Manual & Automated Text Mining & Network Generation: A Brief Analysis

Francis Nimick

Abstract

In performing research and generating networks from historical text, we have multiple options for network generation. We can generate networks manually after reading the text, or we can automate generation of networks through a text mining application such as AutoMap. Both of these methods have significant disadvantages, which will be briefly explained below.

Automated Network Mapping

Automated network mapping involves the generation of networks from raw plain text. The networks generated can be either simple semantic networks that link words based on frequency and proximity, or detailed meta-networks that classify items based on a user-created thesaurus.

Advantages

Automated network mapping has several advantages over manual entry methods. Networks generated through automated mapping are consistent; the same automated mapping procedure will produce identical networks when applied to identical documents in identical contexts. These consistent results, as independent as possible from human bias, provide a solid framework for statistical analysis. In addition, automated mapping scales very well in theory, as the related documents we would be working with contain similar documents and terms. After the initial document has been analyzed and the initial thesauri generated, any additional related documents can then be analyzed with minimal additional time expenditure.

Disadvantages

Despite the many advantages of automated network mapping, it is still not a viable option for large projects such as ours. The output of a text analysis is highly dependent upon the methods employed to analyze it, and these methods must be perfectly consistent if reliable results are to be obtained. This leads to a situation in which only one person can practically work on a document or set of documents at a time; if the documents are to be analyzed as a single network, the terms and representations must be consistent across all documents. In addition, even if only one person is working on a document at a time, the time cost of verifying terminology increases exponentially with document length: even a simple journal article of less than 30 pages resulted in a thesaurus of over 150 terms. As this thesaurus grows, the time required to verify the consistency of terminology throughout the document becomes prohibitive. I would estimate that at a document length of about a hundred pages, the thesaurus would become so large and the time required for verification of consistency would be so great that a continued network analysis would be practically unfeasible.

In addition, even today's most advanced text analysis software has difficulty analyzing complex and high-level text, such as that found in the Journal of Agricultural History. The software is incapable of distinguishing between words that refer to different objects, i.e. last names in context, necessitating that the user manually distinguish between them by modifying the source text. This means that a user can never simply assume that any given analysis is valid, but must first comb through the text to identify instances of errors and duplicate meanings. This adds still more time to that required to perform any basic text analysis. In addition, the software packages currently available cannot adequately resolve ambiguous pronouns,

Manual Network Mapping

Manual network mapping involves the manual creation of a network from the information contained in a document.

Advantages

Manual network mapping has several advantages over automated network mapping. Manual network mapping does not require the user to extract or process text, thus eliminating a large portion of the time required for automated mapping. In addition, manual network mapping eliminates the issues surrounding multiple representations & multiple meanings that frustrate automated mapping techniques. Manual mapping is also more time-efficient, as it eliminates the multiple steps of preprocessing and manual verification involved in automated mapping.

Manual network mapping also produces networks that are more relevant and accurate than those produced by automatic mapping. Manual mapping ensures that the only concepts represented in the network are those that are relevant to the research at hand; automated network mapping depends upon the simple textual relation of words, where manual mapping is capable of including a deeper analysis of the underlying large-scale concepts. Where automated mapping commonly generates networks that assign high importance to common terms such as "not" or "which," manually generated networks include only important pieces of information. In addition, manual network mapping eliminates the tedious step of creating meta-network thesauri. The assignment of categories to terms can be done on a case-by-case basis during the process of network generation.

Disadvantages

Manual network mapping, unfortunately, is not without its disadvantages. Manually generated networks may not be consistent between users, and therefore cannot be reliably analyzed. As the link weights assigned in a manually generated network are arbitrary, and not dependent upon any formalized structure such as the relationship or correlation between words, we cannot rely upon the link weight to accurately represent any measurable aspect of the source document. Even if we were to eliminate link weights entirely and create a binary network, we would be dogged by the analogous question: how strong need a link in the sources be in order to be formally represented in the generated network? Manual network mapping produces networks that may be more accurate to the source material, but may also be inadequate for the purposes of mathematical analysis.

A Proposed Solution – Semi-Automated Mapping

I shall now attempt to propose a solution to the problems of automated and manual mapping described above. I term this solution semi-automated network generation.

Semi-automated network generation will be a two-step process. First, a list of all relevant nodes in the document will be created, and categories will be assigned to them. Then, link weights will be determined by a controlled procedure. While further study will be needed to create a reliable procedure, a sample one would be as follows:

- No relation – link weight 0
- Knowledge of – link weight 10
- Opposition to – link weight 20
- Support of – link weight 30

The spacing of numbers is such so that there will be room for variance in the data. For example, if person A supports person B on one topic and opposes him on another, the link weight between them would be approximately 15. We could also take the frequency that the link is mentioned in the source documents to add another element of impartial analysis. For example, for each mention of the link in the source document, we could add 1 to the link weight.

While this method is far from perfect, it seems at present to be the best method for analyzing the large quantities of text involved in our project. It allows for the flexibility and relative speed of manual network generation, while providing an element of the impartial analysis of automatic network generation.