**VIETNAM GENERAL CONFEDERATION OF LABOR**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**REPORT**

**DESIGN AND ANALYSIS OF ALGORITHMS**

*Instructor*: **NGUYEN CHI THIEN**

*Student*: **BUI ANH PHU - 521H0508**

**BUI HAI DUONG - 521H0220**

**NGUYEN HOANG PHUC - 521H0511**

*Class*: **21H50302**

**HO CHI MINH CITY, 2023**

**VIETNAM GENERAL CONFEDERATION OF LABOR**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**REPORT**

**DESIGN AND ANALYSIS OF ALGORITHMS**

*Instructor*: **NGUYEN CHI THIEN**

*Student*: **BUI ANH PHU - 521H0508**

**BUI HAI DUONG - 521H0220**

**NGUYEN HOANG PHUC - 521H0511**

*Class*: **21H50302**

**HO CHI MINH CITY, 2023**

ACKNOWLEDGEMENT

To complete this essay, besides our own efforts, we have received a lot of help in terms of knowledge, experience, and skills from the school and teachers. First and foremost, we would like to express my special gratitude to Nguyen Chi Thien - the lecturer of Design and Analysis of Algorithms course who has taught us valuable knowledge of the subject. That knowledge is the foundation for us to continue learning and effectively apply it to this essay. Additionally, We would like to thank the teacher for allowing us to complete this essay, which has helped us to further develop my understanding of the subject. Thank you for your guidance and support in helping us to complete this essay to the best of my ability. Moreover, we would also like to express our gratitude to the school and the teachers who have compiled the Design and Analysis of Algorithms materials, providing me with useful resources for research and essay writing. Thank you sincerely!

**COMPLETION OF THESIS**

**AT TON DUC THANG UNIVERSITY**

We here by certify that this thesis is my/our own work and was conducted under the guidance of Nguyen Chi Thien. The research and results presented in this thesis are truthful and have not been published previously in any form. The data presented in tables and figures used for analysis, comments, and evaluations were collected by the author from various sources and are clearly cited in the reference section.

Moreover, this thesis includes some comments, evaluations, and data from other authors and organizations, which are properly cited and referenced.

If any misconduct is detected, I fully take responsibility for the content of my thesis. Ton Duc Thang University is not liable for any copyright infringement that may occur during the thesis completion process.

*Ho Chi Minh City, October 22, 2023*

*Author*

*(signature and full name)*

ACKNOWLEDGEMENT AND EVALUATION SECTION BY INSTRUCTOR

**Instructor's Acknowledgement Section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ho Chi Minh City, 2023

(signature and full name)

**Instructor's Evaluation Section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ho Chi Minh City, 2023

(signature and full name)

SUMMARY

**TABLE OF CONTENTS**

[ACKNOWLEDGEMENT i](#_Toc153060725)

[ACKNOWLEDGEMENT AND EVALUATION SECTION BY INSTRUCTOR iii](#_Toc153060726)

[SUMMARY iv](#_Toc153060727)

[**TABLE OF CONTENTS** 1](#_Toc153060728)

[CHAPTER 1 – EFFICIENT WEIGHTED PROBABILISTIC FREQUENT ITEMSET MINING IN UNCERTAIN DATABASES THEORY BASED ON THE STUDY DOCUMENT 1](#_Toc153060729)

[1. Problem definitions 1](#_Toc153060730)

[2. Definitions 1](#_Toc153060731)

[Definition 1 (Uncertain dataset): 1](#_Toc153060732)

[Definition 2 (Possible world): 2](#_Toc153060733)

[Definition 3 (Expected support-based frequent itemset): 3](#_Toc153060734)

[Definition 4 (Probabilistic frequent itemset): 3](#_Toc153060735)

[Definition 5 (Weighted probabilistic frequent itemset): 3](#_Toc153060736)

[Definition 6 (Itemset weight): 4](#_Toc153060737)

[3. Theory 4](#_Toc153060738)

[Theorem 1 (Equivalence between PFI and w-PFI): 4](#_Toc153060739)

[Theorem 2 (Anti-monotonicity property for PFI): 4](#_Toc153060740)

[Theorem 3 (Anti-monotonicity property for weighted PFI): 5](#_Toc153060741)

[Theorem 4: 5](#_Toc153060742)

[Theorem 5: 6](#_Toc153060743)

[Theorem 6: 6](#_Toc153060744)

[4. Algorithm 7](#_Toc153060745)

[Algorithm 1: 7](#_Toc153060746)

[Algorithm 2: 7](#_Toc153060747)

[Pseudocode given by the paper: 7](#_Toc153060748)

[Algorithm 3: 8](#_Toc153060749)

[CHAPTER 2 – JAVA CODE IMPLIMENTATION FOR EFFICIENT WEIGHTED PROBABILISTIC FREQUENT ITEMSET MINING IN UNCERTAIN DATABASES 10](#_Toc153060750)

[1. UncertainDatabase class 10](#_Toc153060751)

[2. wPFIItem class 10](#_Toc153060752)

[3. wPFIItemset class 10](#_Toc153060753)

[4. wPFIApriori class 10](#_Toc153060754)

[5. Implementation Benchmark 10](#_Toc153060755)

[**REFERENCES** 14](#_Toc153060756)

**LIST OF SYMBOLS AND ABBREVIATIONS**

**Symbols**

Uncertain dataset of size n

A set of transaction identifiers

An itemset of size m

An existential probability for item appearing in the transaction

A weight table for the itemset I. Each item has a real-valued weight

An integer between (0, n], the minimum support for expected support- based FI mining

**msup** An integer between (0, n], the minimum support for probabilistic FI mining

A real value between (0, 1], the probabilistic frequent threshold for probabilistic FI mining

A real value between [0, 1], the scale factor

**Abbreviations**

**PFI** probability frequent itemset

**w-PFI** weighted probability frequent itemset

CHAPTER 1 – EFFICIENT WEIGHTED PROBABILISTIC FREQUENT ITEMSET MINING IN UNCERTAIN DATABASES THEORY BASED ON THE STUDY DOCUMENT

* 1. Problem definitions

Uncertain data mining has attracted so much interest in many emerging applications

over the past decade. Several probability models are presented to measure the frequency of an itemset in the dataset, and it is noted that the frequency itself cannot identify useful or meaningful patterns in some scenarios. Therefore, weighted (importance) frequent itemset mining in uncertain databases has been done in some studies but the result is still inefficient. To overcome this issue, we introduce the new algorithms and some pruning methods for narrowing space and improve the candidate generation for the result.

* 1. Definitions

In this section, there are all the fundamental definitions mentioned in the study document, which are needed before dealing with the problem of finding efficient weighted probabilistic frequent itemset mining in uncertain databases.

* + 1. Definition 1 (Uncertain dataset):

An uncertain dataset DB consists of a set of transactions DB = , Each transaction contains a subset of an itemset I = {}. There is an existential probability for each item in the transaction of DB.

For example:

DB = {}

= {(Milk, 0.4), (Fruit, 0.9), (Video, 0.6)}

= {(Milk, 0.3), (Fruit, 0.7)}

With:

DB is the uncertain database.

is the transaction identifier of DB.

In Itemset contains items Milk, Fruit, Video with the corresponding probability 0.4, 0.9, 0.6.

* + 1. Definition 2 (Possible world):

A possible world is the subset of domain . With the existential probability can be calculated by the formular:

.

For example:

We have DB = {}

= {(Milk, 1.0)}

= {(Milk, 0.3), (Fruit, 1)}

|  |  |  |
| --- | --- | --- |
| W | Tuples in W | Probability |
| 1 | {Milk}{Fruit} | 0.7 |
| 2 | {Milk}{Milk, Fruit} | 0.3 |

Apply the algorithm for calculating the probability:

* + 1. Definition 3 (Expected support-based frequent itemset):

An itemset is an expected support-based frequent itemset if and only if:

With:

) is the support (occurrence) of the itemset in the possible world ..

is the minimum expected support.

For example:

* + 1. Definition 4 (Probabilistic frequent itemset):

An itemset is a probabilistic frequent itemset if and only if:

With:

is the probabilistic frequent threshold.

is the minimum support.

is the support (occurrence) of the itemset

For example:

Set the , =

* + 1. Definition 5 (Weighted probabilistic frequent itemset):

Weighted probabilistic frequent itemset is the probabilistic frequent itemset product with the weight of the itemset :

With:

is the weight of the itemset

is the minimum support.

is the support (occurrence) of the itemset

For example:

* + 1. Definition 6 (Itemset weight):

The weight of X is the average weight of the items in the itemset X:

Itemset weight calculated by the formula:

For example:

We have the itemset = {(Milk, 0.4), (Fruit, 0.9), (Video, 0.6)}

The weight of X:

* 1. Theory

In this section, here are the theorems mentioned in the document which have been used in the research. Some of these theorems are mainly used for the candidate generating and pruning algorithm, which will be present below soon.

* + 1. Theorem 1 (Equivalence between PFI and w-PFI):

Given a probabilistic frequent threshold t, a under if and only if under .

The theorem 1 represent the connection between the and the based on the weight.

* + 1. Theorem 2 (Anti-monotonicity property for PFI):

If an itemset is a , then any itemset is also a .

This is the pruning method for the PFI for pruning the search space by eliminating any candidate itemset that contain a non-PFI subset. By using this method, it can help to decrease items added to the itemset.

* + 1. Theorem 3 (Anti-monotonicity property for weighted PFI):

If an itemset is a , with the length of the set is k, there is at least one itemset with the length of the set is is a .

From which has been said in the paper, the anti-monotonicity property for the weighted PFI is different from the PFI. Therefore, they decided to design the new novel candidate generation items in the set without having the combination of the duplicated items.

Corollary 1:

The Corollary 1 based on the theorem 3:

We have the and where is the set of the wPFI with the size , is the item have the smallest weight in the itemset .

With , is not a if and ,

where and consist of all size 1 item of .

* + 1. Theorem 4:

We have is the random variable, following the Poisson Binomial Distribution.

With:

is the probability of the itemset , where is the subset of transaction

is the mean in the Poisson Binomial Distribution.

is the variance in the Poisson Binomial Distribution.

Frequentness probability of an itemset X calculated by this formular:

With:

F(.) is the cumulative distribution function of Poisson distribution.

* + 1. Theorem 5:

The frequentness probability increases monotonically with

Following the paper (Wang et al., 2010), an itemset X is more likely to be PFI if it has higher .

* + 1. Theorem 6:

We have two itemset and .

~PBD(), where the mean and variance

~PBD(), where the mean and variance

, where the mean and the variance

Based on the theorem:

Corollary 2:

The Corollary 2 based on the Theorem 6:

We have the itemset is not a if by applying from the Theorem 6

With:

The

Corollary 3:

We also have the Corollary 3

We have the itemset is not a if

With:

The .

is the number of transactions.

* 1. Algorithm
     1. Algorithm 1

Pseudocode given by the paper:

The algorithm 1 is implemented by applying the corollary 1.

A computer screen shot of a computer code

Description automatically generated

Understanding the algorithm:

The algorithm starts by finding all candidates. This is done by simply scanning the database and keeping the items that have a support of at least the minimum support threshold.

Once the set of candidates has been found, the algorithm proceeds to generate all possible candidates by joining the size-k candidates with each other. The algorithm then prunes the candidate set by removing any candidates that are not downwardly closed. A candidate is downwardly closed if it contains all its subsets.

The remaining candidates are then scanned in the database to count their support. Any candidates with a support of at least the minimum support threshold are added to the set of . The algorithm then repeats until no more WPFI candidates can be generated.

* + 1. Algorithm 2

Pseudocode given by the paper:

The algorithm 2 is implemented by applying the corollary 2.

A screenshot of a computer program

Description automatically generated

Understanding the algorithm:

Suppose we have a database of transactions, where each transaction contains a set of items, and each item has a weight. We want to find all w-PFI with a minimum support threshold of 0.5.

The algorithm would then iterate over each itemset X in the database and compute the w-PFI support of X using the Poisson distribution and the mean and scale of the weight of the items in X. Suppose that the w-PFI support of a particular itemset X is 0.6.

Since the w-PFI support of X is at least the minimum support threshold of 0.5, X would be added to the set of w-PFI candidates.

The algorithm would then iterate over each w-PFI candidate X in the set of w-PFI candidates and prune any candidates that do not reach the threshold.

The resulting set of w-PFI candidates would contain all itemset that have a w-PFI support of at least 0.5.

* + 1. Algorithm 3

Pseudocode:

The algorithm 3 is based on the corollary 3.

A screenshot of a computer program

Description automatically generated

Understanding the algorithm:

CHAPTER 2 – JAVA CODE IMPLIMENTATION FOR EFFICIENT WEIGHTED PROBABILISTIC FREQUENT ITEMSET MINING IN UNCERTAIN DATABASES

1. **UncertainDatabase class**

**Introduction:** The UncertainDatabase class represent for the uncertain database with existential probabilities. This class is implemented based on the definition 1.

**Detail:**

1. **wPFIItem class**

**Introduction:** The wPFIItem class is implement for the representation of the item in the uncertain database. The class is created based on the definition 1 of the definition mentioned in the chapter 1.

**Detail:**

1. **wPFIItemset class**

**Introduction:** The wPFIItemset class is implemented for the representation of the itemset which is contained in the transaction of the uncertain dataset. The class is created based on the definition 1 of the definition mentioned in the chapter 1.

**Detail:**

1. **wPFIApriori class**

**Introduction:** The wPFIApriori class is implemented for the representation of 3 algorithms mentioned in the theory in chapter 1. This also call 3 java classes definition above in the chapter 2 for the use of the algorithm function.

**Detail:**

1. **Implementation Benchmark**

Here is the result when apply the implementation of the Algorithm 1 and 2.

**With:**

Data:

* Connect: contains 67,557 transactions with 129 distinct items, which are prepared by Roberto Bayardo from UCI Machine Learning Repository

Runtime: 28703ms

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**REFERENCES**