and more

researchers have paid more attention to the latter, and have gained a series of achievements.

On the image database construction, Shi (2016) proposed the construction principles of maturity, standardization and planning based on the analysis of the existing problems in image database of ancient books. Wu et al. (2016) intended to take the ancient literature of ethnic minorities in Yunnan as an example, aiming to construct the virtual digital database of ancient books.

On the edition library construction, Deng et al. (2014) discussed the necessity of edition library construction of ancient books by introducing the ontology technology. Liu and Zhao (2017) constructed computer aided collation repository of ancient editions followed the construction standard of collation database. Gu (2021) focused on the literary works and the related collation, annotation and textual research results so as to research the integration and optimization of ancient literature information resources. All the above studies were carried out on the small- and medium-size corpora. However, the version database construction, in fact, involves a large-scale ancient digital resources, and it is worthy of further discussion about whether the existing technologies and methods can be applied to such situations.

On the knowledge base construction, Jia (2015) discussed how to deal with *Zhouyi* and its annotation by computational linguistics and information technology. Zhao (2020) applied qualitative and quantitative methods to propose a hypothetical index model for evaluating the service efficiency of ancient book digital resources. Most of current studies on knowledge base construction of ancient books are based on theoretical discussion and practical summarization, and lack of quantitative research based on investigation.

On the metadata database construction, Gao et al. (2011) constructed a system to extract metadata from the title page of an ancient book. Inspired by “evidence-based practice”, Xia, Lin and Liu (2017) designed a data model of ancient books that can integrate ancient book catalogues, metadata, full text and knowledge from different sources and formats. Zhang and Wang (2020) extracted features of 300 common ancient book databases based on Dublin core metadata set.

Ancient Book Information Service

Ancient book information service is based on information technology and network technology. The personalized and professional information or solutions are formed for users by searching, sorting, organizing and analyzing ancient book digital resources. Until now, ancient book information services mainly include collation, indexing, translation, retrieval and compilation.

(a) Collation

The collation is to automatically distinguish the differences between different versions

of ancient books with the help of information technology. Chang, Hou and Cao (2007) designed a collating algorithm based on window matching. Xiao and Chen (2010) proposed an automatic method to find the version differences among ancient books. The above studies were rule-based and the performance of them depended on the construction of rule bases. As the number of ancient books has reached 2.6 million words, how to efficiently and automatically build the rule base needs to be further discussed.

(b) Index

The index is a specific form for revealing the contents of ancient books. It extracts and indexes the name, word, sentence, place name, and theme from ancient books, and edits them in a certain arrangement. Huang (2011) used VFP, Word and other tools to get editing indexes of multi-text ancient books. Xiao (2017) took the empirical research on ancient book index from the aspects of resource description, the text fragmentation, data mining and the creation of new data. General Catalogue of Chinese Ancient Books in Korea was published by Korea researchers, in which the 28 catalogues of Chinese ancient books were compiled (Jeon 2005). The ancient book index database was constructed by Oriental Culture Research Center of University of Tokyo, aiming at indexing *Jin, Shi, Zi, Ji* (Oriental Culture Research Center of University of Tokyo 2011). The above studies attempt to explore automatic methods of creating the ancient book index, so as to overcome the inefficiencies of manual index to a certain extent.

(c) Translation

Wang, Zhang and Han (2009) proposed an example-based machine translation system for ancient books based on linguistics and machine translation theory. Ding, Li and Li (2012) described an instance-based method called KNN-based bootstrapping to annotate rhythm of *Gongchepu*, one of the popular ancient Chinese musical scores. Han, Yan and Song (2015) proposed a rule-based machine translation for ancient Chinese under the framework of sentence-focused syntax theory. The corpus scale used in sentence alignment is quite small, which directly affects the training efficiencies and leads to the poor performance of translation.

(d) Retrieval

Zhang et al. (2001) tried to find the characteristics of Chinese ancient books and designed a full-text retrieval system for ancient books. He and Cao (2006) discussed the semantic retrieval mechanism based on domain ontology. Guo and Dai (2001) tried to improve the retrieval ability by ancient book digitization, considering database retrieval by organization and management of ancient books. Bai and Bao (2017) adopted a topic model based on Latent Dirichlet Allocation (LDA) on the corpus of *Gan Zhu Er Jing.* Generally speaking, the information retrieval of ancient books has made a progress and emerged a number of achievements. However, there have been two challenges. One is the retrieval results considering little on personalized needs, the other is current technologies unavailable to massive ancient books.

(e) Automatic compilation

Automatic compilation is to analyze the ancient books by information technology, and extract information related to a subject and compile them into a book. Zhang (2002) focused on the computer compilation software used for ancient books. Chang and Hou (2007) discussed the automatic compilation algorithm of agricultural ancient books. Chang (2009) designed and implemented an automatic compilation system of agricultural ancient books, such as *Qimin Yaoshu* and *Nongzheng Quanshu.* The above studies explored computer-aided compilation, some of which need manual intervention, far from intelligence. Meanwhile, the performance of automatic compilation cannot satisfy user's requirements, and whether domain knowledge can be introduced in deep learning model for compiling ancient books deserves further research.

Ancient Chinese Digital Humanities

The current studies on ancient Chinese digital humanities mainly focused on the platform construction of digital humanities, which has provided abundant information resources and efficient information services. A few projects are worth highlighting here, for example the construction of a Chinese ancient book digital humanities research platform to support historical China studies (Chen and Chang 2019); the creation of the knowledge graph of the Hundred Schools of Thought including *Confucius*, *Laozi* and *Mozi,* which discussed the application value and realization path of knowledge graph in digital humanities (Wei and Liu 2019); and the development of a character social network relationship map tool for supporting an ancient book digital humanities research platform (Chen, Chen and Chang 2019). Recent initiatives include Zhu and Zhang's (2020) propoasl of a new digital humanities cyber-infrastructure conceptual model for ancient China studies based on the concepts of "ancient China studies" and "digital humanities"; and a system named GeLaiGeLai for data analysis of classical Chinese poetry by AP-LSTM-CRF and knowledge graph (Wei et al. 2020). Ouyang and Ren's (2021) reviewed the visualization theoretical basis, summarize the visualization processes, and analyze the visualization technologies of ancient books.

In general, the informationization and ontologicalization processes of ancient books can produce massive data, with the help of data mining and analysis technologies. The implicit knowledge and scientific law can be obtained, which is helpful for users to retrieve, download, share and reuse. The ancient Chinese digital humanities platform provides a bridge for interdisciplinary ancient Chinese information processing research and simplifies the cooperation way between researchers and institutions.

RESEARCH TREND AND PROSPECTS

With the improvement of theories, the maturity of technologies and much more attentions paid to traditional culture, ancient book information processing has appeared an unprecedented opportunity for development. Therefore, it is necessary to prospect the research trends of ancient book information processing, including digital

resource construction, data mining, system construction and information service of ancient books.

(a) Digital resource construction of ancient books

The digital resource construction of ancient books should follow the principles of scientificity, pertinence, universality and security; emphasize the retrieval and interactive service platform construction; highlight the ancient book resource integration by retrieval platform; and provide multi-level retrieval service, personalized consultation and interactive service. In order to share the ancient book resources for the public, it is necessary to establish a standard system suitable for ancient book digitalization based on the information life cycle theory. With the increasing contribution of information technologies, the ancient book digital resource construction should strengthen innovation of equipment and technology in practice. On one hand, we are supposed to pay attention to the development of the latest digital equipment. On the other hand, we should update and upgrade the existing digital software in time, comprehensively upgrading the hardware and software of ancient book digital resource construction. With the advent of big data era, the digital resource construction of ancient books has encountered two contradictions. One is the contradiction between digitalization and data-orienting. The existing ancient book digital resource base transforms the format of ancient books into text format, which destroys the objectivity of ancient books in a certain extent, resulting in impossibility to obtain the original information. The other is the contradiction between convenience and objectivity in the retrieval process. The best way to ensure the objectivity of ancient digital resources is to establish ancient digital resource base by photocopying, which is inconvenient to browse and search. The key to solve the above contradictions is to integrate the strength of ancient book institutions, promote the infrastructure construction, and establish a big data-driven sharing system.

(b) Data mining of ancient books

Future researches on ancient Chinese sentence breaking and punctuation may focus on expanding the corpus scale and improving the efficiencies; highlighting the design of deep learning model on the large-scale corpus; emphasizing on solving the problem of few available features of ancient books; and introducing transfer learning and attention mechanism. Some theoretical problems of ancient Chinese word segmentation need further research, for example, what is a word, how to distinguish word and morpheme, and what is the difference between word and phrase. The key to solve the above problems is to explore the definition of word, refer to modern Chinese word segmentation methods, reversely deduce the difference of word segmentation in ancient books, and conduct theoretical analysis, computability analysis, and compatibility analysis. It is possible to supervise the annotation process of ancient books with the help of literature commentaries. Meanwhile, annotation granularity should be finer, and the POS and its subclasses should be also annotated. The corpus can be annotated by man-machine cooperation mode. In the process of POS tagging, the efficiencies of semantic tagging can be improved by integrating various technologies

such as word segmentation, clustering, classification and pattern matching. The efficiencies of entity extraction can be improved by deep learning model, which extracts more semantic features from ancient corpus. Many efforts have been made to solve the problem of unbalanced distribution of senses and the semantic disambiguation of ancient books based on the linguistic theory. The research in future will focus on further expanding the corpus scale, improving the learning ability of the model, and improving the knowledge mining abilities by constructing a top-level semantic description framework for ancient books. The knowledge organization of ancient books will focus on how to grasp the standardization and standardization of terminologies and how to use the information technology to improve efficiencies. The corpus scale should be further expanded, the learning abilities of the model should be improved, and the abilities of knowledge mining should be enhanced by constructing a top-level semantic description framework for ancient books. The knowledge organization of ancient books will focus on how to standardize the terminologies and how to utilize the information technology to improve its efficiencies.

(c) Information service of ancient books

Current researches on ancient book system are mainly qualitative research, including theoretical discussion and experience summarization. The quantitative research based on investigation and demonstration should be strengthened in future; it focuses on the performance evaluation method and tool, evaluation index system and related empirical research. The ancient book index focuses on solving the problems of integration, standardization and application. Whether the research on ancient book digitization based on index science, or research on index science based on ancient book index, a series of innovative achievements will appear, which will promote further research on ancient book index and guide the practice of ancient book index. The research on automatic translation introduces the latest research achievements such as semantic analysis and knowledge graph; compares the word alignment results of ancient and modern Chinese; and obtains the word pairs. After that, the research revises the translation results of word pairs manually, or imports thesaurus, refines the grammar rule base, so as to improve the translation efficiencies. The semantic retrieval mechanism is explored on the basis of integrating bibliographic database, version database, full-text database and knowledge base. It emphasizes on semantic retrieval, visual retrieval, and semantic web publishing, and makes some transformations from monotonic retrieval to diverse retrieval, from static retrieval to dynamic retrieval, laying theoretical and technical foundations for the future intelligent retrieval.

CONCLUSIONS

There exists few researches on the whole process of ancient book information processing. Most of the existing researches are periodical summarization of the ancient book digitization - these achievements have been made earlier in time. In view of this, this paper reviews and prospects the researches on the ancient book information

processing, and it focuses on the whole process based on the information life cycle theory. It can be seen from the research that with the development of information technologies, especially the natural language processing technologies, the ancient book information processing has attracted more and more researchers' interest and made great achievements. However, in view of the particularity of ancient books, the methods applicable to modern books are not suitable to ancient books, the reason for such appearance is that the expressions between ancient books and modern books are quite different; the available corpus is small-scale and manual annotated; the feature extraction abilities of popular models are limited; the knowledge organization encounters some tough problems; the construction efficiencies of ancient book system is still inefficient; and the personalized and professional ancient book information or solution cannot meet user's expectation. This paper clearly points out the future research direction, including integrating the ancient book information resources and constructing the sharing system; expanding the corpus scale; proposing a series of high-performance models; and conducting the researches on the evaluation method, tool, and index system. The research is beneficial for much more researchers to comprehensively understand the advance of ancient book information processing in China, and it is also convenient for them to research and use the ancient book information resources. Meanwhile, it contributes for other countries to understand China's history and culture and promote interdisciplinary research and international cooperation in ancient book information processing.

ACKNOWLEDGEMENT

This research was supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No. 21JHQ081)

REFERENCES

Bai, S. X. and Bao, Y. L. 2017. LDA-based word image representation for keyword spotting on historical Mongolian documents. *Journal of Modern Information*, Vol. 37, no. 7: 51-54, 88.

Chang, E. 2009. The automatic compilation system construction of the agricultural ancient books. *Researches in Library Science*, no. 6: 10-14.

Chang, E. and Hou, H. Q. 2007. Research on automatic compilation of ancient agricultural books. *Journal of Nanjing Agricultural University (Social Sciences Edition)*, Vol. 7, no. 1: 99-104.

Chang, E., Hou, H. Q. and Cao, L. 2007. Research on automatic version comparison and analysis of ancient book and its realization. *Journal of Chinese Information Processing*, Vol. 21, no. 2: 83-88.

Chang, E., Zhang, C. X., Hou, H. Q. and Hui, F. P. 2013. Automatic word sense disambiguation of ancient Chinese based on vector space model. *Library and*

Information Service, Vol. 57, no. 2: 114-118.

Chang, Y. C., Lu, C. and Zhai, J. P. 2019. Application of knowledge organization of ancient Chinese prose based on linked data. *Library Theory and Practice*, no. 2: 55-59.

Che, C. and Zheng, X. J. 2016. Sub-word based translation extraction for terms in Chinese historical classics. *Journal of Chinese Information Processing*, Vol. 30, no. 3: 46-51.

Chen, C. M. and Chang C. 2019. A Chinese ancient book digital humanities research platform to support digital humanities research. *The Electronic Library*, Vol. 37, no.2: 314-336.

Chen, F. Y., Chen, C. M. and Chang, C. 2019. Development and evaluation of a character social network relationship map tool in an ancient book digital humanities research platform. *Proceedings of the 8th International Congress on Advanced Applied Informatics*, Toyama, Japan: 73-78.

Chen, T. Y., Chen, R., Pan, L. L., Li, H. J. and Yu, Z. H. 2007. Archaic Chinese punctuating sentences based on context n-gram model. *Computer Engineering*, Vol. 33, no. 3: 192-196.

Cheng, Z. and Liu, X. J. 2021. Feature extraction of ancient Chinese characters based on deep convolution neural network and big data analysis. *Computational Intelligence and Neuroscience*, 2021: 2491116.

Deng, Z. H., Huang, X., Lu, Y. J. and Li, M. J. 2014. Discussion about the construction method of ontology library in field of ancient book editions. *Library, Document & Communication*, Vol. 4: 80-87.

Ding, Y. L., Li, R. F. and Li, W. X. 2012. Ancient Chinese musical score translation via instance-based learning. *Proceedings of the 2012 International Conference on Audio, Language and Image Processing*, Shanghai, China: 1035-1040.

Fu, X. J., Yuan, T., Li, X. B. and Wang, Z. G. 2019. Research on the method and system of word segmentation and POS tagging for ancient Chinese medicine literature. *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine*, San Diego, USA: 2493-2498.

Gao, L. C., Zhong, Y., Tang, Y. M., Zhi, T. and Xuan, H. 2011. Metadata extraction system for Chinese books. *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, Beijing, China: 749-753.

Gao, M. 2021. Research on rapid sharing of digital ancient literature resources in cloud computing environment. *Proceedings of the 6th International Conference on Smart Grid and Electrical Automation*, Kunming, China: 248-252.

Gao, S., Jin, P. and Zhang, D. Z. 2019. Research on named entity recognition of TCM classics based on deep learning. *Technology Intelligence Engineering*, Vol. 5, no. 1: 113-123.

Gu, L. L. 2021. Integration and optimization of ancient literature information resources based on big data technology. *Mobile Information System*, Vol. 2021: 6452418.

Guo, W. L., and Dai, Y. Q. 2011. The retrieval research on ancient books digitization. *Library Theory and Practice*, no. 10: 13-16.

Han, F., Yang, T. X. and Song, J. H. 2015. Ancient Chinese MT based on sentence-focused syntax. *Journal of Chinese Information Processing*, Vol. 29, no. 2: 103-110.

Han, X., Wang, H., Zhang, S. Fu, Q., and Liu, S. 2019. Sentence segmentation for classical Chinese based on LSTM with radical embedding. *Journal of China Universities of Posts and Telecommunications*, Vol. 26, no. 2: 1-8.

He, L. and Cao, L. 2006. Research of building and retrieval of ancient agricultural book ontology. *New Technology of Library and Information Service*, Vol. 12: 37-39, 53.

Ho, H. I. B., and De Weerdt, H. 2014. MARKUS: Text analysis and reading platform. Available at: http://dhchinese-empires.eu/beta/.

Huang, J. N. 2011. An experiment of editing multi-text Chinese ancient book indexes based on VFP and Word. *New Technology of Library and Information Service*, no. 10: 85-89.

Huang, J. N. and Hou, H. Q. 2008. On sentence segmentation and punctuation model for ancient books on agriculture. *Journal of Chinese Information Processing*, Vol. 22, no. 4: 31-38.

Huang, J. N. and Hou, H. Q. 2011. An experiment on word segmentation for ancient agriculture books. *Journal of the China Society for Scientific and Technical Information*, Vol. 30, no. 6: 618-625.

Huang, S. Q., Wang, D. B. and He, L. 2015. Research on constructing automatic recognition model for ancient Chinese place names based on pre-Qin corpus. *Library and Information Service*, Vol. 59, no. 12: 135-140.

Jia, F. X. 2015. Construction method of ancient books knowledge base based on knowledge clustering. *Journal of Library Science*, no. 5: 45-48.

Jiang, X., Jiang, Y., Fang, M., and Wang, R. P. 2010. Tree pruning based fast segmentation of classical texts - a case study on *Classic of Tea*. *Journal of Chinese Information Processing*, Vol. 24, no. 6: 10-13, 42.

Jeon, Y. C. 2005. *General catalogue of Chinese ancient books in Korea*. Seoul: Korea Learning Ancient Publishing House.

Li, B., Lu, W., Yuan, W., and Gu, Y. 2017. Discover social relations and activities from ancient Chinese history book Zuo Zhuan. *Proceedings of the 2017 International Conference on Behavioral, Economic, Socio-cultural Computing*, Krakow, Poland: 1-5.

Liang, S. H. and Chen, X. H. 2013. Methodological study of automatic word segmentation in pre-Qin document *Mencius*. *Journal of School of Chinese Language and Culture, Nanjing Normal University*, no. 3: 175-182.

Liu, C. H. 2004. The knowledge representation of ancient Chinese medicine books based on knowledge element. *Proceedings of the third International Convention of Traditional Medicine*, China, 313-314.

Liu, C. L., Huang, C. K., Wang, H. and Bol, P.K. 2015. Mining local gazetteers of literary chinese with CRF and pattern based methods for biographical information in Chinese history. *Proceedings of the 2015 International Conference on Big Data*, Washington, D C, USA: 1629-1638.

Liu, J. Y., and Zhao, X. W. 2017. Construction of computer aided collation repository of ancient editions. *Library Theory and Practice*, no. 3: 54-58.

Liu, Z. B., Dang, J. F., Zhang, Z. J. 2020. Research on automatic extraction of historical events and construction of event graph based on *Historical Records*. *Library and Information Service,* Vol. 64, no. 11: 116-124.

Ma, C. X., Chen, X. H. and Qu, W. G. 2013. Study and design on knowledge network of classical ancient books and commentary literatures. *Library and Information Service*, Vol. 57. no. 9: 124-128.

Mao, J. J. 2006. Development and construction on ancient books digitization in the overseas. *Digital Library Forum*, no. 12: 24-28.

Oriental Culture Research Center of University of Tokyo. 2011. Full text database of rare Chinese ancient books. Available at: http://shanben.ioc.u-tokyo.ac.jp/help.html.

Ouyang, J. and Ren, S. H. 2021. Visualization of ancient texts reading in digital humanities research. *Library Journal*, Vol. 40, no. 4: 82-89.

Qian, Z. Y., Zhou, J. Z., Tong, G. P. and Su, X. N. 2014. Research on automatic word segmentation and POS tagging for *Chu Ci* based on HMM. *Library and Information Service*, Vol. 58, no. 4: 105-11.

Shi, L. J. 2016. Research on common problems and Countermeasures of ancient image database construction. *Library Work and Study*, no. 9: 62-66.

Wang, B., Shi, X., Tan, Z., Chen, Y. and Wang, W. 2016. A sentence segmentation method for ancient Chinese tests based on NNLM. *Proceedings of the 17th Chinese Lexical Semantics Workshop*, Singapore, 387-396.

Wang, B. L., Shi, X. D., Su, J. S. 2017. A sentence segmentation method for ancient Chinese texts based on recurrent neural network. *Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol. 53, no. 2: 255-261.

Wang, C., Zhang, X. H., Han, C. H. 2009. Research on sentence segmentation and punctuation in ancient Chinese. *Journal of Henan University (Natural Science)*, Vol. 39, no. 5: 525-529.

Wang, D. B., Gao, R. Q., Shen, S. and Li, B. 2018. Deep learning-based classification of pre-Qin classics questions. *Journal of the China Society for Scientific and Technical Information*, Vol. 37, no. 11: 1114-1122.

Wang, D. B., Huang, S. Q., and He, L. 2017. Researches of automatic part-of-speech tagging for pre-Qin literature based on multi-feature knowledge. *Library and Information Service*, Vol. 61, no. 12: 64-70.

Wang, P. 2014. Research on information organization and classification of dictionary combined retrieval system based on handed down Chinese Dictionary in Japan, South Korea and China. *The Journal of Chinese Character Studies*, Vol. 10: 1-25.

Wang, S., Xiong, D. L. and Wang, X. X. 2009. The research and implementation of example based machine translation of ancient Chinese. *Journal of Xuchang University*, Vol. 28, no.5: 88-91.

Wang, S. S., Wang, D. B., Huang, S. Q., and He, L. 2018. Research on the automatic word segmentation of The Book of Songs under multi-dimensional domain knowledge. *Journal of the China Society for Scientific and Technical Information*, Vol. 37, no. 2: 183-193.

Wang, Y. C. and Tsai, R. T. H. 2013. Transliteration extraction from classical Chinese buddhist literature using conditional random fields. *Proceedings of the 27th Pacific Asia Conference on Language*, Taipei, Taiwan: 260-266.

Wei, J. Z. and Liu, R. 2019. An approach of constructing knowledge graph of the Hundred Schools of Thought in ancient China. *Proceedings of the 19th ACM/IEEE*

Joint Conference on Digital Libraries, Urbana-Champaign, Illinois: 335-336.

Wei, Y. T., Wang, H. Z., Zhao, J. Q., Liu, Y. T., Zhang, Y. and Wu, B. 2020. GeLaiGeLai: a visual platform for analysis of classical Chinese poetry based on knowledge graph. *Proceedings of the 11th IEEE International Conference on Knowledge Graph*, Nanjing, China: 513-520.

Wu, X., Wu L., Duan, X. T., Ren, T. L. and He, J. 2016. Digital protection and inheritance of culture under the background of "The Belt and Road Initiative". *Proceedings of the 2016 IEEE International Conference on Electronic Information and Communication Technology*, Harbin, China: 231-235.

Xia, C. J., Lin, H. Q. and Liu, W. 2017. Designing a data model of Chinese ancient books for evidence based practice. *Journal of Library Science in China*, Vol. 43, no. 232: 16-34.

Xiao, L. and Chen, X. H. 2010. Automatic detection of version differences among ancient Chinese texts. *Journal of Chinese Information Processing*, Vol. 24, no. 5: 50-55.

Xiao, Y. 2017. Research on the application of the index data of ancient books. *New Century Library*, Vol. 5: 45-48.

Xu, R. H., and Chen, X. H. 2012. A method of segmentation on *Zuo Zhuan* by using commentaries. *Journal of Chinese Information Processing*, Vol. 26, no. 2: 13-17.

Xue, L. G. 1995. Development status and prospect of collected literature digitalization. *Taiwan Branch of National Central Library,* Vol. 4, no. 1: 10-21.

Yao, Y., and Huang, Z. 2016. Bi-directional LSTM recurrent neural network for Chinese word segmentation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Kyoto, Japan: 1197-1206.

Yu, J. S., Wei, Y., and Zhang, Y. W. 2019. Automatic ancient Chinese texts segmentation based on BERT. *Journal of Chinese Information Processing*, Vol. 33, no. 11: 57-63.

Yu, J. S., Wei, Y., Zhang, Y. W., and Yang, H. 2020. Word segmentation for ancient Chinese texts based on nonparametric Bayesian models and deep learning. *Journal of Chinese Information Processing*, Vol. 34, no. 6: 1-8.

Yu, X. J. and Wei, H. F. 2019. A machine learning model for the dating of ancient Chinese texts. *Proceedings of the 2019 International Conference on Asian Language Processing*, Shanghai, China: 115-120.

Yu, L., and Guan, J. W. 2017. A situation and development analysis on the digitalization of ancient books in China. *Digital Library Forum*, no. 11: 41-47.

Yu, L. L., Ding, D. X., Qu, W. G., Chen, X. H. and Li, H. 2009. The ancient Chinese word sense disambiguation based on CRF. *Microelectronics and Computer*, Vol. 26, no. 10: 45-48.

Zhang, K. X., Xia, Y. Q., and Yu, H. 2009. CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua University (Science and Technology)*, Vol. 49, no. 10: 1733-1736.

Zhang, L. Y., and Wang, J. 2020. Design of faceted classification system of ancient book databases. *Library Development*, no. 3: 56-61

Zhang, M., Ma, S. P., Jiang, Z., and Huang, K. 2001. Statistical learning and analyses of Chinese ancient books for information retrieval. *Proceedings of the 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man*

for Cybernetics in Cyberspace, Tucson, USA: 869-873.

Zhang, X., Chen, H., Xu, T. 2013. Deep learning for Chinese word segmentation and POS tagging. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, USA: 647-657.

Zhang, Z. X. 2002. Research on automatic compilation of ancient documents by computer. *Lexicographical Studies*, no. 5: 42-48.

Zhao, H. Y. 2020. Evaluation index system for the service efficiency of digital resources of ancient books. *Library Tribune*, no. 7: 150-160.

Zhou, L. N., Hong, L., and Gao, Z. Y. 2019. Construction of knowledge graph of Chinese *Tang Poetry* and design of intelligent knowledge Services. *Library and Information Service*, Vol. 63, no. 2: 24-33.

Zhu, B. J., and Zhang, J. Z. 2020. Digital humanities cyberinfrastructure for ancient China studies: past, present, and future. *Library Trends*, Vol. 69, no. 1: 319-333.

Zhu, S. L., and Bao, P. 2015. The use of Geographic Information System in the development and utilization of ancient local chronicles, *Library Hi Tech*, Vol. 33, no. 3: 356-368.