Improving Data Excellence Reliability and Precision of Functional Dependency

Tiwari Pushpa[1], K. Dhanasree[2], Dr.R.V.Krishnaiah[3]

[1]M.Tech Student, [2]Associate Professor, [3]PG Coordinator

[1]Dept of CSE, [2]Dept of CSIT, [3]Dept of CSE

[1,2]DRK Institute of Science & Technology, Hyderabad,

Andhra Pradesh, India

[3]DRK Group of Institutions, Hyderabad, Andhra Pradesh, India

ABSTRACT:

Poor quality data is a rising and expensive problem that affects many enterprises across all aspects of their business ranging from operational effectiveness to revenue protection. A novel type of semantic rules extended from traditional functional dependencies is proposed as Conditional Functional Dependencies (CFDs). In this paper, for detecting inconsistencies in data, we present an approach that efficiently and robustly discovers conditional functional dependencies and improves data quality. An expensive process that involves intensive manual effort is to find the quality of CFDs. We develop techniques for discovering CFDs from relations to effectively identify data cleaning rules. The discovery problem is more difficult for CFDs and indeed, mining patterns in CFDs introduces new challenges. For discovering general CFDs two algorithms are developed they are the First is a level wise algorithm that extends TANE, a well-known algorithm for mining FDs. And the second algorithm is a method for discovering FDs which is based on the depth-first approach used in Fast FD. CFD Miner efficiently discovers constant CFDs as verified by our experimental study. In general as CFDs, it does not scale well with the arity of the relation as CTANE works well when a given relation is large. A set of cleaning-rule discovery tools are provided by these two algorithms for the users to choose for different applications.

Keywords: ***Conditional Functional Dependencies, CFD Miner, CTANE, Functional Dependencies.***

1. INTRODUCTION:

The main problem that is getting worse in many organizations is that they are suffering from poor quality data because data is growing at astonishing rates and few organizations have an effective data governance process [1] [3]. Poor data quality can occur along several dimensions and consistency is one dimension that many Organizations struggle with. For data cleaning, Conditional functional dependencies (CFDs) were introduced recently [2] [4] [6]. And b y enforcing patterns of semantically related constants the CDFs extend standard functional dependencies (FDs). In detecting and repairing inconsistencies of data, CFDs have been proven more effective than FDs and are expected to be adopted by data-cleaning tools that currently employ standard FDs [5] [7] [8]. However, it is necessary to have techniques in place that can automatically discover or learn CFDs from sample data for CFD-based cleaning methods to be effective in practice, and to be used as data cleaning rules. Indeed, to design CFDs it is often unrealistic to rely solely on human experts and by using an expensive and long manual process [9] [10].

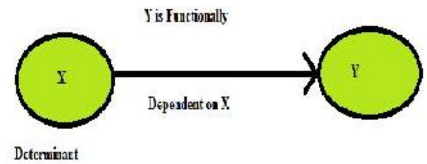

Fig 1: Functional Dependency between X and Y

As indicated to commercial data quality tools, cleaning-rule discovery is critical. Each CFD in the canonical cover should be minimal, i.e., nontrivial and left-reduced to reduce redundancy [11] [13]. The discovery problem is, however, highly nontrivial. Moreover, with constants CFD discovery requires mining of semantic patterns, a challenge that was not encountered when discovering FDs [12] [14] [15]. The possible applications of CFDs in data cleaning show up the need for further investigations of CFD discovery. First constant CFDs are particularly importantfor object identification as remarked earlier, and thus deserve a separate treatment. Without paying the price of discovering all CFDs, one wants efficient methods to discover constant CFDs [16] [19]. Second is a Level wise algorithm which may not perform well on sample relations of large arty, given their inbuilt exponential difficulty. To deal with datasets more effective methods have to be in place with a large arity. Third for association rule mining a host of techniques have been developed, and it is only natural to capitalize on these for CFD discovery [17] [18] [20]. These techniques can not readily be used in constant CFD discovery, but also considerably speed up general CFD discovery.

2. CONDITIONAL DEPENDENCIES RELATED WORK:

Conditional dependencies has mostly focused on the reliability and allegation analyses of CFDs, and repairing methods to localize and fix errors detected by CFDs, propagation of CFDs from source data to views in data integration, extensions of CFDs by adding disjunction and negation or adding ranges, confidence of CFDs, as well as extensions of inclusion dependencies with conditions. CFD discovery was only studied to our knowledge and the previous work assumes that CFDs are already designed and provided. There has been a host of work on minimal FD discovery. However, minimal CFDs are more involved than their FD counterparts: they require both the minimality of attributes and the minimality of patterns. Our algorithms CTANE and Fast CFD extend TANE and Fast FD are respectively used, for discovering minimal CFDs. For a fixed traditional FD, proposed criteria for sensible patterns that, together with the FD, make useful CFDs. It showed that the problem of finding such patterns is NP-complete, and developed efficient heuristic algorithms for discovering patterns from samples. In contrast to, this work studies CFD discovery when the embedded traditional FDs are not given. An algorithm for discovering CFDs is developed, and which aims to find both traditional FDs and patterns in CFDs, the same as what this work does. In contrast, the algorithms of this work are developed to discover minimal k-frequent CFDs. The connection between association rule mining and constant CFD discovery is also observed.

3. RESULTS:

conditional functional dependencies (CFD)Miner only mines stable CFDs which is multiple orders of magnitude faster than the other algorithms which determines both constant and variable CFDs. When database size becomes superior, there are more item sets with large support that require to be considered for constructing the difference sets. This results in a significant performance degradation of NaiveFast. when DBSIZE is less than one million tuples which is reasonably large, then FastCFD outperforms CTANE and NaiveFast. This can confirm the effectiveness of the optimization by leveraging the closed- item sets.

4. CONCLUSION:

For discovering minimal CFDs we have developed and implemented three algorithms: First, CFD Miner for mining minimal constant CFDs, for both data cleaning and data integration a class of CFDs are important; Second is CTANE for discovering general minimal CFDs based on the level wise approach and third is Fast CFD for discovering general minimal CFDs based on a depth-first search strategy, and a novel optimization technique via closed-item set mining. These results provide a set of tools for users to choose for different applications. One can simply use CFD Miner without paying the price of mining general CFDs when only constant CFDs are needed. One should opt for Fast CFD when the arity of a sample dataset is large. One could use CTANE when k-frequent CFDs are needed for a large k and there is naturally much to be done. While we have employed in Fast CFD techniques for mining closed item sets, we expect that other mining techniques may also shed light in improving the performance of discovery algorithms. And we plan to explore the use of CFD inference in discovery, to eliminate CFDs that are entailed by those CFDs already found.

REFERENCES:

[1] W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, “Conditional functional dependencies for capturing data inconsistencies,” *TODS*, vol. 33, no. 2, june 2008.

[2] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, “Improving data quality: Consistency and accuracy,” in *VLDB*, 2007.

[3] M. Arenas, L. E. Bertossi, and J. Chomicki, “Consistent query answers in inconsistent databases,” *TPLP*, vol. 3, no. 4-5, pp. 393–424, 2003.

[4] J. Chomicki and J. Marcinkowski, “Minimal-change integrity maintenance using tuple deletions,” *Information and Computation*, vol. 197, no. 1-2, pp. 90–121, 2005.

[5] J. Wijsen, “Database repairing using updates,” *TODS*, vol. 30, no. 3, pp. 722–768, 2005.

[6] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006.

[7] E. Rahm and H. H. Do, “Data cleaning: Problems and current approaches.” *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[8] Gartner, “Forecast: Data quality tools, worldwide, 2006-2011,” 2007.

[9] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*. Addison-Wesley, 1995.

[10] L. Golab, H. Karloff, F. Korn, D. Srivastava, and B. Yu, “On generating near-optimal tableaux for conditional functional dependencies,” in *VLDB*, 2008.

[11] E.-P. Lim, J. Srivastava, S. Prabhakar, and J. Richardson, “Entity identification in database integration,” *Inf. Sci.*, vol. 89, no. 1-2, pp. 1–38, 1996.

[12] H. Mannila and K.-J. R¨aih¨a, "Dependency inference," in *VLDB*, 1987.

[13] Y. Huhtala, J. K¨arkk ainen, P. Porkka, and H. Toivonen, "TANE: An efficient algorithm for discovering functional and approximate dependencies," *Comput. J.*, vol. 42, no. 2, pp. 100–111, 1999.

[14] C. M. Wyss, C. Giannella, and E. L. Robertson, "FastFDs: A heuristicdriven, depth-first algorithm for mining functional dependencies from relation instances - extended abstract," in *DaWak*, 2001.

[15] P. A. Flach and I. Savnik, "Database dependency discovery: A machine learning approach," *AI Commun.*, vol. 12, no. 3, pp. 139–160, 1999.

[16] S. Lopes, J.-M. Petit, and L. Lakhal, "Efficient discovery of functional dependencies and armstrong relations," in *EDBT*, 2000.

[17] T. Calders, R. T. Ng, and J. Wijsen, "Searching for dependencies at multiple abstraction levels," *TODS*, vol. 27, no. 3, pp. 229–260, 2003.

[18] R. S. King and J. J. Legendre, "Discovery of functional and approximate functional dependencies in relational databases," *JAMDS*, vol. 7, no. 1, pp. 49–59, 2003.

[19] I. F. Ilyas, V. Markl, P. J. Haas, P. Brown, and A. Aboulnaga, "Cords: Automatic discovery of correlations and soft functional dependencies," in *SIGMOD*, 2004.

[20] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 259–289, 1997.

BIOGRAPHY:

TIWARI PUSHPA has completed MCA from Maharaja Sayajirao University, Baroda, and pursuing M.Tech (C.S.E) in DRK Institute of Science and Technology, JNTUH, Hyderabad, Andhra Pradesh, India. Her main research interest includes Data Mining & Databases.

K. Dhanasree M.Tech, (Ph.D) Associate Professor Dept.Of CSIT, DRK Institute of Science and Technology, JNTU, Hyderabad. Research Area Datamining, Computer Networks And Securtiy Computing. Previously Published 3 International Journal Papers And 1 International Conference Paper.

Dr.R.V.Krishnaiah, did M.Tech (EIE) from NIT Waranagal, MTech(CSE) form JNTU, ,Ph.D, from JNTU Ananthapur, He has memberships in professional bodies MIE, MIETE, MISTE. He is working as Principal in DRK Institute of Science and Technology, Hyderabad. His main research interests include Image Processing, Security systems, Sensors, Intelligent Systems, Computer networks, Data mining, Software Engineering, network protection and security control. He has various publications and presentations in various national and international journals.