# Literature Review Report: Discovering Similarity Inclusion Dependencies

## Introduction

At the beginning of the paper, the authors describe the problem with the Inclusion Dependencies (INDs), which needs to be based on clean data.

In order to solve the problem, the authors add similarity measure to the INDs, which is called sINDs.

This literature review aims to provide an overview and analysis of the paper, evaluate its contributions, and discuss its implications in the field of data mining and database management.

## Summary of the Paper

The paper formalized the concept of SINDs and their significance in data analysis. It highlights the limitations of traditional functional dependencies in capturing relationships between similar values and emphasizes the need for a more robust approach. The authors then present their proposed method (Sawfish), which involves defining similarity measures and utilizing them to identify SINDs. They describe the algorithmic steps involved in discovering SINDs and provide a detailed explanation of each step. Sawfish finds all unary sINDs based on the edit-distance and the Jaccard similarity measure. It combines approaches of traditional IND discovery and string similarity joins with a novel sliding-window approach and lazy candidate validation. The paper concludes with an evaluation of the proposed method using real-world datasets and a comparison with existing approaches.

## Analysis and Critique

The presentation and organization of the paper are clear and logical, making it easy for readers to follow the proposed methodology and understand the experimental results. The authors provide sufficient background information on SIDs and related work, allowing readers to grasp the context and significance of the research.

The paper makes several significant contributions to the field of data mining and database management. Firstly, it introduces the concept of SINDs, which extends the traditional notion of functional dependencies and provides a more comprehensive representation of relationships between similar values. This extension is particularly valuable in domains where similarity plays a crucial role, such as text mining or image recognition. Secondly, the authors propose a novel method for discovering SINDs that leverages similarity measures. This approach not only improves the accuracy of SIND discovery but also reduces computational complexity compared to existing methods. The paper also provides a comprehensive evaluation of the proposed method, demonstrating its effectiveness and efficiency on real-world datasets.

However, it would have been beneficial to include a more detailed discussion on the limitations and potential challenges of the proposed algorithm. The proposed method heavily relies on the definition and selection of similarity measures, which can be subjective and domain-specific. Future research should focus on developing standardized similarity measures that can be applied across different domains. Additionally, a comparison with state-of-the-art methods for SID discovery would have further strengthened the paper's contribution.the evaluation of the proposed method could be further strengthened by comparing it with a wider range of existing approaches and conducting experiments on larger datasets.

## Conclusion

Overall, the paper presents a novel approach to discovering SINDs in relational databases. The proposed method leverages similarity measures to identify and extract SINDs, providing a more comprehensive representation of relationships between similar values.

The paper's contributions include the introduction of SINDs, the development of an efficient algorithm for their discovery, and a comprehensive evaluation of the proposed method. The findings have implications for both research and practice, opening up new avenues for exploring the relationships between similar values in databases and enhancing data quality in various industries.The authors' proposed algorithm for discovering SINDs is effective and efficient, and has the potential to improve data quality and accuracy in a variety of applications.

However, further research is needed to explore the full potential of SINDs and to develop more advanced algorithms for discovering them.