Data Brokers: To know them is to hate them, but how consistent are they?

Alexandra Nisenoff
nisenoff@uchicago.edu

Data brokers are shadowy companies that collect user data for profit. While recent articles have shed more light on these companies, there is still a lot that we don't know about these companies. Currently, the process for requesting all of the information collected by data brokers is either time-consuming or costly. Some data brokers offer peeks at the information that they have collected. Their websites offer a way to explore the types of data, consistency, and prevalence of the data types, as well as how the records for an individual varies across sites.

1. Methodology

To explore these questions I wrote code that automatically collects all of the records for a given first and last name from the free previews from various data brokers. This tool primarily used Selenium to access the records from these sites. The choice of Selenium over other tools was made because of ease of use for programming and there was a better chance that the requests wouldn't be blocked.

The records from each site were loaded into a consistent dictionary format. The keys corresponded to the rank of the result on the website to allow for a comparison of correlated records across sites.The dictionary for each user was made up of specific key-value pairs. The types of information that were stored included: the name associated with the record, their current city of residence, all locations they had lived, a list of relatives, a list of aliases, and their age.

The main thing that my code does is provide general insights into the information included in these records through various charts. The graphs include information about the ages included in the records, what percentage of the records included each type of data, the average and max counts of data when a type of data (e.g. a list of relatives) could include multiple values,and how many of the results matched the queried name.

After creating these charts, I attempted to create a system for matching up records between two sites. This is important because it is possible to answer interesting questions about the consistency of data across data brokers. To determine these match ups, I provide a framework that compares every record from one site to every record from a second site assigning a score of 0 to 1 for each type of information that could have been collected by the data scraper. The way records were compared was a combination of edit distance and Jaccard similarity. The code also gives the option of specifying coefficients to change the weight of the individual similarity metrics in the overall similarity calculation. I decided not to determine a standard set of coefficients since there is probably no objectively right choice for these values. When attempting to match searches with very few results you may want very different numbers than if you were trying to match a large number of results. This may also depend on what sites you are comparing as not all sites may provide the same categories of data. Since the choices of coefficients influence how the records are matched up, the biases introduced must be considered in any analysis of these data. While the coefficients are not set here, the framework that is needed to determine these values and match them up is provided.

2. Discussion

While completing this project, I learned that there are a lot of variations in the information that data brokers provide. Some of the most common information that data brokers provided was a person's current city of residence, all locations they had lived, a list of relatives, a list of aliases, and their age. That being said, not all sites included all of those fields for every record or at all. There seems to be quite a bit of variability in the record for a given search regardless of how common the name is. Some may include all of the categories of data and others may only include one or two. Aside from name, age is the most common type of data to be included.

Based on the search results from these sites, there are some interesting differences. Some websites, like infotracer, will top out at 100 records, but other sites will return far more results. Within the categories of data, there can also be a lot of variability. For example, sites like Spokeo seem to limit the number of locations that they display in their free preview to 4, while truthfinder does not seem to have a cut off for the number of relatives or locations that they will list.

In this project, I created code that collects information about search results and provides a framework for approaching linking records across data broker sites, but there is still a lot of work to do in determining the similarity coefficients and exploring the information gathered about search results over a broader sample of queries.