7. AMEENA K NAZEER

An Unsupervised Approach of Truth Discovery from Multi-Sourced Text Data

Abstract

Truth discovery methods aim to identify which piece of information is trustworthy from multi-sourced data. Most existing truth discovery methods, however, are designed for structured data and fail to meet the strong need to extract trustworthy information from raw text data. More specifically, existing methods ignore the semantic information of text answers, i.e., answers may contain multiple factors, the word usages may be diverse, and the answers may be partially correct. In addition, ubiquitous long-tail phenomenon exists in the tasks, i.e., most users provide only a few answers and only a few users provide plenty of answers, which causes the user reliability estimation for small users to be unreasonable. To tackle these challenges, we propose a Graph Convolutional Network (GCN) based truth discovery model to automatically discover trustworthy information from text data. Firstly, Smooth Inverse Frequency (SIF) is utilized to learn real-valued vector representations for answers. Then, we construct undirected graph with these vectors to capture the structural information of answers. Finally, the GCN is utilized to store and update the reliability of these answers, and sums up all the feature vectors of all neighboring answers to improve the accuracy and efficiency of truth discovery. Different from traditional methods, we use vectors to store the reliability of answers which have higher representation capability compared with real numbers, and network is utilized to capture complex relationships among answers rather than simplified functions. The experiment results on real datasets show that though text data structures are complex, our model can still find reliable answers compared with retrieval-based and state-of-the-art truth discovery methods.