Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Ubiquitous Computing Systems (Keiichi Yasumoto (Professor))		
Student ID	2411264	Submission date	2026 / 1 / 22
Name	松本 一晟		
Thesis title	A Framework for Automatically Generating Dataset Profiles from Raw Data Using Large Language Models		

Abstract

The rapid growth of data across diverse domains has created significant opportunities for data-driven decision-making; however, many datasets remain underutilized because effectively understanding and applying them requires specialized technical expertise. While recent advances in artificial intelligence have enabled natural language summaries to improve dataset interpretability, existing approaches primarily focus on localized patterns or statistical relationships and provide limited support for understanding how entire datasets can be applied to concrete tasks.

To address this gap, this research focuses on automated dataset profile generation, defined as the task of producing structured natural language descriptions that summarize a dataset's content, structure, quality, and potential use cases. Building on this concept, we propose and develop an automated dataset profile generation framework that takes raw datasets as input and outputs interpretable use-case descriptions. The framework uses all column names and a limited number of rows to extract structured metadata summarizing the dataset's domain, structure, content, and characteristics. Based on this metadata, large language models (LLMs) generate multiple candidate dataset profiles describing possible dataset applications. These profiles are then automatically evaluated and ranked according to relevance and coherence using few-shot prompting, enabling the framework to recommend profiles that are most likely to be useful and interpretable.

To evaluate the proposed framework, both quantitative and qualitative analyses were conducted. The quantitative evaluation assessed profile quality across four uni-modal and multi-modal datasets by measuring relevance and coherence and analyzing alignment between automatic evaluation scores and human judgments using rank-based correlation metrics. The results show that relevance and coherence exhibit the strongest alignment with human judgment, with Spearman's ρ reaching up to 0.76 and Kendall's τ up to 0.66 for relevance, and $\rho = 0.65$ and $\tau = 0.52$ for coherence, indicating that these criteria serve as effective signals for automatic evaluation.

In addition, a qualitative, human-centered user study was conducted through semi-structured interviews with ten participants, who evaluated dataset profiles generated for both uni-modal and multi-modal datasets. The qualitative analysis examined how users assess dataset profiles in practice, identified evaluation criteria not captured by automatic metrics, and explored the framework's practical usefulness in real-world data exploration scenarios. The findings reveal that users place greater emphasis on criteria such as specificity, interpretability, and trustworthiness, highlighting limitations of current automatic evaluation methods and the need for more human-centered assessment approaches.