Visualization of relationships among historical persons from Japanese historical documents

Fuminori Kimura
College of Information Science and Engineering, Ritsumeikan University, Japan

Takahiko Osaki
Independent Researcher

Taro Tezuka
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan

Akira Maeda
College of Information Science and Engineering, Ritsumeikan University, Japan

Abstract

Digitization of Japanese historical documents has gained much attraction in the field of humanities in Japan recently, and numbers of documents are already available in digitized text format. However, text analysis of these documents has rarely been done mainly due to the lack of natural language processing tools that can handle pre-modern Japanese. In this article, we propose a method to extract and visualize the relationships among persons from Japanese historical documents with an aid of supplementary information such as personal name and place name indices. The goal of the method is to extract dynamics of relationships among historical persons. The method utilizes locational information to obtain latent relationships among persons based on their spatial activities. The proposed method is applied to a Japanese historical chronicle written in the 12th century. Experimental results showed a strong correspondence to the known historical facts, and the results of a user survey completed by researchers of Japanese history demonstrated some potential for the method to serve as a new approach in the fields of humanities.

Correspondence:
Fuminori Kimura, College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan.
Email:
fkimura@is.ritsumei.ac.jp

1 Introduction

In recent years, there has been an increasing use of digital technology in the study of humanities. Many historical documents are now digitally archived, enabling further analysis using computers. There are archives that are accessible on the World Wide Web, including Database of National Institute of Japanese Literature (National Institute of Japanese Literature, 2011) and Perseus Digital Library (Crane, 2011).

Literary and Linguistic Computing, Vol. 28, No. 2, 2013.

doi:10.1093/llc/fqs045 Advance Access published on 20 February 2013

Until recently, the storage of historical documents and data has been the main target of digital archive research. There are, however, many works that go on to analyze the content of the historical documents using text mining techniques. In this article, we propose a method to extract and visualize the relationships among persons from historical documents. The goal of the method is to extract dynamics of relationships among historical persons. The method utilizes locational information to obtain latent relationships among persons based on their spatial activities.

The method has a potential to serve as a new approach in the fields of humanities since it makes it possible to analyze and visualize the contents written in an ancient language that is difficult to analyze using current natural language processing tools, by utilizing partial information such as personal names and place names (Itsubo *et al.*, 2011; Osaki *et al.*, 2011). In order to assess the usefulness of the proposed method for the humanities researchers, we conducted a user survey of the method.

2 Japanese Historical Documents Used in This Study

We use a Japanese historical document, namely 'Hyohanki'. 'Hyohanki' is a personal diary written by an aristocrat Nobunori Taira in the late Heian period between 1112 and 1187. 'Hyohanki' is written in the form of a diary, thus it is easy to obtain the date of each entry, which is important information for temporal analysis.

For this document, there is a personal name index, which is compiled manually by humanities scholars (Hyōhanki Rindokukai, 1999). This index contains all the personal names appearing in the documents, including the names described only by their titles or aliases but not by the real names, along with the positions in the document where they appeared. We also used a place name index of Kyoto in this period, which is 'Index of Kyoto's Place Names' created by Noboru Tani based on 'Place Names of Kyoto' (Hayashiya *et al.*, 1979) as data sources for place names.

In the viewpoint of grammar, pre-modern Japanese used in historical documents, including the one mentioned above, is different from modern Japanese. Natural language processing techniques such as morphological analysis are not applicable for pre-modern Japanese. Since Japanese language does not have explicit word boundary, it is not easy to extract words from a sentence. For that reason, we use personal and place name indices mentioned above in order to extract personal and place names from the text.

3 The Strength of Relationships

Our proposed method uses co-occurrence frequency as an indicator of the strength of the relationship between persons. We define co-occurrence as the case when two words appear in the specific text block. We use a paragraph as the unit of a text block.

We use a co-occurrence as the indicator of a relationship between a person and a location. If a personal name and a place name appear in the same paragraph, we consider it as a co-occurrence between the two. In many cases, a co-occurrence between a person *A* and a location *B* indicates that a person *A* was at a location *B*. There are, of course, cases where *A* and *B* co-occur in an expression such as '*A* has never visited *B*.' It is difficult to exclude such cases, unless applying extensive natural language processing. If we could collect many co-occurrences, however, the result is expected to represent a positive relationship between a person and a place name, due to the law of large numbers.

4 Visualizing Inter-personal Relationships Using the Co-occurrence Information between Personal and Locational Information

4.1 Overview of the method

In this method, we extract personal features using information of their visiting place and create feature vectors of them. Personal relationships are calculated by cosine similarity of their features. Cosine similarity is a commonly used measure to calculate

the similarity between two vectors based on the cosine of the angle between them. This method can estimate personal relationships between people without necessarily having direct co-occurrences. Therefore, this method can also estimate personal relationships between people living in different times. The algorithm of the method is as follows:

(1) Obtain personal and place names from texts.
(2) Count co-occurrence frequencies between each personal name and place names.
(3) For each person, create a vector having the number of co-occurrences with a place name as a component.
(4) Calculate the similarities among persons based on these vectors.
(5) Cluster the vectors using *K*-means method.
(6) Visualize the relationships using the similarities and the clustering results.

4.2 Information extraction

In the first step of our proposed method, we obtain frequencies of co-occurrences between each personal name and place names. We used pattern matching to find place names that were included in 'Index of Kyoto's Place Names', which is explained in Section 2.

We use co-occurrence as the indicator of the relationship between a person and a location. If a personal name and a place name appear in the same paragraph, we consider it as a co-occurrence between the two. Since each paragraph often covers a specific situation or a topic, we considered it to be a better unit than dates.

4.3 Clustering

Each location name is considered as a dimension of a vector space. For each person, we create a vector having the number of co-occurrences with a place name as a component.

K-means is a commonly used method of clustering, which aims to partition objects into a fixed number of groups or clusters. *K*-means method assigns each node to the nearest given centroid. After assigning, the method calculates the new centroid of each cluster. Then, the method re-assigns each node to the nearest new centroids. The method repeats this process until all centroids do not move. In *K*-means clustering, there is a proposal of improving convergence by setting initial values in a probabilistic manner (Arthur and Vassilvitskii, 2007). We used this modified *K*-means method for clustering personal names. The parameters in our method are K and L. K is the number of clusters used in *K*-means clustering. L is the number of repetitions for finding the optimal initial centroids.

4.4 Visualization

We use the similarity measure and the result of clustering for visualization. We used JUNG, a Java open-source library for drawing graph structure. It has a capability of drawing graphs when the relevance measures between each pair of nodes are given. We use 'FR layout' based on Fruchterman–Reingold algorithm, which is a force-directed layout algorithm for drawing a graph. In this algorithm, the position of a node is influenced by forces around it, which are calculated based on the number of edges connected to a node.

5 Visualization Results and Discussions

Using our proposed method, we created graphs that visualize relationships between historical persons. We focused on the time range of the Hōgen Rebellion, which started in early July 1156 and ended in late July of the same year. The Hōgen Rebellion was a short civil war caused by a power struggle between Emperor Goshirakawa and former Emperor Sutoku, who was the elder brother of Goshirakawa.

We chose 78 persons belonging to either the faction following former Emperor Sutoku or the faction following Emperor Goshirakawa (Hyōhanki Rindokukai, 1999). Most of them are aristocrats and samurai warriors. It is distinguishable from historical records to which faction each person belonged to. In 'Hyohanki', 31 of these persons had co-occurrence with place names. We used $K = 3$ for *K*-means clustering and $L = 20$ for initialization.

Figure 1 shows the result of visualization using the similarity of co-occurring place names. The number at the end of the node label and the node shape

indicate to which faction each person belonged to. A node labeled '1' (square node) at the end indicates that he followed Sutoku. On the other hand, a node labeled '2' (circular node) at the end indicates that he followed Goshirakawa. Lines are drawn when similarity is over 0.4. Dotted lines indicate similarity between 0.4 and 0.7. Solid lines indicate similarity over 0.7. The number within each node indicates each historical person, which corresponds to the number attached to a person listed in Table 1.

Figure 2 shows the result of clustering. The number at the end of the node label and the node shape indicate to which cluster the person was allocated to. The members of the Cluster 1 are represented by triangular nodes, the members of the Cluster 2 are represented by circular nodes, and the members of the Cluster 3 are represented by square nodes. Table 1 shows to which cluster and to which faction each person belonged to.

Figure 3 closes up a part of the result of clustering persons using locational information (Fig. 2). The triangular nodes (Nos 2 and 3) belonged to the faction of former Emperor Sutoku. However, they are distinguished from people plotted by circular nodes, because they battled in the front line. This result indicates that this method can visualize the difference of each person's position, which might not necessarily be mentioned clearly in existing literature.

In the experiments, we used persons that we know to which faction he belonged to during the Hōgen Rebellion. The result shows that the obtained clusters significantly correspond to the historically known factions. The Cluster 2 (circle) corresponds

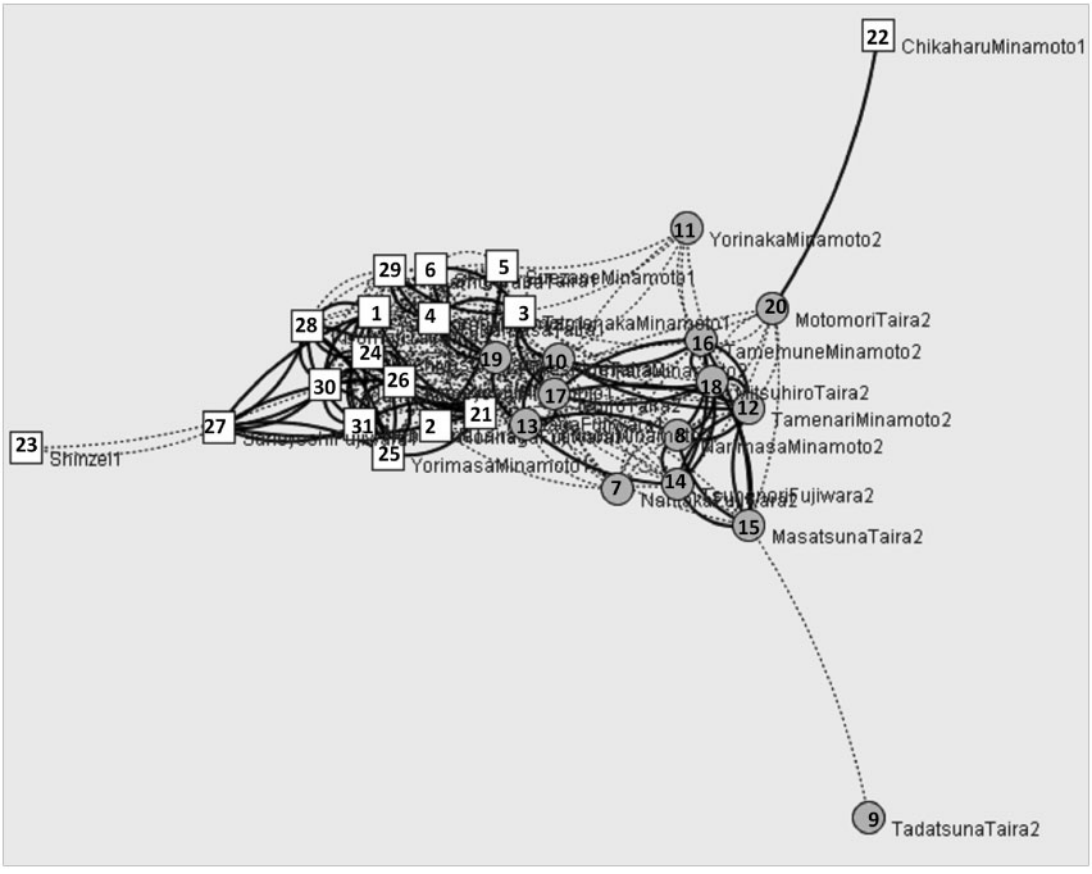

Fig. 1 Relationships among persons and historical factions

Table 1 Comparisons of factions and clusters

	Cluster 1	Cluster 2	Cluster 3
Faction of former Emperor Sutoku	1. Nagamori Taira 2. Norinaga Fujiwara 3. Tamenaka Minamoto 4. Tadamasa Taira	7. Naritaka Fujiwara 8. Narimasa Minamoto 9. Tadatsuna Taira 10. Yorikata Minamoto 11. Yorinaka Minamoto 12. Tamenari Minamoto 13. Yorinori Minamoto 14. Tsunenori Fujiwara 15. Tadatsuna Taira 16. Tamemune Minamoto 17. Iehiro Taira 18. Mitsuhiro Taira	21. Yorinaga Fujiwara 22. Chikaharu Minamoto
Faction of Emperor Goshirakawa	5. Suezane Minamoto 6. Shigemori Taira	19. Koreshige Taira 20. Motomori Taira	23. Shinzei 24. Yoshitomo Minamoto 25. Yorimasa Minamoto 26. Tameyoshi Minamoto 27. Saneyoshi Fujiwara 28. Kiyomori Taira 29. Yoshiyasu Minamoto 30. Tadamichi Fujiwara 31. Nobukane Taira

to the former Emperor Sutoku's faction and the Cluster 3 (square) corresponds to the Emperor Goshirakawa's faction. The Cluster 1 (triangle) is the intermediate group. This is a satisfactory result, considering the fact that we used only co-occurrences with place names and no other external information.

There were some exceptions, though. For example, Yorinaga Fujiwara, one of the main figures in the former Emperor Sutoku's faction, was allocated to the cluster consisting mostly of the members of the Emperor Goshirakawa's faction. To clarify the reason for such strange allocation, a further exploration of the raw data is required. A close investigation on frequently occurring pairs among personal names and place names may reveal the reasons for such allocation.

6 User Survey

We conducted a user survey of the proposed method in order to assess the potential of it in humanities research. Four users participated in the survey, and all of them are researchers of Japanese history. The users are asked to use the web interface of the proposed method, in which the user can select any combination of people in the index and time period within that of 'Hyohanki'. Then, the users are asked to evaluate the method by three evaluation criteria; 'historical credibility', 'novelty for historical research', and 'availability for historical research'. For each evaluation criterion, the users gave a score from 1 to 7, in which 1 is the lowest (strongly disagree), 4 is neutral (neither agree or disagree), and 7 is the highest (strongly agree). Table 2 shows the result of the user survey.

The user survey was conducted in Japanese, and the English translations of the questions are as follows: for 'historical credibility', the question is 'Does the visualization result correspond to the historical facts, and is appropriate and credible?', for 'novelty for historical research', the question is 'Does the proposed method have novelty compared to the existing research in Japanese history and humanities in general?', and for 'availability for historical research', the question is 'Is the proposed method useful for the field of Japanese history and humanities in general?'.

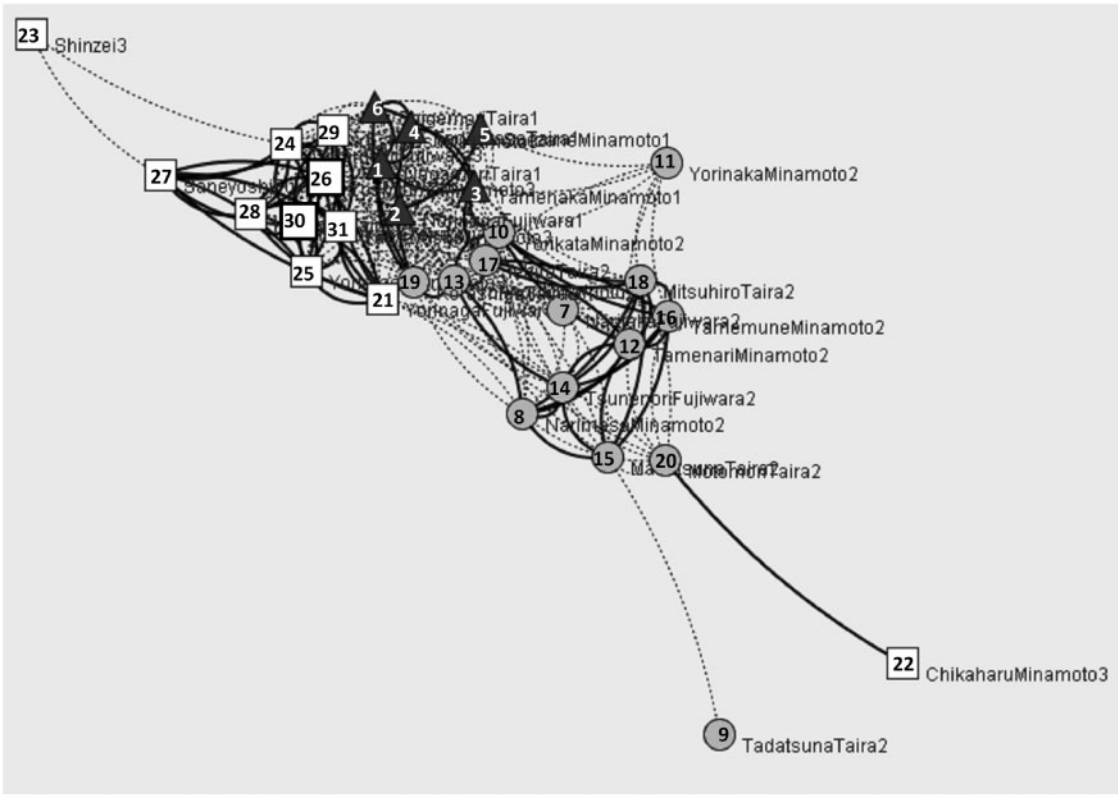

Fig. 2 Result of clustering persons using locational information

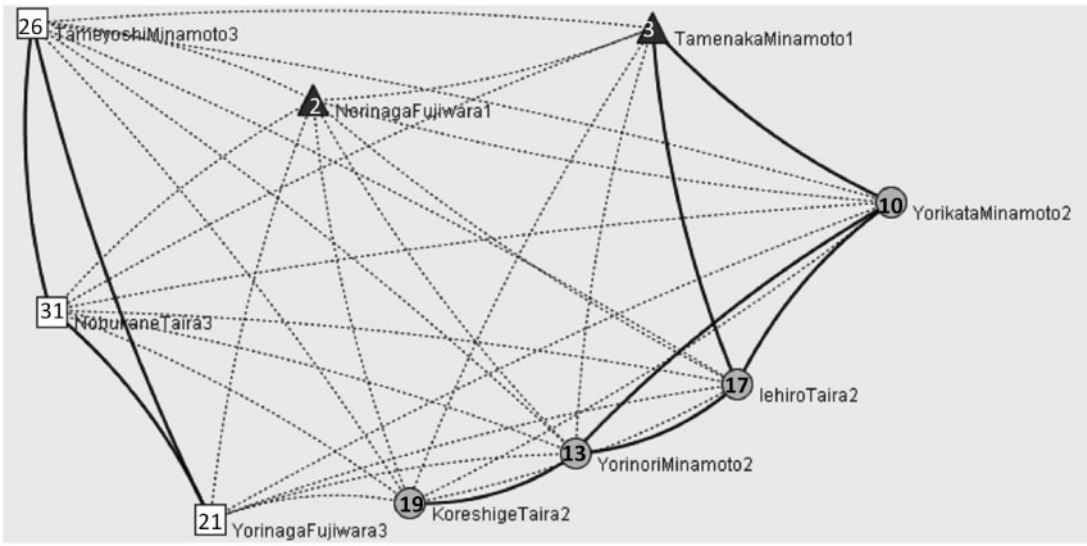

Fig. 3 A close-up of a part of the result of clustering persons using locational information (Fig. 2)

Table 2 The result of the user survey

Evaluation criteria	User				Average
	A	B	C	D	
Historical credibility	5	5	4	4	4.50
Novelty for historical research	5	5	5	6	5.25
Availability for historical research	5	5	5	5	5.00
Average	5.00	5.00	4.67	5.00	4.92

Although the average score of 4.50 for the 'historical credibility' is the lowest of the three evaluation criteria, it exceeded the neutral score of 4.00 and no one rated it negative. This result indicates that the proposed method is able to visualize mostly appropriate relationships among historical persons. 'Novelty for historical research' marked the average score of 5.25 and all of the users rated it positive. This result shows that the proposed method has a certain degree of novelty compared with the traditional ways of research in Japanese history and humanities in general. 'Availability for historical research' also marked a relatively high average score of 5.00, in which all of the users rated it as 5 (somewhat agree). This result indicates that the proposed method has potential to be a useful tool in the field of Japanese history and humanities in general. Although the number of users who participated in the survey is too small to give any conclusions and the scores are generally not very high, 10 out of 12 scores are positive, 2 neutral, and none are negative. It demonstrates that the proposed method gave a positive impression by the humanities researchers.

7 Conclusion and Future Work

In this article, we proposed a method to visualize relationships among historical persons using personal names and place names from Japanese historical documents. Using the proposed method, we obtained experimental results that correspond to the historical facts by visualizing relationships among persons from a historical document. The experimental results of the proposed method indicate a strong correspondence between the factions and the clusters, indicating the effectiveness of using locational information for clustering people. From the result of the experiment, we can estimate to which faction he belonged to during the Hōgen Rebellion.

The unique feature of our proposed method is that it can estimate the relationships among persons based on their spatial activities. Therefore, our method can estimate the relationships among persons who do not necessarily have direct relationships. In such cases, they may have a similar social position, behavioral patterns, family, etc. Besides, our method can also estimate the relationships among persons living in different ages, if their geographical scope of activity is the same or similar. Moreover, our method can be applied to multiple documents, if these conditions are met. Although we conducted experiments only for one document and a particular period of time in this article, our method has the potential to be applied to other documents, as long as we can extract personal and place names from the text and can define an appropriate unit of co-occurrence for that particular document.

Currently, our proposed method requires indices of personal names and place names in order to extract personal and locational information from text. However, there have been some researches on term extraction techniques from pre-modern Japanese text, which have a potential for automatic extraction of personal and locational information from text, though the accuracy of extraction still needs substantial improvements (Kimura *et al.*, 2011).

In future work, we aim to verify the usefulness of the proposed method in humanities research. In this article, we conducted experiments only for a particular person or a period. However, we need to conduct large-scale experiments of our method using various combinations of people and periods. We believe that our proposed method will be a useful tool for humanities researchers, which has possibilities for finding new knowledge which is undiscovered or different from existing theory.

Funding

This work was supported in part by the Grant-in-Aid for the Global COE Program 'Digital

Humanities Center for Japanese Arts and Cultures (DH-JAC)' from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and MEXT-Supported Program for the Strategic Research Foundation at Private Universities 'Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials', and Grant-in-Aid for Young Scientists (B) 23700302 'Information Extraction and Visualizing from Archaic Documents' from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Arthur, D. and Vassilvitskii, S. (2007). *k-means++: The Advantages of Careful Seeding, Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. New Orleans, LA, pp. 1027–35.

Crane, G. R. (2011). *Perseus Digital Library* http://www.perseus.tufts.edu/ (accessed 22 November 2012).

Hayashiya, T., Murai, Y., and Moriya, K. (1979). *Nihon rekishi chimei taikei 27: Kyōto-shi no chimei [Japan's Historical Place Names 27: Place Names of Kyoto]*. Tokyo: Heibonsha (in Japanese).

Hyōhanki Rindokukai. (1999). *Hyōhanki jinmei sakuin [The Index of Hyohanki's Personal Names]*. Kyoto: Shibunkaku Shuppan (in Japanese).

Itsubo, S., Osaki, T., Kimura, F., Tezuka, T., and Maeda, A. (2011). *Visualization of Co-occurrence Relationships Using the Historical Persons and Locational Names from Historical Documents, Conference Abstracts of Digital Humanities 2011*. Stanford, CA, pp. 326–9.

Kimura, F., Yoshimura, M., and Maeda, A. (2011). *Term Extraction from Japanese Ancient Writings Using Probability of Character N-grams, Proceedings of the Second International Conference on Culture and Computing (Culture and Computing 2011)*. Kyoto, Japan, pp. 183–4.

National Institute of Japanese Literature. (2011). *Database of National Institute of Japanese Literature.* http://www.nijl.ac.jp/pages/database/ (accessed 22 November 2012).

Osaki, T., Itsubo, S., Kimura, F., Tezuka, T., and Maeda, A. (2011). *Visualization of Relationships among Historical Persons Using Locational Information, Proceedings of the 10th International Symposium on Web and Wireless Geographical Information Systems (W2GIS2011)*. Kyoto, Japan, pp. 230–9.