Wikipedia Mention Graphs by Example

Armin Hoenen
CEDIFOR
Goethe university Frankfurt
hoenen@em.uni-frankfurt.de

Abstract content

1. Introduction

In this paper, we present the format of *Wikipedia Mention Graph* by example of Japan. Here, we focus on the Edo period of isolation where foreign relations were minimised especially with Western countries, but continued to exist in the Asian proximity (Katō, 1966; Toby, 1984; Morris-Suzuki, 1994; Laver, 2006; Boxer, 1957; Kinski, 2013, on Edo Japan). It was generally forbidden for foreigners to enter and for Japanese to leave the country during that time. Schich et al. (2014) have shown how datbases can be used as sources to visualize larger global migratory patterns; the isolatory period should be visible in these data.

2. Experiment I - Artisan Databases

Like Schich et al. (2014) we look at birth and death places and dates using the databases Freebase.com (FB), Allgemeines Künstlerlexikon (AKL), The Getty Union List of Artist Names (ULAN). The goal is to identify people who were born outside of Japan, but died in Japan during the Edo period and vice versa. Mapping of geo-coordinates and placenames to Japan, we extracted and merged (deleting duplicates) entries adding the larger dataset of the German National Library (GND) ending with roughly $5,000$ historical people with either a birth or deathplace in Japan (mostly either). Quantitatively, Figure 1 shows a steady slow increase was followed by a decline roughly coinciding with the beginning of the isolation and another increase roughly coinciding with the end of the isolation. The decrease towards the end is an artifact of the exclusion of living people. Qualitatively, for examples, see Table 1.[1]

Name	Born	Died	Group	Databases
Koxinga	1624, Japan	1662, Taiwan	merchant	FB
Antoine Fauchery	1823, France	1861, Japan	adventurer	AKL
Henry Bell	1808, USA	1868, Japan	military	FB
Sokuhi	1616, China	1671, Japan	Zen	AKL, GND
Ingen Ryuki	1592, China	1673, Japan	Zen	GND, AKL, FB
Itsunen Shoyu	1601, China	1668, Japan	Zen	AKL, FB, ULAN
Gempin	1587, China	1671, Japan	intellectual	AKL
Shinetsu Toko	1639, China	1695, Japan	Zen	AKL,GND
Mokuan Shoto	1611, China	1684, Japan	Zen	AKL
Dokuryu	1596, China	1672, Japan	Zen	AKL
Lorenzo Ruiz	1616, Phillipines	1637, Japan	Christian	FB
Luis Sotelo	1574, Spain	1624, Japan	Christian	FB
Gonsalo Garcia	1557, India	1597, Japan	Christian	FB
Pedro de Avila	1592, Spain	1622, Japan	Christian	GND
Carlo Spinola	1564, Italy	1622, Japan	Christian	FB,AKL
Caius of Korea	1571, Korea	1624, Japan	Christian	FB
Giovanni Sidotti	1668, Italy	1715, Japan	Christian	GND
Philip of Jesus	1572, Mexico	1597, Japan	Christian	FB
Domingo de Erquicia	1589, Spain	1633, Japan	Christian	FB

Table 1: Edo period people either foreigners who died in or Japanese who died outside of Japan.

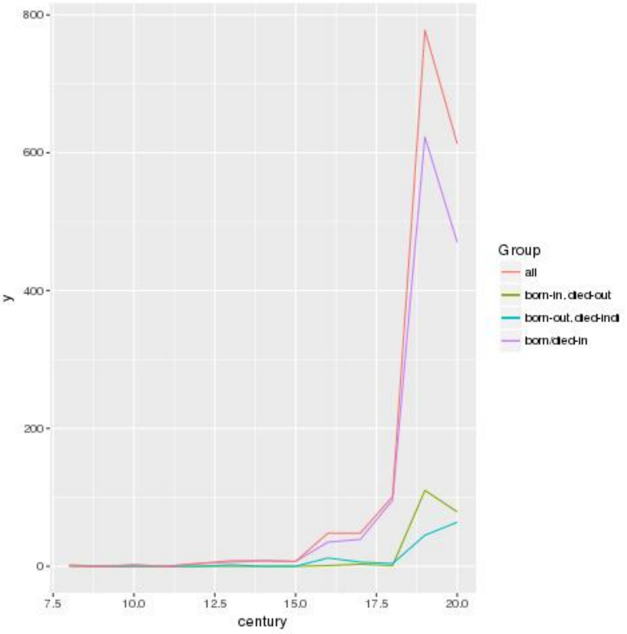

Figure 1: Approximated migration patterns of extracted persons from the artisan databases and the GND.

3. Experiment - Mention Graph

We sought to overcome the limitation that the databases were Western and not tailored for the research question of characterising the isolation period in that they only contained data on (cultural) artisans on the one hand and were only fit to investigate migration. Since the Wikipedia as a free resource provides full textual descriptions of any kind of people, we extracted indicators not only of migration which is limited in isolatory periods, but also of cultural/intellectual exchange and to this end collected all mentions (uniquely) in the Japanese Wikipedia articles *of all Japanese people, where they referred to non-Japanese contemporaries or earlier non-Japanese persons* and vice versa as an approximation of cultural exchange.

4. Preparation

In preparation, we extracted a comprehensive person dataset from the Japanese Wikipedia. The biggest and probably most well-known projects extracting knowledge from Wikipedia is DBPedia (see Bizer (2009; 2015)). DBPedia

[1] For a more exhaustive list, compare articles 明治維新以前に日本に入国した欧米人の一覧, 江戸時代の漂流者. Compare also the map of Edo period trade relations from (Vie, 2002, p.72).

Datapoint	Error Rate	Precision	Recall
birthdate	0.01	**1.0**	0.99
deathdate	0.015	**1.0**	**1.0**
birthplace	0.03	0.98	0.77
deathplace	0.14	**1.0**	0.3
Japanese Y/N	0.03	0.87	0.79

Table 2: Evaluation results, error rate for if the extracted entity was erroneous.

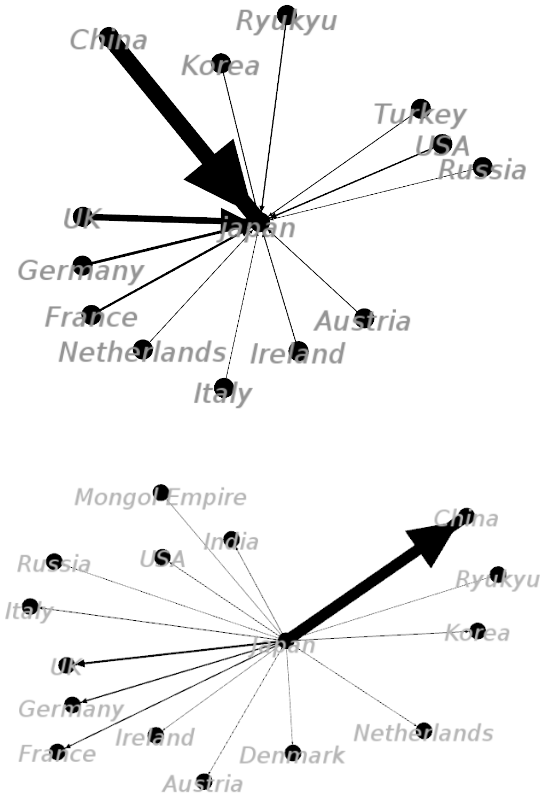

Figure 2: Patterns of cultural exchange during Edo Japan, pruned, visualisation with gephi.

uses the content of info-boxes and there is an ongoing effort to complete these by the Wikipedia, although for a large number of persons, they may still be missing (in our evaluation only roughly 40% featured an infobox). We extracted people with a birth and death date, which had any type of relation to Japan. This resulted in roughly 13, 000 people.

4.1. Datapoint Extraction and Evaluation

We used information from infoboxes, the first sentence of the article, the text body and the categories for the extraction of the datapoints seen in Table 2. For the evaluation, we randomly chose 100 people whereof we made sure that at least 50 were historical people and manually labelled them for the datapoints.

5. Mention Graph

Mapping people to countries (nationalities), we compute relative frequencies of mentions per country, see Figure 2 with a focus on Japan and Figure 3.

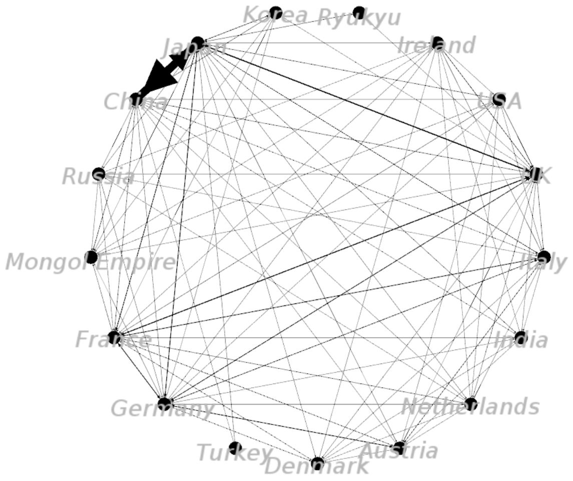

Figure 3: The current Wikipedia Mention Graph.

6. Discussion

While the Asian proximity and Europe are mentioned most and most mentioning Japanese people, other regions rarely occur. Historically, China was important especially on the intellectual (literature, arts, religion etc.) level, where mentions capture contact or perception of historical people such as *Confucius*, likely to appear in many articles. More research is necessary to evaluate the historical dimension.

For the whole graph, the average distance of the top 10 mention countries to any base country (in case all 10 had a value larger 0) was 2871 km where excluding Japan reduced this to 1978 km making a bias towards the Wikipedia home language plausible for Mention Graphs. Additionally, the Jaccard coefficient between the most mentioned and the geographically closest (after (Mayer and Zignago, 2011)) 10 countries was 0.24 which is by no means low considering that Japan and links to (former) colonial powers intervene. Thus, a larger probability of regional neighbours in Mention Graphs is also expectable.

7. Discussion

The results presented here are by all means nothing more than preliminary, since as a short paper, they represent unfinished research. An evaluation set of only 100 units is far too small to be representative. In this light, results are best interpreted as a ”what-if” scenario. The preliminary character holds especially for all interpretational aspects of the current contribution. Preliminary analyses suggest that Thailand had a more significant role and that some nationality extraction needs more fine-grained evaluation and improvement.

8. Conclusion and Outlook

We produced a Mention Graph mapping contact with foreign nations in the case of Japan in an isolatory period. The format of Wikipedia Mention Graph reflects to some degree cultural exchange and may be generated for any language and scenario.

9. Bibliographical References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165. The Web of Data.

Boxer, C. R. (1957). Sakoku, or the closed country, 1640-1854. *History Today*, 7(2).

Katō, H. (1966). Die Abschließung Japans. *Kagami*, IV(1):67–88.

Kinski, M. (2013). *"Riten" beginnen bei "Essen und Trinken". Entwicklung und Bedeutung von Etikettevorschriften im Japan der Edo-Zeit.*, volume 13 of *IZUMI*. Wiesbaden: Harrassowitz.

Laver, M. S. (2006). *The Sakoku edicts and the politics of Tokugawa hegemony*. Cambria Press.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Mayer, T. and Zignago, S. (2011). Notes on cepii's distances measures: The geodist database. Working Papers 2011-25, CEPII.

Morris-Suzuki, T. (1994). *The technological transformation of Japan: From the seventeenth to the twenty-first century*. Cambridge University Press.

Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., and Helbing, D. (2014). A network framework of cultural history. *science*, 345(6196):558–562.

Toby, R. P. (1984). *State and diplomacy in early modern Japan: Asia in the development of the Tokugawa Bakufu*. Stanford University Press.

Vie, M. (2002). *Histoire du Japon, des origines á Meiji*. Presses Universitaires de Franc.