ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2015

Strategies Developed for Effective Discovery of Dependencies: A State-of-the-Art Survey

R.Santhya[1], S.Latha[2], Prof.S.Balamurugan[3], S.Charanyaa[4]

Department of IT, Kalaignar Karunanidhi Institute of Technology, Coimbatore, TamilNadu, India[1,2,3]

Senior Software Engineer Mainframe Technologies Former, Larsen & Tubro (L&T) Infotech, Chennai, TamilNadu, India[4]

ABSTRACT: This paper details about various methods prevailing in literature for efficient discovery of matching dependencies. The concept of matching dependencies (MDs) has recently been proposed for specifying matching rules for object identification. Similar to the functional dependencies with conditions, MDs can also be applied to various data quality applications such as detecting the violations of integrity constraints. The problem of discovering similarity constraints for matching dependencies from a given database instance is taken into consideration. This survey would promote a lot of research in the area of information mining.

KEYWORDS: Data Anonymization, Matching Dependencies(MDs), Object, Similarity Constraints, Information Mining.

I. Introduction

Need for publishing sensitive data to public has grown extravagantly during recent years. Recent days have seen a steep rise in preserving data quality in the database community due to the huge amount of "dirty" data originated from different. These data often contain duplicates, inconsistencies and conflicts, due to various mistakes of men and machines. In addition to the cost of dealing with the huge volume of data, manually detecting and removing "dirty" data is definitely out of practice because human proposed cleaning methods may introduce inconsistencies again. Therefore, data dependencies, which have been widely used in the relational database design to set up the integrity constraints. Hence protecting privacy of individuals and ensuring utility of social network data as well becomes a challenging and interesting research topic.. In this paper we have made an investigation on the attacks by matching dependencies and possible solutions proposed in literature and efficiency of the same.

II. Approximate inference of functional dependencies from relations

In this paper , the author describes the FD inference problem . The FD inferences problem states that , in this given relation 'r' find the set of FD and that is equivalent to the set of all FD holding in r . So the approximate dependency inference is taken over measures the error in a relation . These error value is 0 if the dependency holds and the value is 1 if the dependency don't hold .

During the database design conditions of integrity constraints defines the what database states are allowed . These exist in several classes of dependencies .So the functional dependencies is one of the most important in that class . In this paper only FD are considered and call them just dependencies .

In this paper the another dependency approximate dependency inference is considered . Where the result no need to be accurate . So this paper contains the two different types of results . The problem of inferring the functional dependencies that hold in a given relation 'r' .First shows the three measures of dependency .Secondly demonstrated the output polynomial algorithm with any accuracy . In covers the set of FD that hold in a given relation .The result shows the approximate techniques to achieve the good results in the dependency inference problem.

10.15680/ijircce.2015.0301051

III. METRIC FUNCTIONAL DEPENDENCIES

In the paper the author describes the metric functional dependencies problems while merging the data from differed sources then it will be a small difference in the data format . This will causes the traditional FDs , to be violated , without there being an any of semantics

FDs that defines the functional relationship between the attributes.In FD key relationship are very special kind of the FDs and these will provide database normalization while processing the design. Conditional function dependencies as well as approximation might not giving the exact result while inherent lack of robustness. So to over come these problems the MFDs are introduced .These will used to capture the small variation in the data.

In this paper the exact algorithms are specified to verify the the MFDs .specifically for general metrics as well as Euclidean distance space.

Dom(x) is the domain of an attribute where X is the sequence of attributes X=A1,A2……Ak, then dom(X) = dom(A1) *dom(A2) *….. *dom(Ak).

In this paper, the problem dealing with the robust to data failure and errors. So we introduce the metric FD.The result shows sound and realistic.

IV. DISCOVERY OF FUNCTIONAL AND FUNCTIONAL DEPENDENCIES IN RELATIONAL DATABASE

In this paper describes the study of developing the foundation, efficient methods for approximate functional dependence in the given relational database and this is based on the mathematical theory of partition .The minimal non-trivial functional dependencies can be found using the level wise algorithm. The FD defines the relationship between the attributes of a database in the relation. It states that attribute value is uniquely identify by the some other attribute values

In this paper, the new algorithmic approach is found for the discovery of functional and approximate functional dependencies. This approach is based on the partitions of the rows identification number from the relation and the breadth first or level wise searches are conducted So the partitions and dependencies can be evaluated efficiently

V. IMPROVING DATA QUALITY THROUGH EFFECTIVE USE OF DATA SEMANTICS

In this paper, the author shows the problem of data quality issues. It is the increasing and important problem in the recent year. So the discovery of the many “data quality” or “data misinterpretation” problem i.e problem with data semantics considered in the paper. The COIN(COnetext and INterchange)technology for knowledge storage and knowledge processing approaches are proposed.

COIN is a knowledge based mediation technology .This will enables meaningful use of the heterogeneous database. This COIN is not only for mediation also for wrapping technology and middle ware services. The wrapping is physical and logical gateway to provide the uniform access to the disparate sources over the network.

In this paper , the framework for understanding house holding problem is presented .The COIN techniques are used in this paper to store and apply the capture knowledge .The future work is to collect the data and to determine the types of corporate house holding knowledge. Secondly to explore the COIN techniques in the corporate house holding and to extend the COIN techniques for capturing, storing, maintaining and applying the house holding knowledge.

VI. AUTOMATIC DISCOVERY OF CORRELATION AND SOFT FUNCTIONAL DEPENDENCIES

In this paper we introduce CORDS, an efficient tool for automatic discovery of correlation and soft FD between column. CORDS searches for column that may useful dependency relation by candidate pair and flexible set of heuristic are used by the pruning unpromising candidates. The CORDS can be used as a data mining tool, producing dependency graphs .So we focus on the use of CORDS in query optimization generally. This approach is relatively easy to implement. CORDs can be used in tandem with query feedback system such as the LEO learning optimization.etc.,.

VII. EFFICIENT DISCOVERY OF FUNCTIONAL AND APPROXIMATE DEPENDENCIES USING PARTITION

In this paper the author gives the new approach for finding functional dependencies based on the partitioning the set of rows with respect to their attribute values. These partition makes easy and very efficient and rows are identified easily. The efficient in practice is a new algorithm are used in the experiments. The running time is improved by the several order of magnitude over previous published results . These will applicable for larger database also.

VIII. AN EFFICIENT ALGORITHM FOR DISCOVERING FUNCTIONAL AND APPROXIMATE DEPENDENCIES

In this paper the author defines the discovery of functional dependencies. It is an important database analysis techniques .Tane ,an efficient algorithm for finding functional dependencies from large database. Tane is based on the partitioning the rows, this will makes the vality of FD.T his partitions will makes the discovery of FD more easy and efficient. For benchmark database the running times are improved by several order of magnitude over the previous paper. So this algorithm is applicable for large dataset also.

IX. AN ALGORITHM FOR INFERRING FUNCTIONAL DEPENDENCIES FROM RELATION

In this paper the author describes the dependency inference problem. It is used to find the set of FD that will hold in a given database relations. The problem is exponential in the number of attributes and application database design, in query optimization, in artificial intelligence So we develop the two algorithm one is reduce the problem of computing the transversal of a hypergraph. The another one is based on the repeatedly sorting the relation with respect to set of attribute.

X. CONCLUSION AND FUTURE WORK

This paper detailed about various methods prevailing in literature for efficient discovery of matching dependencies. The concept of matching dependencies (MDs) has recently been proposed for specifying matching rules for object identification. Similar to the functional dependencies (with conditions), MDs can also be applied to various data quality applications such as detecting the violations of integrity constraints. The problem of discovering similarity constraints for matching dependencies from a given database instance is taken into consideration. This survey would promote a lot of research in the area of information mining.

REFERENCES

1. Shaoxu Song, Lei Chen, "Efficient discovery of similarity constraints for matching dependencies", Data & Knowledge Engineering, Elsevier, 2013.
2. S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley, 1995.
3. R. Agrawal, T. Imielinski, A.N. Swami, Mining association rules between sets of items in large databases, SIGMOD Conference, 1993, pp. 207–216.
4. R. Bassée, J. Wijsen, Neighborhood dependencies for prediction, PAKDD, 2001, pp. 562–567.
5. C. Batini, M. Scannapieco, Data quality: concepts, methodologies and techniques, Data-Centric Systems and Applications, Springer, 2006.
6. L.E. Bertossi, S. Kolahi, L.V.S. Lakshmanan, Data cleaning and query answering with matching dependencies and matching functions, ICDT, 2011, pp. 268–279.
7. M. Bilenko, R.J. Mooney, W.W. Cohen, P. Ravikumar, S.E. Fienberg, Adaptive name matching in information integration, IEEE Intelligent Systems 18 (5) (2003) 16–23.
8. D. Bitton, J. Millman, S. Torgersen, A feasibility and performance study of dependency inference, ICDE, 1989, pp. 635–641.
9. P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, Conditional functional dependencies for data cleaning, ICDE, 2007, pp. 746–755.
10. L. Bravo, W. Fan, F. Geerts, S. Ma, Increasing the expressivity of conditional functional dependencies without extra complexity, ICDE, 2008, pp. 516–525.
11. L. Bravo, W. Fan, S. Ma, Extending dependencies with conditions, VLDB, 2007, pp. 243–254.
12. T. Calders, R.T. Ng, J. Wijsen, Searching for dependencies at multiple abstraction levels, ACM Transactions on Database Systems 27 (3) (2002) 229–260.
13. F. Chiang, R.J. Miller, Discovering data quality rules, PVLDB 1 (1) (2008) 1166–1177.
14. W.W. Cohen, Integration of heterogeneous databases without common domains using queries based on textual similarity, SIGMOD Conference, 1998, pp. 201–212.
15. G. Cong, W. Fan, F. Geerts, X. Jia, S. Ma, Improving data quality: consistency and accuracy, VLDB, 2007, pp. 315–326.
16. A.K. Elmagarmid, P.G. Ipeirotis, V.S. Verykios, Duplicate record detection: a survey, IEEE Transactions on Knowledge and Data Engineering 19 (1) (2007) 1–16.

17. W. Fan, Dependencies revisited for improving data quality, PODS, 2008, pp. 159–170.
18. W. Fan, H. Gao, X. Jia, J. Li, S. Ma, Dynamic constraints for record matching, The VLDB Journal (2010) 1–26.
19. W. Fan, F. Geerts, L.V.S. Lakshmanan, M. Xiong, Discovering conditional functional dependencies, ICDE, 2009, pp. 1231–1234.
20. W. Fan, J. Li, X. Jia, S. Ma, Reasoning about record matching rules, PVLDB, 2009.
21. W. Fan, J. Li, S. Ma, N. Tang, W. Yu, Interaction between record matching and data repairing, SIGMOD Conference, 2011, pp. 469–480.
22. W. Fan, S. Ma, Y. Hu, J. Liu, Y. Wu, Propagating functional dependencies with conditions, PVLDB 1 (1) (2008) 391–407.
23. P.A. Flach, I. Savnik, Database dependency discovery: a machine learning approach, AI Communications 12 (3) (1999) 139–160.
24. J. Gardezi, L.E. Bertossi, I. Kiringa, Matching dependencies with arbitrary attribute values: semantics, query answering and integrity constraints, LID, 2011, pp. 23–30.
25. C. Giannella, E.L. Robertson, On approximation measures for functional dependencies, Information Systems 29 (6) (2004) 483–507.
26. L. Golab, H.J. Karloff, F. Korn, A. Saha, D. Srivastava, Sequential dependencies, PVLDB 2 (1) (2009) 574–585.
27. L. Golab, H.J. Karloff, F. Korn, D. Srivastava, B. Yu, On generating near-optimal tableaux for conditional functional dependencies, PVLDB 1 (1) (2008) 376–390.
28. L. Gravano, P.G. Ipeirotis, N. Koudas, D. Srivastava, Text joins in an rdbms for web data integration, WWW, 2003, pp. 90–101.
29. Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen, Efficient discovery of functional and approximate dependencies using partitions, ICDE, 1998, pp. 392–401.
30. Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen, Tane: an efficient algorithm for discovering functional and approximate dependencies, The Computer
31. I.F. Ilyas, V. Markl, P.J. Haas, P. Brown, A. Aboulnaga, Cords: automatic discovery of correlations and soft functional dependencies, SIGMOD Conference, 2004, pp. 647–658.
32. R.S. King, J.J. Legendre, Discovery of functional and approximate functional dependencies in relational databases, JAMDS 7 (1) (2003) 49–59.
33. J. Kivinen, H. Mannila, Approximate inference of functional dependencies from relations, Theoretical Computer Science 149 (1) (1995) 129–149.
34. N. Koudas, A. Saha, D. Srivastava, S. Venkatasubramanian, Metric functional dependencies, ICDE, 2009, pp. 1275–1278.
35. S. Kramer, B. Pfahringer, Efficient search for strong partial determinations, KDD, 1996, pp. 371–374.
36. S.E. Madnick, H. Zhu, Improving data quality through effective use of data semantics, Data & Knowledge Engineering 59 (2) (2006) 460–475.
37. H. Mannila, K.-J. Räihä, Design of Relational Databases, Addison-Wesley, 1992.
38. H. Mannila, K.-J. Räihä, Algorithms for inferring functional dependencies from relations, Data & Knowledge Engineering 12 (1) (1994) 83–99.
39. A. McCallum, K. Nigam, L.H. Ungar, Efficient clustering of high-dimensional data sets with application to reference matching, KDD, 2000, pp. 169–178.
40. G. Navarro, A guided tour to approximate string matching, ACM Computing Surveys 33 (1) (2001) 31–88.
41. B. Pfahringer, S. Kramer, Compression-based evaluation of partial determinations, KDD, 1995, pp. 234–239.
42. T. Scheffer, Finding association rules that trade support optimally against confidence, Intelligent Data Analysis 9 (4) (2005) 381–395.
43. J.C. Schlimmer, Efficiently inducing determinations: a complete and systematic search algorithm that uses optimal pruning, ICML, 1993, pp. 284–290.
44. S. Song, L. Chen, Discovering matching dependencies, CIKM, 2009, pp. 1421–1424.
45. S. Song, L. Chen, Differential dependencies: reasoning and discovery, ACM Transactions on Database Systems 36 (4) (2011).
46. S. Song, L. Chen, P.S. Yu, On data dependencies in dataspaces, ICDE, 2011, pp. 470–481.
47. U. ul Hassan, S. O'Riain, E. Curry, Leveraging matching dependencies for guided user feedback in linked data applications, Proceedings of the Ninth International Workshop on Information Integration on the Web, IIWeb '12, ACM, New York, NY, USA, 2012, pp. 5:1–5:6.
48. H. Wang, R. Liu, Privacy-preserving publishing microdata with full functional dependencies, Data & Knowledge Engineering 70 (3) (2011) 249–268.
49. C.M. Wyss, C. Giannella, E.L. Robertson, Fastfds: a heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances —extended abstract, DaWaK, 2001, pp. 101–110.
50. B.Powmeya , Nikita Mary Ablett ,V.Mohanapriya,S.Balamurugan,"An Object Oriented approach to Model the secure Health care Database systems,"In proceedings of International conference on computer , communication & signal processing(IC^{3}SP)in association with IETE students forum and the society of digital information and wireless communication,SDIWC,2011,pp.2-3
51. Balamurugan Shanmugam, Visalakshi Palaniswami, "Modified Partitioning Algorithm for Privacy Preservation in Microdata Publishing with Full Functional Dependencies", Australian Journal of Basic and Applied Sciences, 7(8): pp.316-323, July 2013
52. Balamurugan Shanmugam, Visalakshi Palaniswami, R.Santhya, R.S.Venkatesh "Strategies for Privacy Preserving Publishing of Functionally Dependent Sensitive Data: A State-of-the-Art-Survey", Australian Journal of Basic and Applied Sciences, 8(15) September 2014.
53. S.Balamurugan, P.Visalakshi, V.M.Prabhakaran, S.Chranyaa, S.Sankaranarayanan, "Strategies for Solving the NP-Hard Workflow Scheduling Problems in Cloud Computing Environments", Australian Journal of Basic and Applied Sciences, 8(15) October 2014.
54. Charanyaa, S., et. al., , A Survey on Attack Prevention and Handling Strategies in Graph Based Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 2(10): 5722-5728, 2013.
55. Charanyaa, S., et. al., Certain Investigations on Approaches forProtecting Graph Privacy in Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 1(8): 5722-5728, 2013.
56. Charanyaa, S., et. al., Proposing a Novel Synergized K-Degree L-Diversity T-Closeness Model for Graph Based Data Anonymization. International Journal of Innovative Research in Computer and Communication Engineering, 2(3): 3554-3561, 2014.
57. Charanyaa, S., et. al., , Strategies for Knowledge Based Attack Detection in Graphical Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 3(2): 5722-5728, 2014.
58. Charanyaa, S., et. al., Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.
59. V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Certain Investigations on Strategies for Protecting Medical Data in Cloud", International Journal of Innovative Research in Computer and Communication Engineering Vol 2, Issue 10, October 2014
60. V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Investigations on Remote Virtual Machine to Secure Lifetime PHR in Cloud ", International Journal of Innovative Research in Computer and Communication Engineering Vol 2, Issue 10, October 2014

61. V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Privacy Preserving Personal Health Care Data in Cloud" , International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014
62. P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, "Investigations on Evolution of Strategies to Preserve Privacy of Moving Data Objects" International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.
63. P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Certain Investigations on Securing Moving Data Objects" International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.
64. P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Survey on Approaches Developed for Preserving Privacy of Data Objects" International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014
65. S.Jeevitha, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Privacy Preserving Personal Health Care Data in Cloud" International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014.
66. K.Deepika, P.Andrew, R.Santhya, S.Balamurugan, S.Charanyaa, "Investigations on Methods Evolved for Protecting Sensitive Data", International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 4, Decermber 2014.
67. K.Deepika, P.Andrew, R.Santhya, S.Balamurugan, S.Charanyaa, "A Survey on Approaches Developed for Data Anonymization", International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 4, Decermber 2014.
68. S.Balamurugan, S.Charanyaa, "Principles of Social Network Data Security" LAP Verlag, Germany, ISBN: 978-3-659-61207-7, 2014
69. S.Balamurugan, S.Charanyaa, "Principles of Scheduling in Cloud Computing" Scholars' Press, Germany,, ISBN: 978-3-639-66950-3, 2014
70. S.Balamurugan, S.Charanyaa, "Principles of Database Security" Scholars' Press, Germany, ISBN: 978-3-639-76030-9, 2014

BIOGRAPHY

R.Santhya and S.Latha are currently pursuing their B.Tech. degree in Information Technology at KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Their areas of research interests include Network Security, Cloud Computing and Database Security.

Prof.S.Balamurugan obtained his B.Tech degree in Information Technology from P.S.G. College of Technology, Coimbatore, Tamil Nadu, India and M.Tech degree in Information Technology from Anna University, Tamil Nadu, India respectively. He is currently working towards his PhD degree in Information Technology at P.S.G. College of Technology, Tamil Nadu, India. At present he holds to his credit **65 papers International Journals and IEEE/ Elsevier International Conferences.** He is currently working as Assistant Professor in the Department of Information Technology, Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India affiliated to Anna University TamilNadu, India. He is **State Rank holder** in schooling. He was **University First Rank holder** M.Tech. Semester Examinations at Anna University, Tamilnadu, India. He served as a Joint Secretary of IT Association, Department of Information Technology, PSG College of Technology, Coimbatore, Tamilnadu, India. He is the **recipient of gold medal and certificate of merit** for best journal publication by his host institution **consecutively for 3 years**. Some of his professional activities include invited Session Chair Person for two Conferences. He has guided 16 B.Tech projects and 2 M.Tech. projects. He has won a best paper award in International Conference. His areas of research interest accumulate in the areas of Data Privacy, Database Security, Object Modeling Techniques, and Cloud Computing. He is a life member of ISTE,CSI. **He has authored a chapter in an International Book "Information Processing" published by I.K. International Publishing House Pvt. Ltd, New Delhi, India, 978-81-906942-4-7. He is the author of 3 books titled "Principles of Social Network Data Security", ISBN: 978-3-659-61207-7, "Principles of Scheduling in Cloud Computing" ISBN: 978-3-639-66950-3, and "Principles of Database Security", ISBN: 978-3-639-76030-9.**

S.Charanyaa obtained her **B.Tech** degree in Information Technology and her **M.Tech** degree in Information Technology from Anna University Chennai, Tamil Nadu, India. She was **gold medalist** in her B.Tech. degree program. She has to her credit **27 publications in various International Journals and Conferences**. Some of her outstanding achievements at school level include **School First Rank holder** in **10th and 12th grade**. She was working as Software Engineer at Larsen & Turbo Infotech, Chennai for 3 years where she got promoted as Senior Software Engineer and worked for another 2 years. She worked at different verticals and worked at many places including Denmark, Amsderdam handling versatile clients. She is also the recipient of **best team player award for the year 2012 by L&T**. Her areas of research interest accumulate in the areas of Database Security, Privacy Preserving Database, Object Modeling Techniques, and Cloud Computing. **She is the author of 3 books titled "Principles of Social Network Data Security", ISBN: 978-3-659-61207-7, "Principles of Scheduling in Cloud Computing" ISBN: 978-3-639-66950-3, and "Principles of Database Security", ISBN: 978-3-639-76030-9.**