Keyword Analysis Visualization for Chinese Historical Texts

Jihui Zeng[1,2]
guaguagua@hnu.edu.cn

Beibei Zhan[1]
beibei_zhan@hnu.edu.cn

Shao Zhang[1]
zhangshao@hnu.edu.cn

Jiajun Bie[1]
ciwei2016@hnu.edu.cn

Sheng Xiao*[,1,2]
xiaosheng@hnu.edu.cn

[1] Hunan University, Changsha, China
[2.] 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Word Segmentation, Data Visualization, Historical Document，Digital History

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.

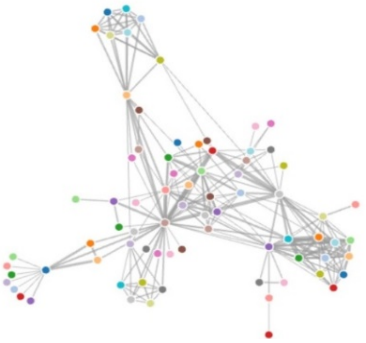


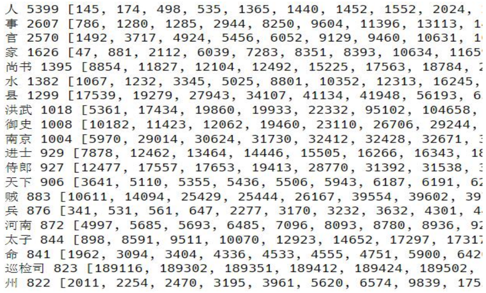
Figure 1: Force-Directed Graph

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
尚书 1395 [8854, 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, 18
侍郎 927 [12477, 17557, 17653, 19413, 28770, 31392, 31538,
天下 906 [3641, 5110, 5355, 5436, 5506, 5943, 6187, 6191, 62
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, 92
太子 844 [898, 8591, 9511, 10070, 12923, 14652, 17297, 1731
命 841 [1962, 3094, 3404, 4336, 4533, 4555, 4751, 5900, 642
巡检司 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 distance is set for analyzing and visualizing its relationship with other nouns.

4 SYSTEM DESIGN

Traditional visualization platforms are normally constructed by desktop application development framework. With the rise of web technologies, JavaScript library provides a pool of visualization tools. Hence our visualization and analysis system are built using node.js.

The algorithm that generates word clouds run iterations over each noun. The noun that ranks in the first place would be put in the center. The size of the graph of (i.e. number of pixels of) each noun is proportion to its number of occurrences. The other nouns are then rendered in the descending order of their numbers of occurrences and their graphs are placed over the areas that are not covered by the stokes of the characters of other nouns.

Wordcloud2.js is a word cloud generating JavaScript library that uses HTML5 's Canvas tags for drawing. Canvas tags provide pixel-level drawing operations that draw large amounts of text or graphics with less system occupancy than text HTML tags.

In the cloud generation, we render all the high-occurrence nouns over a canvas of which the size is 4096x2160 and set the minimum font size to 9 pixels. The result is illustrated in **Figure 3** (Since the weight of "people" (人) is too high to be displayed along with other nouns so it is omitted for a better illustration.)

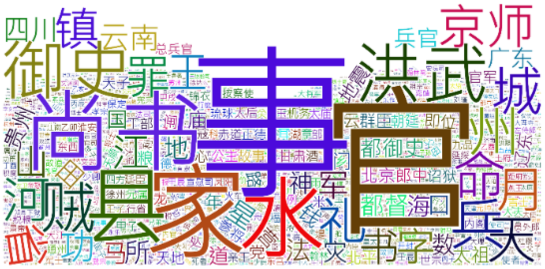


Figure 3: Word cloud of the history of Ming

In order to visualize the relationship between a keyword and the nouns related to them represented by word distances, we choose undirected graph with the nodes to represent the nouns and the connecting lines to represent the relationship. However, when we set the threshold of word distance at 20, the number of nodes in the visualization can vary from none to many and thus an algorithm is needed for the distribution of the nodes and their connecting lines. We employ the force-directed graph algorithm, which adds gravity and repulsion to the nodes through stimulating Hooker's law, so that the positions of the nodes in the kinematics simulation would consistently adjust until the system reaches a stable state. Furthermore, in our case the number of nodes is normally fairly high and the nodes would appear too dense in a 2-d plane so a three-dimensional force-directed graph is rendered for the visualization (See **Figure 4**)

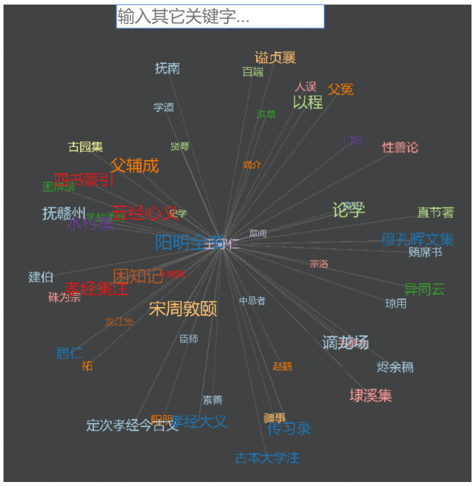


Figure 4: Force-directed graph based keyword analysis tool

5 CASE STUDY

In this section we analyze the History of Ming through our keyword analysis system and demonstrate its utility by conducting two user cases.

5.1 Word Cloud of the History of the Ming (1368-1644)

The word cloud provides an intuitive view of the weight of different words by assigning the font size in proportion to its weights. It not only allows readers to quickly capture important words with larges weights, but also brings a wider range of words to the reader's attention. In the field of journalism, the word cloud is often used to show popular words for a particular period of time.

Inspired by this concept of popular words in the field of journalism, we assigned the number of occurrence of a noun as its weight and thus generate a word cloud of the History of Ming. The high-occurrence nouns to an extent reflect the main contents of the History of the Ming. Those nouns either represent the factual aspect of the Ming, i.e. the eminent figures and important places and events of the dynasty; or embody traditional Chinese thoughts with a strong sense of Confucian values.

The Ming dynasty is widely recognized as being ruled by a highly autocratic government. In our word cloud in **Figure 3** the “official” (官) , "minister" (尚书)， "imperial censor" (御史)， "prince" (王) pop up as high-occurrences nouns. In particular "Hongwu" (洪武), the reign title of the founding emperor of the dynasty, which refers both to the reign (1368-1398) and the emperor appear among the most visible nouns. The founding emperor laid the institutional foundation for the entire dynasty and thus played a significant role in the history of the dynasty.

With a closer inspection we can detect the names of other historical figures such as "Xu Da" (徐达， a famous general of early Ming period)， "Li Wenzhong" (李文忠, nephew of the founding emperor) and "Zhang Cong" (张璁, a eminent official involved in a major political event in the mid Ming) etc. This suggests the History of Ming sets their focus on elite class.

The terms such as "Army of prince of Yan" (燕兵), "prince of Yan" （燕王）and (the city of) "Beiping" (北平) are also fairly visible in the word cloud. These are the keywords of the prince of Yan's (also known as the emperor Chengzu of Ming, reign 1403-1424) usurpation of the throne and his subsequent relocation of the capital from Jinling (modern Nanjing) to Beiping (modern Beijing). This event is one of the turning points of the dynastic history.

The other group of visible nouns includes "family" (家), "heaven" (天), "rituals" (礼) which is highly relevant to Confucian values.

Our word cloud of the History of Ming offers an intuitive and structured insight into its content. To an extent through data analysis it provides important evidence which proofs the general impression regarding the nature of the dynastic history as a history of eminent people, the historical events that were centered with them and the traditional values that were propagated by them.

5.2 Keywords Analysis

Besides the global view of the main elements of the History of Ming provided by the word cloud, our system also offers a close-up view of a chosen keyword and its relevant nouns through the force-directed graph.

Figure 5 illustrates a force-directed graph with the keyword "Wang Shouren"(王守仁, 1472-1529), a famous official-scholar of the Ming dynasty. In this graph a number of nodes stand for his works, for example the "instructions of practical living" (传习录) and the "complete collection of Yangming" (阳明全书) which reflects his scholarly contributions while "demotion to [be a governor] of Longchang" (谪龙场), "protector of the south" (抚南) and "protector of Ganzhou" (抚赣州) refer to his political career. The automatic generation of force-directed graphs over keywords would be especially useful for the construction of knowledge maps of historical words.

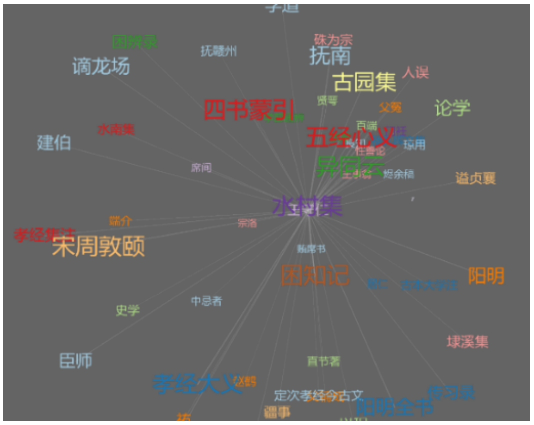


Figure 5: Force-directed graph of Wang Shouren

6 CONCLUSIONS

In this paper we present a system that applies the research method from data science to historical research. We developed tools for data analysis and visualization by constructing word cloud and force-directed graphs over the corpus of the History of Ming. Obviously, these tools can be applied to the corpus of any Chinese historical works. This system is a step towards a cross-disciplinary solution for history research which could replace a large amount of manual work of information extraction from historical texts and analysis with computer-assisted automatic extraction and analysis. Furthermore, the intuitive visualization by our system also enables amateurs to have a better grasp of the main content of historical texts, which suggests its good potential in history education.

ACKNOWLEDGEMENT

This work is supported in part by NSFC-61300217, National “2011” High Performance Computing Collaborative Innovation Center Funds, Hunan University Young Scholar Funds, and Education Reform Project Fund, Hunan Province.

REFERENCES

[1] Welcome to the Digital History Workshop, https://projects.iq.harvard.edu/dhworkshop/home (last accessed 06/08/2019)

[2] Wu Ling, 大数据时代历史学研究若干趋势(Several Trends of Historical Research in the Era of Big Data), http://his.cssn.cn/lsx/sjls/201511/t20151103_2555403_1.shtml

[3] Zhou Bei, Zhang Yan, 清代基层社会聚众案件的量化分析(Quantitative Analysis of Grass-roots Social Gathering Cases in the Qing dynasty), http://www.cssn.cn/zt/zt_xkzt/zt_lsxzt/jjszsxyj/dsjylhyj/201512/t20151224_2798318.shtml

[4] Wang Zhixuan, 古代史料数字化的两种解决方案及其实施(Two Solutions for the Digitization of Historical Materials), http://www.cssn.cn/zt/zt_xkzt/zt_lsxzt/jjszsxyj/gnwszsxyj/201512/t20151221_2791567.shtml

[5] Zhu Benjun, Nie Hua. 跨界与融合: 全球视野下的数字人文——首届北京大学“数字人文论坛”会议综述(Border-crossing and Fusion: Digital Humanities from a Global Perspective，Summary of the 1st Digital Humanities Conference at Peking University) [J].大学图书馆学报（Journal of Academic Libraries）, 2016, 34(5): 16-21.

[6] Saffran J R, Newport E L, Aslin R N. Word Segmentation: The Role of Distributional cues[J]. Journal of Memory and Language, 1996, 35(4): 606-621.

[7] Xue N, Shen L. Chinese Word Segmentation as LMR Tagging[C]//Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17. Association for Computational Linguistics, 2003: 176-179.

[8] Nakagawa T. Chinese and Japanese Word Segmentation using Word-level and Character-level Information[C]//Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, 2004: 466.

[9] Papageorgiou C P. Japanese Word Segmentation by Hidden Markov Model[C]//Proceedings of the Workshop on Human Language Technology. Association for Computational Linguistics, 1994: 283-288.

[10] Sun J. Jieba[J]. Chinese Word Segmentation tool, 2012.

[11] Zhongguo Li, Maosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguistics, vol. 35, no. 4, pp. 505-512, 2009.

[12] Heimerl F, Lohmann S, Lange S, et al. Word Cloud Explorer: Text Analytics based on Word Clouds[C]//2014 47th Hawaii International Conference on System Sciences. IEEE, 2014: 1833-1842.

[13] Fruchterman T M J, Reingold E M. Graph Drawing by Force‐Directed Placement[J]. Software: Practice and experience, 1991, 21(11): 1129-1164.