Advances and Applications in Mathematical Sciences
Volume 20, Issue 11, September 2021, Pages 2937-2945

TOWARDS MORE EFFICIENT 3NF DETERMINATION USING REDUCED FUNCTIONAL DEPENDENCY SETS

MADHU SUDAN CHAKRABORTY*, AMITAVA BONDYOPADHYAY and TAPAS KUMAR GHOSH

Department of Computer Science
Indas Mahavidyalaya
Indas, Bankura
West Bengal, 722205, India
E-mail: mailmschakraborty@rediffmail.com

Department of Computer Science
Mankar College, Mankar
West Bengal, 713144 India
E-mail: amitavabondyopadhyay@mankarcollege.ac.in

Department of Computer Science
Bankura Sammilani College
Bankura, 722102, West Bengal, India
E-mail: tapas_bsc38@yahoo.co.in

Abstract

3NF can reduce redundancies and anomalies from a relational database as far as possible while still observing lossless join and even dependency preservation properties. Obviously in spite of existence of higher order normal forms in theory, normalization into 3NF is widely viewed as necessary and sufficient for practical applications. Accordingly efficient determination of 3NF is a prerequisite for good database design. In this article the 3NF interpretation is strived to be reduced, yielding some more efficient 3NF determination techniques. One of the resulting interpretations is theoretically an optimal. Some prospective implications of the proposed interpretations are also discussed.

2010 Mathematics Subject Classification: 68P15; 97P30.
Keywords: Relational database, 3NF Determination, Prime attribute, Intractable problem, Optimal interpretation.
*Corresponding author.
Received January 25, 2021; Accepted August 16, 2021

1. Introduction

Relational database theory has provided effective guidelines to the developers for using mathematical/logical models efficiently for ensuring good database design ([1]-[8]). These rules are particularly useful for investigating the keys, redundancies and anomalies with reference to given functional dependencies (FDs) and some other constraints. Normalization is an integral part of relational database design where the given relation schema (R) is analysed with respect to keys and FDs, striving to reduce redundancies and anomalies. As per ordinary theory, for a given relation schema (R) defined on a set of FDs (F), expressed as (R, F), initially the normalization status is tested through some normal form (NF) in a particular order which is termed as determination of the normalization level. In case of non conformity with the NF, R is then either decomposed or synthesized into the NF. In theory, there are several NFs and in the ascending order of strength the NFs are 1NF, 2NF, 3NF, BCNF, 4NF, 5NF etc. However, in the standard literature normalization into 3NF is treated sufficient as it may reduce redundancies and anomalies as far as possible while still ensuring some other essential feature of database design, including dependency preservation ([1]-[8]). Clearly an efficient determination of 3NF may be a prerequisite for good database design and this article is focused on exploring more efficient 3NF determination schemes using some reduced interpretations of F.

The rest of the article is organized with four sections as follows. In the preliminaries section, the insight into the problem is intended to be provided with detailed background. In the proposed methods section 3NF is strived to be interpreted in terms of some reduced sets of FDs, immediately resulting in faster 3NF determination techniques and the proofs in favour of their correctness are also given. In the results and discussion section, the comparative merits of the proposed 3NF determination techniques are assessed. Finally the article ends in the conclusion section, outlining some future directions too.

2. Preliminaries

As per the original interpretation (R, F) is in 3NF iff for every non

trivial $FD\ X \to Y$ either X is a super key of R or Y is a prime (or key) attribute of R ([1]-[8]). In several classical texts 3NF is determined directly on the basis of the interpretation without referring any FD set explicitly ([1]-[4]). In some other standard text books the 3NF testing addresses the closure of $F(F^+)$ i.e. (R, F) is in 3NF iff for every nontrivial $FD\ X \to Y \in F^+$ either X is a super key of R or Y is a prime attribute of R ([5]-[6]). The F^+-centric interpretation is also advocated in article [7] but implicitly, as it refers to the set of all FDs implied by F which is again F^+. In some other texts 3NF testing involves only the FDs in F i.e. (R, F) is in 3NF iff for every nontrivial $FD\ X \to Y \in F^+$ either X is a super key of R or Y is a prime attribute of R ([8]-[9]). In addition in a recent work 3NF is supposed to be determinable merely on the basis of an optimal canonical cover (G) of F i.e. (R, F) is in 3NF iff for every nontrivial $FD\ X \to Y \in G$ either X is a super key of R or Y is a prime attribute of R [10]. The interpretation of 3NF in terms of F^+ obviously holds true as $(F^+)^+ = F^+$. However, one problem regarding 3NF determination is that, in the literature it is not known whether the F-centric interpretation and G-centric interpretation are also correct or not. Even if both interpretations hold true, then the question of other interpretations, including possibly an optimal interpretation may arise.

It is needless to say that finding asymptotically more efficient algorithms for 3NF determination is constrained by the intractability of the problem, caused by its sub-problems. Speaking more elaborately, most likely there is no polynomial-time algorithm for candidate keys' determination, attributes' primality detection and obviously 3NF identification as all these problems belong to the NP-complete class ([11]-[13]). However, motivated by some early works ([12], [14]-[15]) which shows capability to determine the candidate key(s) quickly for the problems subject to typical characteristics of their attribute sets, some 3NF-determination algorithms have been proposed ([16]-[17]) which may often run in polynomial time.

Another interesting aspect of the problem is that although 3NF determination problem is NP-complete, there exists a 3NF synthesis algorithm [18] which guarantees lossless join and dependency preservation and acquires polynomial-time if considered in stand-alone mode. So one may

directly consider applying 3NF synthesis algorithm without its testing so that 3NF status of R may be achieved as a whole in polynomial time. However, this idea may ultimately appear to be ineffective owing to two problems. Firstly a prerequisite input for applying 3NF synthesis algorithm is an optimal (or minimal) canonical cover of $F(G)$. G is obtained by eliminating all superfluous attributes as well as all superfluous FDs from F in any order. As determination of G is known to be NP-complete ([19]-[21]), 3NF synthesis algorithm is *NP*-complete as a whole. Secondly as the stand-alone 3NF synthesis algorithm [18] does not involve normalization status checking at any point, it may further decompose R even after achieving the desired 3NF status [17].

3. The Proposed Methods

In coherent form the axial part of the proposed reduced interpretations of 3NF, Theorem 1, is as follows.

Theorem 1. *3NF status of* (R, F) *is equivalently determinable* $\forall H$ *where* $H^+ = F^+$.

Proof. Two declarative statements (propositions), say statement A and statement B, are treated equivalent (or $A \equiv B$) iff simultaneously both hold true or both hold false.

Introduce Statement 1 and Statement 2 where

Statement 1. For every nontrivial *FD* $X \to Y \in H$ either X is a super key of R or Y is a prime attribute of R.

Statement 2. For every nontrivial *FD* $X \to Y \in F^+$ either X is a super key of R or Y is a prime attribute of R.

If possible, suppose that regarding the 3NF status of a given relation schema R statement 1 and statement 2 do not simultaneously hold. It means either case 1 or case 2 might arise where.

Case 1. For R statement 1 does not hold true but statement 2 holds true.

Case 2. For R statement 1 holds true but statement 2 does not hold true.

Without any loss of generality hereafter every FD is supposed to be non trivial and canonical, unless stated otherwise. In case 1, non satisfying of statement 1 means that $\exists$ a $FD\ P \to Q \in H \mid$ neither P is a super key of R nor Q is a prime attribute of R. However, as $\forall FDs \in H$ are also necessarily $\in F^+$, $\exists$ a $FD\ P \to Q \in F^+ \mid$ neither P is a super key of R nor Q is a prime attribute of R. It contradicts statement 2. Therefore case 1 does never arise.

In case 2, non satisfying of statement 2 means that $\exists$ a $FD\ S \to T \in F^+ \mid$ neither S is a super key of R nor T is a prime attribute of R. Consider the inference of $S \to T \in F^+$ with reference to H. As $H^+ = F^+$, either originally $S \to T \in H$ or there is some $FD(s) \in H$, called source FD(s), from which the $FD\ S \to T$ is inferred by applying one or more instances of augmentation rule or transitive rule or their derivatives or compositions. Let the derivation path of the $FD\ S \to T$ is given by. $S \to W_1 \to W_2 \ldots \to W_i \to T$ where $W_1, W_2, \ldots, W_i$ are formed over the attribute(s) of R. It is obvious that $\{S \to W_1, W1 \to W_2, \ldots, W_i \to T\}$ is a subset of F^+ and $\exists V \to T \in H \mid V$ is a subset of W_i. If possible assume that V is a super key of R. It means W_i is a super key of R. Then the FD $W_{i-1} \to W_i$ implies that W_{i-1} is a super key of R. Proceeding in this manner W_1 also appears to be a super key of R. But it is given that S is not a super key of R. Then the $FD\ S \to W_1$ implies that a non-key of R functionally determines a super key of R, which is a contradiction. Therefore V is not a super key of R and case 1 does never arise. Hence the result follows.

Corollary 1. *The* 3*NF status of R is equivalently determinable in terms of* F^+, F *and G.*

Corollary 2. *The* 3*NF status of R is optimally determinable in terms of G.*

Corollary 1 immediately follows Theorem 1. In order to prove corollary 2, if possible, suppose that $\exists$ a subset L of $G \mid$3NF status of R is also determinable in terms of L. It means, although $L^+ \neq G^+$ (given) and $G^+ = F^+$ (known), $L^+ = F^+$ which is a contradiction. Hence corollary 2 immediately holds true.

Let K be the set of FDs obtained from F by removing the superfluous FD(s) only and K' be the set of FDs obtained from F by removing the superfluous attribute(s) only.

Corollary 3. *The* 3*NF status of R is equivalently determinable in terms of K.*

Corollary 4. *The* 3*NF status of R is equivalently determinable in terms of* K'.

Corollary 3 and corollary 4 immediately follow Theorem 1. Let SK and P denote the set of all super keys and prime attributes of R respectively. Then corollary 2 and corollary 3 may air new 3NF testing proposals, say, Method 1 and Method 2, respectively as follows.

Method 1. For all non trivial *FDs* $X \to Y \in G$ test if $X \in SK$ or $Y \in P$.

Method 2. For all non trivial *FDs* $X \to Y \in K$ test if $X \in SK$ or $Y \in P$.

It is needless to recall that the Method 1 and Method 2 represent an optimal and minimum 3NF testing approaches respectively. It may also be noted that although computing G most probably needs exponential time, there exists a polynomial-time solution for K ([19]-[20]).

4. Results and Discussion

For the known 3NF determination methods the various stake holders are computing SK, P and checking the determinant and attribute of every FD. For a given relation, RX, the set of FDs, F_X and the associated attribute set, A_X; without any loss of generality, the time complexities for computing SK, P and checking the determinants and attributes of all FDs may be expressed as $O(f(|A_X|, |F_X|))$, $O(g(|A_X|, |F_X|))$ and $O(h(|A_X|, |F_X|))$ respectively where $|S|$ denotes the number of elements in the set S. Let $O(\Psi(|A_X|, |F_X|)) = O(f(|A_X|, |F_X|)) + O(g(|A_X|, |F_X|)) + O(h(|A_X|, |F_X|))$. As f ( ), g ( ) and h ( ) are all increasing order functions, $\Psi(\,)$ is also an increasing order function in terms of $|A_X|$ and $|F_X|$. For a given R, let the attribute set corresponding to the FDs F^+, F, K and G are

A_{F^+}, A_F, A_K and A_G respectively. As here $|A_G| \le |A_K| \le |A_F| \le |A_{F^+}|$ and $|G| \le |K| \le |F| \le |F^+|$ hold true, $O(\Psi(|A_G|, |G|)) \le O(\Psi(|A_K|, |K|)) \le O(\Psi(|A_F|, |F|)) \le O(\Psi(|A_{F^+}|, |F^+|))$ must also hold true. It means that the proposed 3NF determination methods can outperform the textbook prescribed traditional methods as well as the recently introduced graphical methods for 3NF determination relying on F^+ or F ([16]-[17]). Another interesting point is that if the proposed G-centric or K-centric methods are employed for 3NF determination instead of the traditional or even the recent graphical methods ([16]-[17]), the follow-up queries can also run faster significantly along with less memory consumption, particularly where $|G|$ and $|F|$ are considerably smaller than $|F^+|$ and $|F|$.

Consider (R, F) where $F = \{A \to C, B \to C, AB \to C, AC \to D\}$. Here $\{A\}^+ = \{A, B, C, D\}$ and so A is a candidate key of R. However, $B^+ = \{B, C\}$, $C^+ = \{C\}$, $D^+ = \{D\}$, $BC^+\{B, C\}$, $BD^+ = \{B, D\}$, $CD^+ = \{C, D\}$, $BCD^+ = \{B, C, D\}$. So any other key does not exist. It means A is the only key and B, C and D are the non prime attributes. Then the $FD\ B \to C \in F$ indicates that (R, F) is not in 3NF. Again as $\{A\}^+ = \{A, B, C, D\}$ the attribute B and C are superfluous in the $FD\ AB \to C$ and $AC \to D$ respectively. Removal of the attribute B and C from the respective FDs causes F to reduce to $\{A \to B, A \to C, B \to C, A \to D\}$. As the rest of the FDs do not have composite determinant, they obviously do not contain any superfluous attribute. So $K'\{A \to B, A \to C, B \to C, A \to D\}$. The $FD\ A \to C$ is superfluous in K' as $K' - \{A \to C\} \models \{A \to C\}$. None of the other FDs is superfluous. Therefore $G = \{A \to B, B \to C, A \to D\}$. (R, F) is also not in 3NF owing to the $FD\ B \to C$. This example shows that the 3NF status of (R, F) can be checked merely from the 3NF status of (R, F). In addition in this example, for determining the 3NF status of (R, F), only three FDs of G are sufficient to consider instead of five FDs of F.

5. Conclusion

In this paper, for any given R and associated F, 3NF has been strived to be determined merely using some parts of F^+ (or F). It has been shown that the F-centric, K-centric, K'-centric and G-centric interpretations of 3NF equivalently hold true. The significance of K'-centric interpretation of 3NF has been observed elsewhere [22]. However, the K-centric and G-centric interpretations and in particular, the optimality of G-centric interpretation had never explicitly appeared in the literature to the best of the authors' knowledge and belief. The comparative merits of the proposed 3NF determination methods over their potential contenders have been demonstrated too.

The optimal and minimum interpretations of 3NF proposed in this paper may be immediately extended for BCNF. In future along with better resolutions on generating candidate keys and checking primality of attributes, the proposed methods may continue to lead in determining 3NF with greater efficiency. A following-up of this study suggests reassessing the information theoretic implications of 3NF ([23]-[24]) in view of its reduced interpretations proposed and it is left as an open problem.

References

[1] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, Addison-Wesley (2010).

[2] C. J. Date, An Introduction to Database Systems, Addison-Wesley (2003).

[3] S. Sumathi and E. Esakkirajan, Fundamentals of Relational Database Management Systems (Studies in Computational Intelligence), Springer (2010).

[4] H. Garcia-Molina, Database Systems, The Complete Book, Pearson Education (2014).

[5] R. Silberschatz, H. Korth and S. Sudarshan, Database System Concepts, McGraw-Hill (2010).

[6] B. C. Desai, An Introduction to Database Systems, West Group (1990).

[7] S. Abiteboul, R. Hull and V. Vianu, Foundation of Databases The Logical Level, Addison-Wesley (1995).

[8] R. Ramakrishnan and J. Gehrke, Database Management Systems, McGraw-Hill Education (2002).

[9] M. Arenas, Third Normal Form, Encyclopedia of Database Systems, Springer, US (2009), 3087-3088.

[10] K. V. Iyer, An Introduction to Functional Dependency in Relational Databases (as a part of Review of Relational Databases, February, 2016). http://www.nitt.edu/home/academics/departments/cse/faculty/kvi/KVI-DBFD-2011.pdf (Retrieved 06.08.2020).

[11] C. L. Lucchesi and S. L. Osborn, Candidate keys for relations, Information Processing Letters 17 (1978), 270-279. https://www.sciencedirect.com/science/article/abs/pii/0020019082900345

[12] SL. Osborn, Normal Forms for Relational Data Bases, Research Report: CS-78-06 University of Waterloo Canada (1978). https://www.comp.nus.edu.sg/~lingtw/papers/tods81.LTK.pdf

[13] J. H. Jou and P. C. Fischer, The complexity of recognizing 3NF relation schemas, Information Processing Letters 14 (1982), 187-190. https://link.springer.com/chapter/10.1007/BFb0035004

[14] S. Kundu, An improved algorithm for finding a key of a relation, Proceedings of ACM Symposium on Database Systems Portland (USA) (1985), 189-192. https://doi.org/10.1145/325405.325431

[15] H. Saiedian and T. Spencer, An efficient algorithm to compute the candidate keys of a relational database schema. The Computer Journal 39 (1996), 134-143. https://academic.oup.com/comjnl/article-bstract/39/2/124/580489?redirectedFrom=fulltext

[16] PB. Worland, An efficient algorithm for 3NF determination, Information Science 167 (1-4) (2004), 177-192. https://dl.acm.org/toc/isci/2004/167/1-4

[17] G. Gottlob, R. Pichler and F. Wei, Tractable Database Design and Datalog Abduction through Bounded Treewidth. Information Sciences 35 (2010), 278-298. https://publik.tuwien.ac.at/files/PubDat_186193.pdf

[18] P. Bernstein, Synthesizing third normal form relations from functional dependencies, ACM Transactions on Database Systems 1 (1976), 277-298. https://www.comp.nus.edu.sg/~lingtw/papers/bernstein.pdf

[19] D. Maier, Minimum covers in the relational database model, Journal of ACM 27 (1980), 664-674. https://link.springer.com/article/10.1007/s00778-011-0239-5

[20] D. Maier, The Theory of Relational Databases, Computer Science Press, Rockville (USA) (1983).

[21] X. Peng and Z. Xiao, Optimal covers in the relational database model, Acta Informatica 53 (2016), 459-468.

[22] T-W. Ling, F. W. Tompa and T. Kameda, An improved third normal form for relational databases, ACM Transactions on Database Systems 6 (1981), 329-346. https://www.comp.nus.edu.sg/~lingtw/papers/tods81.LTK.pdf

[23] S. Kolahi and L. Libkin, An information theoretic analysis of the worst-case redundancy in database design, ACM Transactions on Database Systems 35 (2010), 1-32. https://dl.acm.org/doi/abs/10.1145/1670243.1670248?download=true

[24] D. Ungureanu, A new definition for the information content of databases, University Politehnica of Bucharest Science Bulletin C-74 (2012), 35-44. https://www.scientificbulletin.upb.ro/rev_docs_arhiva/full6e3_410662.pdf