Big Data Applications in Historical Studies

Neil Eliason
Indiana Unversity
Anderson, Indiana

ABSTRACT

KEYWORDS

i523, HID 312, Big Data, History

1 INTRODUCTION

1.1 Big Data

Big data attention and success stories. Driven by More's Law.

Big data to date can claim numerous victories in a variety of fields, and promises more. Businesses such as Facebook and Netflix have built corporate empires off of the insights gathered from their big data, and physicists and biologists are learning what makes up the universe and ourselves via big data [1].

Despite all this, the concept itself is rather nebulously defined. A rough description of is data with quantitative factors that require specialized techniques to utilize. The most commonly referenced big data factors are volume (amount of data), variety (number of data source types), and velocity (rate of data collection or input) known as the three vs. While this definition is generally accepted, its application varies based upon the industry or field of study and often changes with developments in information technology [5].

The focus on big data arises partially from the phenomenon of data storage capabilities growing at a faster rate than data processing. This creates a situation where data can be economically stored, but not as economically processed, requiring specialized analytic techniques. As big data progresses through the storage, cleaning, analysis, and interpretation stages of the data life cycle, specialized approaches are required [1].

DIKW Data lifecycle

1.2 History of History

The historian's labor has involved interacting with voluminous and varied data for centuries. Before computers, this process involved searching physical archives for relevant data, and manually copying and organizing it into useful information to be analyzed. Though this method can deliver deep insights, some data sets are too big to be studied in a manual fashion [7].

Around the mid-twentieth century, computers had become sufficiently powerful and usable for historians to begin using them to process larger amounts of information. This facilitated a change towards a more quantitative approach and a focus by some from tracing the rise and fall of political or ideological forces, to developing a more complete understanding of mundane topics, such as the family or economics.

Now as archives become digitized and accessible via the internet, the quantity of data available leads to the appeal to big data analytic methods [4]. The potential of unlocking significant connections and developing big picture historical insights at the scale of the growing digital archives of the world is alluring. This hope has driven the labor of many researchers towards developing more big data informed research methods and has directed funds of many institutions towards investments in data infrastructure. However, many are also concerned that the promises of big data are at best optimistic, and at worst hiding potential pitfalls to the historical process [7].

1.3 Thesis

Big Data Analytics have the potential to provide new insights to the field of historical studies. However, their application will differ due to the nature of historical data, and they will serve as an additional tool for the historian, rather than the only tool.

2 BIG DATA IN HISTORICAL STUDIES

2.1 Data Sources

Source types

Methods to get information from data

Methods different from streaming data It could be argued that the seeds of big data history have long laid dormant in archives and libraries, waiting to be germinated by sufficient computational capabilities to process them. As big data analytics mature, pressure develops to make more data available for analysis by digitizing more archival material. This is evidenced not only by the familiar repositories of e-books, but also by archives containing millions of pages from newspapers [7] centuries of letters [4], and

Sources for big data research consist not only of the content of documents in an archive, but also the bibliographical records.While originally designed to allow individual works to be located in an archive, historians have began to study the bibliographical data themselves, an approach called distant reading. By looking at the data about a document, rather than the document's content, societal or intellectual trends can be identified across large scale factors such as time or geography in a more comprehensive way. This approach has elicited some criticism that collections of bibliographical data are not complete enough to derive such large-scale conclusions. Still, considerable interest exist in targeting these data sets for historical analysis [10].

However, the data from these sources differs from that of other fields which utilize big data analytics. Historical data is not streaming the way that social media or smartphone sensors are. It is data which has already been collected, organized, and often times analyzed for a purpose defined by people from a different time and different needs/constraints from ourselves. This creates data sets which are difficult to compare and often require considerable cleaning and reworking to be used in a larger framework. [4].

2.2 Analytics for Big Historical Data

Analytic Techniques

Due to the natural reliance on documents in historical studies, text analytic techniques are the primary set of big data approach utilized by historians. Text analytics is broad category of related algorithms and statistical techniques, such as artificial intelligence, machine learning, and natural language processing that attempt to extract specific information from the text and identify patterns and relationships within the body of data [7].

Artificial intelligence is "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings"[2]. In the context of historical research, this would include tasks such as extracting relevant content from sources, identifying relationships within the data. A specific type of artificial intelligence is machine learning, which consists of programs which change their actions autonomously in response external input. Their ability to adapt allows them to do decision-making tasks, and thus can search through data sources in a more intelligent way to find relevant data [1]. Natural language processing is another artificial intelligence technique, which aims to create programs that can take human language, and make it machine readable [9]. Historians can use such programs connect archival documents to more complex analytic algorithms.

In order to interpret the results of big data analysis, visualization is critical. This is a challenge, as the large scale of the data makes striking a balance between a sufficiently big picture perspective without losing relevant details difficult. Many approaches attempt to utilize high resolution approaches to avoid losing important information [1]. This process is especially challenging in historical studies, as the data is often incomplete and may have inconsistencies which prevent assuming a uniform set of data. For this reason, historians often use visualizations to identify qualitative, rather than quantitative relationships in the data, to inform further inquiry [4].

2.3 Software Packages and Resources for Big Data History

A variety of software packages have been utilized to assist the process of translating raw data into historical insights, such as such as Tableau, Gephi, R, and ArcGIS. However, a limitation of these tools is their their quantitative focus, which tends to exclude more qualitative approaches [4].

Some software has been developed to provide a more qualitative visualization tool set for researchers.For example Stanford University developed a software package called Palladio, designed to visualize connections in large scale historical data. Their approach focused on visualizations that encouraged exploring data, rather than creating statistical statements about it. Examples of this would be mapping connections between historical actors over geography or creating a visualization of the social network of a particular figure in history. They do not create statistical arguments, rather they give a framework for understanding how the data are connected [6].

Another tool with a qualitative visualization focus is WAHSP. It was developed to extract data from the National Library of the Netherlands' newspaper archive, but has been utilized on a number of other databases as well. It provides a number of useful analyses, such as word frequency cloud visualizations, detecting positive or negative sentiment related to certain terms, and Named Entity Recognition, which can identify people, places, events, etc. and then connect them into a relational or geographical framework. It also provides an interactive histogram where the resolution of the data can be adjusted to quickly move between a big picture and detailed data perspective. A derivative project is BILAND, which is a program that can perform many of the analyses of WAHSP, but applies them across two languages, Dutch and German for comparative cultural studies [7].

Along with these data intensive tools specifically designed for historical studies, there are also resources to help the historian learn some of these methods. For example, The Programming Historian website provides a wide range of tutorials and lessons on how to use digital tools in historical studies. At the time of this writing there were 67 lessons available organized by their target stage of research, including lessons on using R, Python, Java, and GitHub [8].

2.4 Insights from Big Historical Data

A number of studies have used these techniques to approach historical research from a big data perspective. Stanford's Mapping of the Republic of Letters project sought to map the social network of Enlightenment thinkers who actively corresponded with each other. This was accomplished by utilizing big data analytics on the metadata of these letters to see how these thinkers related temporally, geographically, socially. Through the research process, the need for more qualitative approaches to visualization was recognized, and eventually led to the development of the Palladio tool set.

Their analysis revealed a number of interesting points. By mapping the social network of John Locke, they supported previous scholarly contentions that the Enlightenment culture was not homogeneously connected, but was made up of a number of subcultures which had thin social connections. Also, by analyzing Benjamin Franklin's letters, they noted that despite his reputation as cross cultural traveler, the main hub of his correspondence was between Philadelphia and London, which were major centers of British culture. [4] .

The WAHSP tool has also been used to analyze large data-sets. One researcher used the tool to study attitudes found towards drugs in the early 20th century newspapers. It found by using the word cloud analysis tool, that before 1924 that drugs such as heroin and opium were used in the context of health, but after 1924 they were associated with crime.

The related tool BILAND was used to study

3 POTENTIAL ISSUES

Opportunistic research: Data driven, not question driven [3], [4] Over-hyped Gaps in data sources Improperly formated data[4]

4 CONCLUSION

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

REFERENCES

[1] C.L. Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275, Supplement C (2014), 314 – 347. https://doi.org/10.1016/j.ins.2014.01.015

[2] B.J. Copeland. 2017. artificial intelligence (AI). Webpage. (01 2017). https://www.britannica.com/technology/artificial-intelligence

[3] Malte C. Ebach, Michaelis S. Michael, Wendy S. Shaw, James Goff, Daniel J. Murphy, and Slade Matthews. 2016. Big data and the historical sciences: A critique. *Geoforum* 71, Supplement C (2016), 1 – 4. https://doi.org/10.1016/j.geoforum.2016.02.020

[4] Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer, and Nicole Coleman. 2017. Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters ProjectHistorical Research in a Digital Age. *The American Historical Review* 2 (2017), 400. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsoaf&AN=edsoaf.a29ec0ac934f1257030b477fa5986b1cff6def96&site=eds-live&scope=site

[5] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 2 (2015), 137 – 144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[6] Stanford Humanities and Design. 2017. Palladio. Visualize complex historical data with ease. webpage. (2017). http://hdlab.stanford.edu/palladio/about/

[7] Eijnatten Joris van, Pieters Toine, and Verheul Jaap. 2013. Big Data for Global History: The Transformative Promise of Digital Humanities. *BMGN: Low Countries Historical Review, Vol 128, Iss 4, Pp 55-77 (2013)* 4 (2013), 55. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsdoj&AN=edsdoj.6259f58bab47404485225cd4776fcf48&site=eds-live&scope=site

[8] Editorial Board of the Programming Historian. 2017. About the Programming Historian. Website. (10 2017). https://programminghistorian.org/about

[9] Technopedia. 2017. Natural Language Processing (NLP). Webpage. (2017). https://www.techopedia.com/definition/653/natural-language-processing-nlp

[10] Sandra1 Tuppen, Stephen2 Rose, and Loukia Drosopoulou. 2016. LIBRARY CATALOGUE RECORDS AS A RESEARCH RESOURCE: INTRODUCING 'A BIG DATA HISTORY OF MUSIC'. *Fontes Artis Musicae* 63, 2 (2016), 67 – 88. http://proxyiub.uits.iu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=114128249&site=eds-live&scope=site

5 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

5.1 Assignment Submission Issues

Do not make changes to your paper during grading, when your repository should be frozen.

5.2 Uncaught Bibliography Errors

Missing bibliography file generated by JabRef

Bibtex labels cannot have any spaces, _ or & in it

Citations in text showing as [?]: this means either your report.bib is not up-to-date or there is a spelling error in the label of the item you want to cite, either in report.bib or in report.tex

5.3 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues

5.4 Writing Errors

Errors in title, e.g. capitalization

Spelling errors

Are you using *a* and *the* properly?

Do not use phrases such as *shown in the Figure below.* Instead, use *as shown in Figure 3*, when referring to the 3rd figure

Do not use the word *I* instead use *we* even if you are the sole author

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

If you want to say *and* do not use *&* but use the word *and*

Use a space after . , :

When using a section command, the section title is not written in all-caps as format does this for you

```
\section{Introduction} and NOT \section{INTRODUCTION}
```

5.5 Citation Issues and Plagiarism

It is your responsibility to make sure no plagiarism occurs. The instructions and resources were given in the class

Claims made without citations provided

Need to paraphrase long quotations (whole sentences or longer)

Need to quote directly cited material

5.6 Character Errors

Erroneous use of quotation marks, i.e. use “quotes” , instead of ” ”

To emphasize a word, use *emphasize* and not “quote”

When using the characters & # % _ put a backslash before them so that they show up correctly

Pasting and copying from the Web often results in non-ASCII characters to be used in your text, please remove them and replace accordingly. This is the case for quotes, dashes and all the other special characters.

If you see a ffigure and not a figure in text you copied from a text that has the fi combined as a single character

5.7 Structural Issues

Acknowledgement section missing

Incorrect README file

In case of a class and if you do a multi-author paper, you need to add an appendix describing who did what in the paper

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

The paper has more than 6 pages of text, i.e. excluding images, tables and figures

Do not artificially inffate your paper if you are below the page limit

5.8 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2

Do use *label* and *ref* to automatically create figure numbers

Wrong placement of figure caption. They should be on the bottom of the figure

Wrong placement of table caption. They should be on the top of the table

Images submitted incorrectly. They should be in native format, e.g. .graffle, .pptx, .png, .jpg

Do not submit eps images. Instead, convert them to PDF

The image files must be in a single directory named ”images”

In case there is a powerpoint in the submission, the image must be exported as PDF

Make the figures large enough so we can read the details. If needed make the figure over two columns

Do not worry about the figure placement if they are at a different location than you think. Figures are allowed to ffoat. For this class, you should place all figures at the end of the report.

In case you copied a figure from another paper you need to ask for copyright permission. In case of a class paper, you must include a reference to the original in the caption

Remove any figure that is not referred to explicitly in the text (As shown in Figure ..)

Do not use textwidth as a parameter for includegraphics

Figures should be reasonably sized and often you just need to add columnwidth

e.g.

```
/includegraphics[width=\columnwidth]{images/myimage.pdf}
```

re