Library Hi Tech

Managing and mining historical research data

Michael S. Seadle

Managing and mining historical research data

Michael S. Seadle
Humboldt Universität zu Berlin, Berlin, Germany

Received 3 September 2015
Revised 3 September 2015
Accepted 7 September 2015

Abstract

Purpose – The purpose of this paper is to review how historical research data are managed and mined today.
Design/methodology/approach – The methodology builds on observations over the last decade.
Findings – Reading speed is a factor in managing the quantity of text in historical research. Twenty years ago historical research involved visits to physical libraries and archives, but today much of the information is online. The granularity of reading has changed over recent decades and recognizing this change is an important factor in improving acce.
Practical implications – Computer-based humanities text mining could be simpler if publishers and libraries would manage the data in ways that facilitate the process. Some aspects still need development, including better context awareness, either by writing context awareness into programs or by encoding it in the text.
Social implications – Future researchers who want to make use of text mining and distant reading techniques will need more thorough technical training than they get today.
Originality/value – There is relatively little discussion of text mining and distant reading in the LIS literature.

Keywords Case studies, Reading, Data collection, Techniques, Data mining, Knowledge mining
Paper type Viewpoint

1. Introduction

Twenty years ago managing and mining historical research data almost exclusively involved visits to physical libraries and archives, which often meant travel to remote locations to research primary sources and rare works. The digitization of books, journals, manuscripts, and other archival materials has changed the options for historical research, and libraries and archives are still learning how to adjust. The chief issue today is not whether to make digital content available, but how to make it available in ways that facilitate research. The old paper-based paradigm remains in the consciousness of librarians, archivists, and historical researchers, but digital content lends itself to greater malleability. This paper discusses the kinds of changes that would facilitate the discovery of historical research data.

2. Reading as text mining

Reading is the traditional basis of most historical research. Libraries and archives make texts available and scholars work through them. This sounds easy because people are accustomed to the process, but the traditional approach to reading has significant problems that scholars tend to forget.

The sheer quantity of text to read has always been a problem in historical research and the problem has grown in a world in which the number of published books and journals has proliferated, and access to them has improved via better interlibrary loan and digital copies. Commentary on published works has expanded into the online world in the form of web pages, blogs, tweets, e-mails, and a host of social media opportunities for communication. The old excuse that certain material is unavailable applies less and

Library Hi Tech
Vol. 34 No. 1, 2016
pp. 172-179

0737-8831
DOI 10.1108/LHT-09-2015-0086

less, forcing scholars to make choices about potentially relevant content that they may have to ignore because they simply cannot find the time to read it all.

Reading speed is a factor in managing the quantity of text. There are people who naturally read at high speeds and people who choose to learn how to do it, and they have a striking advantage over someone who can only get through 30 or so standard pages in an hour. Doubling reading speed is certainly possible for most adults in my experience, without any loss of and often with improved comprehension, but even a high-reading speed bumps up against practical limitations, such as how long a person can reasonably sit and read in a day. Retention is an issue as well. Most people cannot remember the details they read without making some form of notes, and notes take time. Twenty years ago most reading notes were in paper and that meant that they were not searchable. Today note-taking can be done with electronic tools, even though many scholars still choose not to. Another problem is the vocabulary necessary for effective reading. A language like English is difficult because of the nuances of meaning and its tendency to absorb words from other languages, and historical research often involves multiple languages and specialized meanings. In the past a shelf of dictionaries was a common research tool. Today tools like Google Translate automate the lookup process, if not always the sense-making.

Reading is one of the most traditional ways of mining historical research data and is bumping up against serious limits. Nonetheless researchers can get around these limits with tools that to a large extent already exist and can become better as the academic community learns how to work with new techniques and to structure the content.

2.1 Granularity

The granularity of reading has changed over recent decades, and this needs to be taken into account when structuring information for contemporary access methods. There was a time fifty years ago or so when most serious historical scholarship (and humanities scholarship generally) took the form of monographs, which had a single consistent argument and which readers generally read from start to finish. Primary sources were not monograph based, but often had a monograph-like character in being primarily at one or two archives, where a scholar would work through a mass of documents from a particular place and time. Robert Darnton (2014) described this process as follows:

> After you have committed yourself to a series and started ordering documents, the first box arrives. You undo a faded ribbon on one side, fold back the cover, and pull out the top dossier. You start reading, one document after another, one folder after another, one box after another. The sequence could go on forever.

The page by page, document by document reading process has a character that is not that different than going through a book. Skipping ahead is of course possible, as with a book, but there is no knowing what might be missed because the structure is almost wholly sequential.

In the 1970s journals had reached a point of respectability where major universities in the USA could consider tenuring professors on the basis of journal publications alone. While this is still controversial in some departments and at some universities, the number of journals in general history alone has grown to 412 in JSTOR[1], which covers mainly only English language sources. DigiZeitschriften[2] lists an additional 15 in German. This suggests that scholars are reading articles as well as books. Books are also increasingly not single-author monographs, but collections of chapters by different authors.

This alters the granularity of the reading from one coherent argument built into a multi-hundred page work to separate topics and viewpoints in units of 20 or 30 pages, where the reader can reasonably skip anything that seems less interesting or urgent.

People do not necessarily even read whole articles or chapters today. Often an abstract is the only content that is freely available in the internet and scholars read only that. The pressure of so-called "information overload" leads to scanning just a few paragraphs in a work to get the gist of the contents. Amazon gives access to sample pages and the table of contents of many books. With Google search, a person can find an appropriate "snippet" that may be no more than a sentence. This is not to argue that people should read no more than a sentence or two out of a work, but there is a growing consciousness that text is no different than data, and with the Google Ngram viewer, it is possible to processes thousands (perhaps millions) of books to find the frequency of particular words or phrases. This reduction in the granularity of reading does not imply less comprehension, but rather comprehension across a much broader range of content.

2.2 Traditional content mining

Reading speed is an old concern. Evelyn Wood founded Evelyn Wood Reading Dynamics Institute in Washington, DC in 1959 and taught people to read at 1,500-6,000 words per minute instead of the usual 250-300 (Van Gelder, 1995). This was in effect a pre-computer approach to text mining that enabled people to absorb five times or more written material with equal or better comprehension. Many people were (and still are) skeptical about improving reading speed and choose other methods for mining text-based content.

One very traditional approach to mine large amounts of information quickly is to read reviews instead of the whole book or article. This works well for works by well-known authors where multiple reviews are available from high-quality sources such as the *NewYork Review of Books*, the *Times Literary Supplement*, or *Die Zeit*, but reviews are rarely available for less famous works including dissertations. In those cases readers call fall back on other traditional techniques, one of which is reading abstracts. In the US University Microfilms International (now Proquest) launched "Dissertation Abstracts" in 1951 with support from the Association of Research Libraries[3]. Abstracts had existed long before, of course, but the requirement to publish a standardized abstract influenced generations of US scholars. In March 1976 the *Journal of Modern History* began a new "Pattern of publishing articles" where some articles would be published "in the Table of Contents together with a 150-word precis written by the author." The explanation was that "this policy will allow the editors to reserve a larger part of the Journal pages for articles dealing with broad issues" (McNiell, 1976). The supplements grew less frequent in the 1980s and the last of them appeared in 1995. The precis (abstract) sufficed for readers to know what was available without having to read further.

Other traditional methods to work through a large amount of content in a short time included the use of bibliographies, which organized and classified published works in ways that provided an overview without having to get or read most of the works. The problem was that the classifications were often idiosyncratic and reflected the bibliographer's biases. Bibliographies were very common tools at one time and some still continue in digital form, such as Charles Bailey's "Digital Curation and Preservation Bibliography" (Bailey, n.d.). Following footnotes and references in books and articles are also standard ways in which scholars discover new information sources and decide whether they are useful, even if the information content is hardly more than an author and title. Human interaction at conferences is another well-established form of content

mining, not merely as a form of oral transmission, but as a forum where text-based works are discussed directly and indirectly at the social events and breaks as well as during the formal sessions.

While these traditional methods for mining information content continue to be as valid in the digital world as for print on paper, additional computer-based approaches can enable scholars to find more and more relevant materials at higher speeds. To use these digital approaches, historical research data needs structuring in ways that facilitate access. The following section describes how this works and how to make it possible.

3. Managing historical data today

Recognizing the different levels of granularity used in reading is an important factor in improving access. Some speed readers like to flip through the pages of a paper-based work to get a quick sense of the content, and find no equivalent in digital systems. This is a feature that software could easily implement in the digital environment to serve as a bridge between traditional reading and text mining. Most publishers and libraries currently structure access at the article or chapter level, and do relatively little to facilitate direct access to paragraphs, sentences, and words, except by a full-text search once an article has been located. Full-text search does not mean stripping away the context any more than an index in a physical book eliminates the surrounding paragraphs.

There are two relatively simple steps that publishers and libraries can do to make searching across large numbers of texts easier. One is to provide the texts in ASCII formats rather than in PDF. Simple ASCII is less attractive to the eye, but far more attractive to a computer program, which can process ASCII quickly without stripping away other information. HTML versions of articles and chapters mostly achieve this condition now, when available, and HTML gives some minimal structure to the content in terms of headings and paragraph breaks.

The second desirable step is for authors or publishers to introduce an enhanced XML markup that would make it clear, for example, what is quoted material, what is a date, what is a place, and what is a name. Much of this is implicit in structuring content for the semantic web. Such tags and links would help to identify the roles of specific words, sentences, and paragraphs within the context of the article: tags could also identify topic sentences, for example, or keywords or abstracts, and either tags or semantic links could identify persons and places. If RDFa[4] markup were a normal and integral part of publisher requirements, searching across large numbers of texts would become easier because the researcher could rely on the judgment of the algorithm for accurate results.

Tables benefit especially from tagging, because the individual cells may be meaningless in a full-text search and the captions for the table may be too cryptic to be meaningful. Diagrams, images, and photographs are troublesome because their formats are opaque to an ASCII text search. The quality of captions matters strongly and captions with tags to indicate the context help even more. The situation is similar with voice and video recordings, until there are cheaper and more effective tools for creating machine-readable transcripts at low cost.

If publishers would provide ASCII texts with TEI encoding and basic semantic web style markup, it would be relatively easy to write programs to do more sophisticated context-sensitive searches (e.g. this name within so many words of this other name or word or place), and it would make it easier for researcher-written programs to access multiple articles simultaneously. This leads, however, to an additional problem. Many websites automatically stop programs that access multiple texts at once. One legitimate

concern is content theft, but the restriction is true even for Project Gutenberg, which has only public domain material. An equally genuine if increasingly outdated issue is the load on a server. Server load issues are becoming less and less of a problem as chip speeds and storage access have continued to increase. The actual amount of data in even several hundred journal volumes or books is trivial, and the theft concern could be solved by enabling a mechanism (perhaps Shibboleth[5]) to establish the legitimacy of both identity and access rights.

Open access to content would make searching across platforms easier as well, but open access is not a necessary precondition as long as publishers and hosting platforms cooperate in allowing legitimate access to users from subscribing institutions. All of this is possible with current technology.

3.1 Mining humanities data today

Franco Moretti (2013) called one form of text mining "distant reading," that is, "understanding literature not by studying particular texts, but by aggregating and analyzing massive amounts of data" (Schulz, 2011). Distant reading is generally computing based, and is the opposite of close reading, where a scholar focusses on a single text. Computer-assisted close reading may be different only in the quantity of works examined, or it may have very different algorithms, depending on what scholars are looking for.

Today a significant amount of text mining is possible via Google via its book scanning projects, which the Southern District Court of New York declared to be legal in a 2013 decision by Denny Chin (Chin, 2013). A simple search in Google can in fact encompass the content of millions of published scholarly works, and the results can be presented in three line "snippets" that give a minimum of context at the sentence level. While these snippets are not generally adequate substitutes for reading a whole work, they can offer enough information to suggest whether further reading is necessary, and they may suffice for discovering very focussed kinds of information.

Publishers assist digital search capabilities by making the abstracts of articles available to web crawlers. The idea of an abstract is of course to give a summary of the contents of an article and to show a potential reader why it is worth looking at, but abstract reading as a substitute for looking at a whole article is well-established as one of the ways that humans can efficiently mine content without digital assistance. Making the abstracts available to Google merely speeds the process, because researchers can search Google to find the abstracts.

The Google Ngram Viewer[6] is a tool that allows scholars to search words and phrases throughout the whole of the Google Books corpus. It is imperfectly reliable for a number of reasons, including the uncertainty about exactly what books are part of the search and uncertainty about the algorithm itself. Nonetheless the Ngram viewer in effect searches through millions of pages and allows users to save the results. The context of the search terms is completely absent, since the point is not to find the search terms, but to discover their frequency. A classic Google Ngram search is for "the Great War, the First World War, World War I" in English language books, which shows that people began referring to the 1914-1918 war as the First World War some years before the 1939-1945 war began. The phrase "erste Weltkrieg" appears in German usage at essentially the same time. These are facts that would not be easy to determine without such a tool.

Even with pre-existing tools like Google search and the Ngram Reader, scholars often need to write programs for mining texts. Programs for mining available digital text need not be complicated. The simplest versions can work on ASCII texts that the researcher has downloaded manually, so that the routines need only search for words

or phrases sequentially across the full range of articles. Some knowledge of the contents is of course necessary to have a reasonable idea what to search for, and the search needs to take into account variants in spelling as well as capitalization, which Google searches have already built in.

The initial learning curve for distant reading is non-trivial. It takes time and effort to develop useful search tools that take the context into account, and they need testing. Programs for searching can and should be shared, but sharing also requires some degree of standardization and uniformity, which most researchers do not bother with. Text mining is not an easy solution to the work of gathering information, but once a researcher has developed a useful set of algorithms, they can be applied across larger numbers of texts than could easily be read quickly. The question is often one of priorities: does reading a few works carefully suffice, or does a subject require examining so many sources that the programming effort pays off.

3.2 Mining humanities data tomorrow

Humanities text mining in the future could be much simpler if publishers and libraries manage the data in ways that facilitate the process, as noted above, but a number of barriers are likely to interfere, not the least among them the urge to present content primarily in human-readable form, rather than in formats better designed for machine reading. Dual versions are certainly possible, and many publishers already provide both PDF and html versions that are machine-readable, even though they may require some cleanup. An XML document using a TEI DTD (text encoding initiative document type definition)[7] would do more to facilitate text mining.

Cross-publisher searching is likely to take time to get in place, even though many publishers already use platforms like Atapon[8] to provide access. Platform sharing does not mean that they want to share content, but at least it means common tools for authentication, which could make it easier for a researcher to search their content. Standardizing abstracts and standardizing metadata remains a work in progress, but the variation in both is growing smaller. Abstracts and metadata also matter less in an environment in which a computer program can read the full text equally quickly, but they play an important role in an environment in which access is limited. They are also good crutches to help slow-reading humans get to the right material.

Some features of humanities data mining still need substantial development, including better forms of context awareness. There are two ways to do this. One is by writing context awareness into programs so that they can, for example, distinguish a quote or a footnote from the core text. This is less easy than it sounds, since some journals and some authors use quotation marks, others italics, and some languages use $< >$ signs. Footnote and endnote styles vary even more. Intellectual content is even harder to program, because it is so various. Often it is easier to capture a snippet of text and allow a human to decide. Humans are efficient at this kind of processing. The other way to handle context awareness is to encode it in the text. HTML already has options for a quoted passage, as does LaTeX. It would not be hard to create encoding for $<$humor$>$, or $<$topic sentence$>$ or $<$key point$>$, but the encoding would need to be standardized, used consistently by authors, and enforced by publishers.

This is not likely to happen in the near future, although some qualitative data analysis tools in effect allow researchers to do exactly this kind of markup in, for example, interview texts. In the more distant future it is possible to imagine a form of humanities data mining in which researchers use packages that recognize a form of

universal context encoding and employ enough machine intelligence to know what researchers really want, perhaps via mechanisms such as IBM's Watson uses today. Tomorrow's researchers may experience it, if contemporary developers work to make it happen.

3.3 Curriculum issues

Enabling future researchers to make use of these text mining and distant reading techniques requires a level of technical training not common for humanities students, since they tend to avoid programming classes. Simple programs for text mining do not require a computer science degree and should be taught as "communications tools" not fundamentally different from "natural" human languages. Getting students past the initial psychological hurdles to try writing text mining programs is as essential as teaching them to read. A wide range of programming languages can be used for text mining. Python and Perl are common languages for text searches because they are relatively easy to use and work well with the sophisticated search options embodied in "regular expressions."

Computing is not the only technical skill that humanities students need today. Descriptive and inferential statistics also belong to literacy in the modern world, and humanities students need to understand at a minimum concepts like "population" and "sample" so that that they can make reasonable judgments about the statistical information that they read, and so that they can create basic statistics on their own. Today it is not necessary to do the proofs or to be able to calculate statistical results by hand. Even spreadsheets offer standard inferential tests and can certainly do ordinary descriptive statistics. The important thing for students to understand is under which conditions particular tests are valid, and what the numbers mean that they get as results.

Of course technical tools are not the only essentials. In mining for information, students must understand the social and historical context of the information sources, not merely the words surrounding a phrase or sentence. Context issues are ubiquitous. For example, a search among content from both peer-reviewed articles and blogs may not be reasonable to combine if the researcher is looking for research results, but may be combinable if the goal is to find a spectrum of opinions. Likewise information about social attitudes will likely be different for sources before 1950 and after 2000. It may seem obvious to experienced researchers to take different contexts into account, but it is not obvious to many students, whose experience is limited or involves overly simple assumptions.

4. Conclusion

Reading and research has changed fundamentally in recent decades, not merely because more information is available, but because the tools for accessing information have changed in ways that most people involved in humanities research are only slowly beginning to recognize. The change in the granularity of reading, for example, clashes with basic assumptions about how humanities scholars structure their works and their arguments. A smaller granularity often seems to devalue the logical development of arguments over hundreds of pages, and yet the fact is that readers have always searched book-length works for details, as well as reading them for their overall argument. A key difference is that in the past the tools were missing for any kind of efficient extraction of detailed information without relying on someone else's judgment of what should go into an index or what the subject classifications should be. Today the free choice is greater because the tools are more powerful, but they are powerful only

when scholars learn how to use them and when publishers and libraries structure the contents to allow the tools to work.

Notes

1. This figure comes from searching for all instances of "< td class = 'ctArticle' > Journal < /td > " in the source code of the "history" subject in JSTOR as of February 1, 2015. It does not include history of science or any of the other subjects listed in the history section.
2. www.digizeitschriften.de/en/startseite/
3. See: www.proquest.com/about/history-milestones/
4. See: http://rdfa.info/
5. See: https://shibboleth.net/
6. See: https://books.google.com/ngrams
7. See: www.tei-c.org/index.xml
8. See: www2.atypon.com/at-atypon/company.php

References

Bailey, C.W. (n.d.), "Digital curation and preservation bibliography", available at: http://digital-scholarship.org/dcpb/dcpb.htm (accessed February 3, 2015).

Chin, D. (2013), "The authors guild et al. vs Google", Southern District of NY, Chin, New York, NY, available at: www.wired.com/images_blogs/threatlevel/2013/11/chindecision.pdf

Darnton, R. (2014), *The Allure of the Archives*, New York Review.

McNiell, W.H. (1976), "A new pattern of publication", *Journal of Modern History*, Vol. 48 No. 2, p. 1.

Moretti, F. (2013), *Distant reading*, *Verso Books*.

Schulz, K. (2011), "The mechanic muse – what is distant reading?", NYTimes.com, available at: www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?pagewanted=all&_r=1& (accessed February 7, 2015).

Van Gelder, L. (1995), "Evelyn Wood, who promoted speed reading, is dead at 86", *New York Times*, August 30, available at: www.nytimes.com/1995/08/30/obituaries/evelyn-wood-who-promoted-speed-reading-is-dead-at-86.html?scp=3&sq=%22evelyn%20wood%22&st=cse (accessed February 2, 2015).

Corresponding author
Michael S. Seadle can be contacted at: seadle@ibi.hu-berlin.de