Computational Triangulation: A new paradigm for Grounded Theory

Abstract

Grounded theory (GT) is a qualitative research method for building theory grounded in data. GT researchers use both qualitative and quantitative data for theory building. The GT method adopts a staged approach to coding for sensemaking from data. Machine Learning (ML) techniques, including natural language processing (NLP), can assist the researchers in the coding process. However, current applications of these techniques do not aid triangulation — combining various types and sources of data reconciling the different worldviews. We use NLP and ML to triangulate qualitative and quantitative data. We call this process *computational triangulation* (CT). We posit that our method does not alter the interpretive approach of qualitative methods. Social science researchers may face technical challenges in adopting some of these methods. We present our open-source python package (QRMine) that reduces this barrier of entry. We explain the features of QRMine and demonstrate its use with an illustrative example. We believe that CT will make GT relevant in the realm of big data.

Qualitative Research (QR) is undergoing a paradigmatic transformation with the increasing popularity of big data, machine learning and artificial intelligence (Wiedemann 2013). Qualitative research methods such as Grounded Theory (GT), though heavily reliant on data, have a predominantly interpretive world view with an emphasis on theory building (Glaser and Strauss 1967). Qualitative researchers have traditionally eschewed the computational techniques and tools that are used for objective data analysis.

Both humans and machines can recognize patterns in data (Holland et al. 1989). However, the increasing volume and heterogeneity of data can overwhelm traditional human pattern recognition (Lindberg 2020). Machines in such situations can augment human pattern discovery (Grimmer and Stewart 2013). Researchers can gain insights from multiple data sources in an abductive manner using natural language processing (NLP) and machine learning (ML). We call this process *computational triangulation* (CT) and propose methods for CT that aligns with the GT methodology. Though GT — the theory that emerges from data — and the grounded theory methodology (GTM) — the research process that facilitates theorizing — are distinct (Charmaz 2006; Glaser 1978b), we have used the term GT to represent both.

Social science researchers without a background in computer programming may find CT technically challenging. We introduce QRMine (Hidden 2019) (pronounced Karmine), our python package for CT and offer guidance on using it for analysis and triangulation using an illustrative example. QRMine is an open-source python package that wraps some of the popular NLP and ML libraries into an easy to use command-line tool. QRMine aligns with the philosophical assumptions and the traditional stages of coding in GT. The numerical ML techniques help in the CT of numerical data to corroborate emergent insights from textual data.

In the rest of this article, we first briefly describe GT and its stages of coding, followed by some similarities between GT and ML methods. Next, we explain computational methods in triangulation and how QRMine can facilitate this process. Finally, we describe QRMine's design and an illustrative use case.

Grounded Theory

GT is a qualitative research method with an emphasis on "generating theory grounded in data that has been systematically collected and analysed" (Strauss and Corbin 1990). Some of the characteristics of GT that differentiates it from other qualitative methods are constant comparison by simultaneous collection and analysis of data, theoretical sampling (Glaser and Strauss 1970) based on the emergent theory, and its emphasis on the 'theoretical sensitivity' (Glaser 1978b) — the insights of the researcher. GT aims to converge towards a theory that adequately explains the phenomenon under study.

GT offers precise techniques and procedures for analyzing data and calls for avoiding apriori theory or pre-existing codebooks. GT has been gaining traction due to the recent emphasis on data. The interpretive and subjective nature of GT makes analysis non-reproducible, which is considered by some researchers as its weakness. We posit that the emerging ML methods align with GT's abductive nature (Richardson and Kramer 2006), and their potential applications extend beyond content analysis. Human analysis augmented by ML can improve the validity of GT.

Coding in GT

Coding is the process of tagging various types of data to define what each segment of the data is about, to help in sensemaking and to converge towards analytic