Meta-Scraping:
Two Technological Approaches to Support Meta-analyses

Kim Nimon
University of North Texas
kim.nimon@unt.edu

Cornelia Caragea
University of North Texas
Cornelia.caragea@unt.edu

Frederick L. Oswald
Rice University
fred.oswald@gmail.com

Abstract

Meta-analysis is a principled statistical approach for summarizing quantitative information reported across studies within a research domain of interest. Although the results of meta-analyses can be highly informative for taking a broad conceptual and empirical approach to an existing body of research literature, the process of collecting and coding the data for a meta-analysis is often a labor-intensive effort fraught with the potential for human error and idiosyncrasy, as researchers typically spend weeks poring over journal articles, technical reports, book chapters and other materials provided by researchers in order to retrieve key data elements that are then manually coded into some form of a spreadsheet for subsequent analyses (e.g., descriptive statistics, effect sizes, reliability estimates, demographics, study conditions). In this poster, we identify two technological solutions to support the process of collecting data for a meta-analysis.

Keywords: meta-analysis, information extraction, machine learning

Introduction

Meta-analysis is a principled statistical approach for summarizing quantitative information reported across research studies within a domain of interest (see Hedges & Olkin, 1985). Meta-analysis is important to applied research as it synthesizes data from multiple studies to estimate the magnitude, consistency, and predictors of reported statistics. In the case of social science research, validity and reliability coefficients may be meta-analyzed in order to explain variability across studies for different tests and measures (cf. Schmidt & Hunter, 1977; Vacha-Haase, 1988). In information science, meta-analysis has been used to consider the effects of users' cognitive ability on information visualization systems (cf. Chen & Yu, 2000).

Problem

Although the results of meta-analyses can be highly informative for taking a broad conceptual and empirical approach to an existing body of research literature, the process of collecting and coding the data for a meta-analysis is often a labor-intensive effort fraught with the potential for human error and idiosyncrasy, as researchers typically spend weeks poring over journal articles, technical reports, book chapters and other materials provided by researchers in order to retrieve key data elements that are then manually coded into some form of a spreadsheet for subsequent analyses (e.g., descriptive statistics, effect sizes, reliability estimates, demographics, study conditions).

Consider the case of a meta-analysis of the reliability estimates generated from the Organizational Commitment Questionnaire (OCQ). In support of the popularity of the OCQ, a Google Scholar search on the two seminal publications related to the OCQ — Mowday, Steers, and Porter, 1979; and Porter, Steers, Mowday, and Boulian, 1974 – results in 4,175 and 3,444 citations, respectively. After duplicate citations have been eliminated, the relevant article for each citation must then be imported from an electronic database or scanned in from a print journal. Then the process of examining each article to retrieve reported reliability estimates and other study parameters begins. Given the vast number of

---

Nimon, K., Caragea, C., & Oswald, F. L. (2013). Meta-scraping: Proposing techniques to support meta-analyses. *iConference 2013 Proceedings* (pp. 665-667). doi:10.9776/13311

collaborators is helpful to divide the work, it presents critical issues regarding the nature and articles to review, such a study would likely be conducted by a team of researchers. Although a team of reliability of the coding process. For example, psychometric checks for inter-coder agreement are critical. However, imagine the very real problem of critical pieces of information that were either not gathered or gathered *inconsistently* across research members. Coders can share the same blind spots or opinions of the data, where they may mistakenly agree on the wrong pieces of information that are present or absent. In other words, statistical indices of agreement will not capture all the shared omissions or mistaken consistencies in comparing codes. It would help to determine ways that improve the process and efficiency of coding information or meta-analysis.

Proposed Solutions

In this poster, we identify two solutions to support the process of collecting data for a meta-analysis. To explain these solutions concretely, we consider the aforementioned meta-analysis of the reliability coefficients for the OCQ, where each research article in the corpus of articles to analyze may contain information such as sample size, reliability estimate, and response rate. This information needs to be extracted and stored in a spreadsheet for subsequent meta-analysis. Examples of such information contained in two different research articles are "*A total of 345 questionnaires were distributed in 16 clinics to administrative and medical personnel, and 200 usable questionnaires were returned for a response rate of 58%."* (Cohen & Vigoda, 1999, p. 395), and "*One-hundred and forty-four questionnaires were distributed and 118 completed questionnaires were returned to the authors, for a response rate of 82%. Eleven questionnaires were incomplete, so data from 107 responses were used in the analysis*" (Luttman et al., 2003, p. 118). Thus, from the above text, we are interested in a tool that would extract the number of usable questionnaires and the response rate.

Machine Learning-based Solution

Here, we consider techniques for extracting structured information from semi-structured or unstructured text[1]. Specifically, we propose to employ both supervised and semi-supervised rule-based learning, commonly used in information extraction tasks (Mitchell, 1998). Rule-based learning involves manually identifying a set of rules from a document corpus and applying it to new documents to extract the desired information (i.e., via pattern matching). These techniques can make use of context information surrounding a token of interest from a text document in order to extract that particular token and store it in the appropriate field in the table (e.g., in the first italicized text above, the tokens of interest are the number 200 and the percentage 58%, which could be extracted and stored in a comma delimited file for subsequent analyses).

Based on a collection of already manually coded/annotated articles, we will build field-specific dictionaries, which comprise words or phrases that are most likely to be found or to describe a particular field of interest. For example, the dictionary specific to the sample size field could comprise words such as "usable," "questionnaires," "completed," "responses," "usable questionnaires," and "completed questionnaires." Second, to avoid the extraction of undesired information such as "*Eleven questionnaires*" in the second italicized text above, we will build "positive" and "negative" word lists for each field-specific dictionary, and analyze a window of 5 to 10 words surrounding a word in the dictionary to determine the occurrence of words in this window from both positive and negative lists. We will then define sets of rules based on the dictionaries and the word lists. For example, IF (DIGIT questionnaires, returned) THEN extract, IF (DIGIT questionnaires, incomplete) THEN not_extract.

Journal-driven Solution

In addition to extracting text of existing journal articles for quantitative information that is usable for meta-analysis, we also are urging the editors of journals to provide data in a format more usable for meta-analysis so that in the future, the extraction process and its associated guesswork is less necessary

[1] Note that we first use PDFBox, available at http://pdfbox.apache.org/, to convert PDF research articles to text.

and perhaps even unnecessary. More specifically, we are working with a small number of journal editors, who in turn are discussing with their respective publishers about the feasibility of providing tables in journal articles (e.g., correlation or covariance matrices) as supplemental material for download. We expect that each editor will have a unique set of external practical constraints and internally motivated desires, such that the nature and format of the available data across journals might vary a great deal. We would argue that these forms of diversity are to be welcomed and not feared, because providing data is certainly better than the current state of affairs where no tables are provided and humans are required for all of the coding. Diversity should be welcomed because it is a key component to healthy evolution: Different data-gathering methods and formats can be compared across journals; hybrids across these different approaches might be created; and journals who happen to be lagging in the practice of tabling and data sharing can learn from their predecessors to create new formats that (at least to them) reflect improvements.

Discussion

Although meta analysts have traditionally relied on the power of individuals to extract information from articles, technology and data sharing should be used to improve this process. We expect that combining human intelligence with machine intelligence and data sharing will improve the reliability and validity of data that is collected from meta-analyses and make this line of research available to more applied researchers. Although the current designs proposed do not remedy the issue of unpublished studies, future research should consider implementing the designs proposed and evaluate how well the solutions perform in comparison to the process of hand-coding articles. It will be interesting to determine what level of semantic ambiguity, easily understood by humans, can be accurately coded via technology.

References

Chen, C., & Yu, Y. (2000). Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies*, *53*, 831-866.

Cohen, A., Vigoda, E. (1999). Politics and the workplace: An empirical examination of the relationship between political behavior and work outcomes. *Public Productivity & Management Review*, *22*, 389-406.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning.*

Luttman, S., Mittermaier, & Rebele, J. (2003). The association of career stage and gender with tax accountants' work attitudes and behaviors. *Advances in Taxation, 15*, 111-143.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. *Proceedings of the International Conference on Machine Learning.*

Mitchell, T. M. (1997). *Machine Learning.* New York: McGraw Hill

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529-540.

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. *Proceedings of the IEEE International Conference on Data Mining.*

Tang, J., Zhang, J., Yao, L, Li, J., Zhang, L., & Su, Z (2008). Arnetminer: Extraction and mining of academic social networks. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Vacha-Haase, T. (1988). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.