 

# TDSGenerator: A Tool for generating synthetic Transactional Datasets for Association Rules Mining

**G. S. Bhamra[1], A.K.Verma[2] and R. B. Patel[3]**

**[1] M.M. Institute of Computer Tech. and Business Management, MMU, Mullana, Haryana -133203, India**

**[2] Department of Computer Sc. & Engineering, TIET, Thapar University, Patiala, Punjab-147004, India**

**[3] Department of Computer Sc. & Engineering, DCR University of Sc. and Tech., Murthal, Haryana -131039, India**

**Abstract**

Data Mining (DM) is the process of automated extraction of interesting data patterns representing knowledge, from the large data sets. Frequent itemsets are the item sets that appear in a data set frequently. Finding such frequent itemsets plays an essential role in mining associations, correlations, and many other interesting relationships among itemsets in transactional and relational database.

In this paper we have presented a tool called, Transactional Dataset Generator (TDSGenerator v1.0) for generating a Binary Dataset as well as Transactional Dataset corresponding to the Binary Dataset. Synthetic datasets generated by this tool will be used to find the list of frequent itemsets and thereafter finding the strong Association Rules among those itemsets. This tool can also be used as a demonstrator for experimenting and explaining the concepts of Association Rules Mining (ARM).

***Keywords:*** *Data Mining, Frequent Itemsets, Binary Data Set, Transactional Data Set, Association Rules Mining.*

## 1. Introduction

Hidden or embedded knowledge in the form of interesting trends or patterns can be extracted using an automated process of mining the data in large databases, the web, other massive data repositories, or data streams [1, 2]. Patterns are extracted using techniques such as classification, association rules, clustering, etc.

A huge amount of basket data can be stored based on items purchased on a per-transaction basis as a result of the advancement in the bar code technology [5]. The itemsets that appear more frequently in such data sets are called frequent itemsets. Extracting such frequent itemsets plays an essential role in mining associations, correlations, and many other interesting relationships among itemsets in transactional and relational databases. Frequent Itemset mining is a core DM task. It has an elegantly simple problem statement: to find the set of all subsets of items that frequently occur together in database records or transactions. Although this task has a simple statement, it is CPU and input/output (I/O) intensive, mainly because the large amount of item sets that are typically generated and large size of the datasets involved in the process[2,6,7].

Mining the Association rules from the frequent itemsets requires a transactional database which can be a real transactional data base of any retail industry or can be a synthetic version generated by a tool. Synthetic Data Set generated by a tool can serve a fundamental requirement for experimenting with the DM concepts and mining the Association rules from the frequent item sets. Newly designed algorithms can be experimented and tested on such synthetic data sets and then the concepts can be implemented on a real data set.

This paper presents a tool called, Transactional Data Set Generator (TDSGenerator v1.0) for generating a Binary Dataset as well as Transactional Dataset corresponding to the Binary Dataset. Synthetic datasets generated by this tool will be used to find the list of frequent item sets and thereafter finding the strong Association Rules among those item sets in a distributed environment.

## 2. Definitions and Preliminaries

The frequent itemset mining can be formally defined as follows:

$DB \Rightarrow$ Transactional Database as shown in Fig. 1.

$D \Rightarrow$ Total number of transactions in $DB$ or size of $DB$

$I = \{i_1, i_2, \cdots, i_m\} \Rightarrow$ Set of m items in $DB$

ISSN (Online): 1694-0814
www.IJCSI.org

$T \Rightarrow$ A transaction in $DB$.Each transaction is assigned an identifier called TID.

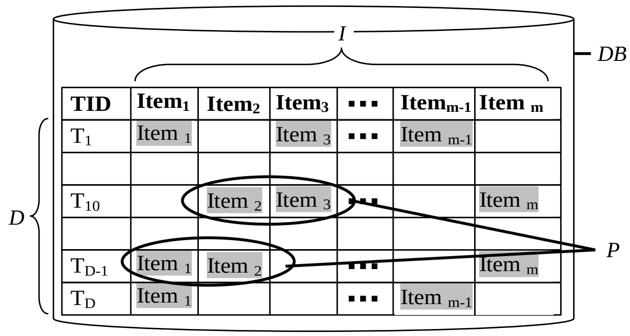


Fig. 1 Transactional Dataset.

$P \Rightarrow$ A set of items (item set or pattern) in a particular transaction $T$, $P \subseteq I$. An item set $P$ containing k items is called **k-itemset**.

$s(P)$ : Support of an itemset $P$ is the frequency of occurrence of $P$ in $DB$

$s(P) = \frac{\#\_of\_T\_containing\_P}{D}$ %,

where $\#\_of\_T\_containing\_P$ is the support count (sup_count) of itemset $P$.

min_sup: given minimum support threshold.

Frequent Itemsets $\Rightarrow$ Frequent itemsets are the itemsets that appear in a data set frequently. and satisfy the minimum support(min_sup) ,i.e., if Support( $P$ ) $\geq$ min_sup.

$C_k \Rightarrow$ Candidate frequent k-Itemsets is the list (or set) of all frequent k-itemsets without the constraint of minimum support threshold, min_sup.

$L_k \Rightarrow$ List (or set) of frequent k-itemsets after pruning the candidate set $C_k$ by applying the constraint of min_sup.

Anti-monotone downward closure property of Frequent Itemsets $\Rightarrow$ if a set cannot pass a test, all of its supersets will fail the same test as well, i.e., all nonempty subsets of a frequent itemset must also be frequent, i.e., any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. This property is used in subset testing of frequent itemsets.

Apriori Algorithm $\Rightarrow$ Apriori is most popular and basic algorithm proposed by R. Agrawal and R. Srikant [4] for mining frequent itemsets for generating boolean association rules. Apriori exploits the basic functionality of a frequent itemset: all subsets of a frequent itemset must be frequent. Starting with singleton itemsets, Apriori computes their supports by scanning the database, and filters out frequent itemsets. At the end of each iteration, only itemsets whose immediate subsets are all frequent at the current iteration are considered at the next iteration.

Association Rule (AR) $\Rightarrow$ An implication of the form $P \Rightarrow Q$, where itemset $P \subset I$, itemset $Q \subset I$, and $P \cap Q = \phi$. Itemset $P$ is the antecedent part and itemset $Q$ is called the consequent part of AR.

Association Rules, first introduced in [5], are used to discover the associations (or co-occurrences) among items in a transactional database. ARs can be used to find the patterns of customer's purchase such as how the transaction of buying some goods will impact on the transactions of buying others. Such rules can be implemented to design the merchandise shelves, to manage the stock and to classify the customers according to the purchase patterns. Support and confidence are two measures to find interesting Association Rules.

Support, $s$, of an AR is the probability that the transaction contains both antecedent and consequent of AR and is given as:

$s(P \Rightarrow Q) = \frac{\#\_of\_T\_containing\_both\_P\_and\_Q}{D}$ %,

we can say that $s\%$ of the transactions support the rule $P \Rightarrow Q$, $0 \leq s \leq 1.0$ or $0\% \leq s \leq 100\%$

Confidence, c, of an AR is the conditional probability that a transaction having antecedent also contains consequent of AR and is given as:

$c(P \Rightarrow Q) = \frac{s(P \Rightarrow Q)}{s(P)}$ =sup_count( $P \Rightarrow Q$ )/sup_count( $P$ ),we can say that when itemset $P$ occurs in a transaction there are $c\%$ chances that itemset $Q$ will occur in that transaction, $0 \leq c \leq 1.0$ or $0\% \leq c \leq 100\%$

min_conf $\Rightarrow$ given minimum confidence threshold.

Strong AR $\Rightarrow$ if $s(P \Rightarrow Q) \geq$ min_sup && $c(P \Rightarrow Q) \geq$ min_conf, then rule $P \Rightarrow Q$ is called strong AR.

Association Rule Mining (ARM) $\Rightarrow$ ARM is the task to find all the strong association rules whose support and confidence are above the min_sup and min_conf, respectively. The ARM can be viewed as a two-step process [8].

1. Find all frequent k-itemsets( $L_k$ )
2. Generate Strong Association Rules from $L_k$

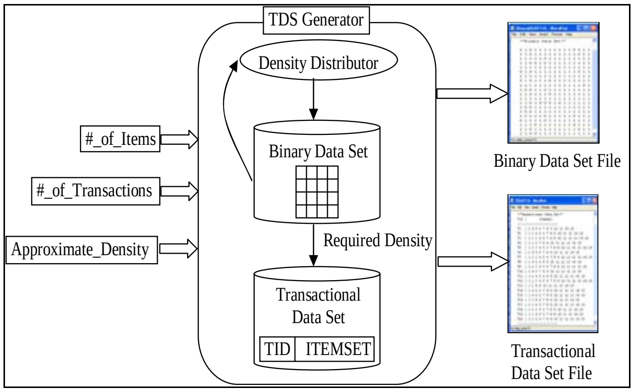


Fig. 2 Block Architecture of TDSGenerator.

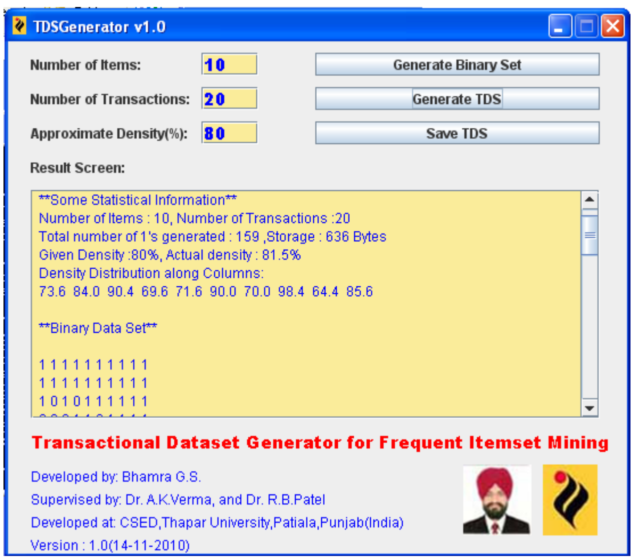


Fig. 3 GUI of TDSGenerator.

## 3. System Architecture

TDSGenerator takes three inputs- 1) total number of items in a transactional dataset 2) total number of transactions, and 3) approximate density of the number of 1's in BDS. A 2-D array of integer values is created with number of columns equals number of items and number of rows equals number of transactions. Array elements can have either of the two binary values '0' or '1' in that '1' represents that the item in a transaction is purchased by the customer and '0' otherwise. Repeated attempts are made to get the desired density array with the help of Density Distributor component. If the BDS of required density is created then a TDS is generated. Each transaction in TDS consists of Transaction Identification (TID) and itemset, i.e., all the items purchased in that transaction. Block architecture of TDSGenerator is shown in Figure 2.

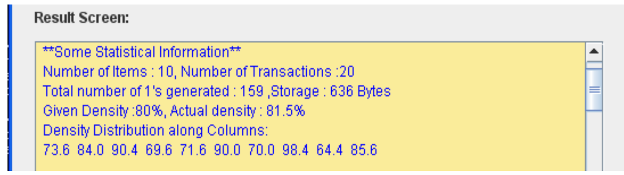


Fig. 4 Statistical Information.

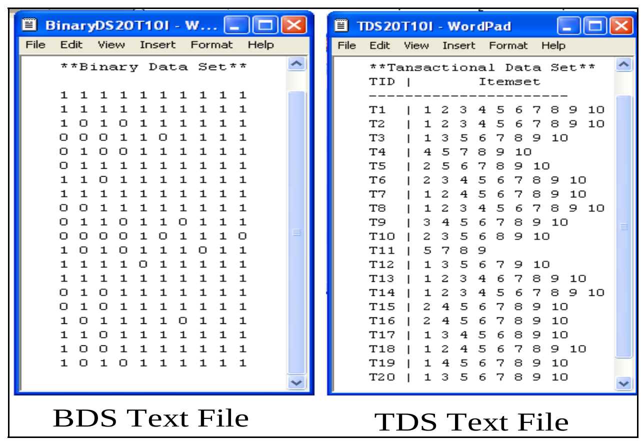


Fig. 5 Output BDS and TDS Text files generated by TDSGenerator.

This tool generates two output text files, as shown in Figure 5, a BDS text file (e.g. BinaryDS20T10I.txt) for BDS and TDS text file (e.g. TDS20T10I.txt) for TDS and stores them at desired location.

## 4. Implementation and performance study

TDSGenerator v1.0 has been implemented in Java language using Java Swing for advanced GUI designing. GUI of TDSGenerator v1.0 is shown in Figure 3.

Some statistical information for the given input (number of items=10, number of transactions=20, and approximate density=80%) is shown in Figure 4 result screen. From this output we have seen that total number of '1' generated in BDS are 159 and if one integer value takes 4 bytes then total 636 bytes are required to store the data set of all the items purchased. The actual density of BDS will be 81.5%.

This tools supports a maximum of $2^{32}$ number of items and a maximum of $2^{32}$ number of transactions in a data set. So with these values it can generate a BDS of $2^{64}$ items.

## 5. Conclusion

Mining the Association rules from the frequent itemsets requires a transactional database which can be a real transactional data base of any retail industry or can be a simulated version generated by a tool. Data Set generated by a tool can serve a fundamental requirement for experimenting with the DM concepts and also mining the Association rules from the frequent itemsets.

In this paper we have presented a tool called, Transactional Data Set Generator (TDSGenerator v1.0) for generating a BDS as well as TDS corresponding to the BDS. This tool can also be used as a demonstrator for experimenting and explaining the concepts of DM.

In our further steps of research the TDSGeneratorv1.0 tool will be used to generate the partitioned data sets at distributed sites and then mining the global frequent itemsets (GFI) from distributed local frequent itemsets (LFI).

### Acknowledgment

Authors sincerely acknowledge Frans Coenen of Department of Computer Science, University of Liverpool, Liverpool L69 3BX, United Kingdom, for analyzing his ARM dataset generator and getting an idea of generating a basic dataset generator.

## References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
[3] Byung-Hoon Park and Hillol Kargupta, Distributed Data Mining: Algorithms, Systems, and Applications, Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, 1000 Hilltop Circle Baltimore, MD 21250
[4] R. Agrawal and R. Srikant, Fast algorithms for mining association rules. In Proc. 1994 Int. Conf. Very Large DataBases, pp. 487-499, Santiago, Chile, September 1994.
[5] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data, pp. 207-216, WA, May 1993.
[6] Matthew Eric Otey, S. Parthasarathy, Chao W., Adriano V., and Wagner Meira Jr., Parallel and Distributed Methods for Incremental Frequent Itemset Mining, IEEE transactions on Systems, Man, and Sybernetics- Part B: Cybernetics, 34(6), December 2004.
[7] J. Han, J. Pei, Y. Yin, Mining Frequent Patterns without candidate generation, Proc. ACM-SIGMOD, Dallas,TX, May, 2000.
[8] You-Lin Raun, Gan Liu, Qing-Hau Li, Parallel Algorithm for Mining Frequent Itemsets, in Proc. of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21, August 2005.

**Dr. A K Verma** is currently working as Assistant Professor in the department of Computer Science and Engineering at Thapar Institute of Engineering & Technology (Deemed University), Patiala. He received his B.S., M.S. and Ph.D. in 1991, 2001 and 2008 respectively, majoring in Computer science and engineering. He has worked as Lecturer at M.M.M. Engg. College, Gorakhpur from 1991 to 1996. He joined Thapar Institute of Engineering & Technology in 1996 as a Systems Analyst in the Computer Centre and is presently associated with the same Institute. He has been a visiting faculty to many institutions. He has published over 30 papers in referred journals and conferences (India and Abroad). He is a MISCI(Turkey), LMCSI (Mumbai), GMAIMA (New Delhi). He is a certified software quality auditor by MoCIT, Govt. of India. His research interests include wireless networks, routing algorithms and securing ad hoc networks.

**Dr. R. B. Patel** received PhD from IIT Roorkee in Computer Science & Engineering, PDF from Highest Institute of Education, Science & Technology (HIEST), Athens, Greece, MS (Software Systems) from BITS Pilani and B. E. in Computer Engineering from M. M. M. Engineering College, Gorakhpur, UP. Dr. Patel is in teaching and Research & Development since 1991. He has supervised several M. Tech, and M. Phil and PhD Thesis. He is currently the chairperson of Department of Computer Science and Engineering, DCR University of Science and Technology, Murthal, Haryana and supervising several M. Tech, and PhD students. He has published more than 95 research papers in International/National Journals and Refereed International Conferences. Dr. Patel received several research awards, few are as under: Providing Security and Robustness to Mobile Agents on Open Networks" received best research paper award at BIS 2003 Colorado, Spring, USA. "Mobile Agents Location Management in Global Networks" had been considered in best five research papers at 8th International Conference on Information Technology, Bhubaneshwar, India and referred to International Journal of Information and Communication Technology, InderSceince Publication. "COMPUTING IN PEER-TO-PEER NETWORKS" is received best Research Paper award in "Challenges & Opportunities in Information Technology" (COIT -2007), 2007, "Dynamic Traffic -Conscious Routing for MANETs", is received best Research Paper award in International Conference on Information & Communication Technology (IICT 2007), Dehradun, UA, India and "An Agent enriched Distributed Data Mining on

Heterogeneous Networks", is received best Research Paper award in "Challenges & Opportunities in Information Technology" (COIT-2008), 2008. Further his contribution in the field of Mobile Agent Technology appreciated by PhD Thesis reviewers (Professor from Indian Institute of Sciences Bangalore best institute in India and Professor from Florida "University, USA).
He has written numbers books for engineering courses (These are "Fundamentals of Computing and Programming in C", "Theory of Automata and Formal Languages", "Expert Data Structures with C," "Expert Data Structures with C++," "Art and Craft of C" and "Go Through C". These books are being followed prominently in all Under Graduate/Post Graduate courses and also for Research, in various Technical Institutions/Universities in India.). Out of which two books namely "Expert Data Structures with C" and "Art & Craft of C" are recommended by Indian Society for Technical Education (ISTE), India.
He is member of various International Technical Societies such as IEEE-USA, Elsevier-USA, Technology, Knowledge & Society-Australia, WSEAS, Athens, etc, for evaluation of research works. He was member of various International Technical Committees and participating frequently in International Technical Committees in India and abroad. Dr. Patel currently actively involved in research activities in collaboration with WSEAS - World Scientific and Engineering Academy and Society, Athens, Greece in the various area of mobile computing.

His current research interests are in Mobile & Distributed Computing, Mobile Agent Security and Fault Tolerance, development infrastructure for mobile & Peer-To-Peer computing, Device and Computation Management, Cluster Computing, etc.

**G. S. Bhamra** is currently working as Assistant Professor in M. M. Institute of Computer Technology & Business Management, M. M. University, Mullana, Haryana. He received his B.Sc. (Computer Sc.) and MCA from Kurukshetra University Kurukshetra in 1995 and 1998, respectively. He is pursuing Ph.D. from Department of Computer Science and Engineering, Thapar University, Patiala, Punjab. He is in teaching since 1998. He has published 05 research papers in International/National Journals and International Conferences. He has received Best Paper Award for "An Agent enriched Distributed Data Mining on Heterogeneous Networks", in "Challenges & Opportunities in Information Technology" (COIT-2008). His research interests are Mobile & Distributed Computing, Distributed Data Mining, and Mobile Agents.