COMPARISON OF ASSOCIATION ALGORITHM' S EFFICIENCIES BETWEEN APRIORI AND FP-GROWTH ALGORITHMS

BANTHITA TIPJAKSU

A THEMATIC PAPER SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015

Thematic Paper
entitled

COMPARISON OF ASSOCIATION ALGORITHM' S EFFICIENCIES BETWEEN APRIORI AND FP-GROWTH ALGORITHMS

……..……..……
Miss Banthita Tipjaksu
Candidate

……..……..……
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Major advisor

……..……..……
Asst. Prof. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Co-advisor

……..……..……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……..……..……
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Program Director
Master of Science Program in Technology of Information System Management
Faculty of Engineering
Mahidol University

Thematic Paper
entitled

COMPARISON OF ASSOCIATION ALGORITHM' S EFFICIENCIES BETWEEN APRIORI AND FP-GROWTH ALGORITHMS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
May 30, 2015

……
Miss Banthita Tipjaksu
Candidate

……
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Chair

……
Asst.Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

……
Asst. Prof Kairoek Choeychuen,
Ph.D. (Electrical and Computer Engineering)
Member

……
Asst. Prof. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Co-advisor

……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……
Lect. Worawit Israngkul,
M.S. (Technical Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

This thematic paper would not be success without support, advices, ideas, consultation and inspection from many people. I would like to thank my kind advisor Asst. Prof. Supaporn Kiattisin and Asst. Prof. Dr. Waranyu Wongseri. They give recommendation and suggestion. This thematic paper is not completed without the support and invaluable help. I am extremely grateful to both of advisors for their assistance here.

Moreover, I would like to thank all of friends, brothers, sisters and My teachers in Master's degree in Information Management Technology Program who always provide assistance and support.

Finally, I am most grateful my father, mother and relatives in my family who are the great background to be what I am today. I do appreciate them all in this opportunity.

Banthita Tipjaksu

COMPARISON OF ASSOCIATION ALGORITHM'S EFFICIENCIES BETWEEN APRIORI AND FP-GROWTH ALGORITHMS

BANTHITA TIPJAKSU 5537931 EGTI/M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THEMATIC PAPER ADVISORY COMMITTEE: SUPAPORN KIATTISIN, Ph.D., WARANYU WONGSERI, Ph.D.

ABSTRACT

This research compares the Apriori algorithm and FP-growth algorithm in terms of the Association algorithm's efficiency outcomes. In order to investigate the processing time for finding the association rules for each algorithm, we assigned the parameters of both support and confidence values to be equal. The Apriori algorithm generates the association rules by searching the frequency of itemsets. Furthermore, the Apriori algorithm is used to replicate the search as close as the level-wise search. In order to reseach itemsets each time, the entire database must be scanned. On the other hand, the FP-growth algorithm generates the association rules by creating the frequent itemsets without the candidate itemsets. Moreover, the FP-growth algorithm uses the data compression of the database in the FP-tree process to avoid duplicated reading data. The data used in the study is the shopping data from Supermarket, which was imported to the Weka program to find the association rules. According to the experiment, it was found that both the Apriori algorithm and FP-growth algorithm are not different in terms of the association rules. Likewise, the FP-growth algorithm is better than the Apriori algorithm in terms of processing time.

KEY WORDS: ASSOCIATION RULE/ APRIORI ALGORITHM/ FP-GROWTH ALGORITHM/ MINSUPPORT/ MINCONFIDENCE

51 pages

การเปรียบเทียบประสิทธิภาพของอัลกอริทึม Association ระหว่าง อัลกอริทึม Apriori และอัลกอริทึม FP-Growth

COMPARISON OF ASSOCIATION ALGORITHM' S EFFICIENCIES BETWEEN APRIORI AND FP-GROWTH ALGORITHMS

บัณฑิฏา ทิพย์จักษุ 5537931 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาสารนิพนธ์ : สุภาภรณ์ เกียรติสิน, Ph.D., วรัญญู วงศ์เสรี, Ph.D.

บทคัดย่อ

การวิจัยนี้ ศึกษาเรื่องการเปรียบเทียบประสิทธิภาพของอัลกอริทึม Association ระหว่างอัลกอริทึม Apriori และอัลกอริทึม FP-Growth โดยจะกำหนดค่าสนับสนุน (Support Value) และค่าความเชื่อมั่น (Confidence Value) ซึ่งเป็นพารามิเตอร์ที่ใช้วัดประสิทธิภาพของกฎความสัมพันธ์ให้มีค่าเท่ากัน เพื่อดูเวลาที่ใช้ในการค้นหาความสัมพันธ์ของแต่ละอัลกอริทึม โดยที่อัลกอริทึม Apriori จะสร้างกฎความสัมพันธ์ด้วยการหาความถี่ของไอเท็มเซ็ต อาศัยการทำซ้ำใกล้เคียงกับวิธีค้นหาแบบ Level-wise ในการหาไอเท็มเซ็ตแต่ละครั้งต้องค้นหาข้อมูลทั้งหมดที่ถูกสร้างขึ้น ส่วนอัลกอริทึม FP-Growth จะสร้างกฎความสัมพันธ์ด้วยการสร้างไอเท็มเซ็ตที่ถูกค้นพบบ่อยครั้ง โดยไม่ต้องสร้างทั้งหมด นอกจากนี้ขั้นตอนวิธี FP-Growth ยังใช้การบีบอัดข้อมูลจากฐานข้อมูลให้อยู่ในรูปแบบของ FP-Tree เพื่อหลีกเลี่ยงการอ่านข้อมูลจากฐานข้อมูลซ้ำหลายรอบ โดยข้อมูลที่ใช้ในการศึกษา คือ ข้อมูลซุปเปอร์มาร์เก็ต ซึ่งเป็นข้อมูลเกี่ยวกับการซื้อสินค้าของลูกค้าในซุปเปอร์มาร์เก็ต นำเข้าโปรแกรม Weka เพื่อหากฎความสัมพันธ์ของอัลกอริทึม Apriori และอัลกอริทึม FP-Growth จาก ผลการวิจัยพบว่า จำนวนกฎของทั้งสองอัลกอริทึมไม่ต่างกันและเวลาที่ใช้ในการหากฎความสัมพันธ์ของอัลกอริทึม FP-Growth น้อยกว่าอัลกอริทึม Apriori

51 หน้า

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

1.1 Statement of the problems

Association Rule is alternative technique, which is used in Data Mining by searching the interesting relations between itemsets. It should lead to find data relations of huge data to apply on analysis and predictions. The Association Rule is widely used in various fields, for instance, scientific experiment, disease treatment, and prediction of natural phenomenon. However, it is mostly popular to concurrently use with Market Basket data in order to conduct the Market Basket Analysis. Moreover, being aware of consumer behavior, whether there is any whether concurrent purchase, type of goods or products were parallel purchase as business implementing plan. It is better utilize to increase sales volume and marketing preparation as well as sales promotions and products arrangement. The Market Basket was the large transaction data and each transaction should be replaced by any current situation. Therefore, efficient algorithm should be utilized to find the Association Rule within limited time period in order to obtain the reliable Association Algorithm that could be implemented on the market analysis to assist on business making decision process.

There are various methods to find Association Algorithm, for instance, Apriori Algorithm and FP-Growth Algorithm which procedure and process of both methods. Both are different as Apriori Algorithm can generate the Association Rule by searching frequency itemsets. Apriori Algorithm requires closing iteration of level-wise search as all database must be scanned in each itemset searching which took time. However, Apriori Algorithm attributes to reduce search space by using its property. It is called anti-monotone, which mean if any set fails the test, all superset should be failed accordingly. While FP-Growth Algorithm is a creation of Association Rule by building frequent itemset without applying a candidate itemsets. On the other hand, data compression of database is utilized in terms of FP-Tree to avoid repeated

searching data from database, minimizing searching time. As search space is smaller but it is not appropriate for enormous database due to high amount of data distribution and FP-Growth structure requires much memory space.

There are many Algorithms application on searching for the Association Rule. For example, both Algorithms mentioned previously, which is the Apriori Algorithm and FP-Growth Algorithm. It is found that both Algorithms have procedures, advantages, disadvantages as well as different limitation and there is no significant conclusion concerning efficiency of both the Algorithms. Thus, a concept to compare efficiency of both Association Algorithms to indicate which Algorithm being the most efficient for function in order that it can utilize the appropriate Algorithm for the data analyzing to find the Association Rule.

1.2 Objectives

1.2.1 To compare the efficiency of Association Algorithm between Apriori Algorithm and FP-Growth Algorithm.

1.3 Expected benefits

1.3.1 Being aware of efficiency of the Association Algorithm by implementing Apriori and FP-Growth Algorithms.

1.4 Scope of the research

1.4.1 To compare efficiency of Apriori and the FP- Growth Algorithms by using support and confidence as a parameter.

CHAPTER II
THEORIES AND RELATED WORKS

2.1 Theories

2.1.1 Data Mining

Data Mining is a procedural process implemented numerous data in order to find forms and relations hidden on data set. Data Mining should explore and analyze data, which is in meaningful form and in form of rule as this relation should reveal useful knowledge in database. The most business organizations currently encounter problem of difficulty in taking abundant raw data to be actually applied. Data Mining is expected equipment as a well-known and widely instrument which applied as Data Mining can reveal collected and hidden knowledge from a great deal of data.

Evolution of Data Mining

- In 1960, Data Collection is an appropriate data collection in reliable equipment and well protection of data loss.
- In 1980, Data Access is the creation of interrelation of collected data utilized in efficient analysis and making decision.
- In 1990, Data Warehouse and Decision Support are the method of data collection stored in large and comprehensive database of the organization to support decisions making.
- In 2000, Data Mining is an analysis and processing of data obtained from database by building models and statistical relations.

Functioning cycle of Data Mining consists of 4 main steps as follows:

1. Identifying business opportunity or business problems as scope of data used in analysis to find the marketing edge or problem solutions should be identified.

2. In part of Data Mining, technique of Data Mining is transfered. It is that raw data should be transformed to be data which could be actually utilized in business.

3. Compliance with data is taking outcome of Data Mining to be actually practiced in business.

4. Measurement of outcome efficiency of Data Mining should be conducted in various methods such as measuring from market share, measuring from quantity of customers, and net profit.

The 4 steps as mentioned above are taking Data Mining to be used on business system as each step should rely on outcome of another step. They become an input of the next step and Data Mining should transform raw data to be applied data. Therefore, the identification of proper data source is very important to the outcome that obtained from such analysis.

Data preparation for Data Mining :

The first thing to be concerned what field that Data Mining technique should be applied, why and what form of knowledge is required by utilizing Data Mining. Suppose the data mining technique is required to apply in education. It is that educational institutes currently have student's data stored for a long period of time. Nonetheless, the most data should be utilized merely during studying period of students; a data was well stored and rarely utilized after their graduation. After that, the Data Mining technique is used to apply in education with student data, the next step is finding the Mining Objective to identify the knowledge. We require searching by applying Data Mining with student data.

If we require utilizing the Data Mining to assist students to select fields of study for students in Faculty of Engineering as there are more than 10 fields of study and most of students who study in Faculty of Engineering. Once, the

time selects fields of study they did not know their abilities to select which academic disciplines to have an opportunity for the most success. Therefore, it is appropriate for us to apply Data Mining in data base of student by using knowledge obtained from Data Mining to assist student to select fields of study. We have an objective to conduct Data Mining. We have to collect student data and it is assumed that we have backdated student data for 10 years. It is devided into two parts, the first part is personal information of students as specified on Table 2.1 and the second part is information of enrollments in each subject of students as specified on Table 2.2.

Table 2.1 Example of student's personal information.

ID	Sex	Name-Surname	Address	SchoolGPA	Major	GPA
1	Mr.	Viroj Pattanakul	86/9 Moo 2	2.5	Electrical	2.3
2	Ms.	Duangporn Aimsook	54/2 Moo 7	3.4	Civil	3.2

From Table 2.1, It is an example of personal information of student; such as ID code, name, sex, nationality, address, birth date, family status, GPA in high school, major and current GPA.

Table 2.2 Example of information of student's enrollment.

ID	Subject	Section	Team	Year	Grade
1	001	1	1	2537	C+
1	002	1	1	2537	D
1	005	1	1	2537	B+

From Table 2.2, It is information of enrollment of students each subject of each semester including sections and grades of each subject of students. After that we collected all data requirement, the next step is data preparation to conduct Data Mining which divided into the following steps:

Steps of Data Mining

It consists of sub-function which transforms raw data to be knowledge as follows:

- Data Cleaning is step of data selection by eliminating unrelated data.
- Data Integration is step of data collection from various sources to be the same data set.
- Data Selection is step of data retrieval form recorded source for analysis.
- Data Transformation is step of data transformation to be appropriate for function.
- Data Mining is step of searching useful pattern from existed data.
- Pattern Evaluation is step of pattern evaluation obtained from Data Mining conduct.
- Knowledge Representation is step of presentation of the obtained knowledge by using comprehensible presentation techniques.

2.1.2 Data Cleaning

The obtained data is merely raw data which cannot be utilized through Data Mining process. Therefore, it is required to manage such data and initial preprocessing data was in the following:

Selecting merely important columns expected to be utilized and columns with rather complete data when compared to number of students specified on Table 2.1. The important and complete data columns were consisted of ID Code of students, address, age, sex, family records, and high school and graduated GPA. While important with less data shall not be considered, for instance, entrance examination scores of each subject and reasons of entrance examination.

For column with the same value in every row such as “Thai nationality” is the data which being not distinguished differences of each row though, therefore, it is not utilized to conduct Data Mining, the columns shall not be considered. For columns with different values on Table 2.1, such as name of parents and telephone numbers,

row with data interrelation. It cannot be found in these data which is not utilized in conducting Data Mining; therefore, unrepeated data should eliminate.

The Data Cleaning is correcting blank value of data (NULL). Correction was made in various methods, such as correcting by eliminating data in row which is NULL. From example on Table 2.2, the data on some rows on column Grade are missing and it is seen merely ID Code and enrolled subjects without grade data, the row is not considered to find an interest relation at all.

Modifying data to have appropriate value for making decision, from the Table 2.1 addressed the data cannot be utilized directly. There is problem as specified in 1.3 because the address of each student is not repeated at all, data modification is required in pattern which can be utilized. In this case, data modification of address column should be Bangkok and Non-Bangkok.

Data grouping is required in order to reduce Binning Data as data of student is not so much but Grades in each subject which should be obtained within 10 Grades {A, B+, B, C+, C, D+, D, F, W, I}. As the results, Binding Data of Grade data of student must be reduced. It is much when comparing to number of students and Grade data grouping of student. It is classified into 3 groups as follow: Grade {A, B+, B} is High, Grade {C+, C} is Medium and Grade {D+, D, F, W, I} is Low. From the Table 2.1 which is personal information of students, we have modified some part of data to be complete as follows:

- Elimination of unnecessary columns in conducting Data Mining, such as student names as each name of students cannot be utilized in Data Mining conduct.
- Selecting merely expected to be conducted Data Mining such as high school as there are lots of schools of each student. We have to modify school data in grouping in balance, which can be utilized in Data Mining by dividing school data into 2 groups which are M.6 Equivalence Test and M.6 graduation. It is defined school = 0 means M.6 graduation and School = 1 means M.6 Equivalence Test.
- Modifying data in some columns which can be implementing in Data Mining such as address column should be modified as this group of students stays in Bangkok or Non-Bangkok. The outcome of Data Cleaning obtained from the Table 2.2 is specified on Table 2.3.

Table 2.3 Example of student's personal information after Data Cleaning.

ID	Sex	Address	School	…	Major	GPA
1	Female	Bangkok	1	…	ELEC	2.3
2	Male	Non-Bangkok	0	…	CIVIL	3.2

From the Table 2.3, which is information of enrollment of student which some data has been conducted Data Cleaning such as:

- Elimination of some unnecessary columns which cannot be utilized in Data Mining such as learning group column.
- Grouping data in Grade column in order to reduce binding data.

The outcome of Data Cleaning obtained from the Table 2.2 is specified on Table 2.4.

Table 2.4 Example of information of student's enrollment after Data Cleaning.

ID	Subject	Team	Year	Grade
1	001	1	2537	Medium
1	002	1	2537	Low
1	005	1	2537	High

2.1.3 Data Selection

We have to select simply students' data which can be utilized such as selecting specifically students of Faculty of Engineering and all subjects that studied by students (in the same subject as we have obtained backdated data for 10 years and data of subjects in the past). It may be different from current subjects due to difference of annual program. As the result, we have to select merely student data in the year which has the same subject.

Selection of student's data in academic disciplines which can conduct Data Mining in 6 disciplines, for example, Chemical Engineering, Civil Engineering,

Computer, Engineering, Electrical Engineering, Industrial Engineering and Mechanical Engineering. The reasons are that we select these 6 academic disciplines

because they are the main disciplines which have both students can be analyzed. While the newly established disciplines that are not selected information may be insufficient for analysis and it may cause error in the test. After we have complied with all previous steps, the obtained data should be more completed.

2.1.4 Data Transformation

It can be seen on Table 2.4 that the obtained data is in level of subjects for being consistent to the objective which we require studying behavior and characteristic of each student. We have to modify data in the level of students by dividing into group of subjects that enrolled by ID Code of student and columns to replace list of subjects and then we combine the Table 2.3 and the Table 2.4 which should become table of preliminary data of student and each row of the table displayed personal information of student and the obtained Grade in each subject of students. This table should adjusted appropriately for other techniques of Data Mining further. All outcomes are displayed on Table 2.5.

Table 2.5 Example of student's preliminary data of students.

ID	Sex	Address	001	002	…	Major	GPA
1	Male	Bangkok	Medium	Low	…	ELEC	2.3
1	Female	Non-Bangkok	High	High	…	CIVIL	3.2

It is believed that preprocessing data in Table 2.5 is in complete pattern and be prepared to conduct Data Mining. In contrast, we have to modify pattern of data appropriately for each technique of Data Mining which is selected to utilize.

Conclusion: Data Mining was searching all relations and patterns which are existed in database that hidden within abundant data. The Data Mining is appropriated for solving merely some problems such as problems that required rational solution or economic or financial problems. Data Mining has several techniques used as solution but there is no technique that can be solved all problems of Data Mining, therefore, diversity of techniques was significant that shall lead to find the best solution of Data Mining.[1]

2.1.5 Component of Data Mining system

Architecture of Data Mining system consists of key components as shown in Figure 2.1.

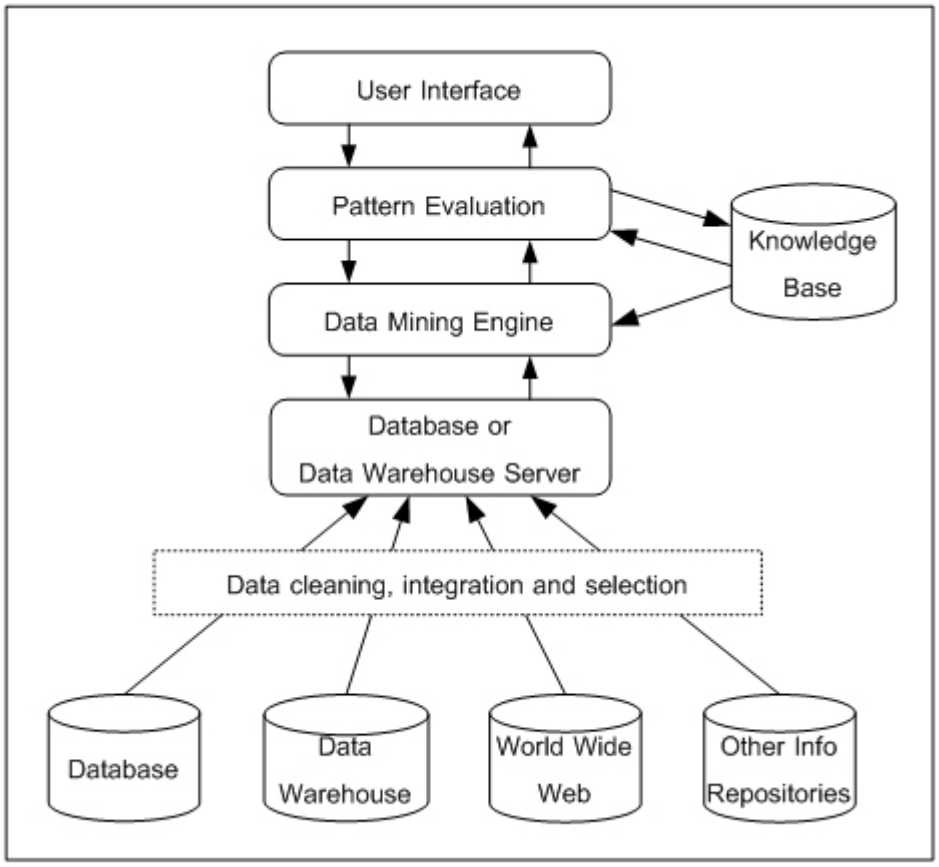

Figure 2.1 Architecture of Data Mining system.[2]

- Database, Data Warehouse, World Wide Web and Other Info Repositories are data sources of Data Mining.
- Database or Data Warehouse Server inputs data due to user's request.
- Knowledge Base is specific knowledge in part which is beneficial for searching or evaluation of the interested outcome pattern that obtained.
- Data Mining Engine is the main component which is consisted of module being responsible for Data Mining in various data such as searching for Association Rule, classification and grouping.

- Pattern Evaluation Module is functioning concurrently with Data Mining server by measuring interest in filtering the obtained outcome pattern that rendered searching specifically interested pattern.

- Graphic User Interface is coordination part between user and Data Mining system which assisted user to identify task required by Data Mining, data retrieval or data storing structure and evaluation of the obtained outcome.

2.1.6 Type of data required by Data Mining

- Relational Database is a database stored in term of table as each table shall be consisted of rows, columns and relations of all data. The data is displayed by using Entity Relationship Model.

- Data Warehouses is data collection from various sources and stored in the same pattern and the same location.

- Transactional Database is consisted of data; each transaction which is replaced by any situation on any moments. For instance, receipt; it is collected data in term of customers' name, and list of purchased products.

- Advanced Database is another database that stores data in other patterns, for example, Object-Oriented data, Text File data, multimedia data and data on Web pattern.

2.1.7 Web Mining

Web Mining is application of Data Mining on searching for the data, interrupting data and information document from web site. It is automatically providing services on the web site to take the acquired knowledge to solve the required problems directly and indirectly. In addition, classification of the Web Mining by consideration of the analyzed data which divided into 3 types such as Web Content Mining, Web Structure Mining and Web Usage Mining:

The Web Content Mining is searching useful data on website; such as messages, or images. The Web Content Mining is divided into 2 types; Information Retrieval and Database and objective. The Web Content Mining from information retrieval is conducting the Web Mining in order to improve searching data or filtering data for users by considering data referred or requested by user. Whereas, an objective

of Web Content Mining in database viewpoint, is mostly trying to reproduce data on the web and collecting data to reduce enquiries being more functional rather than using main words as specific searcher.

The Web Structure Mining is a method that tried to search for a pattern of connection structure which is significant and hidden on the website. The pattern shall rely on a pattern of document connection on the website. The acquired pattern classifies on web page group and generates suitable information data in order to utilize on the web structure adjustment which provided services to users rapidly.

The Web Usage Mining is a method that tried to find the meaning of generated data during each functional period of user or generating from user behavior. It is called Web Log Mining during the Web Content Mining and the Web Structure Mining is operating actual data or fundamental data on web. However, the Web Usage Mining shall collect data from functional records, such as functional record of Proxy (Proxy Server Log), Registration Data or other data arisen from co-function for analysis. Hence, the Web Usage Mining is a functional method that emphasized on utilizing technique which can predict user behavior during operating on the website. Functional system of Web Usage Mining is divided into 2 methods as follows:

1. Matching functional data of web service providers in term of relation tables prior to modifying these existing data on the Web Usage Mining.

2. Utilizing data from functional record directly by using Preprocessing Technique to prepare data before finding Pattern Discovery and conducting Pattern Analysis.

2.1.8 Web Usage Mining Techniques

Web Usage Mining Techniques are as follows:

- Statistical Analysis is a method searching knowledge concerned about web users by analyzing part of file. This method can identify the different types of variables statistical analysis; such as, accessed page, time of access, length of web site route, logical instruments and web traffic. They are mostly in term of statistical information report, for instance, frequency of accessed pages, average time of accessed pages or average length of web route. This report is included limited error analysis, for example, searching item which is not eligible to access or searching for

incomplete URL. Knowledge pattern should be very beneficial for system improvement, increasing security of the system, facilitating for the web site improvement and supporting marketing decision making.

- Association Rule is searching for association rule of data by searching association rule of 2 sets of data or more and storing them together. Measurement of significant rules can conduct by using 2 data; such as Support value is a percentage of the proper rule implementation. The second data used for measurement is called Confidence value which is number of the proper rule implementation related to number of rule implementation. There are many methods used to find these relations but the well-known and widely usage is called the Apriori Algorithm.

- Clustering is a group of data which is similar to data classification, but it is different from classification. It is categorized by analyzing prototype data. On the other word, clustering data is categorized by using step of clustering in order to find acceptable group for clustering without considering existed or well-known type of data. This means a group of material is compared similarity and classified as the same clustering.

- Classification is a category of data by searching for prototype data or functional data set to notify and classify type of data. Its objective is utilizing as prototype data to predict type of materials or data which cannot identify its type. This prototype is generated by analysis of Training Data. It might be a data set which has been identified type and category already. Prototype patter shall be identified in several forms such as Classification Rules, Decision Trees or Neural Networks.

Sequential Pattern is a searching technique in term of sequence which is searching for relation of data during transaction and time is related in term of interested sequence. It demonstrates that if there is the event or the data set it is tended to have event or data set consequentially.

- Dependency Modeling is a dependent model which is another useful form of finding in conducting Web Mining. It is objective that is to develop prototype which can be generated by significant agent of dependency among other variables on the web domain, such as, it may require generating model on different locations for user’s accessibility during purchasing products online. There are several possible

learning techniques utilized as model of searching for user behavior; for instance, Hidden Markov Model and concept of Bayesian Belief Network.

2.1.9 Steps of Web Usage Mining

Web Usage Mining process is consisted of 3 steps as follows:

- Preprocessing is a step of data conversion in web being appropriate for searching for relation such as Data Cleaning, Data Filtering.
- Pattern Discovery is a step of conducting Data Mining such as Association Rule Discovery, Sequential Pattern Discovery in order to searching for relations or pattern from web data which has been already preprocessing.
- Pattern Analysis is a pattern of acquiring outcomes to obtain from searching to be analyzed in order to support business decisions making or planning.

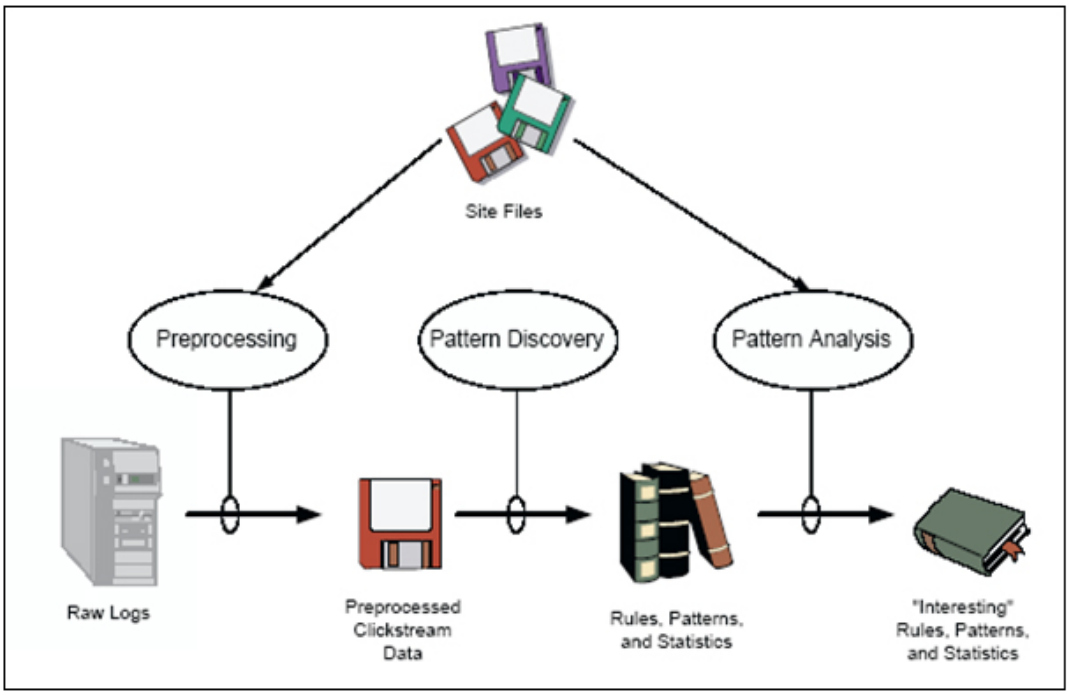

Figure 2.2 Steps of Web Usage Mining (Srivastava, 2000).[2]

Data used to conduct Web Usage Mining shall be obtained from 3 sources as follows:

- Collecting data from record of Server-Side
- Collecting data from record of Client-Side
- Collecting data from record of Proxy Server

Data Mining should apply variously, which is widely classified as 2 main groups for prediction group and explanation group.

Predictive Data Mining is conducted by taking knowledge obtained from existing data to utilize prediction of new data to be found in the future; such as customer's data of Bank Credit Department. Customers have been prioritized who was good customers, intermediate customer or defaulted customer. The Data Mining is aware of these data and searching for model which can notify characteristic of good customers, intermediate customers and undesirable customers. This model can be used to predict new customers who apply for credit. They probably classify as what type of customers.

Descriptive Data Mining is searching for interested pattern form data set. This pattern is usually related or connected to the data which is different from the previous pattern. The users are not expected to use the Data Mining program to search for which pattern or which model but they search for interested pattern form of data.[1]

2.1.10 Association Mining

Database utilized in Association Mining is usually Transaction Databases and Association Rule which can be written in terms of itemsets lead to itemsets resulted from Market Basket Analysis.

Association Rule

Association rule discovery is conducted on huge data to support analysis and business decision making. An example of Association rule discovery is an analysis of sales and purchased products of customers called "Market Basket Analysis".

Pattern of Association rule discovery

General pattern of Association rule discovery is $A ==> B$

Where, A is condition and B is outcome.

As efficiency of association rule discovery relies on Support value and Confidence value. Support value is percentage of conditioned data and outcome which is reliable to the rule per number of total data. as the following equation:

$$\text{Support}\ (A,\ B) = \frac{NumberofTransaction(A,B)}{NumberofallTransaction} \quad (2.1)$$

Confidence value is percentage of conditioned data and outcome which is consistent to the rule per number of all conditioned data which can be written as the following equation:

$$\text{Confidence}\ (A,\ B) = \frac{NumberofTransaction(A,B)}{NumberofTransaction(A)} \quad (2.2)$$

By consideration on selecting the rule Support value and Confident value which is higher than Threshold (Default value) is considered to accept while Minimum support and Minimum confidence must be defined**[**3**]**.

Evolution of Association Rule Patterns

Association Rule is discovered by Agrawal et al. as the interested relation is found between product items which have purchased parallel and leads to the beginning of wide studies and constituted various concepts.

Conceptual Issues

Conceptual research is focused on development of scope of description theory under relations analysis and expansion of handling methods of new patterns.

Gunopoulous et al[4]. indicates relations between problem in searching maximal frequent item sets and hypergraph transversal problem.

Zaki et al.[5,6] and Pasquier et al[7.] has applied to utilize analysis of conceptual issues of creating frequent item sets and Zaki later presented a concept of closed frequent item sets

Friedman et al[8]. studies issues on analyzing environmental relations of bump hunting by using multi-dimension data particularly when there is consideration in creation of frequent item sets by searching for possibility in the area with high density of multi-dimension data.

There are new patterns discovered in many years later, for instance, profile association rules, cyclical association rules, uncertain association rules, exemption rules, negative association rules, weighted association rules, dependence rules, peculiar rules, intertransaction association rules, partial classification rules and other patterns including closed item sets, maximal item sets, hyper clique patterns, support envelopes, emerging patterns and contrast sets.

Relations analysis is successful when sequential, partial, and graph-based data are utilized. It is firstly presented by Hui et al[9]. efficient algorithm (which is called Hyper clique Miner) and it is automatically eliminated by utilizing cross-support pattern.

There are many researches that lead to application of the Association Rule and Attributes: nominal, ordinal, interval and ratio, the main key is a measurement of support value of these Attributes. This method is presented by Steunbach et al[10]. who expands original support concept of general pattern and types of attribute.

Implementation Issues

Research activity of this part focused on integrating ability of Data Mining, the existed database, Developing efficiency and scope of Data Mining steps, Management of users and limited and specific domain and Pattern that obtained after processing.

This was several benefits in integration of relations analysis by utilizing technology of existed database. Firstly, it can apply index and an ability of query-processing of database system. Secondly, it can utilize DBMS support for Scalability,

check-pointing, parallelization. SETM Algorithm, developed by Houtsma et al[11]. which is the simplest step to support the Association rules discovery via SQL inquiries and there are various methods to develop to be able to find the Association Rules in database system, for example, DMQL Language and M-SQL Query in order to expand SQL base and new operator used to conduct Association Rule, Mine rule operator which is indicated SQL operator and manage both groups of attributes and list of sequence.

Tsur et al[12]. develops generate- and- test approach called query flocks for mining association rule and distribution of OLAP-based infrastructure developed by Chen et al.[13] for mining multilevel association rules.

Other Algorithms application discovered frequent item set including DHP (Dynamic hashing and Pruning) algorithm presented by Park et al.[14] and Partition algorithm which is developed by Savasere et al[15]. A sampling-based frequent item set generation algorithm presented by Toivonen[16] algorithm which database is read once and candidate item set is created immediately as required.

Dynamic Item set Counting (DIC) algorithm could read database 1.5 times and it can generate candidate item set less than sampling based algorithm.

Other Algorithm such as tree projection algorithm and H-Mine.

Issue of relations analysis is number of patterns which are too excessive to create acceptable Algorithm. However, its benefit is contributed development of rank, summarize and filter pattern.

Toivonen et al presents concept of elimination of repeated rule by utilizing Structural rule covers and eliminating grouping of the rule by utilizing Clustering.

Liu et al[17]. is applied to utilize statistics of Chi-square Test to eliminate excessive patterns and collecting main patterns by utilizing in form of subset called direction setting rules.

Utilization of objective measure in order to obtain filter patterns is happened to many researchers such as Brin et al[18], Bayardo and Agrawal[19], Agrawal and Yu[20],and DuMouchel and Pregibon[21].

Property of measurement is analyzed by Piatestsky-Shapiro[22],Kamber and Singhal[23], Hilderman and Hamilton[24],and Tan et al[25].

Example of grade-gender by emphasizing significance of rows and columns by Mosteller.

Tea-coffee example is limitation of confidence discovered by Brin et al as limitation of confidence. It is the concept which used interest factor to measure interest.

All measurements of confidence are presented by Omiecinski.[26]

Xiong et al[27]. presents property of Cross-support and indicates that all measurements of confidence can utilize cross-support patterns for elimination and presents efficient Association Rule by utilizing upper bound function the coefficient.

Application Issue

Analysis of relations was applied by various applications; such as, Web mining, Document analysis, Telecommunication alarm diagnosis, network intrusion detection and bioinformatics.

Application of association and correlation patterns analysis is utilized to study global science.

Association pattern should apply on learning issues such as classification, regression and clustering.

Comparison between classification and association rule mining was created by Freit as in his research by utilizing association pattern for clustering which is studied by various researchers, such as, Han et al[28], Kosters et al[29], Yang et al[30] and Xiong et al[31].

2.2 Related Works

2.2.1 Real World Performance of Association Rule Algorithm

This study compares five Algorithms which were well-known by utilizing three sets of real-world datasets and one set of artificial data set. There was a reference of experimental results to confirm the improvement of efficiency of artificial data function previously by Authors. However, there was some inappropriate utilization

applied in real data set which indicated the inappropriateness of IBM artificial data set of Algorithm. Consequently, data set was both artificial data set and real data set. They are taken placed an experiment in order to search efficiency of Association Rule Algorithm, in such experiment, all data set were utilized to create association rule in every Algorithm to be tested, changing parameters of both data set in order to find efficiency of each Algorithm.

The Five Well-known Algorithms were employed in the experiment (Apriori, FP-Growth, Closet, Charm, and Magnum-Opus) and it was found that these Algorithms were not suitable to utilize in artificial data sets, the main reason were because of the artificial data sets have very different characteristics and also it was essential for researchers improving the artificial data sets. Prior to utilizing on the research finding Association Rules or utilizing in real data sets it was also found that the Apriori Algorithm can generate one million rules within 10 minutes as well as when parameter of Support increases. Moreover, number of rules each Algorithm reduces. This experiment was more interesting if the Confidence value was tested by parameter, likewise, in order to find out that this parameter may affect to efficiency and number of rules whether or not.[32]

2.2.2 A Comparative Study of Association Rules Mining Algorithms

This study was compares between the algorithms use candidates set generation and test and the algorithms without candidates set generation. There were comparison between the Apriori, FP-growth and DynFP-growth algorithms. Because the database was much larger, So It had to developed tools to extract useful information for financial forcast, marketing decision and other applications. From experimental with sample data can be conclude that the DynFP-growth was better than FP-growth algorithm. The FP-growth was not affect by the support factor. While the performance of the Apriori algorithm decreases by the support factor. The Apriori algorithm was well only for small database. In the other case is better for the DynFP-growth and FP-growth algorithms.[33]

2.2.3 Mining Association Rules Between Set of Items In Large Databases

This Study was to find the algorithm for generate all association rule between items in the database. Because We get the large database of customer transaction that was a supermarket data. Each transaction consists of items purchase by customer in supermarket. So We needed to present efficiency algorithm for generate all association rule between items in the database. The algorithm could be manage by estimate and pruning technique.[34]

2.2.4 Collection itemsets Technical that occur frequently by consider the minimum confidence to support increasing data.

The Association rule was important that the rule was pass the minimum support value and the minimum confidence value. Research in the past has offer efficient technique for finding the rules were important in large database. However there was still a problem. When the data increase, we had to update rule. Because the data increase not only make the itemset at the minimum support was not pass but the itemset was not pass the minimum support is pass. This study has proposed technique to keep the itemset for accommodate for adding data. Fast Update of Frequent Itemsets (FUFI) considers the minimum confidence and the lower bound criteria for select itemset was expected to pass the minimum support in the next round. So the technique of calculation of FUFI was reduce the procedure for finding in large database when adding data. Moreover it could still improve itemset was pass the minimum support. From the results with dense dataset shoe that technique FUFI has performance in term of time, processing and reduce the number of all itemset are expected to pass the minimum support in the next round when compare with FUP and NUWEP technique.[35]

CHAPTER III
IMPLEMENTATION

The experiment is conducted to compare efficiency of Association Algorithm. The Weka program utilizes on this experiment in order to searching for the Association Rules between items each transaction. By requiring support value and confidence value are equal. It takes to find association rule of each algorithm to see the processing time.

3.1 Complement of Implementation

3.1.1 Data that used on the study:

Data set is used in the experiment called Supermarket Data which is data concerning Transaction which occurred during customer's shopping at one Supermarket. As data is consisted of 4627 Transaction and 217 Attributes.

3.1.2 Computer program that used in this study:

The Weka program applied on this experiment to find Association Rule of the Data set; the program has an order to find Association Rule by using 2 Algorithms such as the Apriori Algorithm and the FP-Growth Algorithm.

3.1.3 Steps of the Experiment:

Acquiring the Data is set on experiment finding the Association Rules by using the Weka program as well as Apriori Algorithm and FP-Growth Algorithm. Parameter values of both Support and Confident were adjusted by increasing gradually from 0.1 to 0.9. In order to study time used for searching the Association Rule of each Algorithm. The number of the Association Rules obtained from each Algorithm which

was comparing the obtained outcomes to find efficiency of the Association Rule Algorithm for further conclusion.

3.2 Association Rule

Association Rule is another process in data mining which is utilized to find relation of two data sets or more within large clusters. The Association Rule displays in terms of equation $X \rightarrow Y$ and Association Rule is interested by value of support and confidence.

Efficiency of Association Rule was measured from:

1. Support value should be comprehensive number of sample according to the rule or comprehensiveness.

The acquired rule is assumed as $X \rightarrow Y$ calculated by $\frac{\sigma(X \cup Y)}{N}$. Number of transaction is provided both at X and Y that is divided by number of all transaction.

2. Confidence used to measure precision and indicated frequency of Item at Y obtained in Transaction which obtained at X.

The obtained rule is assumed as $X \rightarrow Y$ calculated by $\frac{\sigma(X \cup Y)}{\sigma(X)}$. Number of transaction is provided both at X and Y that is divided by number of transaction obtained at X.

As specified in the following table:

Table 3.1 Example of Transaction data.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

It is assumed that Association Rule is {Milk, Diaper} → Beer.

$$S = \frac{\sigma(Milk, Diaper, Beer)}{|T|} = \frac{2}{5} = 0.4$$

$$C = \frac{\sigma(Milk, Diaper, Beer)}{\sigma(Milk, Diaper)} = \frac{2}{3} = 0.67$$

Definition:

Significant Association Rule must have Support >= minsupport and Confidence >= minconfidence (having Support value is more or equal to minimum Support value and having Confidence value more or equal to minimum Confidence value).

In general, the possible number of rules of each data set with d items should be:

$$R = 3^d - 2^{d+1} + 1$$

Form the above Table, there were 6 items, and the rules should be created as follow:

$$3^6 + 2^7 + 1 = 602$$

Which is too excessive so minsupport, minconfidence shall be defined to reduce number of rules.

3.2.1 Steps of creating Association Rule

1. Frequent Item set Generation is created to find all itemsets with Support count is more or equal to minsup, these item sets were called Frequent Item set.

2. Rule Generation is created to find Association Rule with high Confidence value from frequent item sets which was obtained previously. This Association Rule is called Strong Rule.

3.2.2 Frequent Itemset Generation

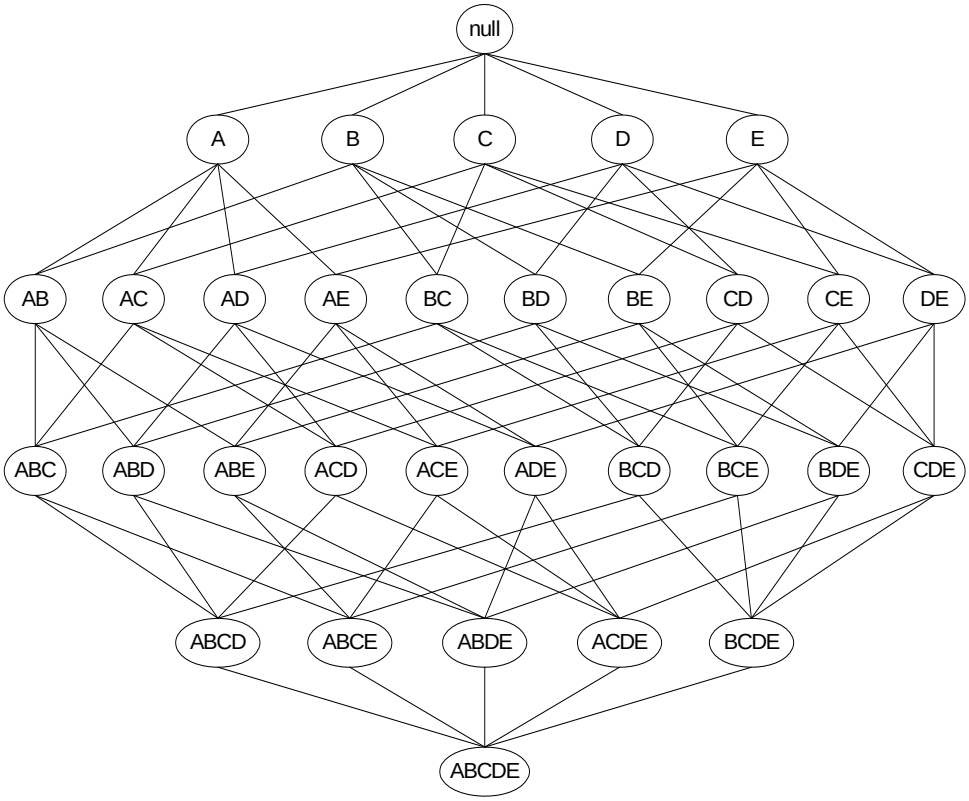

Figure 3.1 An itemset lattice.

Itemsets with *k*-items generally can crate Candidate itemsets (itemsets which are prepared to be Frequent itemsets) totally 2^k-1 which is too excessive to create Frequent Itemsets.

There is a method to reduce complication in creation of Frequent Item set as follows:

1. Reduction number of Candidate item sets by utilizing Apriori principle.
2. Reduction number of Comparisons, instead of matching in every transaction, reduction number of Comparison is replaced.

3.2.3 The Apriori Principle

It is mentioning in measurement of Support value which can reduce number of Candidate item sets in creating Frequent itemsets and utilizing Confidence value to eliminate the Candidate item sets by utilizing Apriori principle.

Definition: if an item set is frequent, then all of its subsets must also be frequent too.

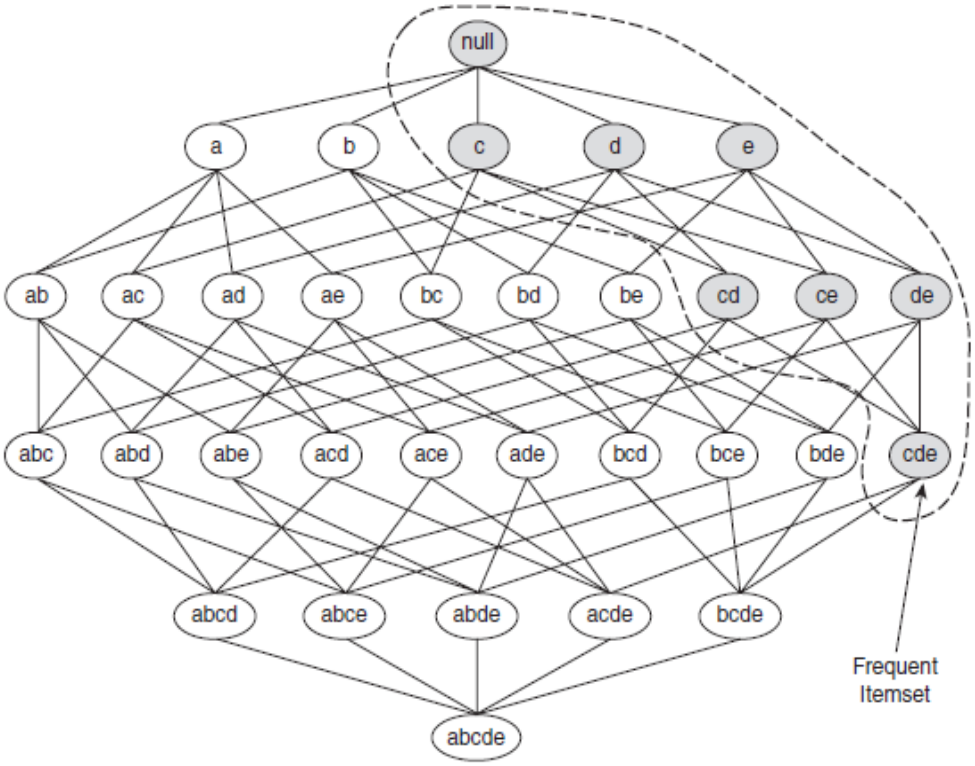

Figure 3.2 An illustration of the Apriori principle. If {*c*, *d* , *e*} is frequent, then all subsets of this itemsets are frequent.

From the Figure 3.2, it is provided {*C,D,E* } as Frequent itemsets, therefore, it is found that any transaction with {*C,D,E* }as Frequent itemsets. Its subsets must be {*C,D*},{*C,E*},{*D,E*},{*C*},{*D*} and {*E*}. If {*C,D,E*} is Frequent

itemsets, all of its subsets such as $\{C,D\},\{C,E\},\{D,E\},\{C\},\{D\}$ and$\{E\}$must be Frequent itemsets too.

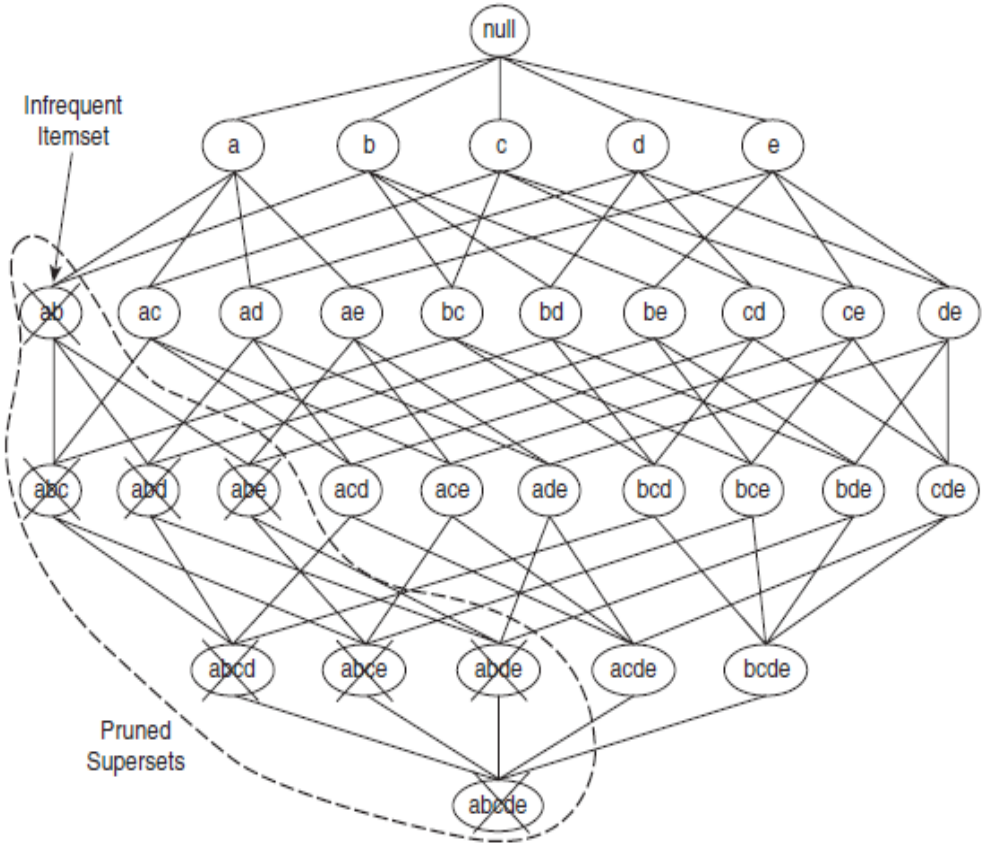

Figure 3.3 An illustration of support-based pruning, if $\{a,b\}$ is infrequent, then all supersets of $\{a,b\}$ are infrequent.

In contrast, it is provided $\{A,B\}$ as Infrequent, therefore, all of its Supersets must be Infrequent too. Super sets of $\{A,B\}$ should be eliminated due to its occasional.

Strategy of this elimination was an exponential searching based on measurement Support value called Support Based Pruning. This eliminated strategy must have significant property in measure Support value. As Support value of item sets must not exceed Support value of its subsets, this property is called Anti-monotone.

3.2.4 Generation of Frequent Itemsets by Apriori Algorithm

Apriori is the first algorithm in generating association rule by utilizing support base pruning to control searching for Candidate item sets.

It is assumed that Support started at 60 % to gain minimum support count = 3.

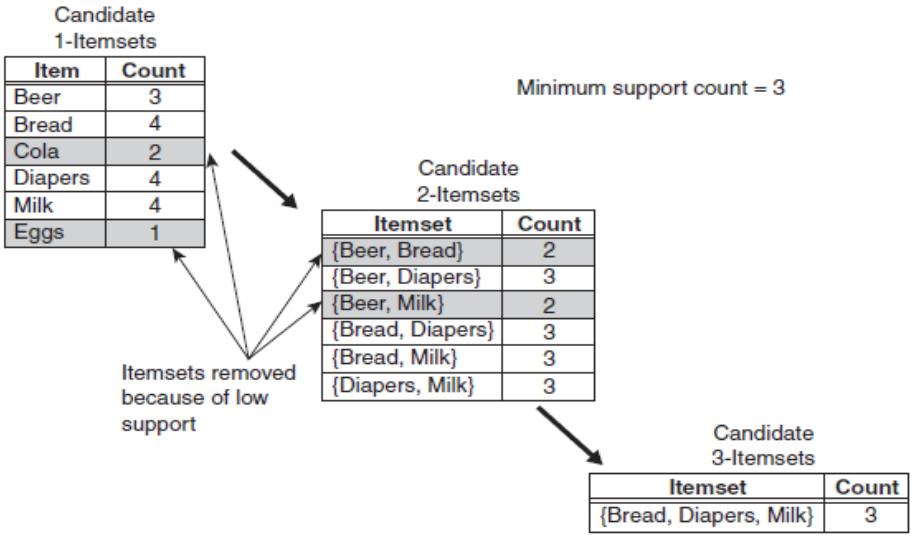

Figure 3.4 Illustration of frequent itemset generation using the Apriori algorithm.

In beginning, it was provided that every item is Candidate 1-itemsets, after counting support of every item, it was found that Candidate itemsets {Cola} and {Eggs} were eliminated. This is because the obtained transaction is less than 3 transactions. Candidate 2-itemsets were created by utilizing Frequent 1-itemsets obtained from the Apriori Principle which is ensured that all Supersets of Infrequent 1-itemsets must be Infrequent too. Therefore, there were merely 4 Frequent 1-itemsets remained and number of Candidate 2-item sets was created by $\binom{4}{2}$ = 6 which is found that 2 Candidate. For example, {Beer,Bread} and {Beer,Milk} were Infrequent. After that calculation to find Support value for another 4 remained Candidates which were Frequent utilizing in generation of Candidate 3-itemsetsand. It was found that if Support base pruning was not utilized, all Candidate item sets $\binom{6}{3}$= 20. Candidate 3-itemsets must be created from 6 items. Moreover, if Apriori principle is utilized, we should have Candidate 3-itemsets from subsets which are Frequent and Candidate. The following property is {Bread,Diapers,Milk}. It was seen that if we found

Candidates normally, it should be $\binom{6}{1}+\binom{6}{2}+\binom{6}{3}= 6+15+20 = 41$. If Apriori principle was utilized, it was reduced to $\binom{6}{1}+\binom{4}{2}+1 = 13$. It is equally to 68%.

3.2.5 Generation of Candidate and Pruning

Anti-monotone property, if such Candidate itemsets has subset which is Infrequent, it must be Infrequent too. Especially it must be ensured that Frequent itemsets are created by Candidate Itemsets which is $Fk \subseteq Ck$. Candidate itemsets must not be similar more than 1 itemset.

It should be mentioned in Candidate generation by utilizing Apriori-gen function which has many methods as follows:

1. Brute-Force Method, this method should consider every *k*-itemsets as number of Candidate itemsets created in level k is equal to $\binom{d}{k}$. Where, *k* is number of items.

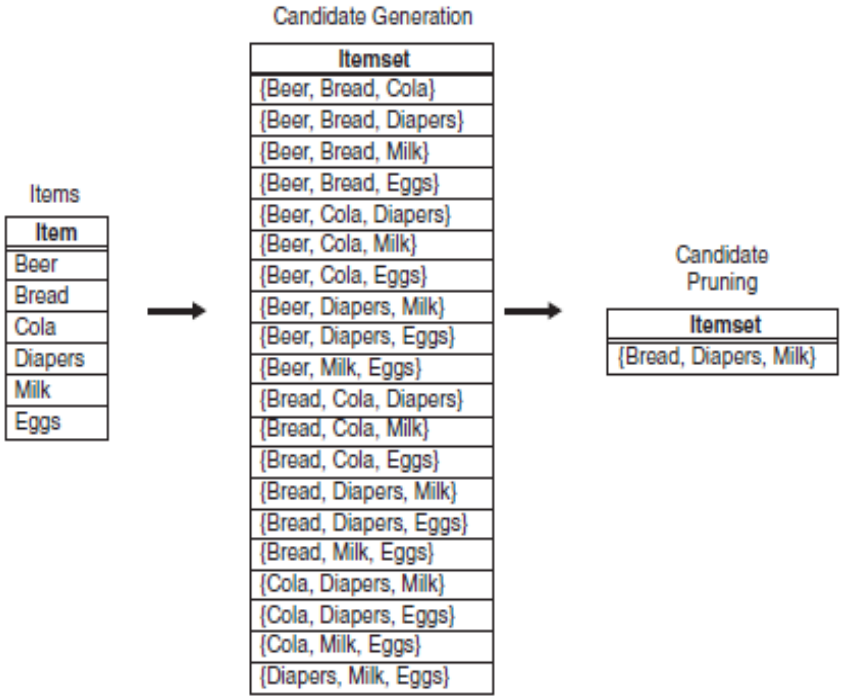

Figure 3.5 A brute-force method for generating candidate 3-itemsets.

2. F*k*-1* F1 Method, this is complete method because every Frequent *k*-itemset was obtained from frequent (*k*-1)–itemset and frequent 1-itemset, therefore, all frequent *k*-itemset was a part of the generated Candidate *k*-itemsets.

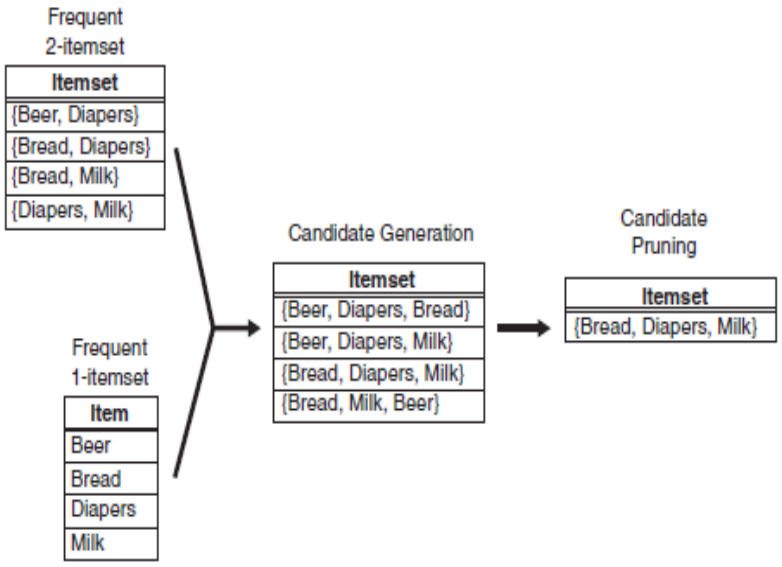

Figure 3.6 Generating and pruning candidate k-itemsets by merging a frequent (*k*-1) – itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

It was not protected the repeated Candidate itemsets such as {Bread,Diapers,Milk} which generated from {Bread,Diapers} and {Milk} or {Bread,Milk} and {Diapers} or {Diapers,Milk} and {Bread}. The best way to prevent the repeated Candidate was allowing item in each frequent item set be prioritized by dictionary order. For instance, {Bread, Diapers} could be matching to {Milk}, due to dictionary, Milk was higher than Bread and Diapers. However, {Diapers, Milk} cannot be matching to {Bread} or {Bread, Milk} could not be matched to {Diapers} because it was not prioritized by dictionary order. However, this method still created unnecessary candidate, such as {Beer, Diapers} and {Milk} were not necessary as its sub set {Beer, Milk} was Infrequent, so it can be deleted.

3. F*k*-1* F*k*-1 Method was acquired by matching between a pair of Frequent (*k*-1)-itemsets. An example of this method shall be frequent item sets {Bread,Diapers} and {Bread,Milk} could be matching. It became {Bread,Diapers,Milk} but algorithm

which cannot be matching, such as {Beer,Diapers}and {Diapers,Milk}. This was because the first item in item sets of them were different, if {Beer, Diapers, Milk} was Candidate obtained from {Beer,Diapers} and {Beer,Milk} should be replaced. The advantage of generating the Candidate prioritized by dictionary order was preventing duplication.

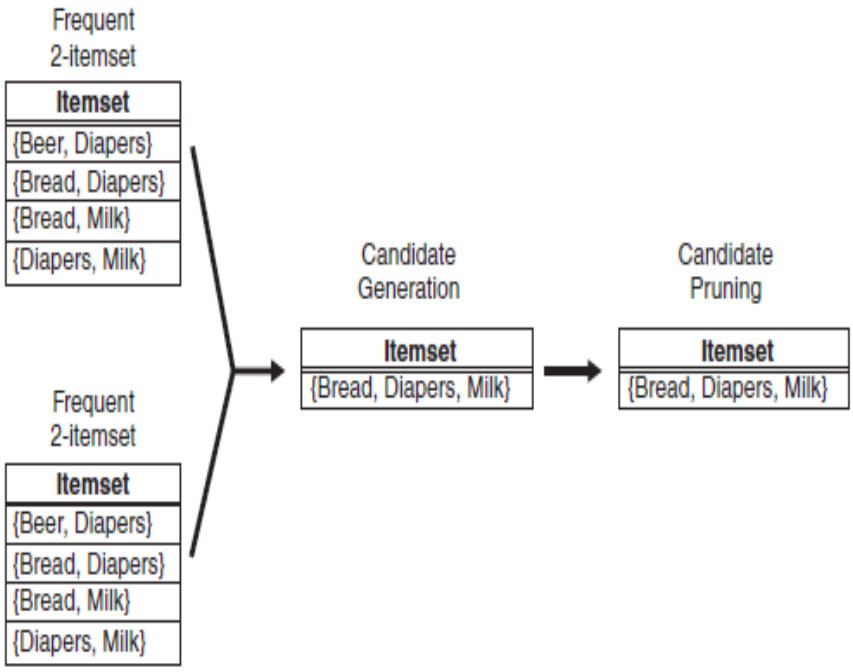

Figure 3.7 Generating and pruning candidate k-itemsets by merging pairs of frequent (k-1) itemsets.

3.2.6 Rule Generation

In each frequent k- itemset, it can generate 2^k-2 rules and can be ensured that every rule had passed initial Support value because they were generated from frequent itemset. such as:

It is provided that X = {1,2,3} is frequent itemset. It consisted of 6 candidate association rule that generated from X which is {1,2}→{3},{1,3}→{2}, {2,3}→{1},{1}→{2,3},{2}→{1,3} and{3}→{1,2}.

3.2.6.1 Confidence-Based Pruning

If a rule $X \rightarrow Y\text{-}X$ does not satisfy the confidence threshold, then any rule $X' \rightarrow Y\text{-}X'$, where X' is a subsets of X, must satisfy the confidence threshold as well.

3.2.6.2 Rule Generation in Apriori Algorithm

Level-wise was utilized in creating Association Rule, such as $\{acd\} \rightarrow \{b\}$ and $\{abd\} \rightarrow \{c\}$. It is the rule with high Confidence value, therefore, Candidate rule is $\{ad\} \rightarrow \{bc\}$ is generated by merge consequent od both rules.

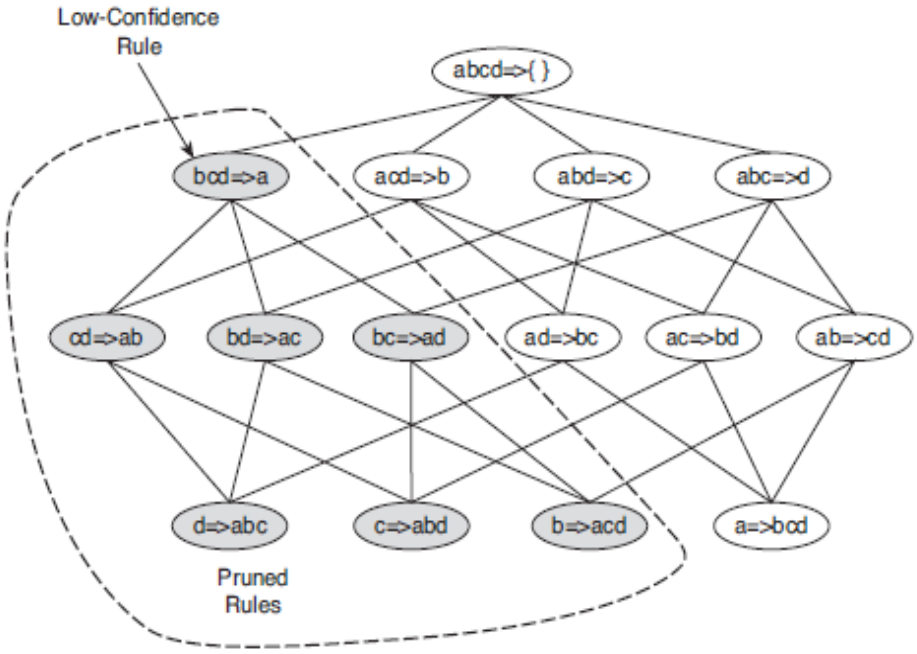

Figure 3.8 Pruning of association rules using the confidence measure.

From Figure 3.8, it was indicated structure of Association Rule generation from frequent item set $\{a,b,c,d\}$. If there was any node in crystal which has a low confidence value , it should be in accordance which is mentioned by theory. It can be seen that area on the red line could be eliminated immediately. It may observed that $\{bcd\} \rightarrow \{a\}$ is low, therefore, all rules were in an outcome side, such as, $\{cd\} \rightarrow \{ab\}, \{bd\} \rightarrow \{ac\}, \{bc\} \rightarrow \{ad\}$ and $\{d\} \rightarrow \{abc\}$ shall be eliminated.

3.2.7 Maximal Frequent itemsets

As frequent itemset of transaction data set might be excessive, Maximal Frequent itemets was utilized to find representative of the frequent itemset which was unnecessary.

Definition: maximal frequent itemsets can be found from frequent itemset which has no superset as frequent. It meant that the frequent item sets which have super set shall be infrequent.

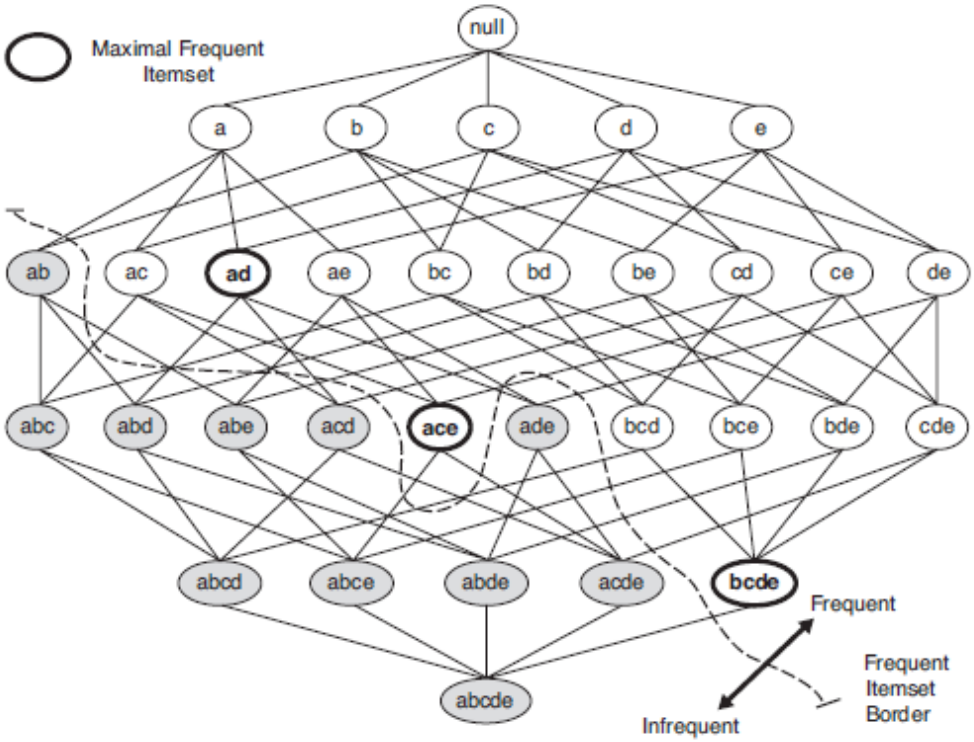

Figure 3.9 Maximal frequent itemset**.**

From the figure 3.9, itemset in structure of crystal was divided into 2 groups as frequent group and infrequent group. The frequent itemset group was above the line and every item set above. The line should be all frequent, while the item sets which were below the line should be infrequent. Itemsets were near the line as $\{A,D\}$,$\{A,C,E\}$ and $\{B,C,D,E\}$. They were considered as maximal frequent itemsets because its superset is infrequent.

The Item set as {*A,D*} was maximal frequent because all of its supersets, such as {*A,B,D*},{*A,C,D*} and {*A,D,E*} were infrequent. In contrast, {*A,C*} was non-maximal as one of its super set was frequent, which was {*A,C,E*}.

It can be seen that the frequent itemset was divided into 2 groups as:

The frequent itemset which was started by item and followed by item *c*, *d* or *e*. It was found that this group has itemset, such as {*a*},{*a,c*},{*a,d*},{*a,e*} and {*a,c,e*}.

The frequent itemset which was started by item *b,c,d* or *e*. It is found that this group had item set such as {*b*},{*b,c*},{*c,d*},{*b,c,d,e*}.

It can be seen that the frequent itemset in the first group was subset of {*a,c,e*} or {*a,d*} while the second group was subset of {*b,c,d,e*}. Consequently, maximal frequent item set {*a,c,e*},{*a,d*} and {*b,c,d,e*} provided to represent as frequent itemset in Figure.

3.2.8 Closed Frequent Itemsets

Definition: An item set *X* is closed if none of its immediate supersets has exactly the same support count as *X*. (Item set *X* was closed if none of its superset has support count exactly the same as *X* but item set *X* was not closed if one of its superset has exactly the same support as *X*.)

Definition: item set should be closed frequent item set if its support count was more or equal to minsup.

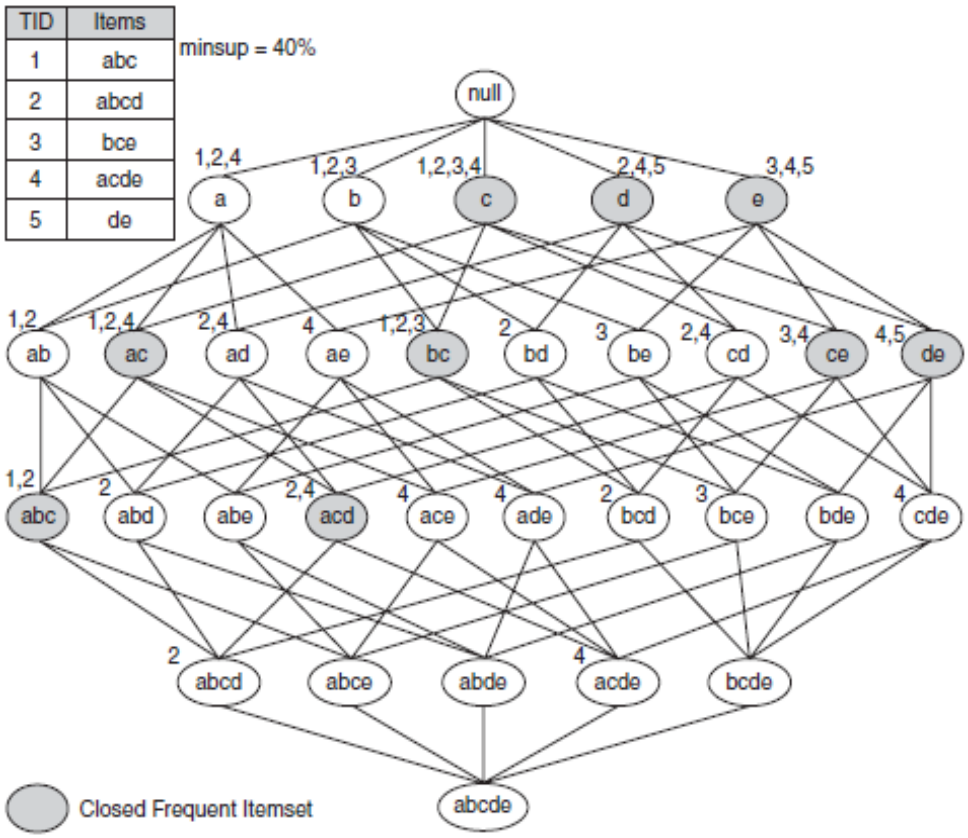

Figure 3.10 An example of the closed frequent itemsets
(with minimum support count equal to 40%).

From the figure 3.10, {*b*,*c*} is closed item set as none of its superset has support count equally as its.

3.3 FP-growth

Finding Frequent Itemsets by utilizing FP-growth Algorithm is consisted of 2 steps as follows:

1. FP-tree generation by utilizing defined database and database should be scanned twice.

❑ The first scan: to count frequency of each items and to eliminate low frequent items and taking the rest items to set priority. It was based on frequency of each items in ascending order in the table called Hash Table.

❑ The second scan: to create FP-tree.

2. Finding Frequent Item sets from FP-tree [3]

Table 3.2 Database in the right column indicates Frequent Itemsets which are in ascending order of frequency in each transaction.

TID	Item	(Orders) Frequent Items
100	1 3 4	3 1
200	2 3 5	2 3 5
300	1 2 3 5	2 3 5 1
400	2 5	2 5

Header Table

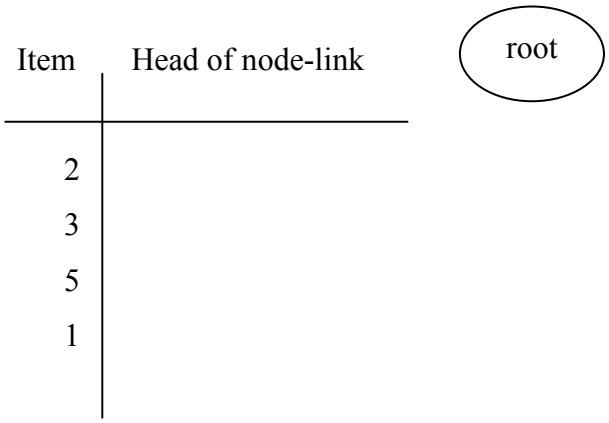

Figure 3.11 a) Header Table and root node.

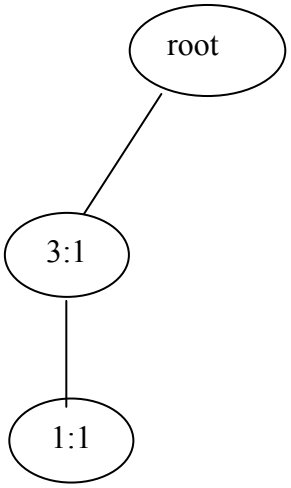

Figure 3.12 b) Adding Items{3,1} from TID 100.

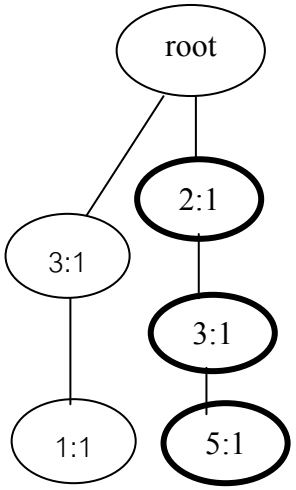

Figure 3.13 c) adding Items{2,3,5} from TID 200.

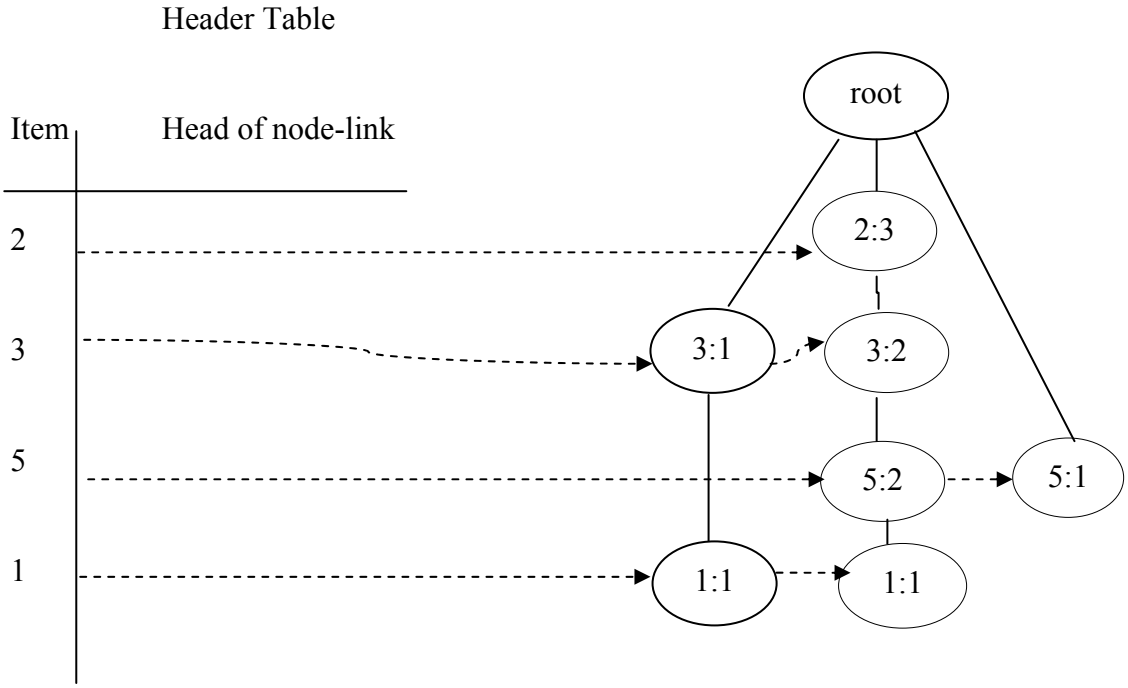

Figure 3.14 f) Complete FP-Tree

Table 3.3 All Frequent Itemsets.

Frequent Itemsets	Support
1	2
2	3
3	3
5	3
1,3	2
2,3	2
2,5	3
3,5	2
2,3,5	2

CHAPTER IV
RESULT

According to the supermarket data which was data concerning about transaction. It occurred during customer's shopping at a supermarket by utilizing the Weka program. In order to find out the Association Rule of data, it was done by adjusting parameter values of both Minsupport and Minconfident to compare efficiency of both Algorithms. For example, the Apriori Algorithm and the FP-Growth Algorithm by counting time and number of acquired Association Rules. The outcome was specified on the following table:

Table 4.1 Example of Association Rule by Apriori Algorithm
(Minsupport=0.5,Minconfidence=0.7)

1	milk-cream=t 2939 ==> bread and cake=t 2337 conf:(0.8)
2	fruit=t 2962 ==> bread and cak =t 2325 conf:(0.78)
3	bread and cake=t 3330 ==> milk–cream=t 2337 conf:(0.7)

Table 4.2 Example of Association Rule by FP-Growth Algorithm
(Minsupport=0.5,Minconfidence=0.7)

1	[milk-cream=t]: 2939 ==> [bread and cake=t]: 2337 <conf:(0.8)> lift:(1.1) lev:(0.05) conv:(1.37)
2	[fruit=t]: 2962 ==> [bread and cake=t]: 2325 <conf:(0.78)> lift:(1.09) lev:(0.04) conv:(1.3)
3	[bread and cake=t]: 3330 ==> [milk-cream=t]: 2337 <conf:(0.7)> lift:(1.1) lev:(0.05) conv:(1.22)

As experimental results, as shown on the table 4.1 and 4.2, it was concluded as follows:

1. The first rule shown that the customers who buy milk-cream to buy bread and cake.
2. The second rule shown that the customers who buy fruit to buy bread and cake.
3. The third rule shown that the customers who buy bread and cake to buy milk-cream.

Table 4.3 Processing Time was results of parameter adjustments of parameters for both Minsupport and Minconfidence for searching the Association Rule.

Minconfidence	**Minsupport**									
	0.1		0.2		0.3		0.4		0.5	
	Ap	FP	Ap	FP	Ap	FP	Ap	FP	Ap	FP
0.1	2:50:56	00:47:70	00:29:06	00:05:55	00:17:21	00:02:86	00:12:01	00:01:76	00:09:98	00:01:24
0.2	2:49:53	00:38:03	00:28:51	00:05:46	00:17:00	00:02:83	00:12:03	00:01:73	00:09:63	00:01:13
0.3	2:43:29	00:31:17	00:28:35	00:05:33	00:16:56	00:02:85	00:12:00	00:01:70	00:09:61	00:01:29
0.4	2:42:19	00:29:66	00:27:96	00:05:30	00:16:50	00:02:80	00:11:93	00:01:70	00:09:56	00:01:28
0.5	2:25:74	00:20:32	00:27:50	00:04:68	00:14:52	00:02:60	00:10:66	00:01:65	00:07:80	00:01:28
0.6	1:40:16	00:11:40	00:27:48	00:04:43	00:09:85	00:02:08	00:07:20	00:01:38	00:04:54	00:00:77
0.7	1:33:47	00:09:03	00:27:45	00:04:38	00:08:91	00:01:81	00:05:54	00:00:86	00:04:35	00:00:58
0.8	1:38:59	00:09:25	00:27:00	00:04:36	00:08:93	00:01:80	-	-	-	-
0.9	1:17:57	00:06:95	-	-	-	-	-	-	-	-

As experimental results, as shown on the table 4.3, it was concluded as follows:

1. Although, parameter values had been adjusted, it was found that the Apriori Algorithm uses processing time for finding the Association Rule longer than the FP-Growth Algorithm.

2. Once increasing parameter values for both minsupport and minconfidence, the processing times for both of algorithms should be reduced. It was concluded that the higher the parameter values,, the less the processing times for finding Association Rules.

3. At the point that Minsupport was equal to 0.6, by the experiment, no Association Rule was found. Therefore, it was concluded that the highest Minsupport of this data set which can find Association Rule of both Algorithm is 0.6.

Table 4.4 Number of Association Rule was results of parameter adjustments of parameters for both Minsupport and Minconfident for finding the Association Rule.

Minconfidence	**Minsupport**											
	0.1		0.2		0.3		0.4		0.5		0.6	
	Ap	FP	Ap	FP	Ap	FP	Ap	FP	Ap	FP	Ap	FP
0.1	183,372	181,292	3,442	3,280	258	246	32	30	4	4	-	-
0.2	170,347	168,537	3,442	3,280	258	246	32	30	4	4	-	-
0.3	137,452	136,215	3,370	3,212	258	246	32	30	4	4	-	-
0.4	101,613	100,809	2,724	2,612	258	246	32	30	4	4	-	-
0.5	69,369	68,841	1,982	1,906	231	220	32	30	4	4	-	-
0.6	43,580	43,294	1,314	1,275	148	147	30	29	4	4	-	-
0.7	24,570	24,467	759	746	67	67	10	10	3	3	-	-
0.8	8,071	8,049	210	206	17	17	-	-	-	-	-	-
0.9	458	455	-	-	-	-	-	-	-	-	-	-

By experimental results, as shown in the Table 4.4, it was concluded as follows:

1. The Apriori Algorithm should display more number of the Association Rule than the FP-Growth.

2. As soon as, increasing parameter values for both minsupport and minconfidence, number of Association rule should be reduced.

3. At the point of Minsupport is 0.6, by the experiment, no Association Rule was found.

The experimental result obtained from the Weka program by utilizing Transaction data of shopping in one supermarket which was experimented by adjusting parameter values. It was provided to increase value from 0.1-0.9, which was found that the point of the Association Rule. It should be utilized in marketing. In fact, it must be the point that both parameter of Minsup and Minconf. They were higher the acquired rules. It should be more reliable and efficient and relation on Transaction of the rule which should have further possibility. In this experiment, it had the point where Minsup was being equal to 0.5 and Minconf was equal to 0.7 and the obtained the Association Rule as it had 3 rules.

According to the study required comparing efficiency of the Apriori algorithm and the FP-Growth algorithm by utilizing data. It was found that the FP-Growth algorithm used longer on searching for the Association Rule than the Apriori algorithm. However, generating more numbers of the Association Rule which was in accordance that the theory mentioned in frequent item sets generation on the Apriori algorithm. The Candidate item-sets must be generating initially and finding frequency of each Candidate item-sets. The database must be scanned every time. In order to find out frequent item-sets at the next level, hence, the more number of frequent item sets the longer time was used. On the other hand, the FP-Growth algorithm and frequent item-sets could be generated immediately without the Candidate item-sets generation.

CHAPTER V
CONCLUSIONS

This research was conducted a comparative study of efficiency of the Association Algorithm by comparing efficiency between the Apriori Algorithm and the FP-Growth Algorithm and the study result was as following:

5.1 Research Result

The Comparison of efficiency of the Association Algorithm was utilizing by the Weka program to finding Association Rule which was compared between the Apriori Algorithm and the FP-Growth Algorithm. The data used on this study, was shopping data of customers in one supermarket. It was consisted of 4627 Transactions as well as parameter values of both Support and Confidence. It had been adjusted from 0.1 to 0.9. Moreover, the obtained outcome was time used on researching for the Association Rule of each Algorithm and number of Association Rule. It obtained from each Algorithm concluding the outcome which was concluded. From the configuration of the support value and the confidence value are equal. The FP-Growth Algorithm was better than the Apriori Algorithm in terms of processing time. Therefore, the results can be utilize to business implementing plan, analyzing customer purchases for the purpose of conducting the promotion, increase sales volume, product price and marketing planning. Futhermore, Association rule also can be applied to store inventory management. For example, Association rule is the customers who buy milk-cream to buy bread and cake. Therefore, In order to sell when stores order milk-cream . Stores should be order bread and cake together. In product placement, Stores should place milk-cream and bread and cake close together. Therefore, If we had limited time and need to facilitate on searching the Association rule. We should use the FP-Growth Algorithm to find association rules because the experiments showed that FP-

Growth Algorithm is more efficient Apriori Algorithm in terms of time of processing time.

5.2 Suggestion

Suggestion for the nest research is as follow:

5.2.1 Data utilized in the study, it should have more numbers of Transaction than this. As the data of transactions, they render more precision and correction on the research result as well as the other aspects outcomes. It can be visible which it cannot be found, if number of transaction was too small.

5.2.2 Values of parameter, parameter must be adjusted for further comprehensive and diversifying, in order that it can obtain more accurate and precise outcome.

REFERENCES

[1] Alaska. Data Mining (in Thai). Available at: (Accessed: 10 September 2014).

[2] Adul Yimgyam. การทำเหมืองข้อมูล (Data Mining) (in Thai). Available at: http://compcenter.bu.ac.th/index.php?option=com_content&task=view&id=75&Itemid=172 (Accessed: 12 September 2014).

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. Introduction to Data Mining. United States of America: Hamilton Printing.

[4] D. Gunopulos, R. Khardon, H. Mannila, and H. Toivonen. Data Mining, Hypergraph Transversals, and Machine Learning. In Proc. of the 16 th Symp. On Principle of Database S Systems, pages 209-216, Tucson, AZ, May 1997.

[5] M. J. Zaki. Generating Non-Redundant Association Rules. In Proc. Of the 6 th Intl. Conf. on Knowledge Discovery and Data Mining, pages 34-43, Boston, MA, August 2000.

[6] M. J. Zaki and M. Orihara. Theoretical foundations of association rules. In Proc. Of the 1998 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Seattle, WA, June 1998.

[7] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In Proc. of the 7 th Intl. Conf. on Database Theory (ICDT'99), pages 398-416, Jerusalem, Israel, January 1999.

[8] J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. Statistic and Computing, 9(2):123-143, April 1999.

[9] H. Xiong, P. N. Tan, and V. Kumar. Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution. In Proc. of the 2003 IEEE Intl. Conf. on Data Mining, pages 387-394, Melbourne, FL, 2003.

[10] M. Steinbach, P. N. Tan, H. Xiong, and V. Kumar. Extending the notion of support. In Proc. of the 10 th Intl. Conf. on knowledge Discovery and Data Mining, pages 259-262, Newport Beach, CA, August 1997.

[11] M. Houtsma and A. Swami. Set-oriented Mining for Association Rules in Relational Databases. In Proc. of the 11 th Intl. Conf. on Data Engineering, pages 25-33, Taipei, Taiwan, 1995.

[12] S. Tsur, J. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Query Flocks: A Generalization of Association Rule Mining. In Proc. of 1998 ACM-SIGMOD Intl. Conf. on Management of Data, pages 1-12, Seattle, WA, June 1998.

[13] Q. Chen, U. Dayal, and M.Hsu. A Distributed OLAP infrastructure for E-Commerce. In Proc.of the 4 th IFCIS Intl. Conf. on Cooperation Systems, pages 209-220, Edinburgh, Scotland, 1999.

[14] J. S. Park, M.-S. Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. SIGMOD Record, 25(2):175-186, 1995.

[15] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In Proc. of the 21st Int. Conf. on Very Large Databases (VLDB'95), pages 432-444, Zurich, Switzerland, September 1995.

[16] H. Toivonen, M. Sampling Large Databases for Association Rules. In Proc. of the 22nd VLDB Conf., pages 134-145, Bombay, India, 1996.

[17] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In Proc. of the 5 th Intl. Conf. on Knowledge Discovery and Data Mining, pages 125-134, San Diego, CA, August 1999.

[18] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Proc. ACM SIGMOD Intl. Conf. Management of Data, pages 265-276, Tucson, AZ, 1997.

[19] R. Bayardo and R. Agrawal. Mining the Most Interesting Rules. In Proc. of the 5 th Intl. Conf. on Knowledge Discovery and Data Mining, pages 145-153, San Diego, CA, August 1999.

[20] C. C. Aggarwal and P. S. Yu. Mining Associations with the Collective Strength Approach. IEEE Trans. On Knowledge and Data Engineering, 13(6):863-873, January/February 2001.

[21] W. DuMouchel and D. Pregibom. Empirical Bayes Screening for Multi-Item Associations. In Proc. of the 7 th Intl. Conf. on Knowledge Discovery and Data Mining, pages 67-76, San Francisco, CA, August 2001.

[22] G. Piatetsky-Shapiro. Discovery, Analysis and Presentation of Strong Rules. In Proc. Piatersky- Shapiro and W. Frawley, editors, Knowledge Discovery in Databases, pages 229-248. MIT Press, Cambrige, MA, 1991.

[23] M. Kamber and R. Shinghal. Evaluating the Interestingness of Characteristic Rules. In Proc. of 2 nd Intl. Conf. on Knowledge Discovery and Data Mining, pages 263-266, Portland, Oregon, 1996.

[24] R. J. Hilderman and H. J. Hamilton. Knowledge Discovery and Measures of Interest. Kluwer Academic Publisher, 2001.

[25] P. N. Tan and V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. In Proc. of the 8 th Intl. Conf. on Knowledge Discovery and Data Mining, pages 32-41, Edmonton, Canada, July 2002.

[26] E. Omiecinski. Alternative Interest Measures for Mining Associations in Databases. IEEE Trans. On Knowledge and Data Engineering, 5(1):57-69, January-February 2003.

[27 H. Xiong, P. N. Tan, and V. kumar. Mining Strong Affity Association Patterns in Data Sets with Skewed Support Distribution. In Proc. of the 2003 IEEE Intl. Conf. on Data Mining, pages 387- 394, Melbourne, FL, 2003.

[28] E.- H. Han, G. Kaeypis, V. Kumar, and B. Mobasher. Clustering Based on Association Rule Hypergraphs. In Proc. of the 1997 AGM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Tucson, AZ, 1997.

[29] W. A. Kosters, E. Marchiori, and A. Oerlemans. Mining Clusters with Association Rules. In the 3 rd Symp. on Intelligent Data Analysis (IDA99), pages 39-50, Amster-dam, August 1999.

[30] C. Yang, U. M. Fayyad, and P. S. Bradley. Efficient discovery of error-tolerant frequent itemsets in high dimension. In Proc. of the 7th Intl. Conf. on Knowledge Discovery and Data Mining, pages 194-203, San Francisco, CA, August 2001.

[31] H. Xiong, M. Steinbach, P. N. Tan, and V. Kumar. HICAP: Hierarchial Clustering with pattern Preservation. In Proc. of the SIAM Intl. Conf. on Data Mining, pages 279-290, Orlando, FL, April 2004.

[32] Zijian Zheng, Ron Kohavi, and Liew Mason. Real World Performance of Association Rule Algorithms. Available at: http://www. robotics. stanford.edu/users/ronnyk/ realWorldAssocPoster.pdf (Accessed: 1 March 2014)

[33] Cornelia Gyorodi, Robert Gyorodi. A Comparative Study of Association Rules Mining Algorithms Available at: http:// citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.123.2771&rep =rep1&type=pdf (Accessed: 23 March 2014).

[34] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. ACM SIGMOD Intl. Conf. Management of Data, pages 207-216, Washington, DC, 1993.

[35] Krisadakorn kongubon, Tanawit Raktammanon, Krissana viyamai. เทคนิคการเก็บไอเท็มเซ็ตที่เกิดขึ้นบ่อยโดยพิจารณาค่าความเชื่อมั่นขั้นต่ำเพื่อรองรับการเพิ่มของข้อมูล (in Thai). Available at : http://www.it.kmutnb.ac.th/journal/pdf/vol6/ch02.pdf (Accessed: 12 March 2014)

[36] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. DATA MINING Practical Machine Learning Tools and Techniques. 3rd ed. United States of America: Morgan Kaufmann Publishers.

[37] Jiawei Han, Micheline Kamber, and Jain Pei. 2012. Data Mining Concepts and Techniques. 3rd ed. United States of America: Morgan Kaufmann Publishers.

BIOGRAPHY

NAME	Miss Banthita Tipjaksu
DATE OF BIRTH	18 MAY 1989
PLACE OF BIRTH	Songkla, Thailand
INSTITUTIONS ATTENDED	Thammasat University, 2007-2011 Bachelor degree of Science (Statistics**)** Mahidol University, 2012-2014 Master of Science (Technology of Information System Management**)**
RESEARCH GRANTS	Association Rule, Apriori Algorithm, FP-Growth Algorithm, Minsupport, Minconfidence
HOME ADDRESS	13 Soi Chakpra14, Talingchan, Talingchan, Bangkok 10170, Thailand Tel. 084 839 3934 E-mail : banthita.tip@gmail.com
EMPLOYMENT ADDRESS	1112 Kpi Tower Phetbureetudmai Road, Makkasun, Rattawee, Bangkok 10400, Thailand Tel. 02 624 1111 E-mail : banthita.t@kpi.co.th
PUBLICATION / PRESENTATION	-