**PANG — System**

<1 table>

**A System for the Textual Inheritance Relation of Chinese Historical Classics ‘Comprehensive Mirror’, ‘Grand Tortoise’ and ‘Standard Histories’**

Pang, W.-H., Hsiang, J., Tu, H.-C. and Chen, L.

In Chinese historical writings, an event may be recorded in different classics and paraphrased in others. How an event was written also reflected the view of the historian (Zhou, 1986). In order to check different versions of the same event, a scholar needs to spend a great deal of time collecting all related writings from different books. This paper describes an automated method to establish the inheritance relationship of texts in different classics about the same event.

The historical writings studied in this paper are Zizhi Tongjian (ZT, 資治通鑑, *Comprehensive Mirror to Aid in Governance*), Cefu Yuangui (CY, 冊府元龜, *Grand Tortoise in the Imperial Treasury of Books*), Zhengshi (ZS, 正史, *Standard Histories*), and Tongjian Jishi Benmo (JB, 通鑑紀事本末, *Narratives from Beginning to End in the Comprehensive Mirror for Aid in Governance*). While these four epics cover about the same period in Chinese history, they represent four most important yet different styles of Chinese historical writings. ZT, written by Sima Guang (1019–1086) in 1084, considered among the greatest Chinese historiographical achievements, is in chronological style and covers the time from 403 BC to AD 959, the end of Wudai.1 CY, compiled by Wang Qinruo in 1013, covers the same period as ZT and was organized in the *leishu* style.2 Although CY is four times as large as ZT, it is held in much less esteem (Sung, 2010). ZS is the collection of Standard Histories, one from each dynasty. Standard Histories are written in the Biographies and Monographs style (紀傳體). ZT used 19 Standard Histories, from Shiji (史記) to Wudaishi (五代史). CY used 17. In 1173 Yuan shu (1131–1205) transformed ZT into an event-based book, JB, by compiling ZT into 239 events. JB opened the door of a third Chinese history-writing style, Narrative of Events (紀事本末體).

Our studies try to build a system to study the textual inheritance relation of these four tomes. After processing the texts (such as identifying entries and aligning dates), we then extract keywords from them. This is done by combining the *Pointwise Mutual Information* method (Pang et al., 2014) with a dictionary (Shi and Shen, 1994). This results in the following number of names extracted:

**Person names Location names Office titles**

ZT 18,893 10,085 13,888

CY 12,351 8,625 8,952

ZS 15,975 10,484 10,846

Estimating textual inheritance requires the comparison of texts, for which we used two different algorithms. Since CY followed the text closely when copying from ZS, we used the *longest common subsequence* method (Nakatsu et al., 1982) to find the textual inheritance from ZS to CY. ZT, on the other hand, *paraphrased* from ZS. In this case we employ the *inverse document frequency* (Salton et al., 1983) method, using common keywords to find similar texts. We discovered that of the 69,906 entries of CY (71,047 including the prefaces), approximately 13,662 (~20%) are taken from ZS. Of the 21,854 paragraphs in ZT, approximately 7,369 (~34%) have similar texts in ZS. Furthermore, about 2,477 paragraphs (~11%) of ZT have similar entries in CY. What might also be interesting is that although JB recompiled the original paragraphs of ZT into 239 events, it has actually only used 16,374 paragraphs, roughly 75% of the material of ZT.

We have built a digital archive system containing these four corpuses. In addition to full-text search and retrieval, we have provided functions such as post-query multifacet analysis, text inheritance analysis, term frequency and co-occurrence analysis, relevance factor of a term to a query, and prefix-suffix analysis. The system is flexible enough to host other similar corpuses easily.3

Chinese history classics often contain several million words. Digital archive systems provide a way to view and compare these classics that would not have been possible before. The method presented in this paper explores the important relation of textual inheritance in historical writings. It should provide a new way for scholars to study how Chinese history writings were constructed, how events were formulated and described by different historians, and the development of Chinese historiography.

**Notes**

1. It was based on 19 Standard Histories of previous dynasties and referenced 322 other historical writings (Du, 1993, 3:79).

2. An ancient Chinese form of reference books, with a classification of knowledge each of whose subjects is attached with related entries copied from older books.

3. We have indeed used the same shell to create context discovery retrieval systems for scores of corpuses.

**References**

**Du Weiyun.** (1993). *Chinese Historiography*.

**Nakatsu, N., Kambayashi, Y. and Yajima, S.** (1982). A Longest Common Subsequence Algorithm Suitable for Similar Text Strings. *Acta Informatica,* **18**(2): 171–79.

**Pang, W., Liu, S.-G., Tu, H. C., Weng, G. A. and Hsiang, J.** (2014). Automated Name-Extraction in Chinese Classics: Applying PMI Segmentation to Zizhi Tongjian. In Hsiang, J. (ed.), *Digital Humanities and Craft: Technological Change*, Taipei: NTU Press, pp. 139–66.

**Salton, G., Fox, E. and Wu, H.** (1983). Extended Boolean Information Retrieval. *CACM,* **26**(11): 1022–36.

**Shi, T. and Shen, Z. H.** (1994). *Grand Dictionary of Zizhi Tongjian*. Jilin People’s Press.

**Sung, C.-F.** (2010). Between Tortoise and Mirror: Historians and Historiography in Eleventh-Century China. Ph.D. thesis, Harvard University.

**Zhou Zhenfu.** (1986). *Writing Techniques of Ancient Writers*. People’s Press, Beijing.