A Study of Data Cleaning in Northwest Folk Songs

Chengqi Dai*, Yan Xu, Xin Wen

School of Information Science, Beijing Language and Culture University, Beijing, 100083, China

*dd04236979@163.com

Abstract. The arrival of the digital modernization era has brought a new paradigm of digital humanities research to the humanities, and the integration of digital data, the construction of corpus and the combination of computer analysis and processing have become an important part of digital humanities research. In order to solve the problems of complicated data format and low information content of the Northwest folk song "Hua'er", Python tools are used to clean the data, and through noise reduction, mean filtering, corrosion and other means, the text recognition degree of the picture material is improved, and cleaner data is obtained, which lays the foundation for the establishment of the database of "Hua'er" and lays the foundation for the establishment of the database of "Hua'er". It lays the foundation for the establishment of the "flower" database, and provides a referable program for the preprocessing of folk song data.

Keywords: digital humanities; northwest folk songs; digitization.

1 Introduction

Databases, fueled by big data and digital humanities are quietly becoming a new way of thinking and approaching the study of traditional music and culture[1]. With the development of information technology, the emergence of digital humanities and big data has led to profound changes in the study of humanities, and the information carried by traditional books, pictures, videos and other carriers urgently needs to be changed in order to adapt to the current new academic research environment.

Known as "an encyclopedia reflecting the folk culture of the Northwest China region", the folk song "Hua'er" has already made a lot of achievements in the field of traditional research mediated by words and paper, with a large number of books and audio-visual materials containing "Hua'er". A large number of books and audio-visual materials have been produced. However, in the information age, these materials, which are mostly in the form of books and songs with scores, are facing problems such as scattered data storage, various data formats and low information content, which need to be solved urgently and require digitized and informatized statistical analyses and researches.

Data is the foundation of information, and good data quality is the basic condition for all kinds of data analysis. When researching the data of "Hua'er", we often feel "rich

Z. Zhan et al. (eds.), *Proceedings of the 2024 10th International Conference on Humanities and Social Science Research (ICHSSR 2024)*, Advances in Social Science, Education and Humanities Research 858,
https://doi.org/10.2991/978-2-38476-277-4_115

in data but poor in information" , the reason for this is the lack of specific data analysis techniques for "Hua'er"; secondly, the quality of the data of "Hua'er" is not high. The reason for this is, firstly, the lack of specific data analysis techniques for "Hua'er"; and secondly, the low quality of data on "Hua'er". Different sources of information bring different forms of "dirty data", resulting in the proliferation of "dirty data" for a variety of reasons, such as data input errors, content identification errors, different sources of data caused by different methods of expression, inconsistencies between data, etc.[1] The reasons for the proliferation of "dirty data" are diverse, such as data entry errors, content recognition errors, different representation methods caused by data from different sources, and inconsistencies between data.

In this paper, we will perform goal-driven data cleaning on the text data of "Hua'er", aiming to provide clean and effective data for the research in various fields of "Hua'er" and reduce the impact of "dirty data" on the research results. The aim is to provide clean and valid data for research in various fields of "Hua'er" and to reduce the impact of "dirty data" on research results. After obtaining cleaner data from data cleansing, this project will use this text to build a specialized corpus according to different purposes and needs, in order to contribute to the research of "Hua'er and Children" in various directions[2].

2 Current Status of Domestic and International Research

With the rapid development of the Internet and big data technology, a large amount of data has been generated in various fields, which can be mined as a carrier of information and can be used to obtain more potentially valuable knowledge. In general, data mining always assumes that the data is "clean" and consistent, but in reality, the perceived data is often redundant, incomplete, contains noise, and there is inconsistency, which we collectively refer to as "dirty data".[3] The quality problems exhibited by these data will seriously affect and hinder the effective data mining. Therefore, as a key technology to improve data quality, data cleaning has gradually become a hotspot for many scholars at home and abroad.

Foreign research on data cleansing first appeared in the United States, beginning with the correction of Social Security number errors across the United States[4] . With the development of the information industry and business in the United States in those years, the research on data cleansing technology was also greatly stimulated, and there were many advances and developments. In recent years, foreign data cleaning technology has developed even more rapidly, and English-based data cleaning research is thriving. There are a large number of data cleaning software on the market[5] There are a lot of data cleansing software in the market, including data cleansing software for commercial trade or service industry, as well as data cleansing programs developed by universities or specific research institutes, such as Data Wrangler, Open Refine and so on.

Literature [6] categorizes data quality problems into four types: single data source schema level problems, multiple data source schema level problems, single data source instance level problems and multiple data source instance level problems. The data

formats, design patterns, and data models of multiple data sources can be very different, so when integrating multiple different data sources, many various data quality problems can arise. They mainly include missing values, wrong values, inconsistent data and similar duplicate records. These problems are solved by data cleansing techniques.

At present, the domestic research on data cleaning technology, especially on featured languages (e.g., dialects and chants), is still in its infancy. Directly targeting data cleaning, especially for Chinese featured languages is even less. Literature [7-11] and other studies are basically based on web data with a single, clean text structure. In view of the fact that the corpus to be cleaned in this study is extremely special, which is the lyrics of the Northwest folk song "Hua'er", there are even fewer related studies, and even fewer theoretical results have been reported.

3 Acquisition and Analysis of Raw Data

3.1 Acquisition of Raw Data

Both text mining and data analysis and research are based on massive data resources, so this study decides to clean the data of "Hua'er" and build a corpus of "Hua'er". In order to ensure the sufficient amount of "Hua'er" corpus, this study decides to collect as much "Hua'er"-related corpus as possible, and the main ways to get it are: writings specializing in "Hua'er", The main ways to get the corpus include: books dedicated to "Hua'er", songbook pictures of "Hua'er", audio and video of "Hua'er" singing, etc.

Books and monographs related to "Hua'er": Searching for resources from the Super Star Digital Library, we have obtained five professional books containing a large number of electronic pdf resources on "Hua'er", including A General Introduction to Chinese Hua'er (Wu Yulin, 2008), A Collection of Hua'er (Longya, 2005), Love Hua'er (Zhu Zhonglu, 2002), Selected Hua'er of Northwest China (Xueli, Ke Yang, 2012), and Selected Hua'er of Minzhou (Ji Xucai, 2005). 2005), Love Hua'er (Zhu Zhonglu, 2002), Selected Hua'er of Northwest China (Xueli and Ke Yang, 2012), and Selected Hua'er of Minzhou (Ji Xucai, 2013).

Pictures of song sheets and lyrics of "Flower Child": We collect pictures of song sheets from websites, information and reports related to "Flower Child", and categorize the pictures according to the content of the noise of the pictures.

Audio and video media resources of "Hua'er": collecting audio and video of "Hua'er" singing from relevant websites, information and reports.

3.2 Analysis and Processing of Raw Data

Although there are many sources, the books and writings related to "Hua'er", songbook pictures, singing audio and video, etc. cannot be used directly as text data resources, and need to be analyzed and processed in order to obtain the most basic and original text data.

Books and monographs related to "Hua'er": Firstly, the text in the books was recognized by OCR using Adobe Acrobat DC software. Secondly, use python to clean the recognized text data, remove most of the conventional "dirty data", and then extract the

text of the "flower child" lyrics from it, divide it into parts according to the first, and record the source and genre of each "flower child", and store it in a table. Then we extracted the text of the "flower child" lyrics, divided them by the first song, and recorded the source and genre of each "flower child" song and stored them in a table. Due to the influence of book format and illustrations in the books, the text data of "Huaer" obtained through books contains a lot of redundant information, which needs to be removed from the format, duplicated content, punctuation and other work. Finally, the text data are subjected to word separation and lexical labeling.

The lyrics of the songbook of "Hua'er": firstly, we use the cv2 toolkit based on python and the Image library of PIL, according to the different noise characteristics of each group of pictures, we use different functions, parameters and order to image process each group of pictures to achieve the purpose of noise reduction. Then use tesseract-OCR toolkit based on python to perform text recognition on the images after noise reduction to obtain the text content in the songbook images. After the image noise reduction process, the text data obtained from the songbook pictures of "Hua'er" still has a certain amount of "dirty data", which needs to be cleaned.

Audio and video media resources of "Hua'er": The unique singing style of "Hua'er" combined with dialect lyrics makes it extremely difficult to recognize the audio of Hua'er's singing, and the data obtained is messy and not highly usable. Even after data cleaning, it is difficult to obtain effective data. Therefore, this study finally discarded this part of the data, and no longer used the audio-video corpus of "flower children".

4 Cleaning for Non-textual Data

4.1 Cleaning Data of Books and Monographs Related to "Hua'er".

"Hua'er" related to the picture form of pdf books converted to word text there will be a large number of redundant content, in order to "Minzhou Hua'er Anthology[12] For example, Figure 1, Figure 2.

Fig. 1. Book's title and picture

Fig. 2. Chapter names and their English

Chapter names and their English, book names and their English, and illustrations are redundant information throughout the book, and can be partially removed using formatting, going to headers and footers, etc., with the remainder being done during text data cleansing.

For the book text data cleaning workload is large, the main use of pandas, numpy, matplotlib and other python extensions library, the cleaning steps are: import the relevant toolkit, import the data set, preliminary exploration of the data, simple data processing, duplicate value processing, outliers processing, missing value processing. After a series of cleaning and processing, you can get cleaner text data.

4.2 Cleaning of "Flower Children" Songbook Lyrics Image Data

4.2.1 Image Noise Reduction

In order to remove the noise, this study adopts OpenCV to process the images of the music score of "Hua'er", and removes as much noise as possible through grayscaling, binarization, morphological corrosion, expansion and other operations, and retains the lyrics.

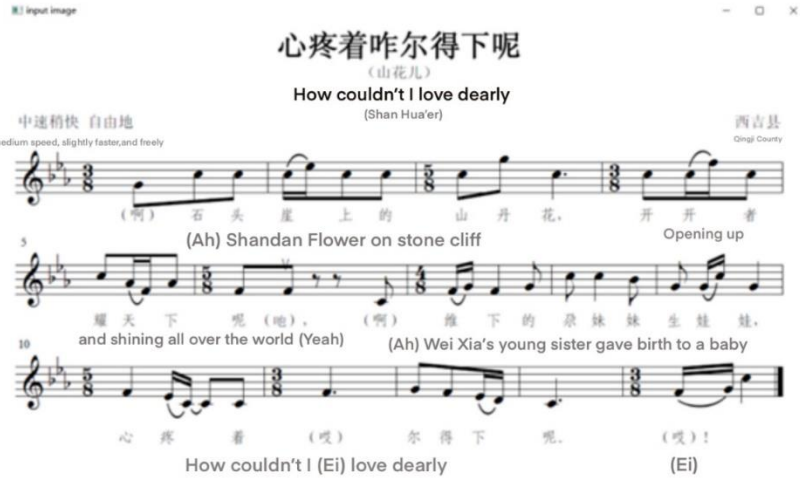

Fig. 3. Original Picture

The input image is the noise information containing the pentameter, song style, and songbook source, as in Figure 3.

Read the "Hua'er" song sheet image function for the cv2 library function, the code is cv2.imread(path), path parameter for the image path storage.

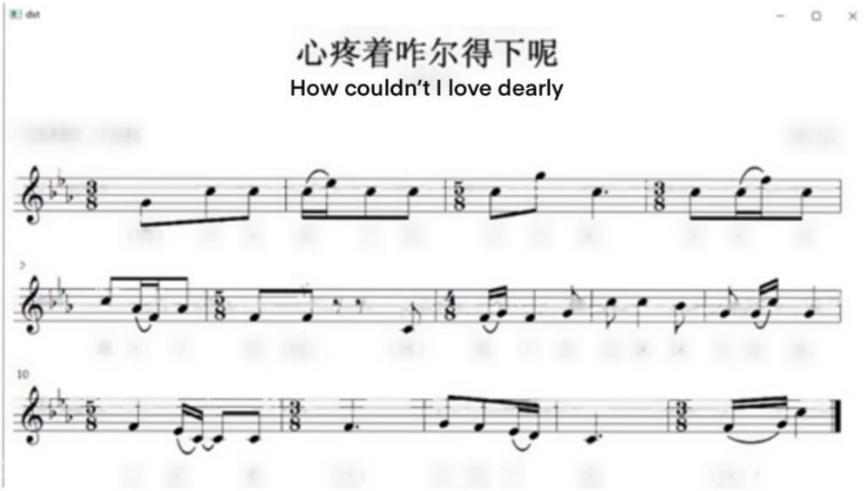

Fig. 4. Filtering, graying

The original input image is filtered and grayscaled and the result is shown in Figure 4.

In order to realize the filtering function using cv2 library function cv2.pyrMeanShift-Filtering, the code for cv.pyrMeanShiftFiltering (image, sp=15, sr=100), parameters: image: set to the input "Hua'er" song sheet image; sp: in order to get a large piece of noise outside the image set to 15; sr: in order to get a more fuzzy results, set to 100. "song score image; sp: set to 15 in order to get a large piece of noise outside the image; sr: set to 100 in order to get a more fuzzy result. use this code to perform mean shift filtering on the input "flower" song score image.

cv2 library function cv.cvtColor function on the "Hua'er" song sheet image grayscale, the code is: cv.cvtColor (dst, cv.COLOR_BGR2GRAY), the parameter dst for the filtered image, the parameter cv.COLOR_BGR2GRAY indicates that this function will convert the color image to grayscale. BGR2GRAY means that this function converts the color image to grayscale.

Mean Filter is a color level smoothing filter for the "Hua'er" songbook image, which can neutralize the colors with similar color distributions, smooth the color details, and erode the smaller color areas. The purpose of using this filter is to obtain the text area of the lyrics in the songbook of "Hua'er" and to prepare for the phase reduction of the image. The purpose of grayscaling the image is to obtain the source file for the next step of expansion and erosion.

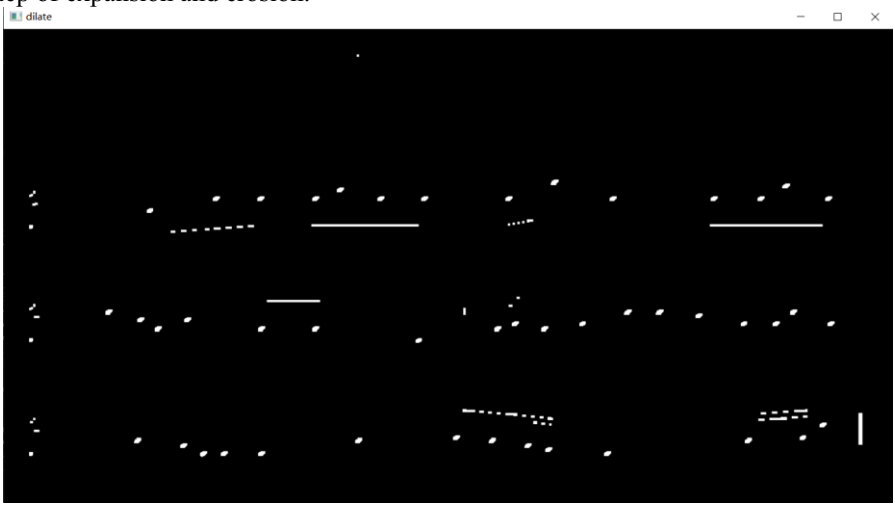

Fig. 5. Expansion

Swelling is performed on the grayscaled image and the result is shown in Figure 5.

The function is cv2 library function erode, the code is cv.erode(binary, None, iterations=2), the parameter binary is the grayscaled image, the parameter None means that there is no anchor point, iterations=2 means that the number of iterations is 2. This function can perform erosion operation on the input "flower The function can corrode the input "Hua'er" song spectrum image with a specific structure element, which determines the shape of the neighborhood during the corrosion operation, and the pixel value of each point will be replaced with the minimum value on the corresponding neighborhood.

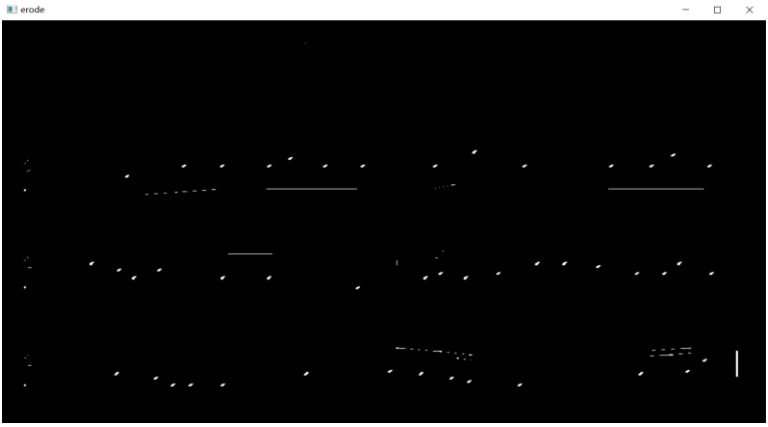

Fig. 6. Corrosion

The grayed out image is corroded and the result is shown in Figure 6.

The function is cv2 library function dilate, the code is cv.dilate(erode, None, iterations=1), the parameter erode is the expanded image, the parameter None means there is no anchor point, iterations=1 means the number of iterations is 1. The function can perform the expansion operation of the input "Hua'er" song sheet image with specific structural elements, the structural elements determine the shape of the neighborhood during the expansion operation, the pixel value of each point will be replaced with the maximum value on the corresponding neighborhood. The function expands the input "flower" image with a specific structure element, which determines the shape of the neighborhood during the expansion operation, and the pixel value of each point will be replaced with the maximum value of the corresponding neighborhood.

The purpose of the two-step operation of expansion and erosion is to obtain the larger noise in the "Hua'er" song sheet image, which is convenient for the subsequent operation to remove.

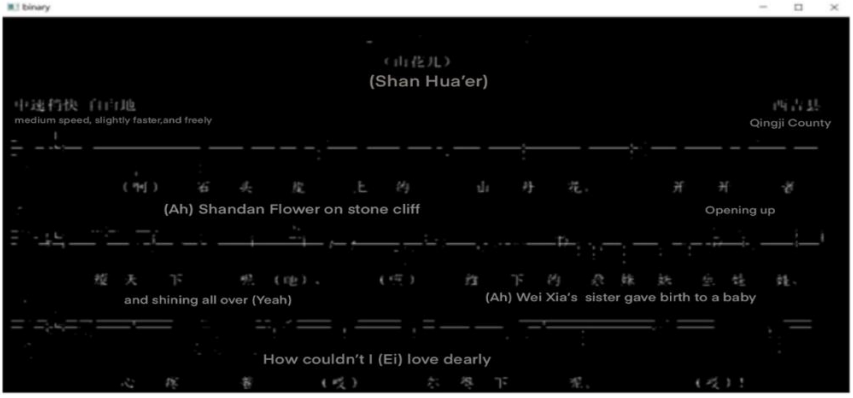

Fig. 7. Binarized picture subtraction

The image of the "Hua'er" song score is binarized and image subtraction is performed, and the results are shown in Figure 7.

Binarization sets the grayscale value of the pixels in the songbook image of "Hua'er" to 0 or 255, so that the whole songbook image presents a clear visual effect of only black and white, and the purpose of this step is to separate the region of interest from the background, and the target of interest is the text area of the lyrics of "Hua'er". The purpose of this step is to separate the ROI (region of interest) from the background, the target of interest in this experiment is the text area of the lyrics of "Hua'er", and the rest can be regarded as the background. Using the cv2 library function threshhold threshold function, the code is cv.threshold(dilate, 0, 255, cv.THRESH_BINARY_INV | cv.THRESH_OTSU), using the threshold function to binarize the image, dilate for the previous expansion of the resulting image, 0, 255 are the two values for binarization, cv.THRESH_BINARY_INV is the binarization thresholding process, and cv.THRESH_OTSU is the OTSU algorithm (Ohtsu method or Maximum Between Classes Variance method).

The last step is image subtraction. The above steps extracted from the "Hua'er" sheet music image noise part of the final image subtraction, the use of the original image to subtract the noise part of the noise, to obtain the noise outside the region. The use of cv2 library function subtract function, code for cv.subtract (binary, image), the parameter binary for the binarized noise image, image for the "Hua'er" songbook binarized image of the original image.

4.2.2 Text Recognition

Image noise reduction processing, the use of pytesseract tools for text recognition of the processed image, the core code for pytesseract.image_to_string (test_message,lang='chi_sim'), image_to_string is pytesseract library text recognition function; test_message and lang='chi_sim' are the two incoming parameters of the function: test_message is the recognition image, and lang='chi_sim' is the qualification of the recognition language as Simplified Chinese (the default recognition language is English). The test image recognition results are shown in Figure 8.

Fig. 8. Recognition results

5 Conclusion

When using other ways to recognize the textual content of folk songs in picture format, pentatonic scores often become the main source of dirty data, in which clefs, beat numbers (often recognized as Arabic numerals), clefs, legato lines, etc. may be recognized, affecting the results of data cleaning.

In this paper, in the process of digitizing the Northwest folk songs, we use the tesseract-OCR toolkit based on python to perform text recognition on the sheet music of "Hua'er" and obtain its text content. The picture (songbook) data of "Hua'er" is cleaned by task-driven data cleaning using pandas, numpy, matplotlib and other program libraries, which reduces the influence of the pentatonic score on text recognition, and the symbol removal rate of the songbook reaches 90.32%, so that a relatively clean "Hua'er" song is obtained. The text data of "Hua'er" is cleaner.

Our next step is to further process the data from other sources to build a more complete corpus and add bricks and mortar to the digitization study of "Huaer", hoping to make contributions to the promotion and development of "Huaer" culture in the field of digital humanities.

References

1. Meng Jian, Hu Xuefeng. Digital Humanities: Media driven Academic Production Mode Transformation [J]. Modern Communication (Journal of Communication University of China), 2019,41 (04): 24-28+54.
2. WANG Yifen, ZHANG Chengzhi, ZHANG Beibei, et al. A review of data cleaning research[J]. Modern Library Intelligence Technology, 2007, (12): 50-56.
3. Li Lei. Research on key technology of similar duplicate record data cleaning in big data environment[D]. Nanjing University of Posts and Telecommunications, 2020. DOI: 10.27251/d.cnki.gnjdc.2019.000328.
4. Mauricio A. Hernández; Salvatore J. Stolfo. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. [J].Data Mining and Knowledge Data Mining and Knowledge Discovery. 1998(1).
5. Wang Ning. Research on Data Cleaning Based on Web Service Information Integration System [D]. Xi'an University of Electronic Science and Technology.2007.
6. Rahm E, Do H H. Data Cleaning: Problems and Current Approaches[J].IEEE Transactions on Knowledge and Data Engineering.2000,23(4):3-13.
7. Hu Yibin, Wu Jingyi, Gao Bo. Digital Protection and Development Strategies for Dongshan Songbooks [J]. Yiyuan, 2023 (03): 98-102.
8. SHEN Zhexu, ZENG Jingjie, DING Jian et al. A study on sentiment categorization of electronic music scores based on pre-trained language models[J]. Journal of Fudan (Natural Science Edition),2022,61(05):581-588. DOI:10.15943/j.cnki.fdxb-jns.20221017.001.
9. Fang Ziling, Kuang Fangjun. Data analysis of Netease ballad lyrics based on Python[J]. Computer and Telecommunications,2018(04):53-56. DOI: 10. 15966/ j. cnki. dnydx. 2018. 04.018.
10. Ma Xuejian Design and Implementation of Score Recognition Software [D]. Nanjing University of Science and Technology, 2022. DOI: 10.27241/d.cnki. gnjgu. 2020. 001928.

11. Yang Yahan Research on Electronic Technology of Paper Spectrum Based on Blocked PCNN [D]. Northwest Normal University, 2023. DOI: 10.27410/d. cnki. gxbfu. 2022. 001066.
12. Ji Xucai. Minzhou flower children anthology. Gansu Culture Press,2013.