Construction and Application of a Word and Entity Annotation-Based Digital Humanities Knowledge Base for Classical Texts: A Case Study of "Zizhi Tongjian: Zhou, Qin, and Han Annals" (Postprint)

Authors: Chang Bolin, Wan Chen, Li Bin, Chen Xinyu, Minxuan Feng, Wang Dongbo

Date: 2023-04-01T16:02:59+00:00

Abstract

[Purpose/Significance] To explore methodologies for constructing humanities knowledge bases that enable word- and entity-based retrieval and knowledge mining.

[Method/Process] Taking "Zizhi Tongjian: Zhou, Qin, and Han Annals" as a case study, after automatic word segmentation and part-of-speech tagging of 68 volumes containing 600,000 characters, we manually annotated entity information such as persons, location GIS, and time in the text to implement word- and entity-based full-text retrieval and map retrieval systems. Utilizing co-occurrence information, we statistically extracted interpersonal relationships and travel itineraries of figures. Subsequently, using the TF-IDF method and time series analysis, we mined results including eventful periods, prominent figures, and significant places.

[Results/Conclusion] Deep information annotation based on words and entities can resolve retrieval challenges such as the lack of word boundaries, homonymy (same name referring to different entities), and synonymy (different names referring to the same entity), and can further provide foundational support for multi-angle knowledge discovery and knowledge services for ancient texts.

Full Text

Construction and Application of a Digital Humanities Knowledge Base for Ancient Books Based on Word and Entity Annotation: A Case Study of the Zhou, Qin, and Han Annals of *Zizhi Tongjian*

Chang Bolin[1], **Wan Chen**[2], **Li Bin**[1], **Chen Xinyu**[1], **Feng Minxuan**[1], **Wang Dongbo**[3] [1]School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097 [2]Department of Chinese Language and Literature, Fudan University, Shanghai 200433 [3]College of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract: *[Purpose/Significance]* This study explores methods for constructing a humanities knowledge base that enables retrieval and knowledge mining based on words and entities. *[Method/Process]* Taking the Zhou, Qin, and Han Annals of *Zizhi Tongjian* as a case study, we first performed automatic word segmentation and part-of-speech tagging on 68 volumes containing 600,000 characters, then manually annotated entity information including persons, locations (GIS), and time references. This enabled full-text retrieval and map-based search systems grounded in words and entities. Using co-occurrence information, we statistically derived character relationships and travel itineraries. Furthermore, employing the TF-IDF method and time series analysis, we identified turbulent periods, pivotal figures, and significant locations. *[Results/Conclusion]* Deep information annotation based on words and entities resolves retrieval challenges arising from ambiguous word boundaries, homonymy (same name, different referent), and synonymy (different names, same referent). More importantly, it provides foundational support for multi-angle knowledge discovery and knowledge services for ancient books.

Keywords: *Zizhi Tongjian*; digital humanities; knowledge mining; ancient book retrieval; classical Chinese information processing

---

China's ancient books and documents, vast in quantity and comprehensive in scope, represent a treasure trove for research on Chinese language, literature, history, and culture. Since the late 20th century, significant progress has been made in the digitization of ancient books and the development of character-based full-text retrieval systems, resulting in a substantial number of usable electronic databases [**?**]. With the rise of digital humanities technologies [**?**], the international historical community has begun transitioning from paper-based historical narratives to structured historical databases. Projects such as the Herodotus historical database [**?**] and the China Biographical Database (CBDB) [**?**] have attempted to describe and correlate historical elements like time, persons, and locations, creating searchable and visualizable historical data platforms. These platforms serve both as foundational infrastructure for academic research and as windows for public education, greatly facilitat-

ing scholarly work—particularly interdisciplinary research—without requiring expert-level classical literacy or extensive historical knowledge [**?**].

However, three major problems remain to be solved in the construction and application of such databases for Chinese ancient books. First, retrieval must evolve from character-based to word-based. Since classical Chinese lacks word boundaries, achieving English-like word retrieval functionality requires word segmentation. For example, when searching for the word "军" (army), results should only include contexts where "军" appears as an independent word, not as part of compounds like "将军" (general) or "护军" (protector-general). Second, annotation must advance from proper name indexing to entity annotation. While many classic ancient books have manually indexed proper names (persons, places, book titles), relying solely on underlines and wavy lines cannot distinguish different types of proper names, nor can it resolve homonymy (multiple entities sharing one name) or synonymy (one entity having multiple names). For instance, searching for "秦始皇" (Qin Shi Huang) should retrieve not only contexts containing this exact string but also all contexts referring to this person, including "嬴政" and "! 政". Comprehensive 梳理 of entity information for different persons, places, and times, with each entity assigned a unique identifier in the text, is necessary to meet the demands of detailed retrieval and statistical analysis. Third, retrieval must move beyond full-text search toward knowledge mining and visualization. Existing retrieval platforms mostly provide character-based results, but after annotating entities like persons, places, and times, data mining techniques can reveal relationships between entities, which can then be visualized intuitively. Therefore, exploring knowledge base construction methods that enable word- and entity-based retrieval and knowledge mining is essential.

2 Research Status

Zizhi Tongjian, a chronicle spanning from 403 BCE to 959 CE, holds immense historiographical and literary value. Research has traditionally focused on editions, punctuation, annotations, and literary-historical analysis. The 1956 punctuation edition published by Ancient Books Press [**?**] was followed by Dong Zhiqiao's identification of punctuation errors in 1988 [**?**]. Annotations fall into three categories: topical, abridged, and complete [**?**]. Chen Shengyong evaluated the political and ethical functions of its historiography [**?**], while Zhao Zhengyang outlined its historical value and contributions from a historiographical perspective [**?**].

Digitization of ancient books and character-based full-text retrieval have matured, yielding numerous databases [**?**]. Notably, in 2014, Zhonghua Book Company launched the high-quality *Zhonghua Classic Ancient Books Database* [**?**], which includes *Zizhi Tongjian* with features for reading full texts, converting chronological eras, and indexing personal names. Proper names are marked with special lines—underlines for persons, places, official titles, and ethnic groups; wavy lines for book titles. However, automatic extraction has high omission

rates and fails to resolve homonymy and synonymy issues. Linguistic annotations (word segmentation, part-of-speech tagging) and GIS information are insufficient, requiring more comprehensive information.

Research on classical Chinese word segmentation and part-of-speech tagging has progressed steadily [**?**]. Although classical Chinese is predominantly monosyllabic, polysyllabic words constitute a significant proportion, particularly in personal names, official titles, and temporal expressions. Word retrieval is impossible without segmentation, and fine-grained part-of-speech distinctions (nouns, verbs, personal names, time expressions) are crucial for classical Chinese research and for distinguishing different word classes in retrieval. Due to high construction costs, only ten-million-character corpora currently exist, primarily Nanjing Normal University's Pre-Qin Corpus [**?**] and Medieval Chinese Corpus [**?**], and the Academia Sinica's Ancient, Medieval, and Early Modern Chinese corpora [**?**].

Given *Zizhi Tongjian*'s massive scale, this study selected the earliest Zhou, Qin, and Han dynasties for initial development, addressing the earliest periods first while enabling comparative analysis with similar texts like *Zuo Zhuan* and *Shiji*. Considering limitations of character-based retrieval and automatic ontology construction, this research adopts a carpet-style full-text annotation approach based on words and entities to integrate more information for knowledge mining and visualization. Current entity annotation is limited to persons and locations. Table 1 illustrates the three-level annotation system: word segmentation (using spaces as boundaries), part-of-speech tagging (nouns, verbs, punctuation, etc.), and entity ID annotation. This enriches each word in every sentence with comprehensive information. By assigning unique IDs to personal and place names, we resolve homonymy and synonymy issues. These IDs are drawn from person and place information tables and remain consistent with entity IDs in the *Zuo Zhuan* and *Shiji* knowledge bases.

3 Construction of the Digital Humanities Knowledge Base for *Zizhi Tongjian · Zhou Qin Han Ji*

3.1 Data Sources

The base text of *Zizhi Tongjian* uses traditional characters, comprising 294 volumes and approximately 3 million characters total. This study primarily references the 1956 Zhonghua Book Company edition [**?**] for collation. To date, we have completed collation and annotation of 68 volumes (600,000 characters) covering the Zhou, Qin, and Han dynasties.

3.2 Word Segmentation and Part-of-Speech Tagging

Manual word segmentation and part-of-speech tagging for classical Chinese is time-consuming and labor-intensive. This study employed machine-automated annotation supplemented by manual correction, significantly accelerating

progress. We adopted the segmentation and tagging system developed by Chen Xiaohe et al. [?] and used Nanjing Normal University's Classical Chinese Part-of-Speech Tagging System [?], which achieves over 85% overall accuracy, followed by comprehensive manual proofreading to create high-quality annotated data.

3.3 Entity Information Annotation

3.3.1 Person Information *Zizhi Tongjian* features multiple names for individuals and frequent name-sharing among different persons, requiring careful disambiguation based on annotations and related materials. Each person entity is assigned a unique ID. If a person appears in *Zuo Zhuan* or *Shiji*, we retain the existing ID; new persons receive new IDs. Person information includes all names, gender, and nationality. Since individuals may have numerous names in classical texts, we designate a "primary name" commonly used in later generations as the standard form for retrieval and visualization. As shown in Table 2 , the person with ID 131, "叔孙州仇," has four names, is male, and belongs to the state of Lu.

3.3.2 Location Information Similar to person annotation, location information builds upon *Zuo Zhuan* and *Shiji* data. New locations in *Zizhi Tongjian* receive new IDs with geographic entity information including location type (state, feudal state, river, mountain, etc.), present-day location, source of textual evidence, and GIS coordinates from Baidu Maps. Primary references include the *Historical Atlas of China* [?] and the CHGIS database [?]. Table 3 provides basic information for the feudal state "邾" (Zhu).

3.3.3 Temporal Information Based on works such as *Chronological Studies of Pre-Qin Masters* [?], each chapter's reign year is mapped to the Gregorian calendar. For example, "Volume 1 · Zhou Ji I · Year 21" corresponds to 381 BCE.

3.4 Database Architecture

Based on the electronic full text, word segmentation, part-of-speech tagging, and entity annotation, we constructed the *Zizhi Tongjian · Zhou Qin Han Ji* database. It comprises six tables: person entities, location entities, text, annotated text, person co-occurrence, and person-location co-occurrence. The structure is detailed in Figure 1 [Figure 1: see original paper]. Using IDs from the entity tables, each person and place in the main text is annotated with its ID. Within the same sentence, different persons co-occur, and persons co-occur with locations. From these co-occurrence patterns in the annotated text table, we extract the person co-occurrence and person-location co-occurrence tables.

4 Full-Text Retrieval Based on Words and Entities

4.1 Retrieval Framework

To serve the public, this study developed an online retrieval system for *Zizhi Tongjian* using web technologies (test version at www.dhbase.com/zztj). The system architecture is shown in Figure 2 [Figure 2: see original paper]. Beyond word-based full-text retrieval, it provides multiple query methods including person, location, and part-of-speech searches, all built upon the underlying structured digital humanities knowledge base.

4.2 Full-Text Entity Retrieval

Unlike traditional string-matching retrieval, full-text entity retrieval operates on annotated text, offering more precise word- and entity-based search that avoids redundancy, omission, and misalignment caused by rigid character matching. For example, Figure 3 [Figure 3: see original paper] shows results for the word "军" (army) that exclude occurrences within compounds like "将军" (general) or "护军" (protector-general), reducing results from 2,098 to 1,872.

4.3 Person Retrieval

The person retrieval function provides basic information including primary name, aliases, gender, and nationality, with cross-database integration showing appearances in *Zuo Zhuan* and *Shiji*. When searching for "汉武帝" (Emperor Wu of Han), the system retrieves contexts containing "武帝" (Emperor Wu) and "刘彻" (Liu Che), not just the exact string "汉武帝." Figure 4 [Figure 4: see original paper] illustrates this functionality. Additionally, the person map retrieval uses person-location co-occurrence data with Baidu Maps to display possible travel locations, and person-person co-occurrence data with ECharts [**?**] to visualize social networks. Figure 5 [Figure 5: see original paper] shows Emperor Wu's primary name, aliases, gender, nationality, and appearances in *Shiji* and *Zuo Zhuan*. Figure 6 [Figure 6: see original paper] approximates Emperor Wu's social circle, where node size indicates co-occurrence frequency and relationship strength.

4.4 Location Retrieval

The location function provides basic information including name, type, and present-day location, with cross-database integration showing appearances in *Zuo Zhuan* and *Shiji*. Using GIS coordinates, it displays locations on modern maps via Baidu Maps. Searching "长安" (Chang'an) reveals its name, type, present-day location, and precise modern map position.

4.5 Part-of-Speech Retrieval

Beyond entity retrieval, all common words have been segmented and tagged, enabling part-of-speech-based search and statistics. The part-of-speech statistics

function provides all entries for a given category and frequency bar charts rendered with ECharts. Figure 7 [Figure 7: see original paper] shows the frequency distribution for verb retrieval ("v"), facilitating classical Chinese research.

5 Quantitative Analysis and Knowledge Mining

Based on the knowledge base and retrieval system, we conducted in-depth quantitative analysis yielding data unattainable through traditional qualitative methods. The *Zizhi Tongjian · Zhou Qin Han Ji* contains 4,588 person entities and 1,451 location entities. Person statistics show an average of 1.95 names per person, with over half having multiple names—50% have 2-5 names, and 3% have 6 or more. The person with the most names is Liu Bang (Emperor Gaozu of Han), demonstrating the necessity of unique ID assignment.

5.1 Word Frequency Statistics

Unlike traditional literary-historical analysis, digital humanities can identify themes and hotspots through keyword frequency. Based on segmented text, we obtained word frequency statistics: 2,610 monosyllabic words and 7,970 polysyllabic words. The most frequent word is "之" (particle) with 5,038 occurrences. Table 4 lists the top 10 polysyllabic words, predominantly content words related to power struggles—"天下" (all under heaven), "诸侯" (feudal lords), "陛下" (Your Majesty), "将军" (general)—reflecting *Zizhi Tongjian*'s focus on dynastic contention. "天下" appears most frequently among polysyllabic words (569 times). This word-based approach also enables diachronic studies of lexical evolution.

5.2 Entity Relationship Mining and Cross-Text Comparison

5.2.1 Most "Socially Connected" Figures Traditional character evaluation relies on subjective assessment of historical roles. Quantitative co-occurrence analysis provides an objective approximation of social networks and historical status—more co-occurring figures indicate broader connections and potentially higher status. Comparing *Zuo Zhuan* and *Shiji* data reveals distinct characteristics. To align with *Shiji*, we limited *Zizhi Tongjian* data to Emperor Wu's period. Table 5 lists the top 10 most connected figures, led by Emperor Gaozu, Emperor Wu, and Xiang Yu. The comparison shows *Shiji* and *Zizhi Tongjian* both emphasize the Qin-Han period.

5.2.2 Character Travel Distance *Zizhi Tongjian · Zhou Qin Han Ji* records extensive temporal, personal, and geographic information. Person-location co-occurrence approximates travel itineraries. Using location coordinates, we calculate spherical distance between two points A (latitude ϕ_1, longitude λ_1) and B (latitude ϕ_2, longitude λ_2) with formula (1):

$$\text{Distance}(A,B) = 111.999 \times \sqrt{(\varphi_1 - \varphi_2)^2 + (\lambda_1 - \lambda_2)^2 \cos^2 \frac{\varphi_1 + \varphi_2}{2}}$$

Summing distances chronologically estimates total travel. Table 6 shows the top 10 travelers: 4 monarchs, 3 military strategists, 2 founding officials, and 1 diplomat. Emperor Gaozu traveled over 140,000 kilometers, reflecting his campaign to establish the dynasty. Cross-database comparison reveals different textual emphases and stylistic tendencies.

5.3 Diachronic Entity Analysis

5.3.1 "Turbulent Times"—Entity Distribution Over Time Analyzing entity frequency density across time reveals period-specific differences. Mapping entities to the Gregorian calendar from 403 BCE to 87 BCE, Figure 8 [Figure 8: see original paper] shows person (blue) and location (orange) counts. Both peak around 207 BCE, reflecting the decisive Battle of Julu, and rise simultaneously around 154 BCE, marking the Rebellion of the Seven States. This spatiotemporal approach quickly locates major historical events.

5.3.2 "Pivotal Figures and Places"—Specific Entity Mining As a chronological history, *Zizhi Tongjian* offers rich temporal data for mining period-specific components. Using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm proposed by G. Salton [**?**], we identify distinctive figures and places for each era. Higher TF-IDF indicates stronger text specificity. Analysis yields era-specific pivotal figures, visualized on a timeline using ECharts' flow diagram in Figure 9 [Figure 9: see original paper], showing historical evolution as figures successively dominate the stage: Wu Qi (active 412-381 BCE), Qin Shi Huang (259-210 BCE), Emperor Wen and Emperor Jing of Han (governing the Wen-Jing era until 141 BCE).

Similarly, Figure 10 [Figure 10: see original paper] maps pivotal locations across eras, revealing how different places successively become focal points—feudal capitals or strategic military sites—reflecting dynastic change.

6 Conclusion and Future Work

In today's landscape of digitized ancient books and ubiquitous full-text retrieval, a critical question is how to integrate digital humanities technologies with China's rich historical literature resources to enable visualization and big data analysis beyond basic search. This study proposes a full-text word annotation solution to address word boundary ambiguity and unclear entity concepts, constructing a digital humanities knowledge base for *Zizhi Tongjian*·*Zhou Qin Han Ji* with word segmentation, part-of-speech tagging, and entity annotation. We developed a word- and entity-based retrieval system visualizing travel routes and relationships using Baidu Maps and ECharts. Quantitative analysis exhaustively counted entities and conducted multi-angle mining: social networks, travel maps, turbulent periods, and pivotal figures/places. Comparison with *Zuo Zhuan* and *Shiji* revealed textual differences.

Future work includes: (1) Expanding annotation to the complete *Zizhi Tongjian*

with thorough verification; (2) Refining entity annotation using latest scholarly research, increasing system openness with error-correction mechanisms, and expanding annotation to official titles, reign eras, artifacts, etc.; (3) Exploring advanced mining methods, optimizing co-occurrence calculations, and adding relationship categories (friend, relative, superior-subordinate); (4) Improving retrieval integration and visualization; (5) Linking with library and museum databases to integrate *Zizhi Tongjian* with other historical documents and artifacts.

References

[1] Ji Peipei. Comparative study of 10 common full-text ancient book databases [J]. *Library Science Research*, 2020(20): 71-80. [2] Liu Wei, Ye Ying. Exploring the technical system and theoretical structure of digital humanities [J]. *Journal of Library Science in China*, 2017, 43(5): 32-41. [3] The Open University. Hestia [EB/OL]. [2021-05-21]. https://hestia.open.ac.uk/. [4] China Biographical Database Project Committee. China Biographical Database (CBDB) [EB/OL]. [2021-05-21]. https://projects.iq.harvard.edu/chinesecbdb. [5] Ouyang Jian. Large-scale ancient text visualization analysis and mining for digital humanities research [J]. *Journal of Library Science in China*, 2016, 42(2): 66-80. [6] Song Yanshen. Exploring *Zizhi Tongjian* research since the founding of the PRC [J]. *Journal of Northeast Normal University*, 1983(5): 88-93. [7] Dong Zhiqiao. Punctuation errors in *Zizhi Tongjian* [J]. *Studies on Ancient Chinese*, 1988(01): 83-87, 36. [8] Lin Song. On Southern Song *Zizhi Tongjian* annotations [J]. *Ancient Civilizations*, 2007(1): 74-81. [9] Chen Shengyong. *Zizhi Tongjian*: Analysis of traditional Chinese historiographical functions [J]. *Historiography Quarterly*, 1995(4): 74-80, 146. [10] Zhao Zhengyang. Overview and historiographical value of Sima Guang's *Zizhi Tongjian* [J]. *Northern Literature*, 2019(9): 41-42. [11] Zhonghua Book Company. Zhonghua Classic Ancient Books Database [EB/OL]. [2021-05-21]. http://publish.ancientbooks.cn/docShuju/platformSublibIndex.jspx?libId=6. [12] Deng Sanhong, Hu Haotian, Wang Hao, et al. Research status and future trends of automatic classical Chinese processing [J]. *Journal of Science and Technology Information Research*, 2021, 3(1): 1-20. [13] Chen Xiaohe, Feng Minxuan, Xu Runhua, et al. *Pre-Qin Document Information Processing* [M]. Beijing: World Book Publishing Company, 2013. [14] Wang Xiaoyu. Design and implementation of a medieval Chinese corpus [J]. *Lexicographical Studies*, 2017(3): 17-26. [15] Academia Sinica Ancient Chinese Tagged Corpus [EB/OL]. [2021-05-21]. http://lingcorpus.iis.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh. [16] Dong Hui, Xu Lei, Wang Fei, et al. Semantic analysis system research (III)—Implementation of Chinese historical semantic analysis system [J]. *Journal of the China Society for Scientific and Technical Information*, 2014, 33(2): 204-214. [17] Sun Xianbin. Ontology-based ancient book analysis system development—Case study of "Zizhi Tongjian Analysis System" [C]//Proceedings of Scientific Data Management, Repository and Application Practice Workshop, 2019. [18] Peng Weiming, Song Jihua. Research on

Zizhi Tongjian historical domain ontology construction and application [J]. *Journal of Chinese Information Processing*, 2010, 24(2): 33-38. [19] China Historical Geographic Information System (CHGIS) [EB/OL]. [2021-05-21]. https://sites.fas.harvard.edu/~chgis/. [20] Yan Chengxi, Wang Jun. Digital humanities perspective: Visualizing Song dynasty political networks through symbolic analysis [J]. *Journal of Library Science in China*, 2018, 44(5): 87-103. [21] Li Bin, Wang Lu, Chen Xiaohe, et al. Ancient text annotation and visualization from a digital humanities perspective—Case study of *Zuo Zhuan* knowledge base [J]. *Journal of Academic Libraries*, 2020, 38(5): 72-80, 90. [22] Li Bin, Li Yaxin, Qian Yue, et al. From history book to digital humanities database: The basic annals of *Shiji* [J]. *Journal of Chinese History*, 2020, 4(2): 528-536. [23] Sima Guang. *Zizhi Tongjian* [M]. Beijing: Zhonghua Book Company, 1956. [24] Shi Min, Li Bin, Chen Xiaohe. Integrated CRF-based pre-Qin Chinese word segmentation and tagging [J]. *Journal of Chinese Information Processing*, 2010, 24(2): 39-46. [25] Tan Qixiang. *Historical Atlas of China* [R]. Beijing: China Cartographic Publishing House, 1982. [26] Qian Mu. *Chronological Studies of Pre-Qin Masters* [M]. Beijing: Commercial Press, 2015. [27] Apache Software Foundation. ECharts [EB/OL]. [2021-05-21]. https://echarts.apache.org/zh/index.html. [28] Han Zhongmin. Calculating precise distance between two points using latitude and longitude [J]. *Public Communication of Science & Technology*, 2011(11): 196, 174. [29] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. *Information Processing & Management*, 1988, 24(5): 513-523.

Author Contributions

Chang Bolin: Software design and development, data statistics, initial draft; Wan Chen: Data annotation and proofreading, initial draft; Li Bin: Overall conceptual design, data verification, paper revision; Chen Xinyu: Data annotation and proofreading; Feng Minxuan: Data organization and proofreading, paper revision; Wang Dongbo: Theoretical discussion, paper structure adjustment and revision.

Note: Figure translations are in progress. See original paper for figures.