

*Dedicated to my parents and late grandparents*

---

## Declaration

I, Carynthia Kharkongor, hereby declare that the thesis entitled “*Compact Representation of Itemsets for Association Rule Mining*” submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering is based on the bona-fide work carried out by me under the supervision of my supervisors. The results embodied in this thesis have not been submitted in part or in full, to any other university or institute for award of any degree or diploma.

Place:

(Carynthia Kharkongor)



Department of Computer Science & Engineering  
Tezpur University

Napaam, Tezpur- 784028, Assam, India.

Phone: +91-3712-275109

Dr. Bhabesh Nath  
Professor

E-Mail: b.nath@tezu.ernet.in

---

## Certificate

This is to certify that the thesis entitled “**Compact Representation of Itemsets for Association Rule Mining**” submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Carynthia Kharkongor** under my supervision and guidance.

All the helps received by her from various sources have been duly acknowledged. No part of this thesis has been submitted else where for award of any other degree.

Signature of Supervisor  
(Prof. Bhabesh Nath)

Professor  
Department of Computer Science and Engineering  
Tezpur University  
Assam, India-784028



## Certificate

This is to certify that the thesis entitled “**Compact representation of itemsets for Association Rule Mining**” submitted by **Carynthia Kahrkongor** to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering has been examined by us on ..... and found to be satisfactory.

The Committee recommends for award of the degree of Doctor of Philosophy.

Signature of Principal Supervisor

Signature of External Examiner

# Acknowledgement

Firstly, I would like to praise God for the goodness and blessings he has bestowed upon me. I would like to express my appreciation and sincere gratitude to my Supervisor, Dr. Bhabesh Nath for his extraordinary support and guidance throughout my Ph.D. Dr. Bhabesh Nath supported my work and showed me the right track and direction toward achieving the goal. I am indebted to Dr. Bhabesh Nath for his insightful suggestions and advice that played an important role in my thesis. I hope I take these lessons learned from him in the next stage of my career.

I want to thank Prof. D. K. Bhattacharyya and Dr. Rosy Sharma who are both doctoral committee members and gave advice and helpful suggestions for my research. I would also like to thank Dr. Arindam Karmakar, Prof. Nityananda Sarma, and Prof. Utpal Sharma, for their beneficial suggestions during my research. I am grateful to my fellow researchers Deena, Piyali, Nirmal, Arundhati, Binayak, Kaushal, Alexy, and Shafiul for helping me understand the field and the challenges during my research. I wish them success in all their future endeavors. I want to thank Prof. Bhogeshwar Borah, CSE Head of Department, and the other faculty members who provided me with the environment and the facilities to do my research. I am grateful to the Director of DSEU Rajokri Campus, H.O.D./Incharge of Diploma and Degree, and other staffs for their support during my leave.

In the end, I thank my wonderful parents, siblings, and relatives for their support, patience, and love.

**Carynthia Kharkongor**

# List of Figures

2-1	An example set of an illustration with a single member 3 . . . . .	9
2-2	The NDI-representation . . . . .	16
3-1	Proposed itemset representation for the itemset $I = \{1, 3, 5, 12, 17, 30, 31\}$ . . . . .	20
3-2	Union operation of the itemset $A = \{1, 3, 5, 12, 17, 30, 31\}$ and $B = \{1, 3, 5, 13, 18, 29, 31\}$ . . . . .	20
3-3	Intersection operation of the itemset $A = \{1, 3, 5, 12, 17, 30, 31\}$ and $B = \{1, 3, 5, 13, 18, 29, 31\}$ . . . . .	21
3-4	Dataset consisting of continuous values . . . . .	22
3-5	Representation of the itemset $I = \{1, 2, 25\}$ using the proposed itemset representation . . . . .	22
3-6	An example of a categorical dataset $I = \{ 'a', 'b', 'j', 'z' \}$ . . . . .	22
3-7	Linked list representation of an Itemset $I = \{1, 3, 5, 12, 17, 30, 31\}$ .	24
3-8	Array representation of an Itemset $I = \{1, 3, 5, 12, 17, 30, 31\}$ . . .	24
3-9	Bitmap representation of an Itemset $I = \{1, 4, 5, 13, 16, 17, 31\}$ . .	25
3-10	Itemset $I = \{ 1, 5, 10, 19, 23, 28, 31\}$ represented using proposed itemset representation . . . . .	26
3-11	Union operation of itemset $I_1 = \{1, 3, 6, 11, 18, 29, 31\}$ and $I_2 = \{1, 2, 6, 13, 18, 30, 31\}$ generates $I_1 \cup I_2 = \{1, 2, 3, 5, 6, 11, 13, 18, 29, 30, 31\}$ . . . . .	28

3-12	The time(millisecond) and memory(kilobits) consumption using the proposed itemset representation and the array representation . . .	31
3-13	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the first dataset with support count =2.5%, 5% and 10% . . . .	33
3-14	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the second dataset with support count =2.5%, 5% and 10% . .	33
3-15	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the third dataset with support count =2.5%, 5% and 10% . . .	34
3-16	Time consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the first dataset with support count =2.5%, 5% and 10% . . . . .	34
3-17	Time consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the second dataset with support count =2.5%, 5% and 10% . . . .	35
3-18	Time consumption of candidate itemset generation represented by array and proposed itemset representation using linear search for the third dataset with support count =2.5%, 5% and 10% . . . . .	35
3-19	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the first dataset with support count =2.5%, 5% and 10% . . . .	36
3-20	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the second dataset with support count =2.5%, 5% and 10% . .	37
3-21	Memory consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the third dataset with support count =2.5%, 5% and 10% . . .	37
3-22	Time consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the first dataset with support count =2.5%, 5% and 10% . . . . .	38

3-23	Time consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the second dataset with support count =2.5%, 5% and 10% . . . . .	38
3-24	Time consumption of candidate itemset generation represented by array and proposed itemset representation using binary search for the third dataset with support count =2.5%, 5% and 10% . . . . .	39
4-1	Flowchart of the proposed graph-based algorithm . . . . .	43
4-2	Time(millisecond) consumption of Different Rule Mining Algorithms for Car dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	45
4-3	Memory(kilobits) consumption of Rule Mining Algorithms for Car dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	45
4-4	Time(millisecond) consumption of Different Mining Algorithms for Bitcoin Heist dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	46
4-5	Memory(kilobits) consumption of Different Mining Algorithms for Bitcoin Heist dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	46
4-6	Time(millisecond) consumption of Different Mining Algorithms for Spatial dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	47
4-7	Memory(kilobits) consumption of Rule Mining Algorithms for Spatial dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	47
4-8	Time(ms) consumption of Rule Mining Algorithms for Hydraulic dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	48
4-9	Memory(kilobits) consumption of Different Mining Algorithms for Hydraulic dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	48
4-10	Time(millisecond) consumption of Rule Mining Algorithms for Cancer dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	49
4-11	Memory(kilobits) consumption of Rule Mining Algorithms for Cancer disease dataset of support count of 1%, 2.5 %, 5% and 10% . . . . .	49
4-12	An example of the proposed graph-based algorithm . . . . .	51



5-1	Memory(kbs) consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence=1% and support count= 1%, 2.5%, 5% . . . . .	56
5-2	Memory(kilobits) consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence 2.5% and support count 1%, 2.5 %, 5 % . . . . .	57
5-3	Memory(kilobits)consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence 5% and support count =1%, 2.5 %, 5 % . . . . .	57
5-4	Time(ms) consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence 1% and support count 1%, 2.5% and 5% . . . . .	58
5-5	Time(ms) consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence 2.5% and support count 1%, 2.5 % and 5% . . . . .	58
5-6	Time(ms) consumption of rule generation algorithms using proposed itemset representation for Hunington's dataset with confidence 5% and support count 1%, 2.5% and 5% . . . . .	59
5-7	Memory(kilobits) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence 1% and support count 1%, 2.5% and 5% . . . . .	59
5-8	Memory(kilobits) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence of 2.5% and support count 1%, 2.5% and 5% . . . . .	60
5-9	Memory(kilobits) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence 5% and support count 1%, 2.5% and 5% . . . . .	60
5-10	Time(ms) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence 1% and support count 1%, 2.5% and 5% . . . . .	61

5-11	Time(ms) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence 2.5% and support count =1%, 2.5% and 5%	61
5-12	Time(ms) consumption of rule generation algorithms using array itemset representation for Hunington's dataset with confidence of 5% and support count= 1%, 2.5% and 5%	62
5-13	Memory(kilobits) consumption of rule generation algorithms using array representation for 1 <sup>st</sup> dataset with support count=1%, 2.5%, 5% and 10%	62
5-14	Memory(kilobits) consumption of rule generation algorithms using proposed itemset representation for 1 <sup>st</sup> dataset with support count =1%, 2.5%, 5% and 10%	63
5-15	Memory(kilobits) consumption of rule generation algorithms using array representation for 2 <sup>nd</sup> dataset with support count=1%, 2.5%, 5% and 10%	63
5-16	Memory(kilobits) consumption of rule generation algorithms using proposed representation for 2 <sup>nd</sup> dataset with support count= 1%, 2.5%, 5% and 10%	64
5-17	Memory(kilobits) consumption of rule generation algorithms using array representation for 3 <sup>rd</sup> dataset with support count=1 %, 2.5%, 5% and 10%	64
5-18	Memory(kilobits) consumption of rule generation algorithms using proposed itemset representation for 3 <sup>rd</sup> dataset with support count=1%, 2.5%, 5% and 10%	65
5-19	Time(ms) consumption of rule generation algorithms using array representation for 1 <sup>st</sup> dataset with support count=1%, 2.5%, 5% and 10%	65
5-20	Time(ms) consumption of rule generation algorithms using proposed itemset representation for 1 <sup>st</sup> dataset with support count=1%, 2.5%, 5% and 10%	66

5-21	Time(ms) consumption of rule generation algorithms using array representation for $2^{nd}$ dataset with support count= 1%, 2.5%, 5% and 10% . . . . .	66
5-22	Time(ms) consumption of rule generation algorithms using proposed representation for $2^{nd}$ dataset with support count= 1%, 2.5%, 5% and 10% . . . . .	67
5-23	Time(ms) consumption of rule generation algorithms array representation for $3^{rd}$ dataset with support count =1%, 2.5%, 5% and 10% . . . . .	67
5-24	Time(ms) consumption of rule generation algorithms proposed itemset representation for $3^{rd}$ dataset with support count=1%, 2.5%, 5% and 10% . . . . .	68

# List of Tables

3.1	Time Complexity of Array and proposed itemset representation . . .	26
3.2	Number of generated itemsets using Array and Proposed itemset representation . . . . .	30
3.3	Number of itemsets generated with maximum length using Array and Proposed itemset representation from different Dataset . . . . .	31
4.1	Time(millisecond) and Memory(kilobits) consumption using Apriori, FP-Tree, and Proposed graph-based algorithm for the different datasets . . . . .	50
5.1	Number of generated rules represented by Array and Proposed itemset representation using NBG's Algorithm . . . . .	70

# Glossary of Terms

FP	Frequent Pattern
SETM	Set-oriented mining
ARM	Association Rule Mining
CSA	Crow Search Algorithm
DHP	Direct Hashing and Pruning
ITARM	Incremental Temporal Association Rules Mining
OCD	Offline Candidate Determination
NDCSA-CAR	New Discrete Version of the Crow Search Algorithm (CSA)
GA-PPARM	Genetic Algorithm Privacy Preserved Association Rule Mining
IFTARMFGT	Incremental Fuzzy Temporal Association Rule Mining using Fuzzy Grid Table
TM	Transaction Mapping
DIC	Dynamic Itemset Counting
NDI	Non Derivable Itemset
EHR	Electronics Health Record
CAG	Cytosine-Adenine-Guanine
BD	Behcet Disease
FSFIM	Fuzzy Set-Based Frequent Itemset Mining
UB	Upper Bound
LB	Lower Bound

# Symbols and Notations

$C_k$	Candidate itemset
$L_k$	Large itemset
$I_k$	Itemset
$t_k$	Transaction
$v_k$	Vertex
$e_k$	Edges
$G$	Graph