

# Keyword Analysis Visualization for Chinese Historical Texts

Jihui Zeng<sup>1,2</sup>  
guaguagua@hnu.edu.cn

Beibei Zhan<sup>1</sup>  
beibei\_zhan@hnu.edu.cn

Shao Zhang<sup>1</sup>  
zhangshao@hnu.edu.cn

Jiajun Bie<sup>1</sup>  
ciwei2016@hnu.edu.cn

Sheng Xiao<sup>\*,1,2</sup>  
xiaosheng@hnu.edu.cn

<sup>1</sup> Hunan University, Changsha, China

<sup>2</sup> National “2011” High Performance Computing Collaborative Innovation Center, China

## ABSTRACT

Historical texts form the basis of the study of antiquities. In the case of Chinese historical texts different genres exist, e.g. chronological and biographical works etc. The contents of these texts normally consist of complex and interrelated information which covers long time period. Traditional history research relies heavily on information extraction and analysis by human researchers. With the recent development of the internet, data science and visualization technologies, digital history gradually attracts more and more attentions and in turn significantly impacts the field of historical study through altering the accessibility of the source materials, the narrative strategy and the analytical methodologies. This paper provides a system that enhances the Chinese historical research using word segmentation, texts analysis and visualization technologies. We can improve the workflow of traditional historical research via automatically detecting important keywords in Chinese historical texts and extracting, analyzing and visualizing the relations between a keyword and other words. This does not only accelerate the text based historical study but also to a great extent increase the scope of the search and analysis of the keywords in Chinese historical texts which used to be limited by the capacity of human researchers.

## CCS CONCEPTS

- Applied computing~Arts and humanities
- Applied computing~Document management and text processing

## KEYWORDS

Word Segmentation, Data Visualization, Historical Document, Digital History

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
VINCI'2019, September 20-22, 2019, Shanghai, China  
© 2019 Association for Computing Machinery.  
ACM ISBN 978-1-4503-7626-6/19/09...\$15.00  
<https://doi.org/10.1145/3356422.3356441>

## 1 INTRODUCTION

Digital history emerged as a cross-disciplinary approach that combines computer technology and historical research. The prospect of digital history does not only lie in its technologies for the fast storage, access and utilization of the historical texts, but also in its ability of visualizing different types of information in novel ways. As a result of the increasing interest to this approach, the American Historical Association (AHA) launched a Digital History Workshop at its annual meeting in 2014 [1]. However, most of the digital history projects, especially those regarding Chinese history, are directed by historians who lack awareness of the variety of technologies that could greatly enhance their work.

Information visualization focuses on representing data and information through the display of graphic forms such as, bar charts, plot charts, pie charts etc. It is widely applied in the fields of business, finance, journalism, administration and digital media. With the rise of digital history, we look to employ and develop information visualization technology for humanistic research.

In this paper, the research methods in the field of data science are applied to the study of Chinese historical texts, where the history texts are regarded as the research data, the data are pre-processed, the nouns in the historical materials are extracted and analysed, and then the visualization techniques such as word cloud and force-directed graph are employed to show the relations between the keywords and other nouns. This development enables the Chinese historians to study the relations between historical figures and events in a much wider range than that they used to.

## 2 RELATED WORK

With the general rise of digital history Chinese humanities scholars were also appealed to apply digital technology to their studies. Wu Ling summed up six trends in the study of history in the age of big data, including the digitalization of historical materials, the improvement in the accuracy of the study, the innovative searching and indexing system of new historical materials, the diversification of historical research, the application

of text analysis methods, the influence of data analysis and cloud technology on history research [2]. Zhou Bei et al. applied quantitative analysis to historical events [3]. Wang Zhixuan proposed two solutions for the digitization of ancient historical materials [4]. In the summary of the Conference on "Cross-border and integration: Digital Humanities in a global perspective", the Library of Peking University reviewed a number of available digital methods for humanities research [5]. It further advocated the combination of digital technologies with traditional humanistic research methods which free humanities researchers from mechanical, painstaking work of textual comparisons and analyzation in order to enable them to explore their research questions in a broader scope.

Word segmentation is widely employed in the study of Chinese and Japanese corpuses [6, 7, 8, 9]. Word segmentation technologies include dictionary-based, HMM-based, CRF-based and deep learning-based algorithms. These algorithms have different performance regarding word segmentation speed, accuracy and new word generation. Among a number of common Chinese word segmentation tools, Jieba [10] and thulac [11] have a faster word segmentation speed and accuracy. However, the training samples of these tools are in modern Chinese. We compared the tools by feeding them with classical Chinese texts and chose THULAC as the word segmentation tool for the study of this paper.

The word cloud, also known as the text cloud, is a visual representation of text data which illustrates words in a shape resembling clouds [12]. The word cloud can visually show the weight of different words and highlight important words.



Figure 1: Force-Directed Graph

By simulating Hooker's law, the force-directed graph can automatically generate graphics for complex relationships [13]. It configures nodes in two-dimensional or three-dimensional spaces, with lines connected between nodes, which are almost equal in length and do not intersect. Each node produces gravitational and repulsive forces, the intensity of which are determined by the properties of the nodes. Starting with a randomly ordered initial state, moving under the co-effect of gravity and repulsion, the system would finally reach a stable state. A clearer visualization of relationships between words is thus generated by the force-directed graph algorithm.

### 3 DOCUMENT PREPROCESS

We use the book *History of Ming* (Ming dynasty, 1368-1644) as our experimental corpus. On one hand, since the *History of Ming* is the last one of the standard histories of imperial China, its language is comparatively closer to modern Chinese, which would contribute to a better performance of our word segmentation. On the other hand, as a historical work compiled in the last dynasty of imperial China, the content of the *History of Ming* is generally richer and more detailed than those compiled in previous dynasties.

We first run Thulac over the *History of the Ming* to convert its text body into a word sequence. The words in the sequence are labelled by the parts of speech (i.e. grammar groups such as nouns, verbs, etc.). Then the total 92411 nouns are selected for our analysis where their occurrences are counted, and locations recorded. Table 1 shows the list of nouns of which the number of occurrences rank top 20.

人	5399	[145, 174, 498, 535, 1365, 1440, 1452, 1552, 2024, ...]
事	2607	[786, 1280, 1285, 2944, 8250, 9604, 11396, 13113, 1 ...]
官	2570	[1492, 3717, 4924, 5456, 6052, 9129, 9460, 10631, 1 ...]
家	1626	[47, 881, 2112, 6039, 7283, 8351, 8393, 10634, 1165 ...]
南	1395	[8654, 11827, 12104, 12492, 15225, 17563, 18784, ...]
水	1382	[1067, 1232, 3345, 5025, 8801, 10352, 12313, 16245, ...]
县	1299	[17539, 19279, 27943, 34107, 41134, 41948, 56193, 6 ...]
洪武	1018	[5361, 17434, 19860, 19933, 22332, 95102, 104658, ...]
御史	1008	[10182, 11423, 12062, 19460, 23110, 26706, 29244, ...]
南京	1004	[5970, 29014, 30624, 31730, 32412, 32428, 32671, ...]
进士	929	[7878, 12462, 13464, 14446, 15505, 16266, 16343, ...]
侍郎	927	[12477, 17557, 17653, 19413, 28770, 31392, 31538, ...]
天下	906	[3641, 5110, 5355, 5436, 5506, 5943, 6187, 6191, 6 ...]
贼	883	[10611, 14094, 25429, 25444, 26167, 39554, 39602, 39 ...]
兵	876	[341, 531, 561, 647, 2277, 3170, 3232, 3632, 4301, 4 ...]
河南	872	[4997, 5685, 5693, 6485, 7096, 8093, 8780, 8936, 9 ...]
太子	844	[898, 8591, 9511, 10070, 12923, 14652, 17297, 1731 ...]
命	841	[1962, 3094, 3404, 4336, 4533, 4555, 4751, 5900, 642 ...]
巡按司	823	[189116, 189302, 189351, 189412, 189424, 189502, ...]
州	822	[2011, 2254, 2470, 3195, 3961, 5620, 6574, 9839, 175 ...]

Figure 2: The first column contains the nouns extracted from the text, the second column shows its number of occurrences and the rest of the columns indicates the locations of the nouns.

The chosen keyword is placed in the center of the canvas with its relevant nouns distributing in the three-dimensional space surrounding it. The word distance between the keyword and a relevant noun is represented by both the color index and the length of the line connecting them. By hovering over a line the exact distance between the keyword and the noun at the other end will appear. An input box is provided to facilitate the interaction, which enables the drawing of the force-directed graph of any chosen keyword.

We classify the nouns into high-occurrence nouns (those which appear over 50 times) and low-occurrence nouns (those which appear under 50 times). Among the high-occurrence nouns the words "people" (人) and "event" (事) are at the top of the ranking list, which reflect the fact that the History of Ming is organized around historical people and events. The other high-occurrence nouns include official titles, names of places and government bodies and names of eminent people. We can also find nouns that highly related to Chinese traditional thoughts, such as family (家) and ritual (礼).

The index of a noun is defined by its distance to the very first word of the sequence. The distance between two words is thus calculated by the numerical difference of their indexes. When a high-occurrence noun is selected, a threshold of 20 of word



involved in a major political event in the mid Ming) etc. This suggests the History of Ming sets their focus on elite class.

## 5.2 Keywords Analysis



**Figure 5: Force-directed graph of Wang Shouren**