Imagine having the ability to search the entire canon of Western literature quickly and easily for the use of a specific metaphor, references to a particular place or instances of an exact sequence of words or phrases. Last year’s publication in the journal Science of research using preliminary results from Google’s book digitization project drew attention to the potential of such data mining for exploring a variety of fields in the social sciences and the humanities [1]. At its simplest, data mining is the process of extracting new knowledge (usually in terms of previously unknown patterns) from sets of data already in existence. For instance, Shakespeare scholars have used data mining techniques to identify patterns of word usage in his plays, the texts of which have already been digitized. Similarly, there is a long history of researchers making use of U.S. census data to identify demographic trends or correlations with other datasets. Data mining is inherently an exercise in quantitative analysis, the results of which are subject to qualitative analyses that link the newly discovered patterns back to particular, representative examples from the original set of data.

In the humanities, data mining necessarily entails an interdisciplinary and collaborative practice because it combines tools, techniques and methodologies from computer science and the humanities. As a consequence, data mining is often associated with the term digital humanities, which includes using cutting edge technology both to present the results of research and to conduct the research itself. Data mining is one example of the latter, and at its best a data mining project involves active collaboration between humanities scholars and information professionals to design and carry out the research program. In addition, because data mining is a relatively recent practice, the research project is often as novel for the computer and information sciences as it is for the humanities. Therefore, data mining projects are driven as much by the information professional as the humanities scholar.
One critically important aspect of data mining is negotiating the method’s relationship with traditional research practices within the humanities. Data mining is certainly a promising tool for investigation and data collection. However, data mining also has the potential to influence the formulation of research questions. That is, while projects can certainly use data mining to test pre-existing hypotheses or answer questions already raised by the researcher, humanities scholars are beginning to see the value in using data mining to develop the hypotheses or questions themselves. Returning to the definition of data mining presented above, the phenomenon of data-driven questions largely stems from the fact that data mining identifies previously unknown patterns that necessarily entail the existence of questions that, in a similar way, would have otherwise remained unasked. Therefore, data mining may join critical reading, comparative analysis and other traditional research methodologies in the humanities as tools for initiating and shaping inquiry.

Technical Components of Data Mining

Within the field of computer science, data mining constitutes part of knowledge discovery in databases (KDD) [2]. While there are several multi-stage versions of this process, the simplest is (1) pre-processing, (2) data mining and (3) results validation. As noted above, humanities scholars often simplify the pre-processing stage by making use of existing datasets. In a specific example, historian Sharon Block and computer scientist David Newman have published research based upon data mining article abstracts from widely used databases [3]. Similarly, economic historian Jeremy Atack has discussed the possibilities for researchers available in the Bateman-Foust sample, which is a database of agricultural and population census data from 1850 to 1880 first begun with machine-readable punch cards in the late 1960s and expanded in the 1990s [4]. The increased number of digitization projects for a variety of texts means that humanities scholars are discovering access to a wealth of potential sets of data.

Many different tasks make up the process of data mining itself, which authors and disciplines have categorized, labeled and prioritized differently. For data mining projects in the humanities, the following are the four most important tasks to consider:

- **N-gram Identification.** When analyzing text, data mining projects often search for instances of a given “n-gram,” a sequence of n items that can be either characters or words. For example, a project might search for instances of the word *cosmopolitan* within a set of newspaper articles from the early 1900s. A critical component of such an analysis is identifying the relevant n-grams either *a priori* or during the project. That is, data mining has the potential of revealing n-grams whose significance were previously unknown to the research team.

- **Classification.** With data mining, items within a dataset often need to be assigned to one or more classes or categories. Strictly speaking, this task is part of the pre-processing of the dataset; but in practice early results from data mining often suggest additional classification. A feedback loop is therefore created. For example, a project to classify the public speeches of President Franklin D. Roosevelt might begin with a particular categorization scheme but may discover new themes through the process of data mining.

- **Dependent Modeling.** Ultimately, the goal of most data mining projects is to identify dependency and correlation among variables (be they n-grams or classes). Standard tools of quantitative analysis can then evaluate the relative strength or weakness of discovered dependencies. Also, preliminary analysis of portions of a dataset can lead to hypotheses that the project can then test on the rest of the data. For example, a data mining project may discover that a selection of Romantic poetry exhibits a strong correlation between the use of n-gram X and theme Y and an inverse relationship with theme Z. The project may then turn to a larger sample of work to test these correlations.

- **Clustering.** Once a data mining project identifies n-grams, classes and dependencies, further analysis can reveal sub-populations of the dataset. For example, data mining a body of texts that include novels, short stories and essays may reveal clusters of themes, word usage, etc. that transcend the categories of genre and format. For example, although approaches to rendering clusters visible depend upon the specific project, common methods include the use of word clouds, depicting the relative weight of multiple dependencies visually as in a network
Diagram, and plotting a pair of variables in two-dimensions while coding the data with a third variable (for example, the number of main characters vs. the average length of monologues might reveal clusters that do or do not correspond directly to genre).

Data mining projects in the humanities can automate each of these tasks to varying degrees depending upon the skills, interests and time available to the particular members of the project team. Also, because the interdisciplinary nature of data mining necessarily involves collaboration between information professionals and humanities scholars, data mining also brings up questions about displaying and communicating data to readers with differing expectations, learning styles and levels of technical literacy. Therefore, the collaborative and interdisciplinary nature of data mining as a practical research exercise means that the discussions that take place within interdisciplinary teams while carrying out the research itself foreshadow the ways in which scholars from diverse disciplines will receive and interpret a project’s published findings. Finally, any survey of the current state of digital humanities research underscores the fact that these interdisciplinary teams frequently include humanities scholars who are pursuing non-traditional and non-tenure-track careers, such as many of the contributors to #alt-academy (http://mediacommons.futureofthebook.org/alt-ac/).

Examples of Data Mining Projects

As the research using Google Books demonstrated, the most efficient way to get a better sense of the potential of data mining is to examine the results of research projects that have made use of data mining methods. Here are some notable examples:

- Literary scholar Eric Gardner constructed a dataset of “subscribers” and “readers” of the Christian Recorder, the principal publication of the African Methodist Episcopal Church, from the lists of acknowledgments published between November 1864 and November 1865 [5]. The set of 834 items included information that allowed Gardner to interrogate the question of readership, particularly how the editors of the Christian Recorder understood the concepts of subscription, dissemination, readers and reading.

- In October 2011, Michael Witmore, the director of the Folger Shakespeare Library, presented early results of a project that data mined excerpts from Shakespeare’s First Folio using software called DocuScope [6]. Through this project, Witmore discovered evidence of variance within vocabulary and syntax among Shakespeare’s comedies, historical plays and tragedies [7]. In particular, the analysis suggests that Othello, which scholars have traditionally categorized as a tragedy, has more in common stylistically with Shakespeare’s comedies.

- Computer scientists David Elson and Kathleen McKeown and literary scholar Nicholas Dames have worked on a project that attempts to recover the social networks within 19th-century novels through data mining of the texts [8]. Their main approach is to identify instances of dialogue between characters, and in this way the project takes an existing database (texts of 19th-century fiction) and extracts a new database (conversations between characters) that is itself then mined for new patterns.

Publicly Available Tools

Most data mining projects are very open about the tools used to undertake the research. Although the outcomes are often the result of significant customization and development, software tools for data mining are available online:

- Google Ngram Viewer (http://books.google.com/ngrams/) allows the user to run searches on Google’s datasets, which are also available for download.
The Machine Learning for Language Toolkit (MALL ET) (http://mallet.cs.umass.edu/) is a downloadable piece of software for document classification, sequence tagging, topic modeling and numerical optimization of sets of textual data.

The Metadata Offer New Knowledge (MONK) Project (www.monkproject.org/) provides online access to a suite of textual analysis tools and publicly available texts from American literature and the works of Shakespeare.

The Stanford Natural Language Processing Group (http://nlp.stanford.edu/software/) has made portions of its software available for download and incorporation into applications that analyze language.

Voyeur (http://hermeneuti.ca/voeyer) is a web-based tool for studying the frequency and distribution of data within user-provided texts.

These examples demonstrate the variety of data mining tools available that perform similar functions yet vary in terms of access (via a web interface or as a download), the coding expertise necessary to make effective use of them, the inclusion of publicly available data and the ability of users to apply these tools to their own datasets.

Conclusion

Information professionals and humanities scholars interested in data mining projects should begin by establishing collaborative and interdisciplinary relationships as early as possible in the development of the project. Research interests and questions from both computer science and the humanities can drive either original or ongoing research. Such projects have the potential to re-examine existing assumptions and theories within particular fields and to develop new lines of research.

Resources Mentioned in the Article