Effective Identification of Citations in the Kanseki Repository

Christian Wittern
cwittern@gmail.com
Kyoto University, Japan

Introduction

The Kanseki Repository is a large repository of premodern Chinese texts. Currently it holds more than 9000 texts, covering all periods of Chinese history from early antiquity to the beginning of the 20th century. The repository is organized into 6 top-level categories and offers full-text search across all textual variants.

Since its opening to the public in March 2016, a frequent request from users was to be able to find texts related to a certain text, especially to investigate and evaluate textual dependencies. This poster reports on some experiments to find an effective way to respond to this requests.

What is needed to solve this problem is an efficient way to identify text passages that are derived from other, earlier texts (based on the assumption that the texts in question can in fact be reliably dated). Such passages will be called **citations** here, although the usage here is not limited to true citations, but also includes quotations from memory, paraphrases and allusions – cases where the reference does not follow the exact wording of the source. As an additional complication, we need to take into account the possibility that the received text differs from the text available to the author of the text that contains the reference.

Related work has examined plagiarism detection (Gipp and Meuschke, 2011 and Schultz, 1999), but the approach taken here makes direct use of some of the unique features of the repository and the index built for it and seems thus to be more efficient than general purpose algorithms. Admittedly, this has not been verified empirically and thus may be reasonably rejected as not relevant. However, the purpose of this presentation is not to compare algorithms and their efficiency when applied to the material here, but rather to collect some low-hanging fruit that became available due to the way the full-text index is constructed.

Identification of Citations

Index

Since a complete index has already been built for the full-text search, all experiments make use of this index (Mandoku 2016). The index is constructed by moving an n-gram window over the text and saving entries at appropriate locations. The resulting raw index is then read by a grep-like program to generate the display. The search display is designed to show the **keyword in context** (KWIC) so some characters are needed in front of the match character; these are appended after a comma character in the index entry. We built the index with a window between 10 and 25, since larger indexes considerable increase the required space and smaller builds will have too little information in the KWIC display. The index also contains information to identify the text, the location of the index excerpt and some information about the context of the match. Figure 1 shows a typical example of such an index for a 21-gram window.

```
索而得男故謂之長男巽一索而得,故稱乎母震        KR1a0002_009:7b:2:16:2
索而得女故謂之長女坎再索而得,謂之長男巽        KR1a0002_009:7b:3:6:2
周易雜卦傳第十一乾剛坤柔比樂,夏易傳卷十        KR1a0002_011:1a:2:7:0
乾剛坤柔比樂師憂臨觀之義或與,雜卦傳第十        KR1a0002_011:1a:4:6:2
夏易傳卷十        KR1a0002_011:5b:8:7:2
卦是謂一卦含四卦繫辭謂之中爻,體交互各成        KR1a0003_000:1a:3:2:0
卦含四卦繫辭謂之中爻所謂八卦,成一卦是謂        KR1a0003_000:1a:3:6:0
握鳥笑之握讀爲屋其說近乎鑿學,之皆讀爲斛        KR1a0003_000:2a:4:12:0
握爲&KR0015;之握讀爲屋其&KR0003;近乎鑿學,之皆讀爲斛   KR1a0003_000:2a:4:12:0  SBCK
周易鄭康成註易類提要等謹案周,庫全書經部        KR1a0003_000:1a:1:9:0   WYG
卷宋王應麟編應麟字伯厚慶元人,易鄭康成註        KR1a0003_000:1a:1:30:2  WYG
卷所存者僅文言序卦說卦雜卦四,文總目惟載        KR1a0003_000:1a:1:161:10        WYG
绂可謂篤志遺經研心古義者矣近,以存漢易之        KR1a0003_000:1b:4:40:24 WYG
作乾儀禮注蛇龍君子之頪)    ,      (        KR1a0003_001:1b:2:4:3   n
成其定稅三百家故三百戶也不易,大夫采地方        KR1a0003_001:3a:5:16:8
易之田休一歲乃種再易之田休二,之田歲種之        KR1a0003_001:3a:6:16:8
歲乃種再易之田休二歲乃種言至,一易之田休        KR1a0003_001:3a:7:1:8
陽小人極盛君子不可有所之故不,之剝也五陰        KR1a0003_001:8b:9:20:23
歲曰菑二歲曰新田三歲曰畬大畜,六二不菑畬        KR1a0003_001:9b:1:18:25
作民)    ,(寘讀爲鳴       KR1a0003_001:6a:8:3:17  n
年而赦不得者不自思以得正道終,年而赦下罪        KR1a0003_001:11a:8:4:29
U:%%- *te00 KR1a.idx    1% (123,0)    (Fundamental +1) Mon Oct 31 5:13PM 0.26
```

Figure 1. Effective Identification of Citations in the Kanseki Repository

Algorithm

To find citations in the indexed files, a window of the same size as the index window is moved over the text passage under investigation. A search is initiated for a string of **n** characters, starting at the first index position. In the example in Figure 1 this would be starting at position 6, since there are 5 characters after the ","; these characters are preceding the indexed characters in the text and have therefore been moved to the back. A query expansion is used for this search, in order to catch character variants in this initial selection of index lines. A large value of **n** will increase the probability that citations are not found due to

slight positional variations in the text, while a small value of **n**, such as 1, will select many lines that are not relevant and will thus increase the processing times. Experiments have shown that a value of 2 or 3 for **n** gives a good optimum for precision and recall. Positional values are also registered to better demarcate the citation boundaries.

The selected lines have to be post-processed to restore the original sequence as found in the text. The line will then be compared to the window of the text passage, with scores given for each match; high scores are taken to be a citation and are marked for further processing. The best results so far have been achieved for a cumulated score of n-gram matches for values of **n** from 1 to 3, but additional experiments are planned. Conclusive results will be reported in the poster presentation.

Additional expected results

With the method introduced here, it becomes feasible to investigate potential citations for whole texts. We plan to build a heat-map of a text with passages that have been cited coloured according to their frequency. This will enable new ways of exploring the intertextuality of texts and will provide new evidence for the history of ideas and flow of intellectual debt in the history of Chinese thought. For the presentation of the poster, we show a preliminary investigation of some key texts of Chinese philosophy as a proof of concept.

We also hope to identify a set of key phrases and look at their usages over time, and in different schools of thought, to see what kind of trends can be seen there.

Bibliography

Gipp, B. and Meuschke, M. (2011). "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence". In: *Proceedings of the 11th ACM symposium on Document engineering* (DocEng '11), Mountain, View, CA, USA, 2011. doi: 10.1145/2034691.2034741.

Mandoku (2016). Mandoku project (Source code) https://github.com/mandoku/mandoku (accessed 2016-10-31).

Schultz, R.L. (1999) *The search for quotation : verbal parallels in the prophets,* Sheffield, Sheffield Academic Press, 1999.

Wittern, C. (2014). "Kanripo and Mandoku: Tools for git-based distributed repositories for premodern Chinese texts", in *Digital Humanities 2014 Book of Abstracts*, 2014, p. 408-409.

Wittern, C. (2016). Special issue Kanseki Repository, *CIEAS Research Report* 2015, Kyoto 2016.