Extracting Characteristics of Classic Literature Using a Data Mining Method

Mayumi Higa

Kwansei Gakuin University

Abstract

This study analyzed the writing habits of three famous Japanese novelists by focusing on the frequencies with which specific characters occur before commas. Clear differences in their writing styles were seen, and specific rules were discovered. The analysis used a data mining method that constructs rules from databases and a visual mining method that the author developed.

Introduction

Researchers have conducted quantitative analyses of Japanese text data using sentence length, the frequency with which specific words appear, the number of words and periods; and the relationships among punctuation and other parts of the text. Conventional statistical methods have been used for these quantitative analyses.

Studies that have attempted to identify the authors of different works include: (1) an attempt to determine the authors of the Uji Jujou (Uji Appendix) in the *Tale of Genji* (Yasumoto, 1958); (2) identifying authors in "Yura Monogatari" (Nirasawa, 1965); and (3) determining the authenticity of writing ascribed to Nichiren (Murakami, 1994).

Studies hoping to determine authorship from characteristic writing styles (Jin, 1994; Jin *et al*., 1993) have focused on the use of commas. To generate data, the studies used main component analysis, cluster analysis, dinstinction analysis, and quantitative theory Class 3, methods that require a significant amount of time to execute. These studies have been able to distinguish authors and detect

characteristic sentence structures. Although those studies have recognized different writing styles, they have not determined what the exact differences are.

This study, however, identified the specific styles that three famous Japanese novelists use in creating sentences. The analysis used a data mining method, and a visual mining method developed by the author.

Data

This research converted the data used in "Position of commas in sentences and the classification of text" (Jin, 1994) for analysis. That study drew data from short novels written by three famous Japanese novelists: Yasushi Inoue, Yukio Mishima, and Atsushi Nakajima. To avoid biased sampling, the dataset was carefully composed by dividing the novels into several parts,. For example, Inoue's ***Koi To Shi To Nami To*** (Love, Death and Waves) was divided into two sections, Nakajima's ***Deshi*** (The Disciple) into three, and ***Riryo*** (Li Ling) into four. In all, 21 sections were analyzed: 8 for Inoue, 4 for Mishima, and 9 for Nakajima.

In the 21 sections of the novels, 26 types of characters appeared before commas. These characters included [*to*], [*te*], [*ha*], [*ga*], [*de*], [*ni*], [*ra*], [*mo*], [*shi*], [*wo*], [*ri*], [*no*], [*ku*], [*toki*], [*ka*], [*ba*], [*ta*], [*i*], [*nochi*], [*zu*], [*re*], [*ki*], [*ru*], [*e*], [*u*], and others—all the sounds of single Japanese characters.

The following equation was used to determine the relative frequency with which a character appeared before a comma in each novel section:

$$C_{ij} = \frac{X_{ij} \cdot 100}{S_i} \qquad \begin{matrix} (i = 1,2,3,\ldots,21), \\ (j = 1,2,3,\ldots,26) \end{matrix} \qquad (1)$$

Where X_{ij} is the number of times a character occurs before a comma, $S_i = \Sigma X_{ij}$ is the total for each novel section, and C_{ij} is the percentage of characters in a paragraph that occur before a comma (%).

According to a Japanese dictionary (*Nihon Kokugo Daijiten*: Great Dictionary of the Japanese Language), a comma is "a mark that shows breaks in the meaning of a sentence in order to clarify breaks/continuation in writing". Commas are commonly used in two ways: to clearly state parallel elements and to break sentences into two or more sections that are somewhat related in meaning. Writers do not differ in the former usage, while they do in the latter, since there are no specific rules on where to put breaks in meaning (Jin, 1994).

Extracting characteristics of novelists from characters before commas
Data mining method: Part 1 – Generalized rule induction

Generalized rule induction (GRI) discovers association rules for data. The association rules are given in the form of an "IF (proceeding condition) THEN (conclusion)" statement. For example, the first rule "IF '[*to*]' > 7.78 THEN

novelist=Yasushi Inoue" was 100% validated in Inoue's eight sections (representing 38% of the data).

GRI extracts a set of rules from the data and deduces rules that give the best information. The following list contains 11 rules obtained by GRI:

1. Novelist = Yasushi Inoue ← [*to*] > 7.78 (8: 38%, 1.0)
2. Novelist = Atsushi Nakajima ← [*nochi*] > 0.415 (9: 43%, 1.0)
3. Novelist = Atsushi Nakajima ← [*to*] < 7.78 & [*toki*] > 1.895 (9: 43%, 1.0)
4. Novelist = Atsushi Nakajima ← [*no*] < 0.7 & [*nochi*] > 0.415 (8: 38%, 1.0)
5. Novelist = Atsushi Nakajima ← [*no*] < 0.7 & [*ri*] < 6.055 & [*te*] < 12.515 (8: 38%, 1.0)
6. Novelist = Atsushi Nakajima ← [*no*] < 0.7 & [*wo*] < 2.12 & [*to*] < 6.605 (8: 38%, 1.0)
7. Novelist = Atsushi Nakajima ← [*no*] < 0.7 & [*te*] < 12.515 & [*to*] < 6.605 (8: 38%, 1.0)
8. Novelist = Atsushi Nakajima ← [*no*] < 0.7 & [*to*] <6.605 & [*te*] <12.515 (8: 38%, 1.0)
9. Novelist = Atsushi Nakajima ← [*de*] < 4.35 & [*ha*]< 16.02 (8: 38%, 1.0)
10. Novelist = Atsushi Nakajima ← [*to*] < 7.78 & [*ha*] < 12.805 (8: 38%, 1.0)
11. Novelist = Yukio Mishima ← [*toki*] < 0.53 (4: 19%, 1.0)

Calculation Conditions:

The maximum value of a rule: 11 Preconditions of the maximum rule: 3
Precision of minimum rules: 50%

Rule 1 characterized Yasushi Inoue and rule 11 characterized Yukio Mishima. Yasushi Inoue tends to put commas after [*to*] 7.78% of the time, while Yukio Mishima rarely uses a comma after [*toki*] (0.54% or less). By contrast, there were many rules for Atsushi Nakajima's writing style. One example is rule 9; in 8 of his 9 sections, no more than 4.35% and 16.02% of the instances of [*de*] and [*ha*], respectively, were followed by commas. This rule clearly identified 38% of the total data as Nakajima's writing. Rules 1 to 11 all have 100% certainty. The advantages of GRI are that (1) it is easy to interpret rule sets; (2) compared to a decision tree, more general rules can be found, since several records can overlap leading to more than one rule; and (3) more than one output field can be processed (novelists are "output fields" in this study).

Data mining method: Part 2—C5.0

This method uses the C5.0 algorithm to generate decision trees or rule sets. It divides the sample according to the field that offers the maximum amount of

information. After the first division, each sub-sample is further divided into smaller sub-samples. The dividing process continues until sub-samples can no longer be divided. Finally, the divided samples at the lowest level are re-evaluated. Levels that do not contribute to the model value are deleted (trimmed).

The C5.0 decision tree is as follows:

[*to*] ≤ 6.67 [mode: Atsushi Nakajima] (13)

[*toki*] ≤ 1.02 [mode: Yukio Mishima] (4, 1.0) → Yukio Mishima

[*toki*] > 1.02 [mode: Atsushi Nakajima] (9, 1.0) → Atsushi Nakajima

[*to*] > 6.67 [mode: Yasushi Inoue] (8, 1.0) → Yasushi Inoue

This decision tree indicates that in 13 sections of data [*to*] is followed by a comma 6.67% of the time or less. Many of these sections are by Atsushi Nakajima. Of these 13 sections, if [*toki*] is followed by a comma less than 1.02% of the time, Yukio Mishima wrote the section with 100% certainty. Four cases fit this situation. Otherwise, the section was written by Atsushi Nakajima with 100% certainty. There are nine such cases. Finally, in eight cases, [*to*] is followed by a comma at least 6.67% of the time; all these sections belong to Yasushi Inoue with 100% certainty. The decision tree thus classifies the works by these three novelists completely using [*to*] and [*toki*].

The advantages of C5.0 are that (1) it can handle problems such as small datasets or large fields, (2) it does not require a long learning period to make conjectures, and (3) the rules are readily interpreted and easier to understand than those in other models.

Visual Data Mining Method

While analyzing this research, a constellation graph (Oyama, 2001) was developed as a method of visual data mining. A constellation graph is an effective way of mapping multi-variable data systems on a 2-dimensional plane to make conjectures. The graph distinguished the three novelists.

Table 1 shows the parameters used to generate the chart for each novelist. These parameters were calculated by processing data for the constellation graphs.

Table 1

	[*to*]	[*ha*]	[*de*]	[*toki*]	[*nochi*]	[*e*]	J_{min}
Yukio Mishima	7	0	0	0	23	0	0.04
Yasushi Inoue	2	13	4	2	9	0	0.01
Atsushi Nakajima	10	10	0	7	1	1	0.13
Total	1	2	5	14	7	1	0.85

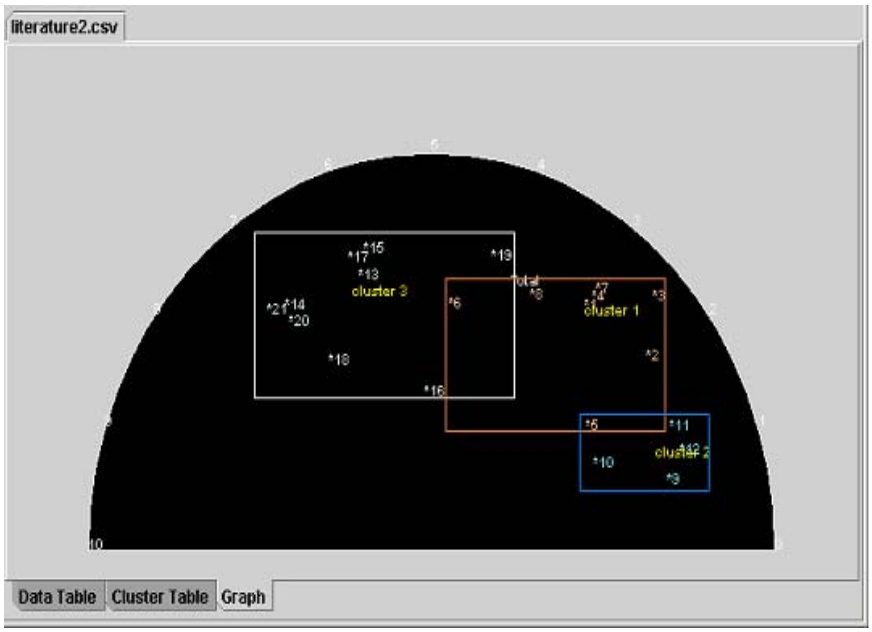

Figure 1: The Constellation graph after incorporating the optimum weighing factors

Cluster 3 (left): Novelist: Atsushi Nakajima
Cluster 1 (center): Novelist: Yasushi Inoue
Cluster 2 (right): Novelist: Yukio Mishima

Constellation graphs feature (1) color-coding to distinguish groups with the same values for dependent variables, (2) a dynamic search involving interactive changes in the weight of each variable, (3) weight calculation of independent variables to better distinguish the values of dependent variables, (4) extraction of parts of the graph for further mining when identification is difficult due to overlapping data or similar values of the dependent data, and (5) possible identification of data that is difficult to distinguish from the values of dependent variables or singular data.

Figure 1 shows the results. After calculating the weight of each variable, a variable that is not necessary for classifying the data can have its significance set

interactively on the screen. The constellation graph shows the characteristics of the novelists.

Conclusion

This study characterized the writing styles of three novelists according to their use of commas. Determining how frequently [*to*] and [*toki*] precede commas enables us to conjecture on the author. The use of [*ha*], [*de*], [*nochi*], and [*e*] also characterized the authors. An unexpected outcome of this research was the suggestion that rules can be extracted by turning all of the attributes of writing into electronic data. Visual mining further demonstrated its utility for visual verification. Future studies will include applying the data mining technique to data in various disciplines in the humanities. Such work will verify rules discovered in the past through human experience, and should discover new knowledge.

References

Jin, Ming-Zhe
1994 Use of commas and sentence classification, *Mathematical Linguistics Society of Japan, 19*, No.7.

Jin, Ming-Zhe, Tadao Kabashima and Masakatsu Murakami
1993 Use of commas and characteristics of writers, *Mathematical Linguistics Society of Japan, 18.*

Murakami, Masakatsu
1994 Influence and outcome of mathematical writing style research, *Linguistics, 23*, No.2.

Nirasawa, Tadashi
1965 Inference in the authorship of "Yura Monogatari", *Mathematical Linguistics Society of Japan,* No.33.

Niwa, T., K. Fujikawa, Y. Tanaka, M. Oyama (Higa)
2001 Visual data mining using a constellation graph, ECML/PKDD-2001, Freiburg.

Oyama (Higa), Mayumi
1999 Data mining with structural analysis tree, *Cultural Science and Information Process, 20.*

Oyama (Higa), Mayumi, Takashi Okada, and Eijun Li
1995 Knowledge discovery method applied to Iris data analysis, Kwansei Gakuin University, *Information Science Study*, Vol.10.

Yasumoto, Yoshinori
1958 Assumption of author by statistics of writing style: Authorship of Uji Appendix in the *Tale of Genji*, Shinrigaku Hyoron. *Psychology Review 2*, No.1.