What makes big data distinct in the social sciences

ABHINAV GARG
MS, Computer Science
University of California, San Diego
A53095668

Introduction -

Sociology is the study of human social relationships and institutions. The subject matter is diverse, ranging from crime to religion, from the family to the state, from the divisions of race and social class to the shared beliefs of a common culture, and from social stability to radical change in whole society. Sociology is an exciting and illuminating field of study that analyzes and explains important matters in our personal lives, our communities, and the world.

Big data, on the other hand, is a term that describes the large volume of data – both structured and unstructured – that affects us on a day-to-day basis. But amount of data is not important, how we visualize it matters. Big data can be analyzed to make better decisions and strategic moves. It is being generated by everything around us at all times. Every digital process and social media exchange produces it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, we need optimal processing power, analytic capabilities and skills. Though courses are being offered to prepare a new generation of big data experts, it will take some time to get them into the workforce. Meanwhile, leading organizations are developing new roles, focusing on key challenges and creating new business models to gain the most from big data.

The digital revolution incorporates tools that greatly improve the efficiency of tasks performed by the human brain. These tools have been named as probabilistic learning, machine learning and data mining, or collectively big data tools. One benefit of the study of history is that we may gain insight into the thoughts and behavior of our ancestors that could have been affected by social factors. Learning where we came from can help us better prepare for where we are going. At the launch of this age of Big Data, we are now well positioned to accomplish that ambition. Millions of gene sequences can be matched now in matter of minutes, thanks to the development of High Performance Computing and analysis tools.

Armed with Big Data Analytics, we can improve the well-being of ourselves and people around us, particularly in Social Sciences Research. Outcomes are very difficult to predict with a traditional reductionist approach when performing research. To gain the most meaningful insight from such huge datasets, we need to make full use of large scale and high performance computing (HPC). Additionally, the most complicated example of an emergent system is the human consciousness. All social phenomena exhibit properties that essentially emerge from various interactions between a large numbers of human brains. Social scientists need to leverage HPC to process the Big Data generated from these complex, intricate social phenomena. In this report, we'll look closely to see how social scientists are using Big Data Analytics.

Research -

Pierre Bourdieu's work in the field of sociology has been remarkable. He predicted that class distinctions exist among humans and different classes have different tastes. In this era, we can verify his theory using Big Data analysis. We can also verify the role of sequence analysis or optimal matching in different social science career developments. We can also see how social network analysis helps sociologists in finding collaboration networks among researchers. We explore the distinct role of Big Data in social sciences on the basis of three projects that we have completed.

In the first project, we tried to verify Bourdieu's theory on class preference. Our aim was to classify headphone buyers on Amazon website as either audiophiles or fashionists for which we used big data analysis. Our group scraped about 10000 - 20000 reviews from the site using HTML parser (Beautiful Soup package) in python library. We selected headphones in order to highlight features such as bass, clarity, noise reduction, sound quality, style, color, durability, material etc. that normally influence people's decision to buy an item. We selected 5 brands of headphones (Bose, Skullcandy, Sennheiser, Beats and Hellokitty) in order to collect reviews of people belonging to different sections of society, be it super rich, middle class or low income classes. Among these sections, we also wanted people with different social characteristics, be it frugal, spendthrift, purist, pragmatist, yuppies, disenfranchised etc. to have a diverse and robust data. After scraping, we prepared data for unsupervised topic modelling. We followed data pre-processing steps to discard reviews with less than 30 words as these reviews might have skewed our results. We also discarded stop words like a, an, the to emphasize more on the words that describe people's traits belonging to the two distinct categories chosen. Using LDA analysis on unigrams, we were able to generate 15 topics that could help us classify people as audio driven or fashion driven. We tried Bigram analysis to capture more specific characteristics but topics obtained from unigrams reflected better traits, when compared in the word cloud. With the help of big data knowledge, we were able to generate good clusters within a week with high accuracy, using fast computing power and large storage in laptops we have these days. Otherwise, manual work for such an analysis would have taken months to accomplish and scope for mistakes would have been very limited. If given more time, we could have gathered more reviews, even in range of millions or billions where we would have actually exploited the power of big data.

Another similar example where Big data plays a distinct role in sociology is Personality Assessment. The idea here is to determine the dimensions along which personalities differ. Big Data analysis is used now but this experiment was performed for the first time by Allport and Odbert in 1936, in the era where Computer Science was just a myth. They sat down with the English dictionary and extracted all terms that could be used to distinguish one person's social behavior from another's. They collected roughly 18000 words, of which 4500 could be described as personality traits. They grouped these words into synonyms through manual clustering. They came up with the categories like -

Spirit - Jolly, merry, witty, lively, peppy
Talkativeness - Talkative, articulate, verbose, gossipy
Sociability - Companionable, social, outgoing
Spontaneity - Impulsive, carefree, playful, zany
Boisterousness - Mischievous, rowdy, loud, prankish

Adventure - Brave, venturous, fearless, reckless
Energy - Active, assertive, dominant, energetic
Conceit - Boastful, conceited, egotistical
Vanity - Affected, vain, chic, dapper, jaunty
Indiscretion - Nosey, snoopy, indiscreet, meddlesome
Sensuality - Sexy, passionate, sensual, flirtatious

To collect this data, they asked lot of people to what extent each of these words described them. They generated this matrix of data where 1 means strongly disagree and 5 means strongly agree.

	shy	merry	tense	boastful	forgiving	quiet
Person 1	4	1	1	2	5	5
Person 2	1	4	4	5	2	1
Person 3	2	4	5	4	2	2

It was difficult to analyze such huge data manually. But now, if we want to extract the important dimensions, we can use big data techniques such as k-means clustering and Principal Component Analysis. We will find high correlation among these traits, to the point where each person either answers 1 to both or 5 to both. We can discard one of them to reduce our data size and scale it to a dimension where our brain can visualize it. Personality Assessment has many applications these days, it is widely used by online dating and match-making sites.

In the second project, we used big data analysis to find out more about social scientists who have similar career transition. Our plan was to use the university tier information in which these social scientists had published their research journals. First step was to collect data that can highlight this relationship for which we used “Web of Science” website where all data related to publications, such as author names, university names, topics, date, document id etc., is available. We scraped about 10000 pages from the website using HTML parser. We obtained multiple author names and their universities from each page. We then separated and arranged each author and university to obtain a list with authors and their corresponding university of publications using our coding skills. Each university was classified into a tier, 1 being the highest and 4 being the lowest. After these pre-processing steps, we used sequence analysis to compare one author’s career with another based on the university tier information. We used different weights while adding, deleting or changing an entry in our list. More emphasis was given to an author who moved from lower tiers to upper tiers during his career. Finally, we used k-means clustering to group authors having similar career transition. We again see how the knowledge of data mining, machine learning and algorithms helped us analyze such a huge data set. We observed that most of the authors had multiple publications in the same university and career transition from lower tier to upper tier universities was limited. Similar projects have been undertaken by researchers around the globe. One of the projects used

3/11 data in New York City and applied edge detection algorithms to census maps in order to measure the distinctiveness of neighborhood boundaries.

In the third and final project, we used big data analysis to check whether the social scientists in same university tier tend to collaborate with each other or not. We scraped about 7000 pages with pre-decided sociology research topics [Social forces, American journal of sociology etc.] from the "Web of Science" website. We extracted author's information and universities of publication from journals on every page we scanned. We separated and arranged universities according to their tiers and assigned them to corresponding author in our list. This time we made use of some data analysis tools like Sci2 and Gephi. Data collected from the journals like authors, universities, abstract, date of publication and accession number etc. were kept in a single file and fed to Sci2 tool for processing. Gephi is a data visualization tool that generates graph with the nodes and edges obtained from the co-author network generated by Sci2. Graph generated gives us a clear indication that researchers from same tiers do tend to collaborate with each other. Graph also showed some outliers that indicates that our hypothesis is not 100% true. But, these analysis tools coupled with the fast processing speeds have given us a platform to visualize large chunks of data within a short span of time. One of the similar projects showed that Google searches can be used to detect racism, given the access to a history of searches. It was based on the theory that the incidence of Google searches for racist terms was associated with increased racial discrimination across American cities. It gives our society a head start in reducing crime rate and narrowing down the potential criminal.

The relation between big data and sociology is also visible with Recommender systems used in Google search, Netflix and YouTube etc. The aim here is to match customers having similar social traits with products. Here, data available is of prior purchases and interests of the registered users, collected over many years and scaled down to make processing fast. The system recommends further products of interest to the new users. A successful approach to achieve this is collaborative filtering - Neighborhood methods and Latent factor methods. These methods match your recent views with those users who had seen similar shows and recommends you what they had seen or showed interest in later on. The suggestions we get nowadays might not be useful but recommender system is only going to improve in the future when the organizations consider more social traits and attributes to compare.

After going through all these projects, I learnt a lot about the unsupervised learning techniques employed in order to get meaningful results from the raw data we gather. I was able to put theory to practice, especially when it comes to k-means clustering, LDA and optimal matching. The projects made me aware that Data Mining is a big challenge in front of us, and we'll not obtain the same results every time we use these algorithms. Also, apart from coding and scraping, finding out which data is skew or relevant plays a key role in how close we get to the desired hypothesis.

Conclusion -

Scholars from a wide range of disciplines in the social and natural sciences have talked about the future of "big data" in social science research. Major technical advancements have given social scientists access to new forms of data and sophisticated analytical tools, but the full potential of these resources has not yet been realized. Sociology in the urban areas is more focused on poverty, neighborhood effects, and unemployment. These are critical issues, but researchers often rely on old methods and fail to take advantage of cutting-edge data and analytic techniques. Big data holds immense promise, but in order to take advantage of it, the technical experts and the subject-matter experts must bring these two distinct fields into conversation. Big Data is here to stay and its distinct role in social science will be visible even in the future.

Acknowledgement –

I thank Professor Juan Pablo Pardo-Guerra for introducing "Theories/Practice of Big Data" to us. I thank my team mates [Abhijit, Amit, Bryan and Tania] for their co-operation and support, they guided me through the roadblocks encountered in the projects.

Appendix -

[1] http://blogs.lse.ac.uk/impactofsocialsciences/2015/10/13/ideological-inheritances-in-the-data-revolution/

[2] https://en.wikipedia.org/wiki/Sociology

[3] http://citiespapers.ssrc.org/bringing-social-science-back-in-the-big-data-revolution-and-urban-theory/

[4] http://www.psychometric-assessment.com/the-lexical-hypothesis-and-factor-models/

[5] lectures and slides from SOCG 290 class