A RegEx Machine

Elif Yamangil (Harvard University, Computer Science)

Shih-pei Chen (Harvard University, China Biographical Database Project)

Peter Bol (Harvard University, EALC)

Introduction

Valuable pieces of information – dates, offices, person names, place names – may begin as anecdote, but if we can extract them from a text corpus we can create data for further analysis. Regular expressions [1][2] are an effective way of extracting such information from texts, provided humans can define the pattern to which a set of strings will conform. RegEx have the advantage of allowing for a great deal of variability and making use of the language that is being targeted. If humans can write the patterns, machines can match them against the text and do the extraction.

In the China Biographical Database project (CBDB) [3], we have been using regular expressions to extract information from thousands of Chinese biographies in order to add data to our database. However, the historians and literary scholars who observe and describe the patterns in which the information they seek is embedded rarely possess the programming skills required to apply this technique. In this paper, we demonstrate the CBDB Regex Machine, a graphical user interface (GUI) that enables people who do not have programming skills to design patterns and extract matches from texts.

The Information Extraction Task in CBDB

Biography has been a major component of Chinese historiography since the first century BCE and takes up over half the contents in the twenty-five dynastic histories. Figure 1 shows the rich information contained in the biography for Lv Zuqian (呂祖謙, 1137-1181, a famous scholar and politician in the Song China). Lv's biography is like most Chinese biographies in that it contains various kinds of information about a historical figure which we would like to collect and add into our database — his/her alternative names, addresses (where his/her family was from and where he/she lived), his/her kinship relations and social relations, how he entered the government (for males), official titles received (for males and females) and offices held.

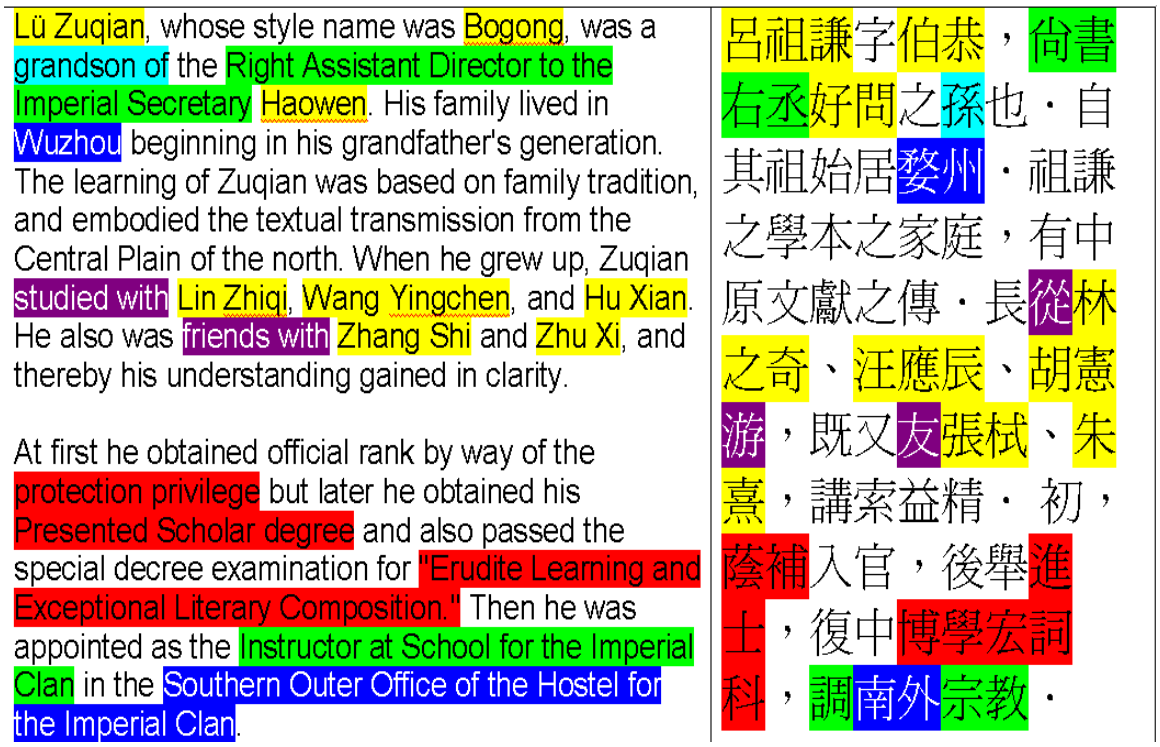

Figure 1 The biography for Lv Zhuqian demonstrates the rich information contained in Chinese biographies.

Once it became clear that most of the "factoids" in biographies are expressed in regular patterns, making it possible to extract them through regular expressions, we began to populate the database largely through writing regular expressions and matching them against thousands of biographical texts via procedures written in Python. For example, we found text strings in biographies that matched reign periods titles, numbers, years, months, and days to capture dates with 99% recall and precision.

However, the CBDB editors who observed and described the patterns with information to be extracted, like most of humanist scholar, did not possess the programming skills required to apply this technique. Our editors were thus always dependent on IT technicians to write the programs to apply their regular expressions. Moreover, the CBDB information extraction task had a workflow that involved editors at different institutions taking responsibility for finding patterns and checking results, making the process inefficient. When the programmer did not know Chinese and the editors did not know programming, it took time for the editors to see the result of applying the regular expressions to the texts, and by the time corrections reached the programmer time had to be spent getting reacquainted with the original problem.

CBDB RegEx Machine

In order to enable people without programming skills to write patterns and match them against texts, we developed the RegEx Machine. The RegEx Machine, which is built

within the Java Swing library, allows a user to graphically design patterns for a corpus of texts, match them against the texts, and see results immediately via a user-friendly color-coding scheme. It consists of a workspace of three main components: (1) a view that displays the textual data currently used, (2) a list of active regular expressions that are matched against the text via color-coding, and (3) a list of shortcut regular expression parts that can be used as building-blocks for composing active regular expressions in (2).

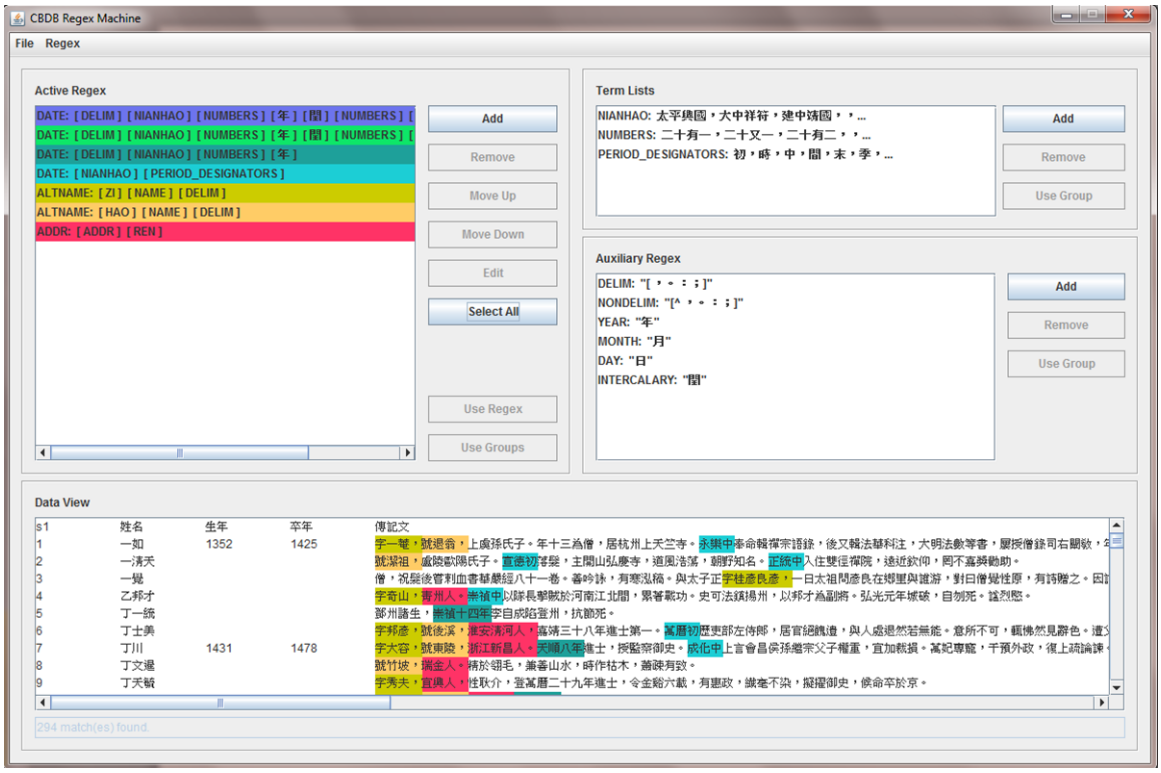

Figure 2 The graphical user interface of the CBDB Regex Machine

Figure 2 is a screenshot of the RegEx Machine in action. We can see from the upper left zone that each of the active regex is color-coded and that the matching strings in the text are shown in the same color. With this color coding scheme a user can see how the text is matched against a pattern immediately and with this instantaneous feedback can adjust the pattern to get better matches.

In order to make the writing of regular expressions as easy and intuitive as possible, we also brought in the concept of modularity for composing regular expressions. While a real working regular expression can be very long and complicated, they often can be broken into smaller building blocks like delimiters, Chinese characters for digits, or a list of reign period titles. Using the building blocks to compose a complicated regular expressions, rather than writing a regular expression from scratch, makes writing regular expressions more intuitive. We designed two facilities in the RegEx Machine to allow users store two types of commonly used components when building regular expressions: Term Lists and

Auxiliary RegEx (see the upper right zone of the interface). Term Lists allows users to import a list of terms and to use it for composing active regular expressions. For example, the list of reign period titles is a major component for composing various patterns of Chinese dates. Auxiliary RegEx, on the other hand, allows users to store commonly used pieces of regex, like delimiters and Chinese digits. When composing an active regular expression in the RegEx Machine, a user only needs to find suitable components from the predefined term lists and/or auxiliary regular expressions, supply the missing parts by typing in free style regular expressions, arrange all the parts in a correct order, and then he/she can apply the regular expression to the text and see the result immediately. Figure 3 shows the composition of an active regular expression: three of its components are from the preloaded term lists and four from the predefined auxiliary regular expressions. Yet the active regex remains readable and intuitive, while one of the term list component contains 170 reign period titles.

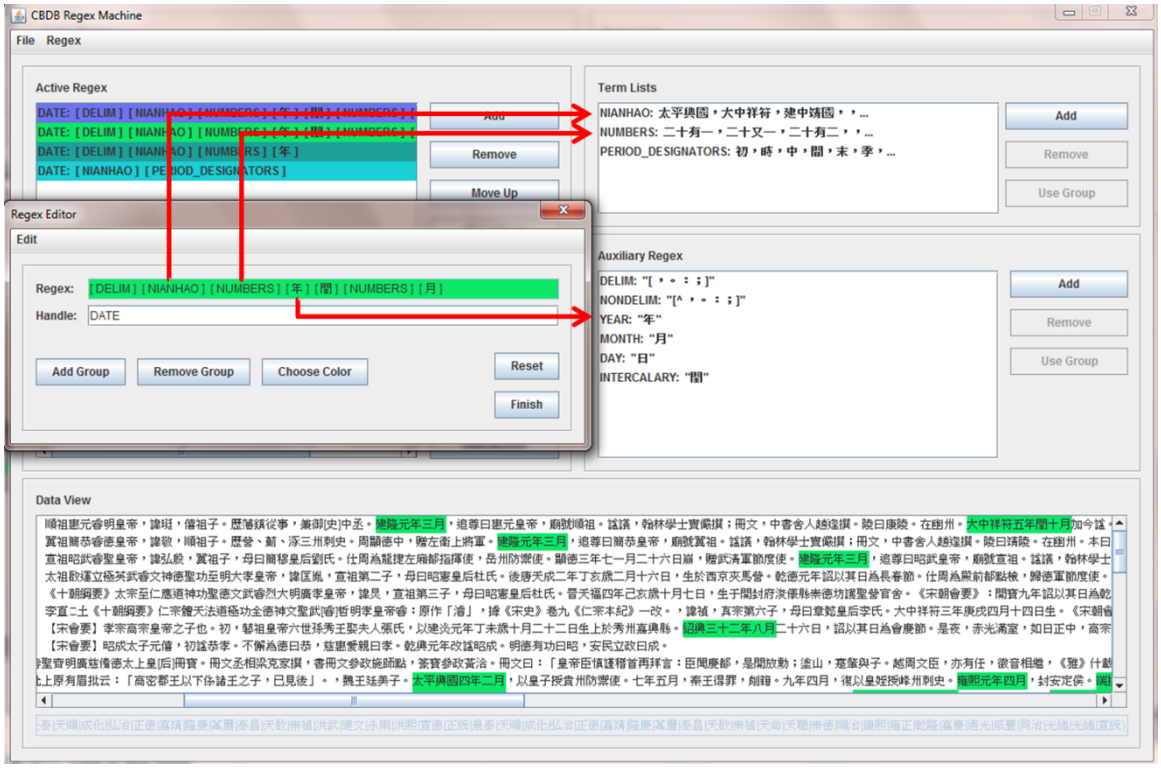

Figure 3 The composing of active regular expressions remains intuitive and readable via our design

Three additional facilities have been built into our product. (1) An XML export ability: users can create XML files that flatten the current workspace (data and regular expression matched against) at the click of a button. This facilitates interfacing to other programs such as Microsoft Excel and Access for database input. (2) A save/load ability: users can easily save/load the workspace state which includes the list of regular expressions and shortcuts and their color settings. (3) A handy list of pre-made regular expression examples: numerous date patterns that can be added instantly to any regular expression

using the GUI menus.

Concluding Remarks

The point of building this application is to allow users with no prior experience with programming or computer science concepts such as regular expression scripting to experiment with data mining Chinese biographical texts at an intuitive template-matching understanding level only, yet still effectively. We hope that this RegEx Machine will not only be helpful to the CBDB editors, but also to historians who wants to take advantage of the computing power of machines and the large quantity of texts that only exhibit in this current age.

References

1. Hopcroft, John E.; Motwani, Rajeev; Ullman, Jeffrey D. (2000). Introduction to Automata Theory, Languages, and Computation (2nd ed.). Addison-Wesley.
2. Sipser, Michael (1998). "Chapter 1: Regular Languages". Introduction to the Theory of Computation. PWS Publishing. pp. 31–90.
3. The China Biographical Database, http://isites.harvard.edu/icb/icb.do?keyword=k16229 .