The Correlation Between Online and Offline Activity

Programming for Social and Cultural Data Analysis
Prof. Joachim Scharloth
1M160301-2
Masaki Minamide

Index of Content:

---

1. Introduction

As the concentration of population into urban areas has been a problem not only in Japan but in many developed countries in the world, Abe regime putting a number of regional creation policy in action but the concentration in Tokyo is not showing any sign to stop yet. Likewise, local governments in Japan are also making a great deal of effort to attract more visitors and migrants, whereas rapidly shrinking population and the aging society problem project the future of regional areas becoming extinct. This current situation cast many implications of the future of Japan and needs for new and novel policy to confront these problems. One of the measures is online marketing to enrich young people and foreigners with information about great places which were hardly

known before. However it is not an easy task to convince local governments to engage with online marketing, it is important to statistically analyse the impact of online activities to the real world consequences. Therefore, this paper puts a research question as “Does the frequency of city name being mentioned online have a correlation with actual number of foreign visitors to those cities?”.

2. Methods

I divided the procedure to ultimately project relation between online and offline activities through statistical analysis into three parts:

1) Acquire data of mentions of city name from large and credible enough travel platform or SNS.

2) Acquire data of actual visitors from any trusted data provider, hopefully the government itself.

1) Acquire data of mentions of city name

Japan-guide.com, one of the oldest platforms out these since the emergence of the Internet, appears to be a good choice for two reasons: 1) large amount of resources has been accumulating with no major suspension of the platform confirmed. All together, from 1997 to 2019, 56,082 threads are archived in travel topic. 2) category based branches provide with trustable posts about traveling to Japan by potential travelers. I considered Twitter as a data resource option due to its ubiquitousness in daily life, but I chose not to use Twitter because there are post spamming of city name with irrelevant content.

Scrape procedure

a) “japanguide_spider.pl” create list of URLs leading to each thread. This is necessary because the platform is deeply structured, meaning that one get request can access only one thread. For this process, “japan_guide.pl” keeps record of URLs starting with basis in hash format, then finally outputs “follow_list.txt”.

b) “scrape_corpus.pl” iterates each URL in previously created list in “follow_list.txt”, then scrapes year, body of thread, and replies out of html by using regex matching. This outputs “pots_$year.txt”, which will be our main corpus. This process of scraping 56,082 threads took about one and a half day to finish with single terminal thread.
c) “calculate.pl” counts appearance of city name in a hash containing all city names in each corpus text, then output “mentions_$year.txt”, which has a structure: Tokyo: num\n Kyoto: num\n…. Since using relative frequency is more relevant to compare with actual data later on, I created “calculate2.pl” which calculate relative frequency and put then in one single “r_f.csv” file.

2) Acquire data of actual visitors

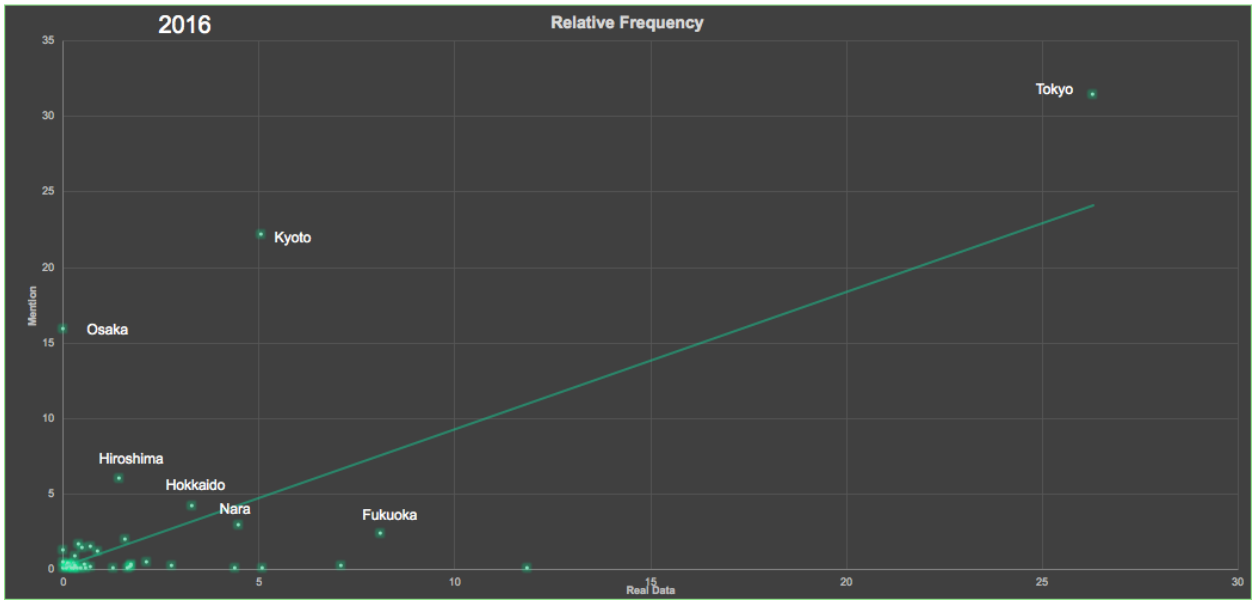
Japan Tourism Agency has a standardised survey that keeps record of various types of data regarding travel both Japanese and foreigners. However, since this survey started in 2010, I had to disregard mention data on japan-guide.com before 2010. In addition, some cities have not yet introduced the survey like Osaka, collected data contains a few blanks. Other than these limitations, their data seems relevant because survey targets people visiting tourist spots in prefectures and visiting festivals/events. All csv files are publicly open on their website, therefore there was no need for preprocessing, but just copy and paste.

3. Findings

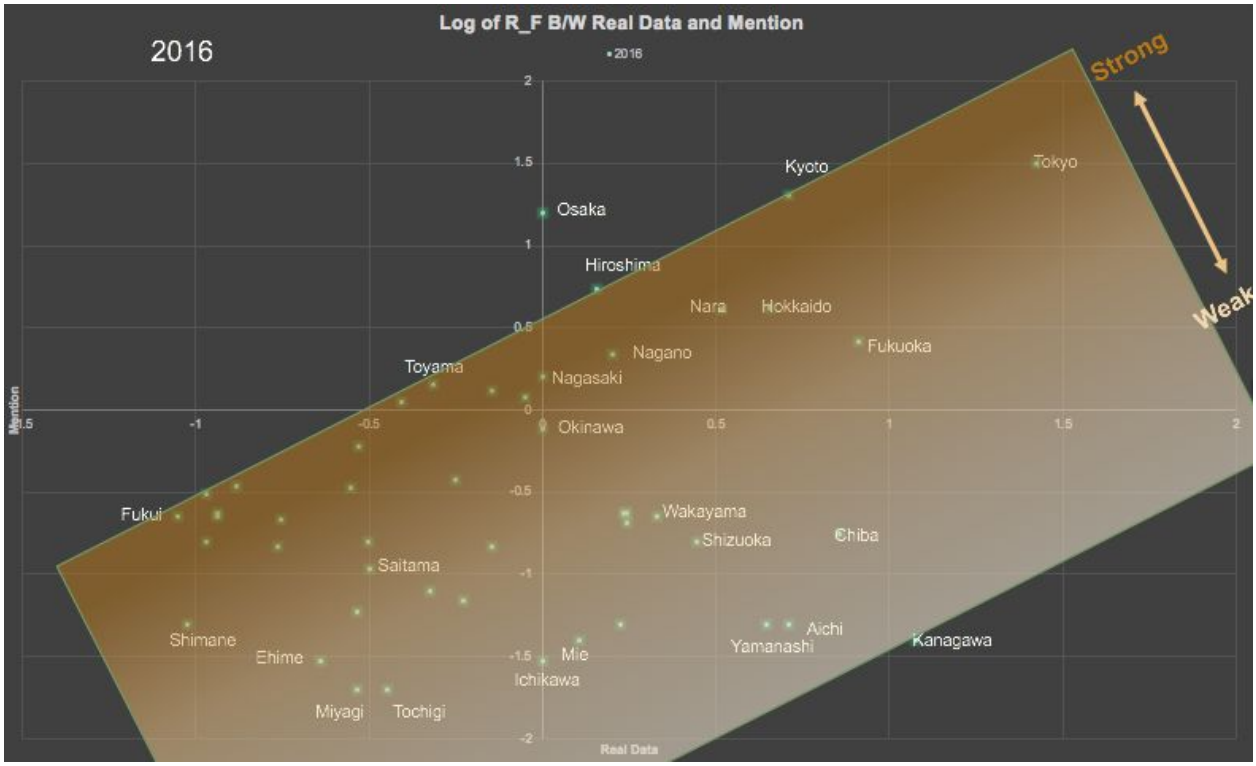
Compare data of mentions and visits by cities and years. Then create charts to confirm if any trend line exists to see the relation.

Before comparing data, I gathered all relative frequency data of mention and visits from 2010 to 2017 to one single file named “comparison.csv”. For comparing data, excel’s primitive function to create a chart by selecting a range of data. First, the chart of 2016 with x-axis as actual visit data, and y-axis as mentions showed Tokyo at the top right corner and Kyoto at the top right, with other cities stacking on each other at the

bottom left. This is typically called a “long tail” graph, with few items dominate other items, which is frequently observed in economics.



In order to display a relevant chart with disperse data, I rescaled data with logarithm of x to the base 10. This outputs a chart below.



This chart now not so clear but slightly shows a tendency of location where items are plotted. I drew a skewed rectangle as a trend line, which shows that there is a correlation between actual visits and mention data. Correlation can be calculated as abs(visits-mentions), which namely means that the smaller the difference is, the more correlated both types of data are. The rectangle box in the chart is colored with gradation of brown to white brown, showing correlation level. The result shows cities like Nara, Hokkaido, Tokyo reflect the correlation well, on the other hand, cities like Kanagawa, Aichi, Yamanachi shows the weakest correlation.

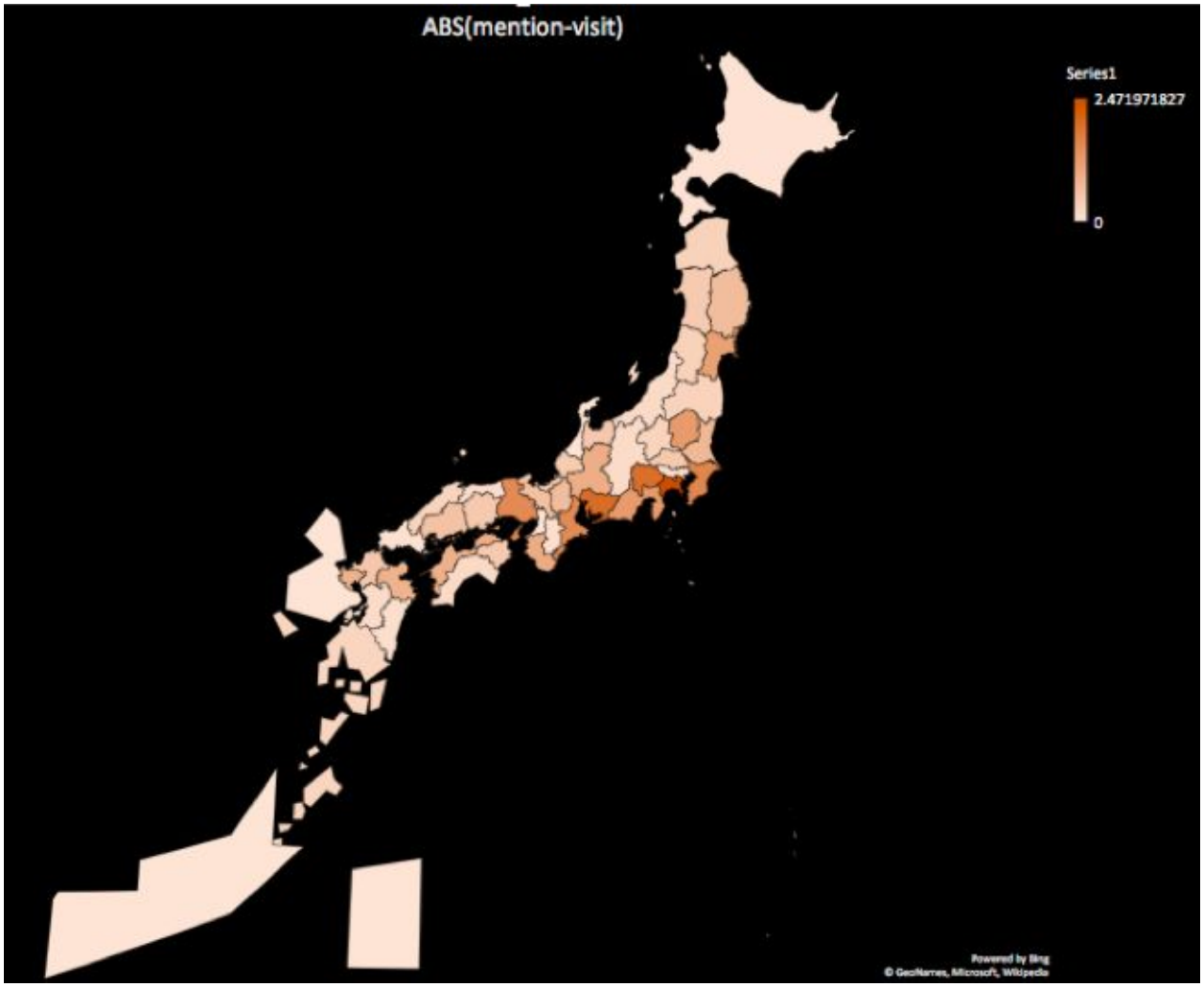
4. Extensive Research

2016	Diff	Visitor	Mention	Bigger
Kanagawa	2.47197183	1.07403182	-1.39794	Visitor
Aichi	2.00821052	0.70718053	-1.30103	Visitor
Yamanashi	1.94260156	0.64157156	-1.30103	Visitor
Chiba	1.59686156	0.85213407	-0.7447275	Visitor
Hyogo	1.52105333	0.22002334	-1.30103	Visitor
Mie	1.50323156	0.10529155	-1.39794	Visitor
Tochigi	1.2494611	-0.4495089	-1.69897	Visitor
Shizuoka	1.23984762	0.4439676	-0.79588	Visitor
Miyagi	1.16285584	-0.5361142	-1.69897	Visitor
Wakayama	0.96773162	0.32945946	-0.6382722	Visitor
Kagawa	0.92576712	-0.2291348	-1.154902	Visitor
Gifu	0.91731464	0.23953394	-0.6777807	Visitor

Through the previous comparison of visit and mention data set, the chart visualized that some cities have weak correlation and some have strong correlation. This led me to a

new question: What could be a possible reason to explain this consequence? Without any clue, I compared the difference of visits and mention, then sort then by order of diff large to small. This led to a noticeable result that top 20 cities with large difference is caused because visit data overwhelms mention data, in other words many cities are not mentioned online but successfully attract foreign visitors. To answer to my next question: how do these cities have more visits than mentions, I visualized plotted difference data to geographical map of Japan by using Microsoft Geometrics.

What an eye-catching point in this geography is that those cities with large gap are concentrated in Kanto(east) area towards Kansai(west) area. Two possible reasons could be considered: 1) Given Tokyo is the most popular city with the most visits, cities with large gap are easily accessed due to the fact that they located adjacent to Tokyo. 2) Under an assumption that it is unlikely that a foreign traveler visits only one city, it is highly possible that they also visit other popular sightseeing cities such as Kyoto and Osaka. This might be a reason why cities between Tokyo and Kyoto show higher contrast.



5. Limitations

Having many findings aside, there are some limitations and rooms to make this research more meaningful. First, the incompleteness of dataset provided by Japan Travel Agency and inconsistency in the method of the survey potentially hindered accurate result. Second, even though the correlation is confirmed(the purpose of this research), it is not easy to determine the probable cause and effect: does more mention bring more visitors or the other way around?

6. Conclusion

As my initial purpose of this research is to find a relationship between online and offline activity by comparing both datasets, so many meaningful insights were obtained by visualizing geographical metrics in 4. Extensive Research. But still as mentioned in 5.Limitations, even though the correlation is confirmed, figuring out the cause and effect would make this research more meaningful data and practical to solve actual cases of concentration of population in urban areas.