



MINING ASSOCIATION RULES FROM CLUSTERING MODELS

ALI AL-JIBOURI

AUGUST 2015

MINING ASSOCIATION RULES FROM CLUSTERING MODELS

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY**

**BY
Ali AL-JIBOURI**

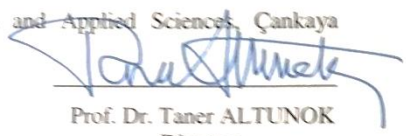
**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF
MATHEMATICS AND COMPUTER SCIENCE**

AUGUST 2015

Title of the Thesis : Mining Association Rules from Clustering Models.

Submitted by Ali AL-JIBOURI

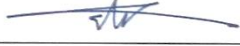
Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

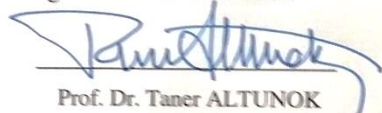

Prof. Dr. Taner ALTUNOK
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.


Prof. Dr. Billur KAYMAKÇALAN
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.


Assist Prof. Dr. Shadi Al-SHEHABI
Co-Supervisor


Prof. Dr. Taner ALTUNOK
Supervisor

Examination Date: 04.08.2015

Examining Committee Members

Prof. Dr. Taner ALTUNOK

(Çankaya Univ.)

Assist. Prof. Dr. Abdül Kadir GÖRÜR

(Çankaya Univ.)

Assoc. Prof. Dr. Fahd JARAD

(UTAA)

STATEMENT OF NON-PLAGIARISM PAGE

I wish to declare that it has been getting the information in this document is in accordance with the rules, academic and moral behavior. I would also like to announce that the conduct and rules that are not original have labeled to sites and sources.

Name, Last Name: Ali AL-JIBOURI

Signature :



Date

: 04.08.2015

ABSTRACT

MINING ASSOCIATION RULES FROM CLUSTERING MODELS

AL-JIBOURI, Ali

M.Sc, Department of Mathematics and Computer Science\ Information Technology
Program

Supervisor: Prof. Dr. Taner ALTUNOK

Co-Supervisor: Assist Prof. Dr. Shadi Al-SHEHABI

August 2015, 56 pages

Association rules used to discover the interested relationships between variables (items) in large database. These relationships have been hidden in large database. Symbolic models consider the most common to extract association rules. Unfortunately, these models suffered of very serious limitations. Association rule generation is a process that take long time to generate a huge number of rules, including large amount of redundant rules. This problem appears obviously specially when dealing with high dimensional description space. To cope with these problems unsupervised approach has been made. This approach has been proposed to extract numerical association rules by establishing interesting links between numerical and symbolic models. Numerical models keep only the important relations between data, so it can extract the most potential and important association rules. This numerical approach can extract only simple association rules that mean each rule has two items, one item in each part of the rule. So it generates a lot of simple rules. Therefore, the

aim of this thesis is to develop this approach to be able to extract complex and important association rules in order to discover relationships between itemsets of different sizes instead of only two items. This approach increases the possibility of numeric methods to extract important association rules. Neural network model with multiple topographic considers the main model in this thesis. One of the strongest features from this model is generalization mechanism that allows association rule extraction from only one database to be performed. The extraction of association rules is itself based on quality measures which evaluate to what extent a numerical classification model behaves as a natural symbolic classifier.

Keywords: Knowledge Extraction, Unsupervised Learning, SOM Model, Multi-SOM Model, Symbolic Model, Numeric Model, Association Rules, Clustering Model.

ÖZ

KÜMELEME MODELLERİNDEN İLİŞKİLENDİRME YÖNTEMLERİ MADENCİLİĞİ

AL-JIBOURI, Ali

Yüksek Lisans, Matematik ve Bilgisayar Bilimleri Bölümü \ Bilgi Teknolojisi
Programı

Danışman: Prof. Dr. Taner ALTUNOK

Ortak Danışman: Assist.Prof. Dr. Shadi Al-SHEHABI

Ağustos 2015, 56 sayfa

İlişkilendirme yöntemleri, büyük veri tabanlarındaki değişkenler (öğeler) arasındaki ilişkileri keşfetmek için kullanılır. Bu ilişkiler, büyük veri tabanında saklıdır. Sembolik modeller, ilişkilendirme yöntemlerini ortaya çıkarma hususunda en çok karşılaşılan modellerdir. Maalesef, bu modeller, çok ciddi sınırlamalara maruz kalmıştır. İlişkilendirme yöntemi oluşturmak, büyük miktardaki fazla olan yöntemler de dâhil olmak üzere, devasa sayıda yöntem oluşturmak için uzun bir zaman alan bir süreçtir. Bu problem, yüksek boyutsal tanım alanı ile uğraşılırken, bariz bir şekilde gözükür. Bu problemlerle başa çıkabilmek için, daha önce kullanılmamış bir yaklaşım oluşturulmuştur. Bu yaklaşım, sayısal ve sembolik modeller arasında ilginç bağlantılar kurarak, sayısal ilişkilendirme yöntemlerini ortaya çıkarmak için önerilmektedir. Sayısal modeller, sadece veriler arasındaki önemli ilişkileri muhafaza

eder, bu nedenle bu en potansiyelli ve önemli ilişkilendirme yöntemlerini ortaya çıkartabilir. Bu sayısal yaklaşım, sadece yöntemin her parçasında bir öge olmak üzere, her yöntemin iki ögeye de sahip olduğu anlamına gelen, temel ilişkilendirme yöntemlerini ortaya çıkartabilir. Bundan dolayı, bu tezin amacı, sadece iki öge yerine, farklı büyüklüklerdeki öge dizileri arasındaki ilişkileri keşfetmek maksadıyla, kompleks ve önemli ilişkilendirme yöntemlerini ortaya çıkartabilecek bu yaklaşımı geliştirmektir. Bu yaklaşım, sayısal metotların önemli ilişkilendirme yöntemlerini ortaya çıkarma olasılığını artırır. Çoklu topografik sinir ağır modeli, bu tezdeki ana modeli göz önünde tutar. Bu modelin en güçlü özelliklerinden birisi, sadece bir veri tabanından ilişkilendirme yönteminin oluşturulmasına izin veren genelleme mekanizmasıdır. Bu birleşme yöntemlerinin çıkarılmasının ta kendisi, sayısal bir sınıflandırma modelinin, doğal-sembolik bir sınıflandırıcı olarak hangi ölçüde hareket edeceğini değerlendiren nitelik ölçümlerine dayanır.

Anahtar Kelimeler: Bilgi Çıkarılması, Denenmemiş Öğrenim, SOM Modeli, Çoklu-SOM Modeli, Sembolik Model, Sayısal Model, İlişkilendirme Yöntemleri, Kümeleme Modeli.

ACKNOWLEDGEMENTS

Thanks to God the most compassionate and the most merciful. My Allah's mercy and peace be upon our leader Mohammed, who invites us to science and wisdom, and members of his family and his followers.

I would to express my deep gratitude after God almighty in the completion of this research to my Co-Supervisor Assist Prof Dr. Shadi Al-SHEHABI who suggested this project and gave me a lot of his time. I am indebted for his suggestions and valuable remarks. Also I thanks my Supervisor Prof. Dr. Taner ALTUNOK for his wonderful help in carrying out of this thesis. Finally, my thanks go to the members of my family for their help and encouragement, and thanks continued to my friends whom help in one way or another in bringing out this work.

My God bestow health and happiness to all of them.

TABLE OF CONTENTS

STATEMENT OF NON PLAGIARISM.....	iii
ABSTRACT.....	iv
ÖZ.....	vi
ACKNOWLEDGEMENTS.....	viii
TABLE OF CONTENTS.....	ix
LIST OF FIGURES.....	xii
LIST OF TABLES.....	xiv
LIST OF ABBREVIATIONS.....	xv

CHAPTERS:

1. INTRODUCTION.....	1
1.1 Research Methodology.....	1
1.2 Thesis Objects.....	2
1.3 Research Steps.....	3
1.4 Literature.....	3
2. DATA MINING AND KNOWLEDGE DISCOVERY.....	7
2.1 Introduction.....	7
2.2 Data Mining Concept.....	7
2.3 Teaxtual Data.....	8
2.4 Knowledge Discovery Process of Database.....	8
2.4.1 Data Transformation to Discover Knowledge from Database.....	9
2.5 The Main Objects for Data Mining.....	10
2.6 Data Minining Techinques.....	11
2.6.1 Classification.....	11
2.6.2 Clustering.....	12
2.6.3 Association Rules.....	12

2.7	Text Mining.....	13
2.7.1	Objects of Textual Mining.....	13
2.7.2	Essential tips of Text Mining.....	13
2.7.3	Text Mining Framework.....	14
2.8	Conclusion.....	14
3.	TEXTUAL DATA AND THEIR REPRESENTATION METHODS.....	15
3.1	Introduction.....	15
3.2	Documents Indexing.....	15
3.3	The Representation of Text Data.....	16
3.3.1	Representation of Text Data using the Vector Model.....	16
3.4	Terms Frequency and Weighting.....	17
3.4.1	Inverse Document Frequency.....	17
3.4.2	Weighting Function.....	19
3.5	Conclusion.....	19
4.	CLUSTERING ANALYSIS	20
4.1	General Trends of Clustering Algorithms.....	22
4.1.1	Partitioning Method.....	22
4.1.2	Hierarchical Methods.....	23
4.1.3	Unsupervised Neural Network Algorithms	24
4.1.3.1	Self-Organizing Map.....	24
4.1.3.2	Multi Self-Organizing Map.....	27
4.1.3.2.1	Generalization Mechanism.....	28
4.2	Evaluation of Clustering Methods.....	29
4.2.1	Evaluating Clustering Based on the Characteristics of Distribution Data.....	29
4.3	Conclusion.....	31
5.	KNOWLEDGE DISCOVERY BY ASSOCIATION ANALYSIS	32
5.1	Introduction.....	32
5.2	Knowledge Discovery using Association Rules.....	33

5.2.1	Symbolic Methods.....	33
5.2.1.1	Apriori Symbolic Method.....	38
5.2.1.1.1	Generating Frequent Itemsets by using Apriori Method.....	38
5.2.1.1.2	Generating Association Rules Based on Apriori Algorithm.....	41
5.2.1.2	Close Symbolic Method.....	42
5.2.1.2.1	Generating Closed Frequent Itemsets.....	43
5.2.1.2.2	Generating Association Rules Based on Close Algorithm.....	46
5.2.1.3	Symbolic Methods Defects.....	47
5.3	Numeric Methods of Knowledge Discovery.....	47
5.3.1	Discover Simple Numeric Association Rules	48
5.3.2	Discover Complex Numeric Association Rules.....	49
5.4	Experimental Results.....	51
5.5	Conclusion.....	53
6.	CONCLUSION AND RECOMMENDATION	55
6.1	Conclusion.....	55
6.2	Recommendations.....	56
	REFERENCES.....	R1
	APPENDICES.....	A1
A.	CURRICULUM VITAE.....	A1
B.	MULTIPLE INDEXATION FOR PATENTEES.....	A2
B1.	The Analysis Phase.....	A2
B2.	The Technical Realization.....	A3

LIST OF FIGURES

FIGURES

Figure 1	Knowledge Discovery Process of Database.....	9
Figure 2	Different Way of Clustering The Same Set Points.....	21
Figure 3	Poor Starting Conditions for K-Means algorithm.....	22
Figure 4	Example of the Agglomerative Hierarchical Clustering.....	23
Figure 5	Self Organizing Map Patterns.....	24
Figure 6	Algorithm to Self-Organizing Map.....	25
Figure 7	Neighboring Neuron Consequence (Gaussian and Bubble).....	25
Figure 8	Self-Organized Map Representation Stages to Distribute Specific Points Exist in Two Dimensional Space.....	26
Figure 9	Map generalization Mechanism.....	28
Figure 10	An Itemset lattice.....	36
Figure 11	An Illustration Example of Apriori Approach.....	39
Figure 12	An Illustration of Support Based Pruning According to Apriori	40
Figure 13	Generating Frequent Itemsets According to Apriori Algorithm.....	41
Figure 14	Clipping of Association Rules According to Apriori Principle.....	42
Figure 15	An Example of Closed Frequent Itemsets.....	43
Figure 16	Algorithm Support counting using Closed Frequent itemsets.....	44
Figure 17	Generating Frequent Itemsets According to Closed Algorithm.....	45
Figure 18	Lattice of Closed Itemsets Generates from Lattice in Figure 15.....	46
Figure 19	Algorithm for Extracting Simple Association Rules from Clustering Model.....	49
Figure 20	Algorithm for Extracting Complex Association Rules from Clustering Model.....	50

FIGURES

Figure 21	Some Extract Complex Numerical Association Rules from USE Database.....	53
Figure 22	Example of a Patent Abstract with Its Generated Multi-Index.....	A4

LIST OF TABLES

TABLES

Table 1	The Difference Behavior to Collection Frequency and Document Frequency in Reuter's Collection.....	18
Table 2	Inverse Document Frequency and Document Frequency Values of a Set of Terms that Taken of Reuters Agency Documents.....	18
Table 3	An Example of Market Basket Transaction.....	32
Table 4	Number of Complex Association Rules Generated using Algorithm..	51
Table 5	Extract Symbolic Association Rules from USE Database.....	52
Table 6	Summary of the Results of Patent Indexation and Map Building.....	A5

LIST OF ABBREVIATIONS

CF	Collection Frequency
CONF	Confidence
CONFSUP	Confidence Support
FCI	Frequent Close Item
FG	Frequent Generator
IDF	Invers Document Frequency
IE	Information Extraction
IF	Term Frequency
IM	Information Mining
IR	Information Retrieval
KDD	Knowledge Discover in Database
MINSUP	Minimum Support
MSOM	Multi Self-Organizing Map
Prec	Precision
Rec	Recall
SOM	Self-Organizing Map
SUP	Support

CHAPTER 1

INTRODUCTION

1.1. Research Methodology

The large increase in the databases information represent a major challenge for data analysts, where the general content of these databases become more complex to understand, and thus the knowledge discovery in useful and suitable way becomes more complex. In addition, the size and various of data prevent any knowledge discovery processes manually, which led to rise of need to suggest methods and active tools for this purpose. The study of these methods called data mining or knowledge discovery in databases. The discovery process is not vulgar process to discover implicit, unknown and useful information from huge database [1,2]. This mission required completion analysis of dimensions of inspect data.

Symbolic Models are the most used methods in knowledge discovery of database, but these methods suffer of important defects, so the process of association rules generation [3], which represents one of knowledge models types, it is an expensive and long process of being generate a huge number of redundant association rules. This prevents the statistic rules and choose what kind of them, where data are huge and represented in numerical descriptive space with many dimensions. Often this last case found in process of text data which we are going to work on it in this thesis. In order to solve these problems due to of using symbolic methods, unsupervised numerical classification models (clustering models) have suggested to extract useful knowledge [4,5] including clustering methods that characterized by their huge capability to collect and summing up data. Therefore the separation of the different data from each other, so the different properties will separate by separation of weak associations between the properties. In spite of these tries remained based on knowledge discovery evolution on norms of symbolic methods that based on the database only which consider data and its properties have the same degree of

importance. This leads to the same problem that exist in symbolic methods, where it does not process the strength of numerical association rules between different properties. Since all these properties are equal in importance. Numerical symbols can solve the problems that related to symbolic models, an alternative algorithm is proposed in this thesis for extracting complex numerical association rules instead of extracting simple ones by a previous algorithm.

This thesis is structured as follows:

In **Chapter 2**, the concept, objects of data mining and the most important using techniques in knowledge discovery including clustering methods and association rules analysis.

In **Chapter 3**, shows the process of pretreatment of text data and methods of data representations in purpose of doing knowledge discovery.

In **Chapter 4**, will show the concept of artificial neural networks and neural clustering in order to reach to the correct extract knowledge

In **Chapter 5**, different ways of knowledge discovery in symbolic and numerical types and we will propose on alternative algorithm for extracting complex numerical association rules.

1.2. Thesis Objects

Thesis objects are the following:

1. Solving the problem of symbolic models which produce huge number of association rules based on clustering models.
2. Extract important and complex association rules to discover relationships between itemsets from different sizes instead of two items.

1.3. Research Steps

This research requires depth knowledge about unsupervised numerical classification methods (clustering) of data, in addition knowledge of symbolic and numerical methods to produce association rules and mathematical methods to evaluate these rules according to the following steps:

- 1- Theoretical study of data mining or knowledge discovery of database.
- 2- Study of text data and clustering analysis.
- 3- Study of association rules extraction using symbolic and numerical methods.
- 4- Propose new alternative algorithm for extracting complex association rules.
- 5- Determine text database to work on it.
- 6- Apply clustering neural network model (models) which are Self-Organizing Map (SOM), Multi Self-Organizing Map (MultiSOM) using its generalization mechanism in order to classify data through working on them by optimal number of cluster [6,7].
- 7- Extract complex numerical association rules by applying one alternative algorithm on Multi Self-Organizing Map.

1.4. Literature

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is how to make the database a competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of the information simply to remain competitive. A mixture of new techniques and technology is emerging to help sort through the data and find useful competitive data.

Fayyad U, Shapiro G, Smyth P, et al (1996) by knowledge discovery in database, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large database has been

recognized by many researchers as a key research topic in database systems and machine learning. Companies in many industries also take knowledge discovering as an important area with an opportunity of major revenue. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications [1].

Association rule clustering is useful when the user desires to segment the data. Lent B, Swami , Wisdom J, et al (1997) proposed a clustering association rule in which they measure the quality of the segmentation generated by association rule clustering system. It uses the minimum description length principle of encoding the clusters on several databases including noise and errors. scale-up experiments show that association rule clustering system using the BitOp algorithm scales linearly with the amount of data [8].

Agrawal R, Srikant R, (1994) most of the procedure focuses on finding the frequent itemset, but several algorithms are available for finding rare items efficiently. Working of that study, motivation for proposal of that algorithm, advantage and their limitations are briefly described here. They are as follows: Existing rare itemset mining approaches are based on level wise approach similar to Apriori algorithm which uses a single minimum support value at all levels to find frequent itemsets. Before generating frequent itemsets, algorithm generates all candidate itemset having number of item from that level. If the support of the subset of candidate itemset is greater than or equal to the user defined minimum threshold it said as frequent item. With use of this algorithm, we can classify frequent patterns not rare patterns. It inherits drawback of many frequent itemset generation and also takes large time, space and memory for candidate generation process [9].

Jadhav et al. (2011) in this study, implementation of a system for pattern discovery using association rules is discussed as a method for web usage mining. Different transactions that are closely related to each other are grouped together by the use of clustering approaches on the preprocessed dataset. The analysis of such clusters will lead to discovery of strong association rules. They obtained all significant association rules between items in the large database of transactions. The relation between different page requests was found. The support and the confidence values of extracted rules are considered for obtaining the interest of the web visitors [10].

Discovering frequent patterns from large datasets using association rule mining algorithms has been extensively employed in the literature (Goethals, 2003, Sotiris and Dimitris, 2006) due to great capabilities of association rule mining algorithms in improving business profit. Over the past decade, a variety of research efforts improved the processing time of association rule mining algorithms like the one introduced by Ceglar and Roddick (2006). Goethals et al (2005) have reduced the output set size that is generated by association rule mining. Overviews of methods and techniques for improving the efficiency of association rule mining are presented below [11].

The disadvantage of Apriori algorithm made the researchers to think about new techniques to mine frequent patterns. The 2 main negative sides are the possible need of generating a huge number of candidates if the number of frequent 1-itemsets is high or if the size of the frequent pattern is big, the database has to be scanned twice repeatedly to match the candidates and determine the support. Mining the frequent patterns without candidate generation would be a big improvement over Apriori. That is what the frequent pattern growth (FP-growth) algorithm does (Han et al, 2000). Wang et al (2002) presented PRICES, an efficient algorithm for mining association rules. Their approach reduces large itemset generation time, which dictates most of the time in generating candidates by scanning the database only once. Another algorithm called Matrix Algorithm developed by Yuan and Huang (2005) generates a matrix which entries 1 or 0 by passing over the cruel database only once. The frequent candidate sets are then obtained from the resulting matrix. Association rules are then mined from the frequent candidate sets [12].

The task of mining association rules consists of two main steps. The first involves finding the set of all frequent itemsets. The second step involves testing and generating all high confidence rules among itemsets. In this study we show that it is not necessary to mine all frequent itemsets in the first step, instead it is sufficient to mine the set of closed frequent itemsets, which is much smaller than the set of all frequent itemsets. It is also not necessary to mine the set of all possible rules. We show that any rule between itemsets is equivalent to some rule between closed itemsets. Mohammed J. Zaki and Ching-Jui Hsiao et al (2002) thus many redundant rules can be eliminated. Furthermore, we present closed association rule mining, an

efficient algorithm for mining all Closed frequent itemsets. An extensive experimental evaluation on a number of real and synthetic databases shows that closed association rule mining outperforms previous methods by an order of magnitude or more. It is also linearly scalable in the number of transactions and the number of closed itemsets found [13].

Kohonen et al. (1990) in the article, a utilization of modern methods of data mining is described and especially the methods based on neural networks theory are pursued. The advantages and drawbacks of applications of multiplayer feed forward neural networks and Kohonen's self-organizing maps are discussed. Kohonen's self-organizing map is the most promising neural data-mining algorithm regarding its capability to visualize high-dimensional data [14].

Lamirel, Al Shehabi, François, and Polanco (2004) present a new approach whose aim is to extent the scope of numerical models by providing them with knowledge extraction capabilities. The basic model which is considered in this study is a multi-topographic neural network model. One of the most powerful features of this model is its generalization mechanism that allows rule extraction to be performed. The extraction of association rules is itself based on original quality measures which evaluate to what extent a numerical classification model behaves as a natural symbolic classifier [4].

CHAPTER 2

DATA MINING AND KNOWLEDGE DISCOVERY

2.1. Introduction

In the last years the number of data that collected by advanced information system have increased. In order to analysis this huge data on mechanism has suggested which called knowledge discovery of database. The main approach for this mechanism is data mining. Data mining applied an active algorithms to discover important models to arrange data. In addition the complexity of data has increased as result of resources diversity. Then, the new data mining methods are necessary to be benefit of these data. We will concentrate in this thesis on data mining and knowledge discovery to extract association rules. The text data consider most important types and wide broad. We will show in this chapter, the concept and objectives of data mining which represent pivotal stage in data knowledge discovery process from database. Then we will expose the most important techniques used in knowledge discovery, including clustering methods and association rules, that are used in this thesis.

2.2. Data Mining Concept

Data mining can be define as applied the statistical methods on a group of different data to clarify the relationships between data, or to gather specific features about real or virtual assets such as (person, or group of people, or commercial project, or other events). Then the results could be used to give statements (or information) about the real or virtual features of these events. The broad definition for using data mining is unintuitive, implicitly and unknown previously for potential and useful information

of data [1]. Data mining depend on many of assumptions which include the following:

The data that are discovered, the data that are collected in order to access to effective knowledge by doing many tasks that we will come to study and detailed later in order to reach to the optimal use of these data in all fields [15].

2.3. Textual Data

As we have seen that most of the previous data patterns are text data collected from multiple sources, according to the purpose of use, and then get a huge amount of these data. Therefore, it become necessary to classify these data and analysis information that exist in documentry data in purpose of knowledge discovery from them. Text data is the most important type of data at all because the huge information that stored in documents and because of the difficulty of analyzing and interpreting this information. This is led us to choose the text data to work on it in this thesis because it have great benefit in society and scientific research.

2.4. Knowledge Discovery Process of Database

Knowledge Discovery in Databases is not a vulgar process to discover an implicit information, unknown previously, useful, new and correct of a huge database [1, 2]. The knowledge discovery of databases is an interactive and iterative process, and includes a lot of steps and decisions set by the user. BRACHMAN and ANAND offered in 1996 and applied point of view to KDD method with assurance on the interactive nature of method [16]. We will mention here the main steps for knowledge discovery process of database [17]:

- 1- The expansion in understanding application field and understanding previous knowledge, determine the goal of KDD process by user.
- 2- Establishment a group of target data: through data group or partial group selection of features of data where knowledge discovery of this group.
- 3- Data Cleaning and Preprocessing: these processes include the following: remove noise if appropriate, and determine strategies to address the lost data.

- 4- Data Reduction and Projection: includes find useful features to represent data according to target of this mission.
- 5- Suitability between aims of KDD process and practical data mining method such as Summarization, Classification, Regression and Clustering.
- 6- Exploratory Analysis and Model Selection: where algorithm has chosen to data discovery and method to search models of data, and connect between data mining methods and KDD process.
- 7- Data Mining: in data mining is looking for useful models of data that have specific representation.
- 8- Interpretation and Evaluation of models: In this step, the interpretation of models that have been discovered with a possible return to any of the steps from (1) to (7) in every iterative process and evaluate it after each iterative.
- 9- Dealing with knowledge discovery: knowledge used directly and integrated with other system.

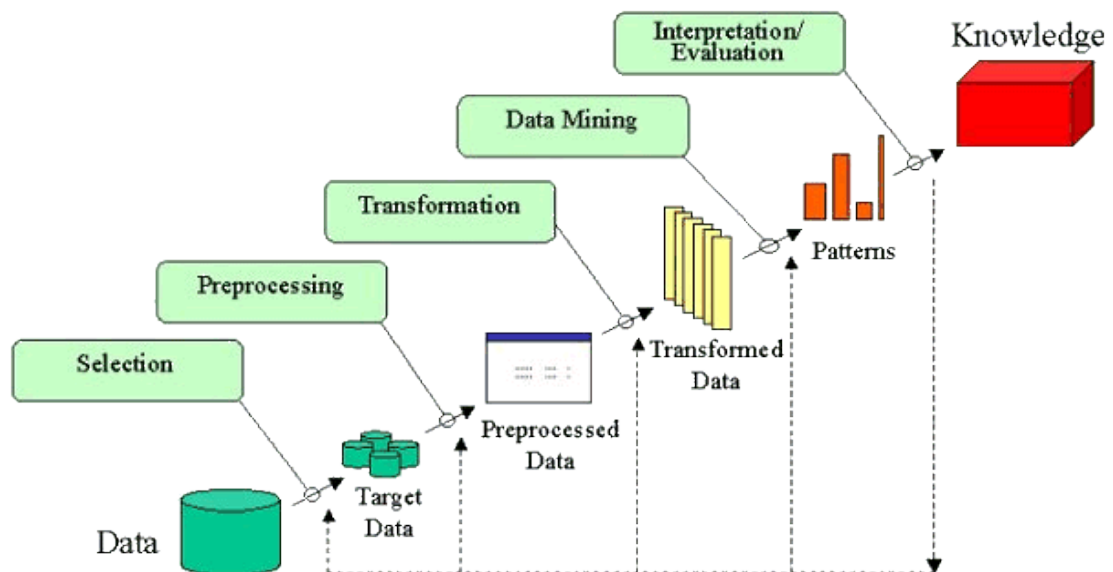


Figure 1 Knowledge Discovery Process of Database

2.4.1. Data Transformation to Discover Knowledge from Database

The most applications of knowledge discover from database, the elements of data does not available in the way that can be processed by using of algorithms of data

mining. So the elements of data should transfer to more clear representation to reflect directly the coherent possible properties. For example the photos convert to features vectors describe figure, structure and color, also text documents convert to vectors words [18], as we will see in chapter three.

2.5. The Main Objects for Data Mining

The goals of data mining or knowledge discovery have been determine according to the purpose of system utilizing. We can distinguish between two types of these goals [17]:

- 1- **Verification:** The system has been designed to prove or check assumption submitted by user.
- 2- **Discovery:** The system finds new important models automatically.

The goal of discovery can be divided for two goals:

- **Prediction:** The system finds models to predict the future behavior of some elements.
- **Description:** The system finds models that will be presented to the user in an understandable way.

Manual analysis which concentrates on specific set of data become insufficient because of the huge amount of data which can be up to one billion element of the data. And there is no alternative to carry such a heavy burden of computer software only to find a group that models itself. Fayyad has described the huge amount of available data which found within raging storm of data [19]. He said that (We are in urgent need of information technology process and the ways in which can manage this huge amount of data in order to obtain quickly and significant results. The expansion in KDD, learning machine and fields which related with them give us at last some solutions, the future of data scientific process depend on create an effective and integral tool for these methods). It is not easy to find important models or search for a specific structure of data, where the determine of important models depend on applied field. It is possible to define the important probabilistic models as those models with higher conditional probability models (according to norm of specific

verification in basket shopping analysis). The conditional probability in analysis of shopping basket to buy specific goods to other purchased goods has studied. So the require is: Firstly put future conclusions similar to occurrence of items in the basket, secondly, placed causal statements for procurement models in ideal way. If you have ability to convince someone to buy item *A*, then probably he may buy item *B*, and this is what will be detailed in the chapter five.

2.6. Data Mining Techinques

Data mining includes compatibility ways with data or identify model figure. The compatibility ways play important rule in knowledge discovery, so the models that reflect the important and useful knowledge consider part of KDD process at all, then the judge of personal user is required. Then, two main mathematical methods have used in the compatibility process: the first one is statistical and the second one is logical. Where the statistical approach allows to inevitable effects in the model, while the logical methods completely absolutly viable. We will focus in our thesis on the statistical approach in data mining becasue it is the widest using approach in different scientific applications in data mining. Most of data mining methods are based on known and tested techniques from machine learning and statistic such as classification, clustering, association rules and some other techniques [20].

2.6.1. Classification

The tools that mining data must conclude style of mining from database, so the user should know one object at least. Database contains one feature or more refer to class. The classification process is a subsequent learning which drop elements of data in one class or several classes and these classes are defined previously. After training setp, classification methods can drop unknown new elements in these classes [3]. Several methodes are used for this purpose sach as:

- Discriminate Analysis
- Rule Induction Methods
- Decision Tree Learning

- Neural Network
- K Nearest Neighbor
- Case-Based Reasoning
- Genetic Algorithms

2.6.2. Clustering

Clustering is the process of creating divisions of data, so the elements in each cluster are similar depending on a specific scale. The cluster is a set of data elements combined with each other because of their similarity. Often, the elements are analyzed to comprehensive group or contradictory group of clusters [7]. When the learning is unsupervised as in clustering case so the system has to create its own clusters by itself. That is mean the system put the available data in database within clusters. The system must produce partial sets of sets. It will be discussed later in chapter four.

2.6.3. Association Rules

By taking the set of items and a set of data elements and each data element contains a number of items taken from these itemsets. Association sub sequence is a reverse process to set of data which return associations or models that found between itemsets. These models can be expressed by such rules: 72% of records that contain items A, B, C also includes items D, E . The specific percentage in case 72% called confidence factor to association rule. Items A, B, C in this rule form a reverse part to items D, E part, that is mean the rule form of the figure $\{A, B, C\} \rightarrow \{E, D\}$. The association rules can include any number of items in any one of these parts. Association analysis will be detailed in chapter five in purpose of discovered knowledge[8].

2.7. Text Mining

Text mining is defined as the mining of textual data [21], or knowledge discovery from textual database [22], generally text mining refers to knowledge discovery process or the useful models of unstructured textual documents. The field may consider as an extension of data mining or knowledge discovery of structured database in order to extract numerical association rules [1,23].

2.7.1. Objects of Textual Mining

- 1- Discovery and using knowledge that contained in available documents set and extraction important information from sets of documents and different resources.
- 2- Data mining allows enquiry about questions that based on texts, and discover information and gained answers cannot be imagine about these questions.

2.7.2. Essential tips of Text Mining

There are three main tips of text mining:

- 1- **Preprocessing:** Preparation text to produce documents with organized figure. Where the text prepared and transformed to a rich text with information and expressed by document matrix. This huge matrix explains every item frequency include set of data. During this stage features have discovered throughout determine positions of limit parts of information such customers' names, classifications and addresses [18].
- 2- **Reducing:** Reducing results to less size to facilitate dealing with it. Then, the use of a mathematical technique called analysis of singular value of original documents items to smaller matrix, during this process most of unimportant words neglected and choose more important words and used the new matrix.
- 3- **Mining:** Data mining by using previous data mining techniques, where clustering, classification and prediction methods are applied on reducing data by using of common data mining techniques.

2.7.3. Text Mining Framework

Text mining is related with the following fields [24]:

- 1- **Information Retrieval (IR):** Retrieve documents that can be consider related to a specific subject [25].
- 2- **Information Extraction (IE):** The information could be extracted from selected documents. This extraction is a process of filling out specific molds with expected information by the user. Information extraction is most important in text mining process.
- 3- **Information Mining (IM):** Database becomes actively organized for every document when the document mold filled in accordance with data mining techniques.
- 4- **Interpretation:** Chosen documents can be interpreted here. This interpretation can be in natural language.

2.8. Conclusion

The concept of data mining has exposed in this chapter which represents the main core for knowledge discovery steps. The knowledge discovery is necessary as mentioned previously in case of huge data with high dimension of data textual as data, these data required set of operations to show data in form that can be processed by data mining techniques, clustering represents the essential technique to apply on data in this thesis, to keep only the important relations between data from these clusters we can extract important association rules, which is one of knowledge required forms.

CHAPTER 3

TEXTUAL DATA AND THEIR REPRESENTATION METHODS

3.1. Introduction

In this thesis we focus on the text data which consider the most difficult data in terms of analyzing and processing for the purpose of classification clustering and knowledge discovery to extract important and complex association rules. There are two types of documents representation. The first one is the Boolean representation and the second one is the vector representation. Where we are going to concern with vector model to represent documents in order to achieve good performance on a range of word similarity tests. In addition to convert this information into a model that can be dealt automatically while maintaining of association information with each other. Document can be consider as a unit of association information, these information are useful and not useful. This documents group formed documentary database, what is needed is to extract the existent relationships between knowledge on one hand, and thereby the relationships between documents from another hand. We will expose in this chapter the process of pretreatment of text data throughout information indexation process, and then we will show the representation of data methods for the tasks of clustering and knowledge discovery to extract numerical association rules.

3.2. Document Indexing

The main object of documents indexation is extract the most important items, and represent these documents in shape of formal model. This process passes by stages to analysis and interpretation content of documents. Documents indexation considered difficult process in indication of the formal model that will used to

represent documents. Formal model is used to perform the documents analysis. Since the difficulties could be process by explain documents content through applied specific systems in order to discover knowledge [25].

3.3. The Representation of Text Data

If we have N document, and the set of these documents contain M items, then the natural representation of these documents will be through documents matrix items with its dimension $N \times M$. So the documents represent rows and items represent columns. Usually the items return to their language before indexation process. For example the words *jealousy* and *jealous* represent in one item that is mean with one dimension in document vector. There are two types of documents representation. The first one is the Boolean model and the second one is vector model [25]. We are going to concern with vector model to represent documents in this chapter.

3.3.1. Representation of Text Data using The Vector Model

The documents representation set by vectors in radial space called the vector model to represent text data [25,26]. Vector model considers a basis to clustering and classification documents. In addition to data mining processes. The term language uses to represent document by vector. This process occurs through neglect of items order for each other in this representation as founded in document. The words weighting approach followed in this case as explain in the next paragraph. So the vector stored items appearance number in document without interest of organized or arranged between each other. Throughout this representation we found the represent of phrase “marry is quicker than john” is similar to “john is quicker than marry”. So axiomatically we conclude that similarity of two documents in contains if they are similar in their vectors. This model will be used in thesis, where we need to distinguish between the different documents, in addition to distinguished between the different items inside this document in purpose of classified in categories and get the active knowledge by using the artificial neural methods for clustering.

3.4. Terms Frequency and Weighting

We can put weight to every terms in document. This weight depends on term appearance number in document, where the simplest approach to determine terms weights that equal in repetition of appearance every term in document. Weight scheme refers to term frequency that symbolizes by $TF_{t,d}$ so these letters refer to term and document respectively. It is possible to represent weights group by vector with one factor separately for every term. Towards this end, we assign to each term in a document a weight for that term, that depends on the number of occurrences of the term in the document. We would like to compute a score between a query term t and a document d , based on the weight of t in d [25].

3.4.1. Inverse Document Frequency (IDF)

The consideration that all terms in document are equal in importance make data mining process useless and contained a lot of mistakes. For example, if we take documents of Reuter agency about assurance industry we notice that assurance mostly founded in every document. For that, there is mechanism to reduce the effect of terms appearance which repeated in documents sets, in order to have meaning in data mining. As primary idea, the terms weights that have high frequency should reduce through repetitions appearance of these terms in the set. The idea accomplished throughout reduction of term frequency by using factor related with term frequency in the set. Instead of, it is useless to use document frequency as a set document that contain the term. When we trying to distinguish between documents so the using of statistical on level of documents better than utilize statistical on level of comprehensive set during utilizing of collection frequency. Collection frequency is number of term repetition appearance in document set [18]. The table 1 displays through simple example the reason behind preference document frequency to collection frequency where the difference behavior to document frequency and collection frequency can be distinguish. In particular, the value of collection frequency for “try” and “insurance” almost equal but document frequency value is

difference. The weight of “insurance” word bigger than of “try” word in spite of “insurance” can be found much less of “try”.

Word	<i>CF</i>	<i>DF</i>
Try	10422	8760
Insurance	10440	3997

Table 1 The Difference Behavior to Collection Frequency and Document Frequency in Reuter’s Collection

Document frequency of a term is used in order to find weight as follows:

Where N represent the number of documents in the collection thus the inverse document frequency of a term t is given as follows [25]:

$$IDF_t = \log \frac{N}{DF_t} \quad (3.1)$$

So inverse document frequency of seldom term (less repetition) might be high while inverse document frequency value of frequent term is low. As shows in table 2 inverse document frequency values document set with 806791 in Reuter’s agency. In addition to different terms frequencies values that taken from these documents.

Term	<i>DF_t</i>	<i>IDF_t</i>
Car	18165	1.65
Auto	6723	2.08
Insurance	19241	1.62
Best	25235	1.5

Table 2 Inverse Document Frequency and Document Frequency Values of a Set of Terms that Taken of Reuters Agency Documents.

3.4.2. Weighting function (TF-IDF)

The phrase of term frequency and inverse document frequency can be merged to get combined weight to every term in the document. The term has been weighted in document according to the following equation:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (3.2)$$

In another way $TF - ID_{t,d}$ refers to weight of term in document. Then throughout the previous equation (3.2) we notice the following:

- 1- The weight is being high when the repetition of term is in a small number of documents (this gives high distinguished power to these documents).
- 2- The weight is being low when the repetition of term is in a few document.
- 3- The weight is being lower when the term found actually in the entire documents. Therefore every document can be represent by vector, and specialized one content within for every item in the dictionary and in addition to its weight according to the previous equation (3.2).

3.5. Conclusion

In this chapter the mechanism that processed text data has been exposed which is represented by indexation. These documents have presented by a vector model. We had seen that the vector method is the best representation method. We had chosen this method as a style to represent data in this thesis because it distinguishes between the different information inside document throughout show the important of every information in compared with counterparts by use specific mathematical function, and that is cannot be represent by boolean model. The selection of a vector model is an important and active in analysis data processes (during studying similarity between data, then could be classified and clustered this will exposed in detailed in chapter four and chapter five). We use in this thesis one database called USE (as explained in appendix B). We will apply on this thesis data analysis processes and knowledge discovery to extract numerical association rules as we will see in chapter four and chapter five.

CHAPTER 4

CLUSTERING ANALYSIS

Cluster analysis groups data aims depend only on information exist in the data that describes the aims and their relationships. The goal is that the targets within a set be similar (or related) to one another and different from (or unrelated to) the aims in other groups. The great similarty within a group and the greater the difference between groups, the better or more distinct the clustering [27]. In different applications, the idea of a cluster is not well defined. To better understand the difficulty of deciding what constitutes a cluster, consider figure 2, which shows twenty points and three different ways of dividing them into clusters. The markers shapes refer cluster membership. Figures 2 (b) and 2 (d) divide the data into two and six parts. However, the apparent division of each of the two larger clusters into three sub clusters may simply be an artifact of human visual system. Also, it may not be unreasonable to say that the points from four clusters, as shown in figure 2 (c). This figure indicates that the cluster definition is imprecise. The best definition based on the data nature and the wanted results. However, clustering derives labels only from the data. In contrast, classification in the sense is supervised classification, new, unlabeled objects are assigned a class label utlize a model improve from objects with known class labels [28]. For this reason, analysis of cluster is sometimes illustrates to as unsupervised classification. When the term classification is utlize without any qualification within data mining. It typically indicates to supervised classification [27]. Also, while the terms dissection and partitioning are sometimes utlize as synonyms for clustering. These terms are frequently utlize for styles outside the traditional bounds of cluster analysis. We find through the figure that the description of clustering is imprecise. So the best definition based on the data nature and the wanted results. There are differnet other names for clustering such as unsupervised learning.

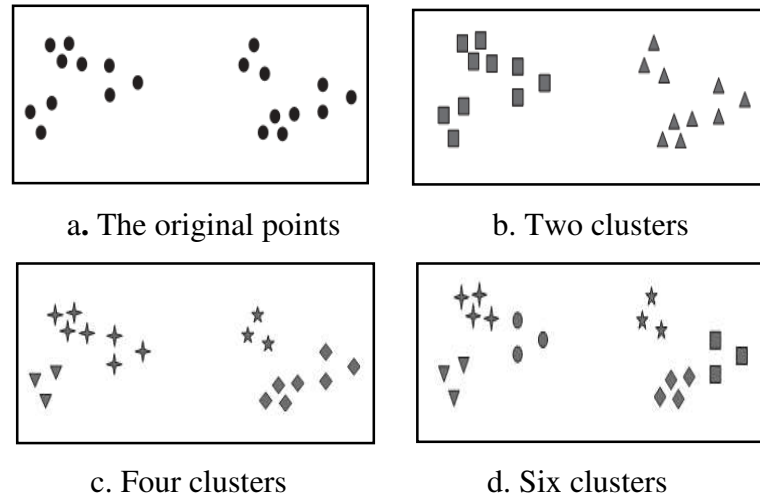


Figure 2 Different Way of Clustering the Same Set Points

Clustering utilize to gain a notion of the existing data set, thus clusters often be enough to give us a notion of the distribution of data within a data set. We are mention several useful examples about clustering such as gathering costumers for assembling documents and shopping. Also in purpose of set the order of the answers in the search engines. And the discovery of models in spatial and temporal distribution of the disease. In this chapter, we'll demonstrate the dissimilar kinds of clustering algorithms, which constitute a central base to carry out investment groups. This involves a clustering way that represent the data correctly and accurately. This process needs evaluation and analysis of clustering objectively through mathematical models. It will be showing in this chapter. We have to begin with ideal cluster in discover knowledge. We are going to display in this chapter artificial neural networks which are the most important classifications and clustering method. It is used in active way in data mining processes. The strength and important of artificial neural networks such as (neural gas, Multi gas, self-organizing map and Multi self-organizing map) due to their topographic structure that means throughout the related and effective relationship between neurons in the network that imitate human brain processes. In this thesis we are going to work on multi self-organizing map, different of learning functions that exist in data when classification processes, through the ineffective by noisy data that distort classification processes and lack to sufficient accuracy to investment in discover knowledge. We will concentrate in this chapter on using of unsupervised learning method that represent neuron clustering which will be

used specifically in knowledge discovery processes to in order extract complex association rules.

4.1. General Trends of Clustering Algorithms

Clustering algorithms can be classified into categories depending on the different styles of each algorithm as follows [30]:

4.1.1. Partitioning Method

The partitioning clustering algorithms dividing data base that contain n element within separated cluster k , where each cluster contains at least one element, and each element belongs to one cluster only. In order to get a good clustering, partitioning method dividing the data set to the initial parts. Then working to improve the quality of clusters through multiple duplicates, some data elements will transfer from cluster to another in every duplicate. The partitioning clustering algorithms often help in guidance but does not guarantee to get high-quality clustering. One of the most important partitioning clustering algorithms is *k-means* as in the figure 3. It is useful to find k which is a circle cluster describes group of data. We note that the partitioning clustering algorithms needs to determine prior to the number of clusters that must be found [31].

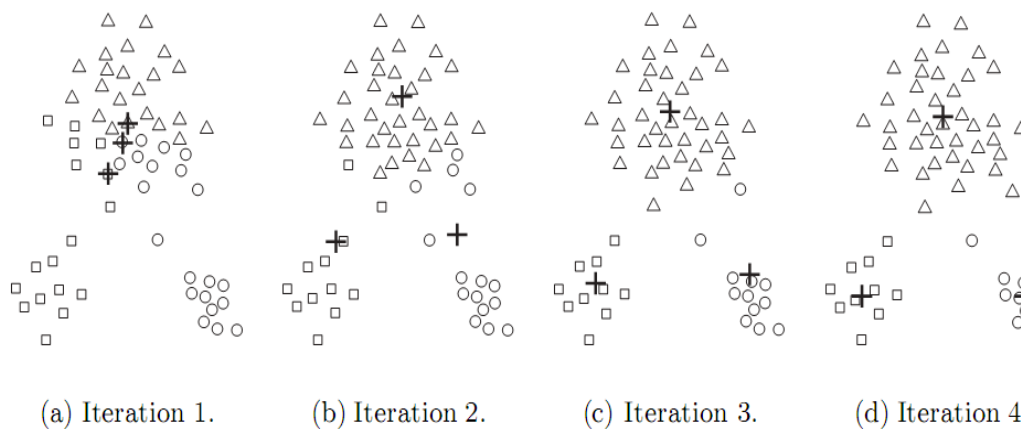


Figure 3 Poor Starting Centroids for K-Means Algorithm

4.1.2. Hierarchical Methods

Hierarchical method sets up hierarchical analysis to data set. Where hierarchical clustering allows to smaller clusters to be part of largest clusters and more general. There are two styles of hierarchical method which are agglomerative clustering and divisive clustering [32]. The agglomerative hierarchical clustering tracking merger approach starting from the bottom to up. It combines data gradually or step by step to reach to one cluster. At the base (bottom) of each element is to be a cluster in itself, and then at the top level is combination of all adjacent clusters to each one cluster, and combination continue at higher levels to completed construction clusters tree, or to limit access to a predetermined condition. Tree clusters calls dendrogram, it contains different size clusters, as in figure 4. For the divisive hierarchical clustering methods are tracking dividing approach starting from top to bottom, and so on successively generating clusters are divided into smaller clusters, setting off from one cluster representative and divide it to be the representation of each element of the data one cluster, starting from one representative cluster and divided it to be every element of data represent by one cluster. Consequently hierarchical clustering is very expensive when forming the matrix distances they deal with many of the data dimensions. At the end of process division to become each cluster consists of only one item, or when we reach to stop condition. There are other kinds of hierarchical clustering algorithms such as cure [33], Chameleon [34] and Birch [35].

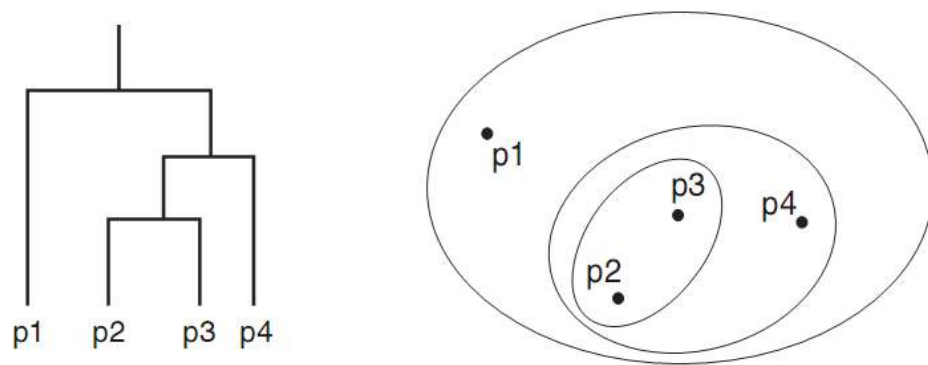


Figure 4 Example of the Agglomerative Hierarchical Clustering

4.1.3. Unsupervised Neural Network Algorithms

4.1.3.1. Self-Organizing Map

Kohonen self organizing map [36] is the most important methods that related with unsupervised learning methods to organize data with various dimensions in structure of self-organized neuron which neighboring relationships between neurons are known previously. The process based on classification approach and drop data on limit dimension network. Self-organizing map has used successively and actively in many applications. Self-Organizing Map used in texts mining field. Self-organizing map consist of several neural processing elements that called neurons, and organized according to specific topographic. Self-Organizing Map is vector and with one dimension or usually organized in two rectangular dimension map or six dimension or three dimension as in figure 5.

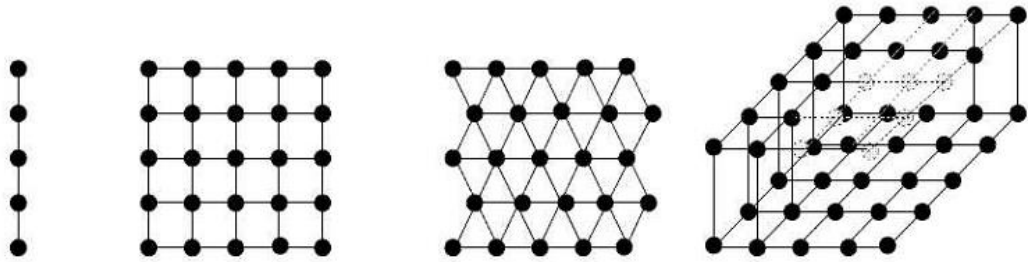


Figure 5 Self-Organizing Map Patterns

Later every c unit will refer to W_c vector where $W_c \in \mathfrak{R}^n$. It is important to notice that these vectors have the same number of dimensions [37]. The distance on this network used as norm in order to conditioning the neuron $r=a_{km}$ when the neuron $s=a_{ij}$ represent the winner. This distance is “Manhattan”. The complete algorithm to Self-Organizing Map as in figure 6.

$$d(r, s) = |i - k| + |j - m| \quad (4.1)$$

Preparing set A with neural c_i which its number is $N = N_1 \cdot N_2$, $A = \{c_1, c_2, \dots, c_N\}$ With vectors its support $W_{c_i} \in \mathfrak{R}^n$ that chosen randomly. N_1 and N_2 Represent two rectangular dimensions. Preparing time medium with primary value. Note that h_{rs}


```

1. Initialize the weight vectors,  $W_{c_i} \in \mathcal{R}^n$ 
2.  $t = 0$ 
3. for epoch = 1 to  $N_{epochs}$  do
4.   for input = 1 to  $N_{inputs}$  do
5.      $t = t + 1$ 
6.     for  $x = 1$  to  $X$  do
7.       Compute distance  $d_k$ ,  $s(x) = \arg \min_{c \in A} \|x - W_c\|$ 
8.     end for
9.     Compute BMU for current input,  $d_c(t) = \min_k d_k(t)$ 
10.    for  $k = 1$  to  $k$  do
11.       $\Delta W_r = \epsilon(t) \cdot h_{rs} \cdot (x - W_i)$ ,  $r = 1, 2, \dots, N$ 
12.    end for
13.  end for
14. end for

```

Figure 6 Algorithm to Self-Organizing Map

neuron neighboring consequence used to define the strong relative to condition neuron r of neural network and has several types. Figure 7 mentions to Bubble consequence the simplest kinds and gaussian consequence that given in the relation: $h_{rs} = \exp\left(\frac{-d(r,s)^2}{2\sigma^2}\right)$. It gives $\sigma(t)$ and learning network average $\epsilon(t)$ according to following relations: $\sigma(t) = \sigma_i(\sigma_f/\sigma_i)^{t/t_{max}}$, $\epsilon(t) = \epsilon_i(\epsilon_f/\epsilon_i)^{t/t_{max}}$.

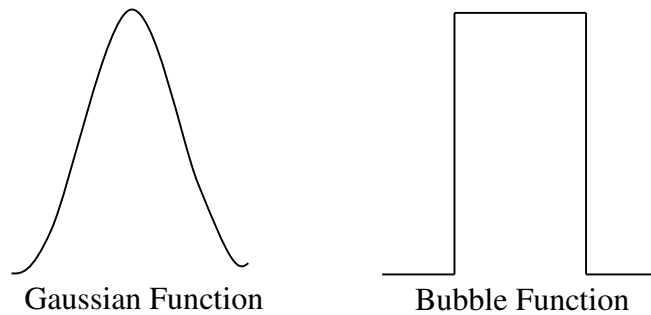
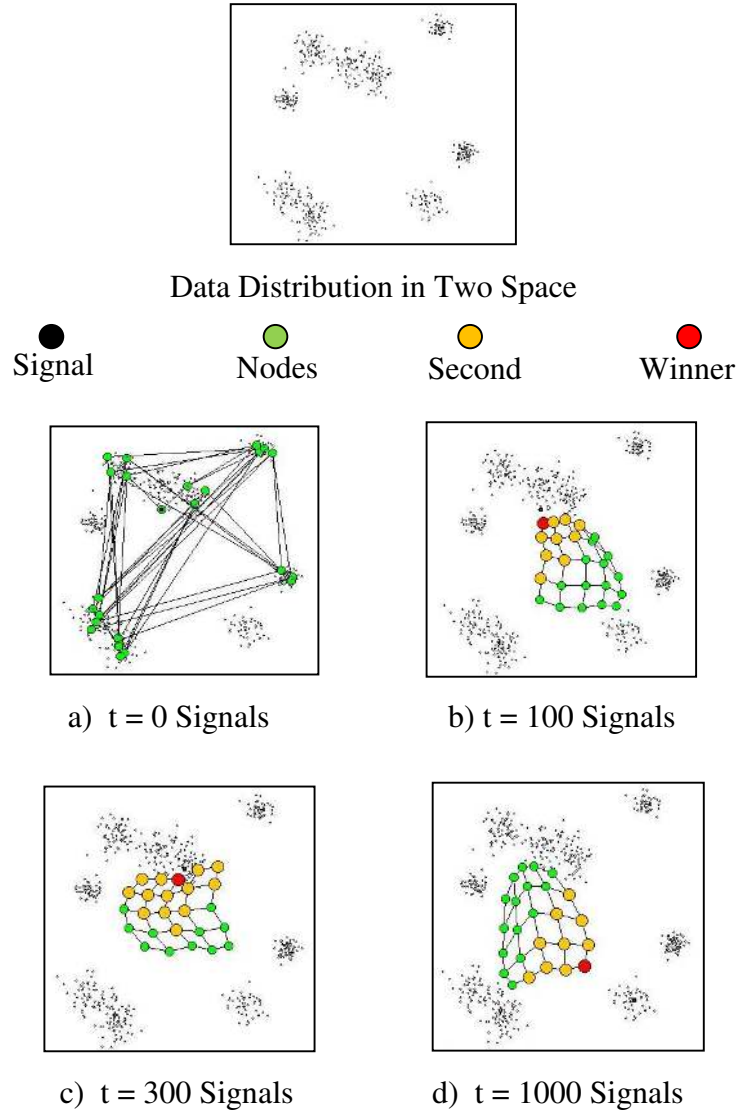


Figure 7 Neighboring Neuron Consequence (Gaussin And Bubble)

It should choose appropriate primary values time mediums consequences (σ_i, ϵ_i) and final values (σ_f, ϵ_f) . Figure 8 shows some of self-organized map to specific distribution exist in two dimensional space, it is a tiny state to text data that exist in space with n dimensions, where (a) represents the initial case of the network, (from b to f) medium cases after learning repetition group, (g) the final case to network, (h) voronoi that agree with of final cases (Voronoi region represents set of space points for which these points near to neuron that agree with other neurons) where:

$$\sigma_i = 10 \quad , \quad \sigma_f = 0.01 \quad , \quad \epsilon_i = 0.5 \quad , \quad \epsilon_f = 0.005 \quad , \quad t_{max} = 40000 \quad , \quad N_1 = N_2$$



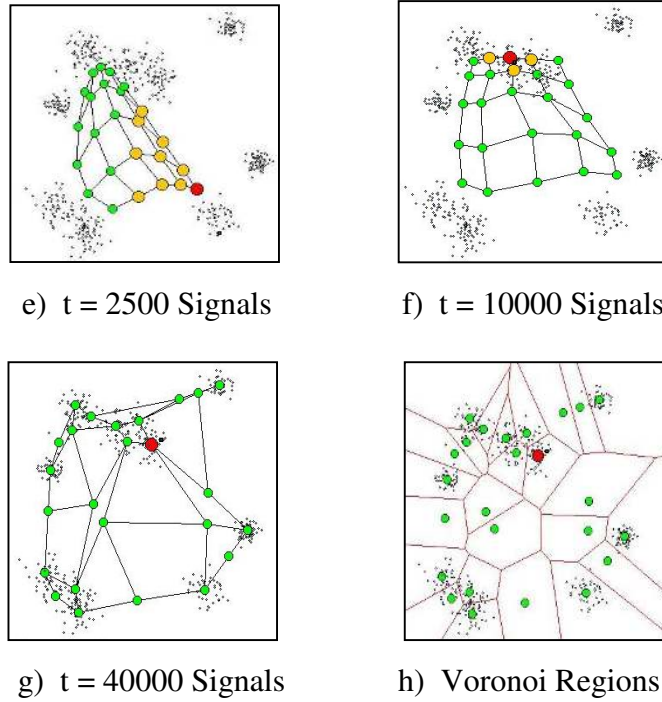


Figure 8 Self-Organized Map Representation Stages to Distribute Specific Points Exist in Two Dimensional Space

4.1.3.2. Multi Self-Organizing Maps (MultiSOM)

The MultiSOM model has been suggested in [38]. It represents a significant extension of the SOM model. The principle of this model is to utilize various point of views each one being represented by a single SOM map. All that is in order to improve both the granularity and the quality of the data analysis. And also to decrease the noise which is inevitably generated in an all classification approach. The conservation of an overall view of the analysis is accomplished by the utilize of a communication mechanism between the maps. The benefite of the multi-viewpoint analysis supplied by MultiSOM as compared to the wordly analysis supplied by SOM has been clearly showed for precise mining tasks such as patent analysis. Another significant mechanism supplied through the MultiSOM model is the generalization mechanism. This mechanism includes starting from the original map and introducing new [38]. Nevertheless, special care must be taken account of a well-known problem associated to the SOM trained structure, namely the border effect. It means that units on edges of the network do not stretch out as much as they

should towards the outliers data, last but not least, the neurons of SOM do not necessarily get close to the structure of the data because of the fixed topological structure of the grid. We are going to work on generalization mechanism in this thesis.

4.1.3.2.1. Generalization Mechanism

An important functionality for information analysis is briefly the map content into more generic clusters through an on-line generalization process. In order to achieve that aim the mission that the system run is to decrease the cluster number of the map in a cohesion method. The method contains in starting from the original map and introducing new clustering levels of synthesis (maps) by progressively decreasing the nodes number. Then the original map has been depend on the basis of a 2D square neighborhood between nodes, the transition from one level to another is achieved by choosing a new node set in which each new node will represent the average composition of a square of four direct neighbors on the original level. Figure 9 shows how the map generalization process operates [31].

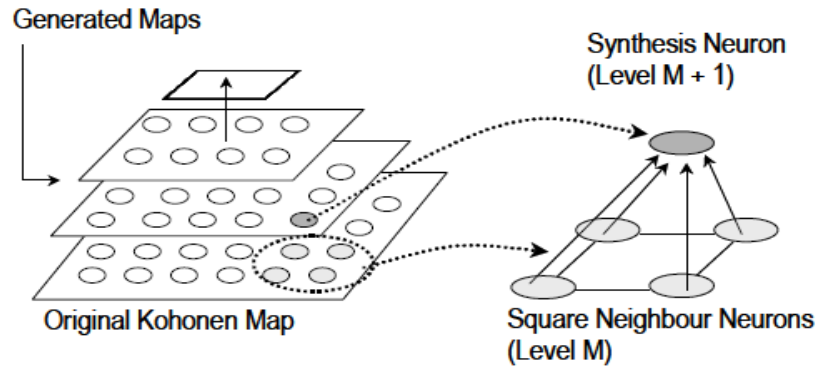


Figure 9 Map generalization Mechanism

This figure shows how the profiles of neighborhood node of one level (map) are utilize to generate the profile of each node of the next more synthetic level. This way could be also consider as the determination of all the square centers of a source level for building a new level. This operation has the benefit of keeping the original neighborhood structure on the new generated levels. In addition it ensures the conservation of topographic features of the map nodes vectors, and consequently the

keeping of the closeness of the nodes areas in the generalized maps [39]. Let $n \times m$ ($n, m \geq 2$) be the dimensions of the map associated to a given level, the generalization process will then produce a next more general level in the form of a $(n - 1) \times (m - 1)$ map. For each new level node n the vector computation formula then applies:

$$W_n^{M+1} = \frac{1}{4} \sum_{n_k \in V_n^M} W_{nk} \quad (4.2)$$

where V_n^M is the square neighborhood set on the map M associated to the node n of the new map $M + 1$. This method could be considered as an implicit and distributed form of a hierarchical clustering method based on neighborhood reciprocity (for this point, see De Rham, 1980).

4.2. Evaluation of Clustering Methods

As we mentioned earlier, there are many clustering methods, and they require the use of a group of mathematical functions to see which of these methods is the best for a particular type of data. Furthermore, the number of generated clusters by using each method cannot be determined optimally using only these mathematical functions. This process leads us to choose the best model (the optimal number of clusters) to represent the data and help us to carry out the processes of knowledge discovery to extract numerical association rules from this model, and this is why the best model to represent data helps to correct discover and precise knowledge, while the non-optimal model produced inaccurate and incorrect knowledge in general. There are several ways to evaluate clustering results and choose the best model or the optimal model. We are going to mention to ways of these methods. Evaluation of clustering based on the features of data distribution that can solve the problems that concerned with symbolic models, an alternative algorithm is proposed in this thesis for extracting complex numerical association rules.

4.2.1. Evaluating Clustering Based on the Characteristics of Distribution Data

Suppose that we have a group of cluster generated from clustering method was

applied on a set of documents, then the recall is (*Rec*) and precision (*Prec*) to a particular feature (*P*) in the cluster can be expressed as follows [7, 40]:

$$Prec(p) = \frac{|c_p^*|}{|c|} \quad (4.3)$$

So the feature precision in cluster represents the ratio between numbers of documents which contain feature in cluster and between the numbers of documents that found in the same cluster. The criterion of precision measures to what extent the contents of generated cluster of clustering are homogeneous. (This means that all the data that found in one cluster have the same features)

$$Rec(p) = \frac{|c_p^*|}{|c_p^*|} \quad (4.4)$$

Thereby the recall feature of cluster represents the ratio between the documents which contain feature in cluster and between all clusters which have the same feature. So the recall measures to what extent the contents of clusters have independent features from one cluster to another.

$$c_p^* = \{d \in c \mid W_d^p \neq 0\} \quad (4.5)$$

Where the notation c_p^* represents the restriction of the set c to the set members having the features p . So we can express the clustering precision measurement as:

$$P = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} \frac{|c_p^*|}{|c|} = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Prec(p) \quad (4.6)$$

Where S_c is the set of features which are peculiar to the cluster c that is described:

$$S_c = \{p \in d, d \in c \mid \bar{W}_c^p = \text{Max}_{c' \in \bar{C}} (\bar{W}_{c'}^p)\} \quad (4.7)$$

$\bar{C}$ represents the peculiar set of clusters extracted from the whole clusters of C :

$$\bar{C} = \{c \in C \mid S_c \neq \emptyset\} \quad (4.8)$$

$$\bar{W}_c^p = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in \bar{C}} \sum_{d \in c'} W_d^p} \quad (4.9)$$

Where W_d^p represents weight of the features p for element d . Clustering recall measurement can expressed as follow:

$$R = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} \frac{|c_p^*|}{|c_p^*|} = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Rec(p) \quad (4.10)$$

The more precision increase, the closer of documents topics which belong to the same cluster to be one, this is help the user to explain the cluster content. So the recall measures the comprehensiveness of the mentioned clusters, and evaluate the extent of existences the special features in an individual clusters. Moreover, we will have an optimal situation for clustering when the precision and recall reach to the same value or one value. Also the precision and recall measurements used to compare between different clustering methods.

4.3. Conclusion

We have showed in this chapter, data clustering methods which represent one of techniques to help discovering of knowledge from database, we exposed the general approach of clustering methods and mentioned in every one of them set of methods. Each clustering method generates different clusters when applied on the same database. This is mean that the same distribution of data in its descriptive space according to difference method. Therefore we have to evaluate the clustering method in order to know the optimal method to represent data. This evaluation is important because of the process of discovering the knowledge. This evaluation is based on descriptive space of data. This would be a strong basis which helps to carry out the functions of the interpretation and evaluation of knowledge discovery. The using of artificial neural network has been exposed in this chapter. Also functions and network learning method has used unsupervised learning which represent an important approach to clustering . It has proved previously that neural clustering is better than classical clustering because of its structure and internal using functions. For this purpose we use this kind of clustering in this thesis. We exposed unsupervised neural network methods such as self-organized map that is utilized for the purpose of clustering and mapping data, but this kinds of clustering effected by topology network. Another important mechanism provided by the MultiSOM model is the generalization mechanism. This mechanism consists in starting from the original map. In addition the generalization mechanism will be invested to control the number of extracted complex numerical association rules. This will be mention in chapter five.

CHAPTER 5

KNOWLEDGE DISCOVERY BY ASSOCIATION ANALYSIS

5.1. Introduction

Many business enterprises accumulate large quantities of data from their day to day operations. For example, large number of customer purchase data are collected every day at the examination counters of grocery shops. Table 3 illustrates an example of such data, commonly known as market basket transactions. Every row in table corresponds to a transaction, which has a unique identifier labeled *TID* and item set bought through a given customer. Retailers are dealt with analyzing the data to learn about the purchasing behavior of their customer. Like important information that could be utilized to support a different of business related applications such as, inventory management, marketing promotions and customer relationships management. This chapter displays a methodology which is famous by association analysis, which is helpful for finding out important relationships. These relationships are hidden in big data sets [27]. The uncovered relationships can be represented in the shape of association rules or sets of frequent items. For example, the next rule might be extracted from the data set shown in

<i>TID</i>	Items
1	{Bread , Milk}
2	{Bread , Diapers , Cheese , Eggs }
3	{Milk , Diapers , Cheese , Cola}
4	{Bread , Milk , Diapers , Cheese}
5	{Bread , Milk , Diapers , Cola}

Table 3 an Example of Market Basket Transaction

table 3: $\{\text{Diapers}\} \rightarrow \{\text{Cola}\}$. The rule suggests that a powerful relationship that find between the diapers sale and cola because many customers who buy diapers also buy cola. Retailers could be utilize this kind of rules to contribute them to determine modern opportunities for cross selling their products to the customers. We will display in this chapter the various methods in both types of knowledge discovery: symbolic methods and numeric methods. Symbolic methods consider most common in knowledge discovery process, but this method is very expensive when processing a huge data rules like text data. Numeric methods proved high activity in knowledge discovery but it depends on some tools and mathematical methods that are used in symbolic methods to extract the knowledge. Thus, we will depend on this thesis on numeric methods and suggest new numeric tools to accomplish evolution and analysis knowledge discovery tasks.

5.2. Knowledge Discovery using Association Rules

There are two main approaches to Knowledge discovery from data, the first one is based on concepts of symbolic methods, the second one based on the concept of numeric method that classify data according their similarities.

5.2.1. Symbolic Methods

Symbolic methods is the most common in active knowledge discovery of huge data. Several method are used such as Apriori and close methods that can extract association rules through using of a lattice. Before starting with detailing these two methods we should introduce some essential conceptions that are related with symbolic methods to knowledge discovery from data.

- **Special conceptions of frequent items**

Count of Support and Itemset: Let $I = \{i_1, i_2, \dots, i_d\}$ be the set of total items in data of market basket and $T = \{t_1, t_2, \dots, t_N\}$ be the set of all transactions. Every transaction t_i consist of a subset of items select from I . In association analysis, a zero group or more items is named by an itemset. If an itemset posses k items, it is named a

k-itemset. For instance, {Cola, Diapers, Milk} is an example of a 3-itemset. The empty set is an itemset which are not include any items. The transaction width is known by the items number present in a transaction. A transaction t_j have been told to includes an itemset X if X is a subset of t_j . For example, the second transaction offer in table 3 consist the itemset {Bread, Diapers} but not {Bread, Milk}. An important features of an itemset is its support count, which illustrate to the transactions number which consist of a particular itemset. Mathematically, the support count, $\sigma(X)$, for an itemset X might be stated as followings: $\sigma(X) = |\{t_i \mid X \subseteq t_i, t_i \in T\}|$, where the symbol $|\cdot|$ refere to the elements number in a set. In the data set shown in table 3, the support count for {Cola, Diapers, Milk} is equal to two because there are only two transactions that contain all three items [27].

Association Rules: An association rule is an subscription of the form $X \rightarrow Y$, where X and Y are disjoint itemsets, $X \cap Y = \emptyset$. The association rule strength could be measured in terms of its confidence and support. Support limits how often a rule is applicable to a given data set. While confidence refere how frequently items in Y appear in transactions that consist X . The formal definitions of these metrics are

$$\text{sup}(X \rightarrow Y) = \frac{\text{sup}(X \cup Y)}{N} \quad (5.1)$$

$$\text{conf}(X \rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)} \quad (5.2)$$

For Example take in account that the rule {Milk, Diapers} $\rightarrow$ {Cola}. Since the support count for {Milk, Diapers, Cola} is 2 and the whole number of transactions is 5, the rule's support is $2/5 = 0.4$. The rule's confidence is gained by separating the support count for {Milk, Diapers, Cola} by the support count for {Milk, Diapers}. Since there are 3 transactions that consist milk and diapers, the confidence for this rule is $2/3=0.67$.

Support and Confidence: Support is an remarkable measure because a rule that has very low support may occur simply by chance. A low support rule is also likely to be uninteresting from a business perspective because it may not be profitable to promote items that customers seldom buy together. For these reasons, support is often used to eliminate uninteresting rules. As will be shown in Section 5.1 and 5.2, support also has a wanted features that could be investement for the efficient discovery of association rules. Confidence, on the other hand, measures the reliability of the

inference made by a rule. For a given rule $X \rightarrow Y$, the higher the confidence, the more likely it is for Y to be present in transactions that contain X . Confidence also supply evaluate of the conditional probability of Y given X . Association analysis results should be explained with caution. The inference made by an association rule should not be imply causality. Instead, it offers a strong co-occurrence relationship between items in the antecedent and consequent of the rule. Causality, on the other hand, the need to knowledge about the causal and effect attributes in the data and typically requires relationships occurring over time. So, strategy of active knowledge discovery could be import by association rules discovery which illustrate linking between various items in database. Given a set of transactions T , find all the rules having support $\geq \text{minsup}$ and confidence $\geq \text{minconf}$, where minsup and minconf are the corresponding support and confidence threshold. A brute force stander to mining association rules is to calculate the support and confidence for each possible rules. This approach is prohibitively expensive because there are exponentially many rules that can be extracted from a data set [41]. More specifically, the whole number of possible rules extracted from a data set that cosist d items is

$$R = 3^d - 2^{d+1} + 1 \quad (5.3)$$

Even for the small data set show in table 3, this approach involves us to calculte the support and confidence for $3^6 - 2^7 + 1 = 602$ rules. More than 80% of the rules are discarded after applying $\text{minsup} = 20\%$ and $\text{minconf} = 50\%$, thus making most of the computation become wasted. In order to avoid performing needless computations, it would be useful to clip the rules early without compute values of support and confidence. An primery step toward developing acting of association rule mining algorithms is to decouple the support and confidence requirements. Notice that the support of a rule $X \rightarrow Y$ based only on the support of its corresponding itemset, $X \cup Y$. For example, the following rules have identical support because they require items from the same itemset, $\{\text{Cheese}, \text{Diapers}, \text{Milk}\}$:

$$\begin{array}{ll} \{\text{Cola}, \text{Diapers}\} \rightarrow \{\text{Milk}\}, & \{\text{Cola}, \text{Milk}\} \rightarrow \{\text{Diapers}\}, \\ \{\text{Diapers}, \text{Milk}\} \rightarrow \{\text{Cola}\}, & \{\text{Cola}\} \rightarrow \{\text{Diapers}, \text{Milk}\}, \\ \{\text{Milk}\} \rightarrow \{\text{Cola}, \text{Diapers}\}, & \{\text{Diapers}\} \rightarrow \{\text{Cola}, \text{Milk}\}, \end{array}$$

If the itemset is frequent, then all six candidate rules can be clip at once without computing confidence values. Therefore, a common strategy by several association rule mining algorithms is to decompose the problem into two main subtasks [41]:

- 1- **Frequent Itemset Generation:** The aim behind is to find all the itemsets that satisfy the *minsup* threshold. These itemsets are called frequent itemsets.
- 2- **Association Rules Generation:** The aim behind is to extract all the high confidence rules from the frequent itemsets found in the previous step. These rules are called strong rules.

Firstly- Frequent Itemset Generation: A lattice structure might utilized to recital the total possible itemsets list. Figure 10 shows an itemset lattice for $I = \{a, b, c, d, e\}$. Generally, a data set which has k items could primary generate up to $2^k - 1$ frequent itemsets, limit the empty set. Since k might be huge in different applications. The itemsets search space that involves to be searched is exponentially large [27].

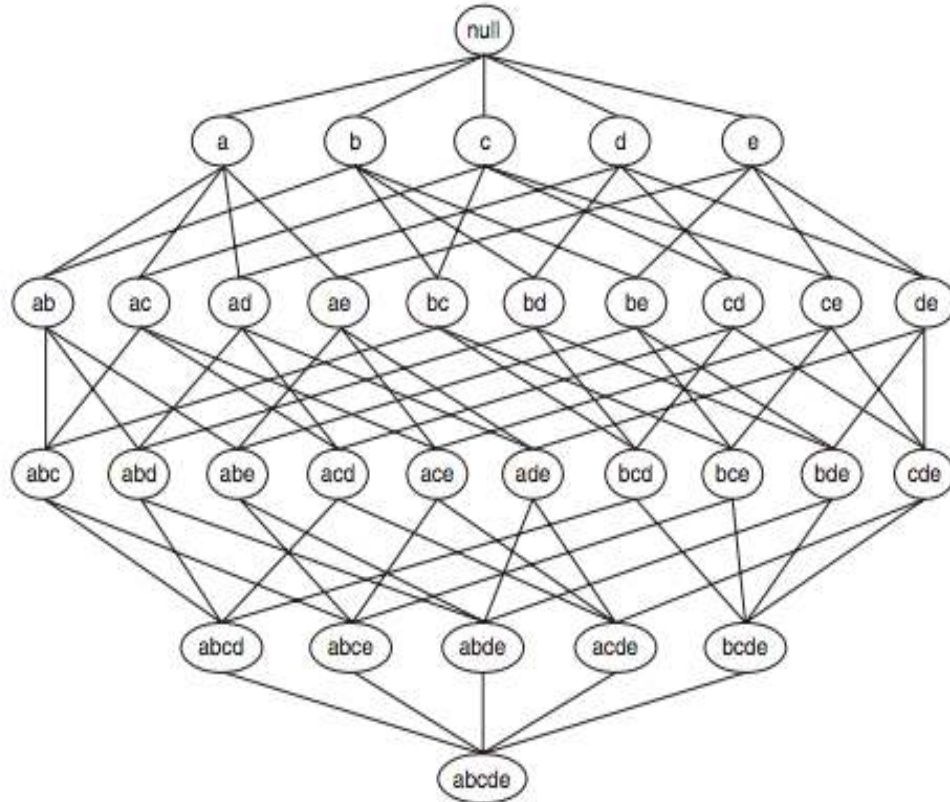


Figure 10 an Itemset Lattice

There are many methods to decrease the computational complexity of frequent itemset generation.

- 1- Decrease the candidate number itemsets (M). The Apriori standard has been explained in the next section.
- 2- Decrease the comparisons number. Instead of comparing each candidate itemset against each transaction, we are able to decrease the comparisons number by utilizing more advanced data structures.

Order feature definition: Let I be a set of items, and $J = 2^I$ be the power set of I . A measure f is monotone (or upward closed) if $\forall X, Y \in J : (X \subseteq Y) \Rightarrow f(X) \leq f(Y)$, which means that if X is a subset of Y , then $f(X)$ must not exceed $f(Y)$. On the other hand, f is anti-monotone (or downward closed) if $\forall X, Y \in J : (X \subseteq Y) \Rightarrow f(Y) \leq f(X)$, which means that if X is a subset of Y then $f(Y)$ must exceed $f(X)$. Any measure that have an anti-monotone features can be contribute directly into the mining algorithm to effectively clip the exponential search space of candidate itemsets, (as shown in figure 14 generate frequent itemsets by using Apriori method).

Secondly - Rules Generation: This section demonstrate how to extract association rules effectively from a given frequent itemset [42]. Each frequent k -itemset, Y , can produce up to $2^k - 2$ association rules, ignoring rules that have empty antecedents or consequent ($Y \rightarrow \emptyset$ or $\emptyset \rightarrow Y$). An association rule can be extracted by partitioning the itemset Y into two non-empty subsets, X and $Y - X$, such that $X \rightarrow Y - X$ satisfies the confidence threshold. Notice that all such rules should be met the support threshold because they are generated from a frequent itemset. For example, let $X = \{1, 2, 3\}$ be a frequent itemset. There are six candidate association rules that can be generated from X : $\{1, 2\} \rightarrow \{3\}$, $\{1, 3\} \rightarrow \{2\}$, $\{2, 3\} \rightarrow \{1\}$, $\{1\} \rightarrow \{2, 3\}$, and $\{2\} \rightarrow \{1, 3\}$, $\{3\} \rightarrow \{1, 2\}$. As each of their support is identical to the support for X , the rules must satisfy the support threshold. Calculating the confidence of an association rule does not need additional scans of the transaction data set. Consider the rule $\{1, 2\} \rightarrow \{3\}$ is generated from the frequent itemset $X = \{1, 2, 3\}$. The confidence for this rule is $\sigma(\{1, 2, 3\}) / \sigma(\{1, 2\})$. Because $\{1, 2, 3\}$ is frequent, the anti-monotone features of support to make sure that $\{1, 2\}$ must be frequent, too. Since the support

counts for both itemsets were already find during frequent itemset generation, there is no need to read the whole data set again [42].

- **Confidence based on pruning**

The different support measure, confidence does not have any monotone features. For example, the confidence for $X \rightarrow Y$ can be larger, smaller, or tied to the confidence for another rule $\tilde{X} \rightarrow \tilde{Y}$, where $\tilde{X} \subseteq X$ and $\tilde{Y} \subseteq Y$. Nevertheless, if we compare rules generated from the same frequent itemset Y , the following theorem holds for the confidence measure. If a rule $X \rightarrow Y - X$ does not satisfy the confidence threshold, then any rule $X' \rightarrow Y - X$, where X' is a subset of X , must not satisfy the confidence threshold as well. In order to prove this theorem, consider the following two rules: $X \rightarrow Y - X$ and $X' \rightarrow Y - X$, where $X' \subset X$ [41].

5.2.1.1. Apriori Symbolic Method

In this section we find out how support measurement is helpful in decrease number of candidate itemsets during the generation of frequent itemsets process:

5.2.1.1.1. Generating Frequent Itemsets by using Apriori Method

To illustrate the idea behind the Apriori principle, consider the itemset lattice has been displayed in figure 11. Suppose $\{c, d, e\}$ is a frequent itemset. Clearly, any transaction that include $\{c, d, e\}$ should be consist its subsets, $\{c, d\}$, $\{c, e\}$, $\{d, e\}$, $\{c\}$, $\{d\}$, and $\{e\}$. As a result, if $\{c, d, e\}$ is frequent, then all subsets of $\{c, d, e\}$ should be frequent. Since determine itemsets in figure 11 are frequent [41].

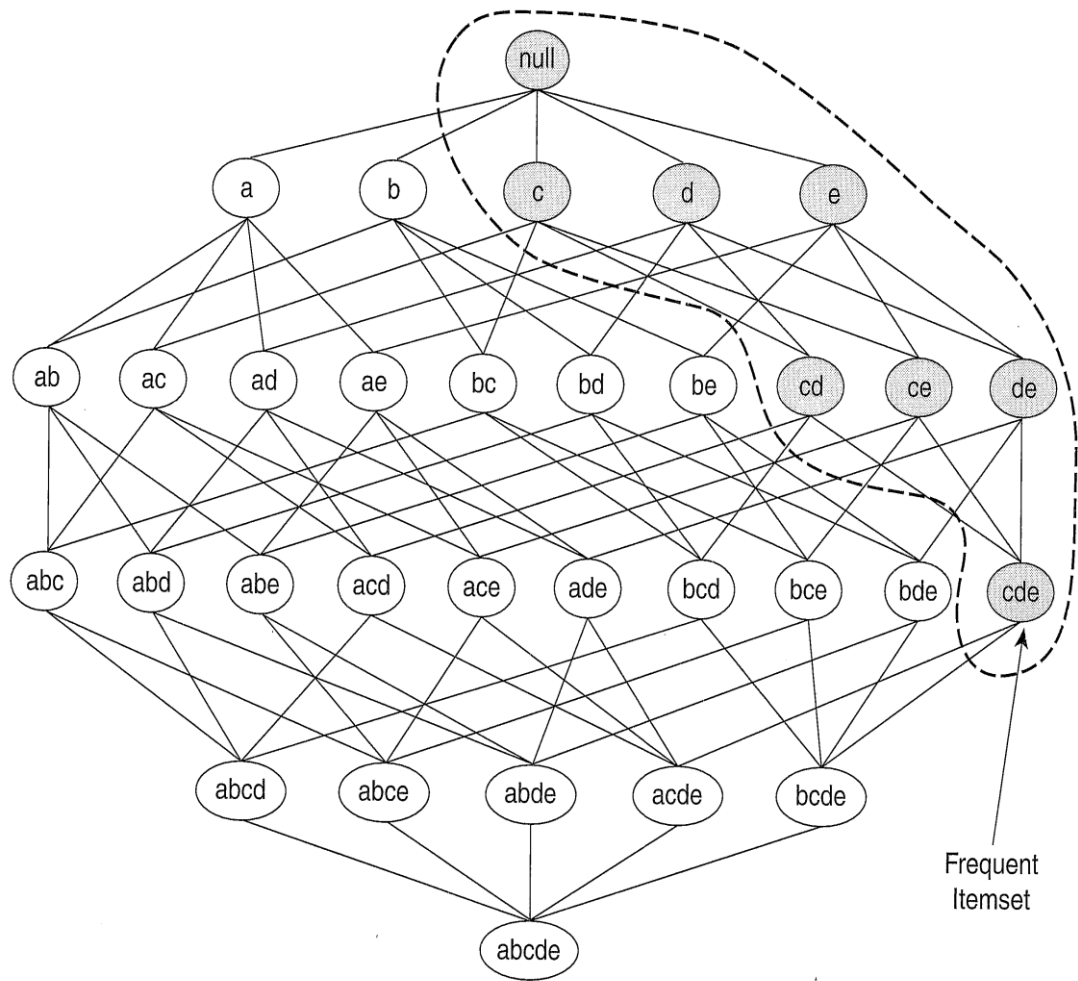


Figure 11 an Illustration Example of Apriori Approach

Conversely, if an itemset such as $\{a, b\}$ is infrequent, then all of its supersets have to be infrequent. As has been shown in figure 12, the whole subgraph consisting of the supersets of $\{a, b\}$ can be clipped directly once $\{a, b\}$ is found to be infrequent. This strategy of trimming the exponential search space depending on the support measure is known as support-based pruning. Like a clip strategy is possible by a key feature of the support measure that the support for an itemset does not exceed the support for its subsets. This feature is famous as the anti-monotone support measure feature [43].

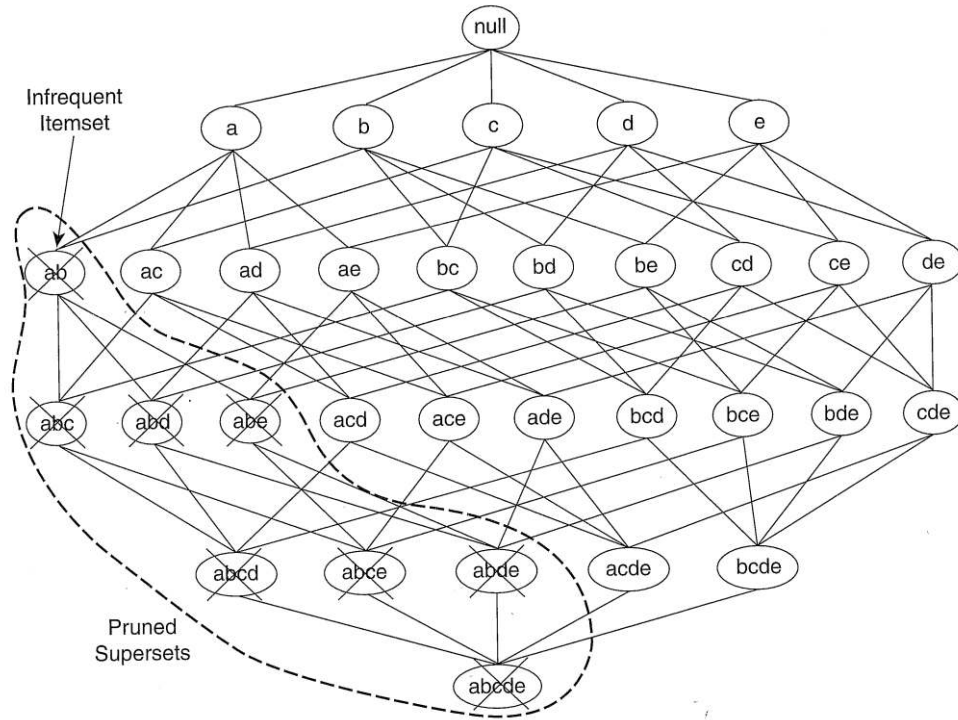


Figure 12 an Illustration of Support Based Pruning According to Apriori

Now, we are going to show Apriori way to generate frequent itemsets through the example that given in table 3. Apriori algorithm is the first search algorithm about association rules. It has been utilized to support based clipping to systematically control the exponential growth of candidate itemsets. We have been assume that the support threshold is 60%. It is equivalent to a minimum support count equal to 3. Primary each item is regarded as candidate 1-itemset. After counting their supports candidate itemsets {Cola} and {Eggs} are discarded. Because of they appear in lower than three transactions. In the next iteration, candidate 2-itemsets are generated by using only the frequent 1-itemsets because the Apriori principle ensures. all supersets of the infrequent 1-itemsets should be infrequent. This is waythere are four frequent only 1-itemsets, the number of candidate 2-itemsets generated by the algorithm is $\binom{4}{2} = 6$. Two of these six candidates, {Cheese, Milk}and {Cheese, Bread}, are subsequently exist to be infrequent after computing their support values. The remaining four candidates are frequent, and thus would utilized to generate candidate 3-itemsets. There are $\binom{6}{3} = 20$ candidate 3-itemsets that could be shaped using the six items given in this example. With the Apriori principle, we only need to

keep candidate 3-itemsets whose subsets are frequent. The only candidate that has this features is {Bread, Diapers, Milk}, as shown in figure 13.

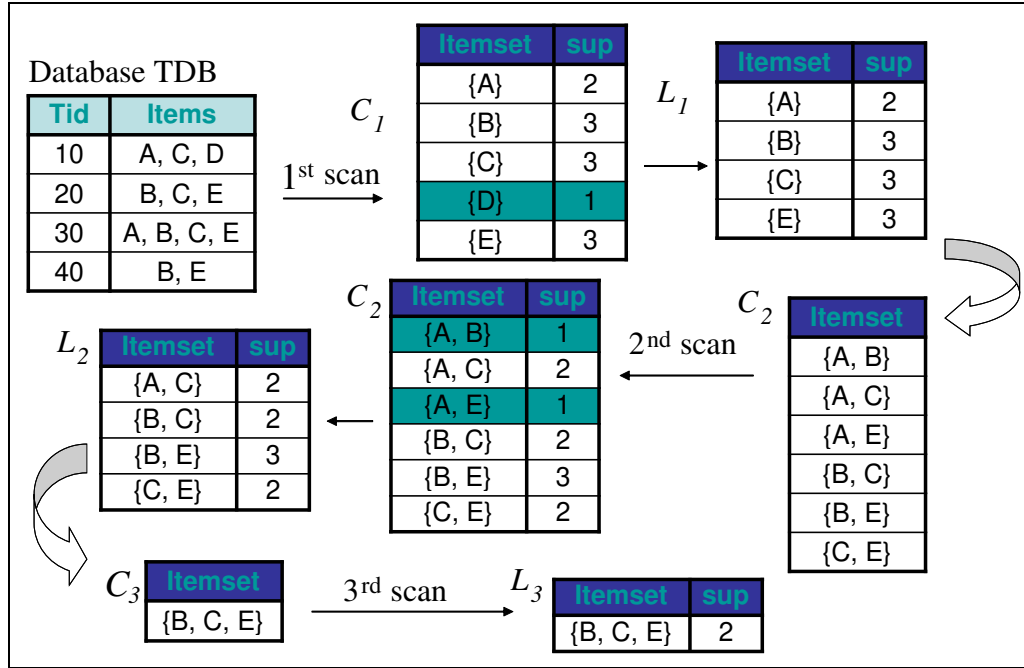


Figure 13 Generating Frequent Itemsets According to Apriori Algorithm

5.2.1.1.2. Generating Association Rules Based on Apriori Algorithm

The Apriori algorithm utilizes a level-wise approach to generate association rules. since every level harmony to the items number which belong to the rule consequent. Primary, the whole rules of the high-confidence which posses only one item in the rule consequent have been extracted. These rules are later utilizes to generate fresh candidate rules. For example, if $\{acd\} \rightarrow \{b\}$ and $\{abd\} \rightarrow \{c\}$ are high-confidence rules, then the candidate rule $\{ad\} \rightarrow \{bc\}$ is generated by merging the consequents of both rules. figure 14 shows a lattice structure for the association rules generated from the frequent itemset $\{a, b, c, d\}$. If any node in the lattice has low confidence, then according to theorem 5.2, the whole subgraph has been extended by the node could be clip at once. Suppose the confidence for $\{bcd\} \rightarrow \{a\}$ is low. All the rules consisting item a in its consequent, including $\{cd\} \rightarrow \{ab\}$, $\{bd\} \rightarrow \{ac\}$, $\{bc\} \rightarrow \{ad\}$, and $\{d\} \rightarrow \{abc\}$ can be discarded [43].

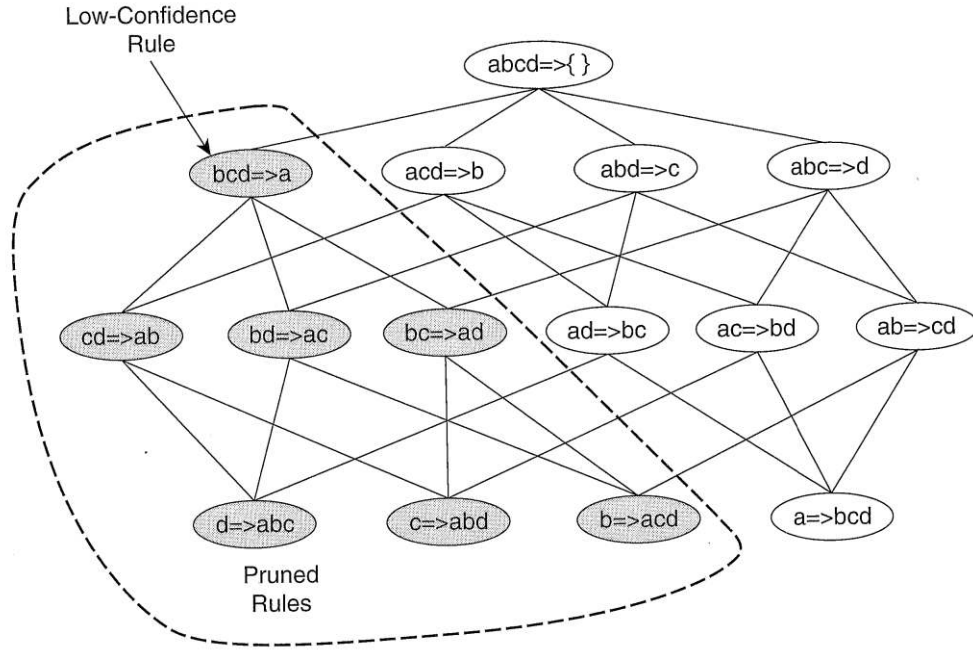


Figure 14 Clipping of Association Rules According to Apriori Principle

5.2.1.2. Close Symbolic Method

Closed itemsets supply a minimal representation of itemsets without wasting their support information. A formal definition of a closed itemset has been shoewd resented below.

Definition of Closed Itemset: An itemset X is closed if none of its immediate supersets has exactly the same support count as X . Put another way, X is not closed if at least one of its immediate supersets has similar support count as X . Examples of closed itemsets has been dispayed in figure 15. We have associated each node (itemset) in the lattice with a list of its corresponding transaction IDs. For example, the node $\{b, c\}$ is associated with transaction IDs 1, 2, and 3, its support count is equal to three. From the transactions given in this diagram, notice that every transaction that consist b also contains c . Frequently, the support for $\{b\}$ is identical to $\{b, c\}$ and $\{b\}$ have not be considered a closed itemset. Since c occurs in every transaction that consist both a and d , the itemset $\{a, d\}$ is not closed. On the other hand, $\{b, c\}$ is a closed itemset because it does not have the same support count as any of its supersets [42].

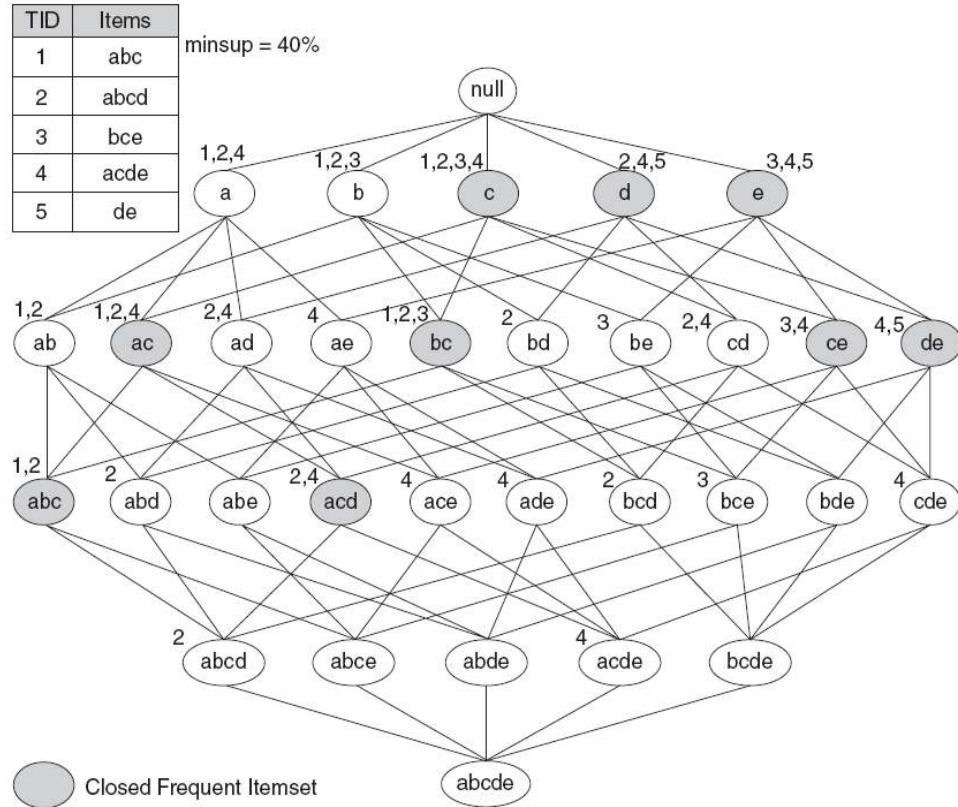


Figure 15 an Example of Closed Frequent Itemsets

5.2.1.2.1. Generating Closed Frequent Itemsets

An itemset is a closed frequent itemset if it is closed and its support is bigger than or equal to minsup [41]. In the previous example, assuming that the support threshold is 40%, $\{b, c\}$ is a closed frequent itemset because its support is 60%. The total of the closed frequent itemsets are illustrated by the shaded nodes. Algorithms are available to extract closed frequent itemsets from a given data set. Readers may illustrate to the bibliographic notes at the end of this chapter. We might be utilize the closed frequent itemsets to identify the support counts for the non-closed frequent itemsets. For example, consider the frequent itemset $\{a, d\}$ shown in figure 15. Because of the itemset is not closed, its support count should be identical to one of its immediate supersets. The key is to identify which superset (among $\{a, b, d\}$, $\{a, c, d\}$, or $\{a, d, e\}$) has exactly the same support count as $\{a, d\}$. The Apriori principle states that any transaction which includes the superset of $\{a, d\}$ must also includes $\{a, d\}$. However,

any transaction that consist $\{a, d\}$ does not have to contain the supersets of $\{a, d\}$. For this reason, the support for $\{a, d\}$ must be equal to the largest support among its supersets. Since $\{a, c, d\}$ has a larger support than both $\{a, b, d\}$ and $\{a, d, e\}$, the support for $\{a, d\}$ must be identical to the support for $\{a, c, d\}$. Using this methodology, an algorithm can be improved to calculate the support for the non-closed frequent itemsets. The pseudocode for this algorithm is shown in algorithm figure 16. The algorithm proceeds in a specific-to-general fashion, from the largest to the smallest frequent itemsets. In order to find the support for a non-closed frequent itemset, the support for all of its supersets must be known. Closed frequent itemsets are helpful for dropping some of the redundant association rules. An association rule $X \rightarrow Y$ is redundant if we have another rule $X' \rightarrow Y'$, where X is a subset of X' and Y is a subset of Y' , such that the support and confidence for both rules are identical.

```

1: Let  $C$  denote the set of closed frequent itemsets

2: Let  $k_{max}$  denote the maximum size of Closed frequent itemsets

3:  $F_{k_{max}} = \{ f | f \in C, |f| = k_{max} \}$    {Find all frequent itemsets of size  $k_{max}$ }

4: for  $k = k_{max} - 1$  downto 1 do

5:    $F_k = \{ f | f \subset F_{k+1}, |f| = k \}$        {Find all frequent itemsets of size  $k$ }

6:   for each  $f \in F_k$  do

7:     if  $f \notin C$  then

8:        $f.support = \max \{ f'.support | f' \in F_{k+1}, f \subset f' \}$ 

9:     end if

10:  end for

11: end for

```

Figure 16 Algorithm Support Counting using Close Frequent Itemsets

Is therefore redundant because it has the same confidence and support as $\{b, c\} \rightarrow \{d, e\}$. Such redundant rules are not generated if wasted frequent itemsets are used for rule generation. At last, we have to notice that all maximal frequent itemsets are closed because none of the maximal frequent itemsets can have the same support count as their immediate supersets. The relationships among frequent, maximal frequent and closed frequent itemsets are shown in figure 17. Closed frequent itemset is a closed frequent itemset if it was closed and has support greater or equal to the definite support threshold. In the former example in figure 15, by assumed that support threshold is 40% which means equal 2, whereupon $\{b, c\}$ set is closed frequent itemset due to its support 60% which means equal 3 [42].

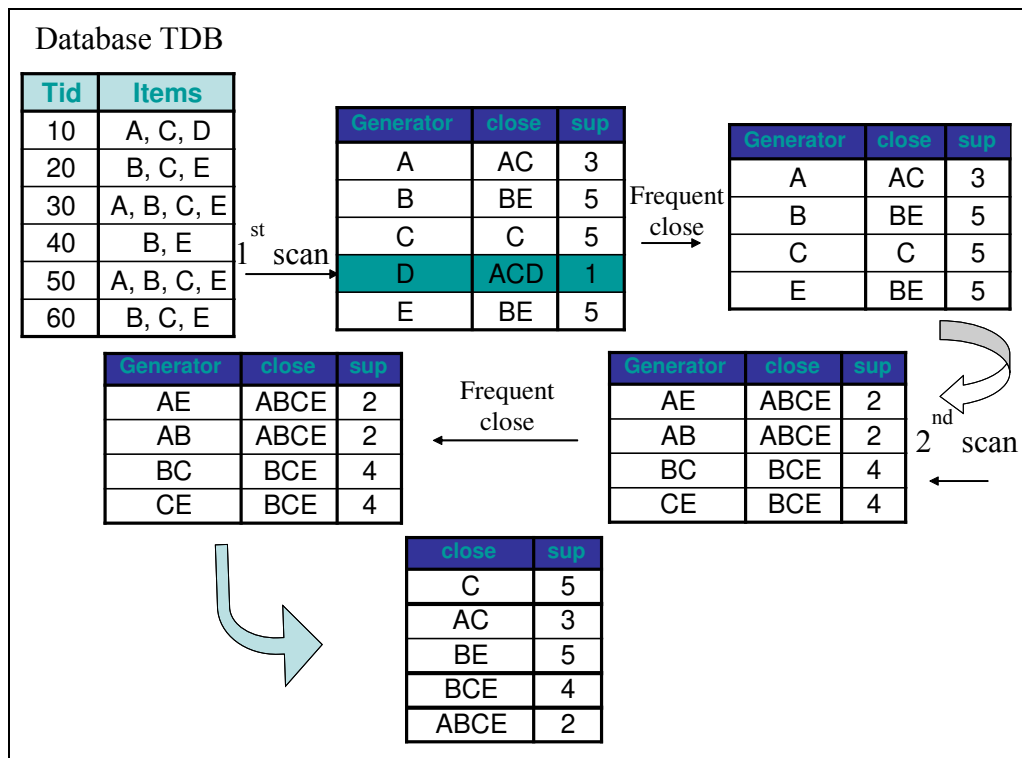


Figure 17 Generating Frequent Itemsets According to Closed Algorithm

Refers to Closed frequent itemsets in lattice by highlight nodes. The Closed frequent itemset are shown in figure 18.

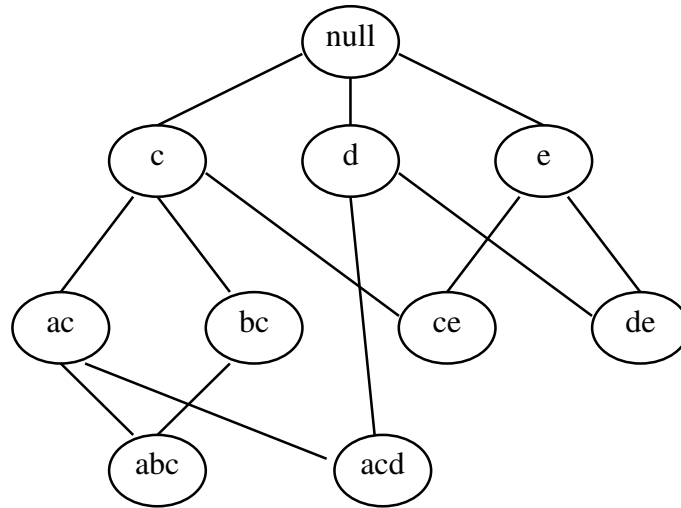


Figure 18 Lattice of Closed Itemsets Generates from Lattice in Figure 15

5.2.1.2.2. Generating Association Rules Based on Close Algorithm

Deferent types of association rules could be generated by utlizing closed itemset and generators set as followed:

If *FCI* sets of closed frequent itemset, *f* is closed frequent itemset, *FG* is frequent generators set and *FG_f* is set of frequent generators in *f* equivalence raw.

- **Finding Exact association rules:** Generative basis to find exact association rules is

$$GB = \{r : g \Rightarrow (f \setminus g) \mid f \in FCI \wedge g \in FG_f \wedge g \neq f\} \quad (5.4)$$

- **Finding Approximate association rules:** Informative basis to find Approximate association rules is

$$IB = \{r : g \Rightarrow (f \setminus g) \mid f \in FCI \wedge g \in FG \wedge g'' \subset f\} \quad (5.5)$$

g'' is closed of g set.

- **Transitive Reduction:** Transitive reduction to IB defined as following:

$$RIB = \{r : g \Rightarrow (f \setminus g) \mid g'' \text{ is maximal proper subset of } f \text{ in } FCI\} \quad (5.6)$$

g'' is the bigger set in f . It is frequent closed item.

- **Minimal Non-Redundant:** Is defined through the following:

$$MNR = GB \cup IB \quad (5.7)$$

- **Transitive Reduction of Minimal Non-Redundant:** Is defined as followed.

$$RMNR = GB \cup RIB \quad (5.8)$$

5.2.1.3. Symbolic Methods Defects

These methods suffered of different problems. it produces a huge number of association rules. it includes high ratio of redundant rules. This matter leads to big cost in extract and selection of on important rule. Therefore this selection occurred by the way of utlizing of support and confident (in addition to using of the other standards). Each one of them should give limit value. This is one of the main defect of symbolic methods. Then we do not know how to define the correct limited values. in addition, important rules could be neglected in case of give these limited values. Also we cannot say if the association rule has high confident they are correct rule [4]. In order to solve this problem new approach has been used to knowledge discovery but throughout numeric clustering methods [4,5].

5.3. Numeric Methods of Knowledge Discovery

There are several ways to use clustering. In order to solve these problems due to use symbolic methods, unsupervised numerical classification models (clustering models) have suggested to extract useful knowledge such as (neural gas, Multi gas, self-organizing map and Multi self-organizing map). In this thesis we are going to work on multi self-organizing map. which will be used specifically in knowledge discovery processes. We have suggested alternative algorithm to solve these problems by extracting complex association rules by using different generalization levels in MultiSOM. So we have developed the algorithm for extracting simple association rules from clustering model as in figure 19. Then we get an algorithm for extracting complex association rules from clustering model as in figure 20.

5.3.1. Discover Simple Numeric Association Rules

A reliable unsupervised neural model, like a MultiSOM, represents a natural candidate to overcome with the associated problems of rule expansion and rule selection. These problems are inherent in symbolic methods [7]. Thus, its synthesis abilities that can be utilize both for decreasing the rules number. Also for extracting the most significant rules. We are going to depend on our class quality principles for extracting rules from the classes of the original MultiSOM and its generalizations, that is the Precision and Recall measures depend on the class members features. The Precision criterion measures in which proportion the content of the classes generated by a classification method is homogeneous. The bigger the Precision, the closes the intensions of the data belonging to the same classes will be one with respect to the other. The more homogenous will be the classes. In a complementary way, the Recall criterion measures the exhaustiveness of the content of said classes, evaluating to what range single features are associated with single classes. We have showed that if both Recall and Precision values reach the unity value, the peculiar set of classes represents a Galois lattice. A class returns to the peculiar set of classes of a given classification if it possesses peculiar features. At last, a features is considered as peculiar for a given class if it is maximized by the class members. As compared to classical inertia measures, averaged measures of Recall and Precision present the major benefits to be independent of the classification method. They could be utilize both for comparing classification methods and for enhasment the results of a method relatively to a given dataset. They also represent a sound basis for extracting simple association rules are shown in figure 19. Where let C being a class, PC being the set of features associated to the members of C , and PC^* being the set of peculiar features of C , with $PC^* \subseteq PC$:

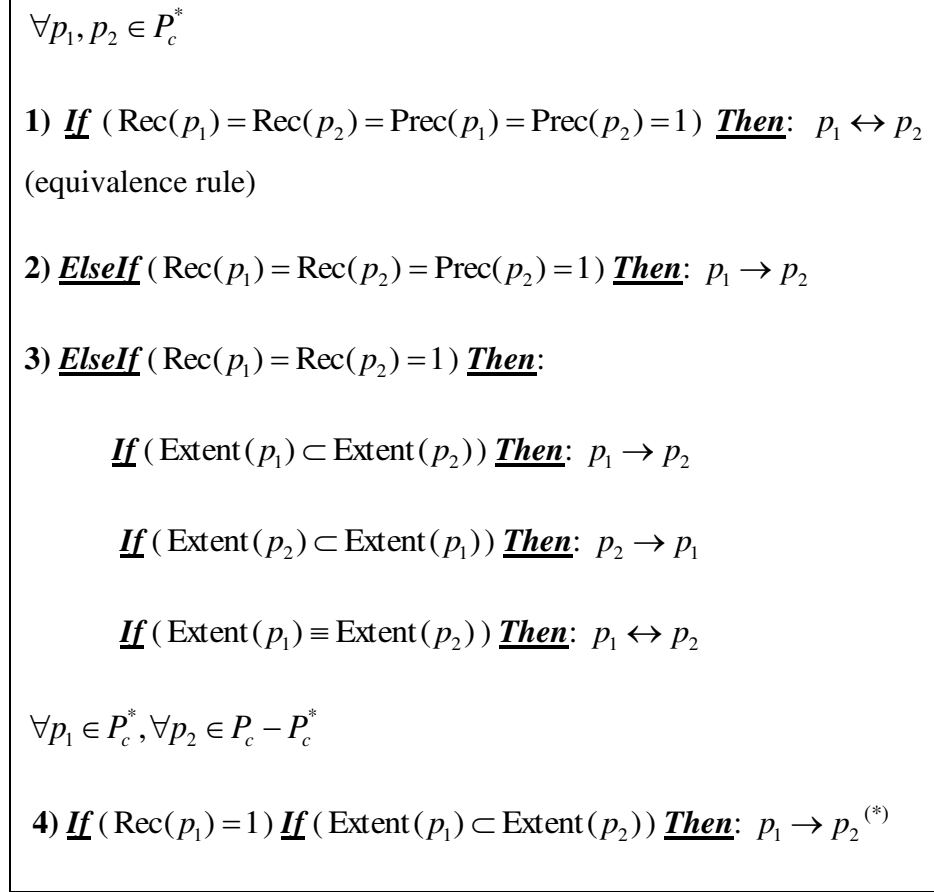


Figure 19 Algorithm for Extracting Simple Association Rules from Clustering Model

5.3.2. Discover Complex Numeric Association Rules

Numeric methods could solve the related problems with symbolic methods as we have seen previously. Extract simple association rules from clustering model so it has partially solved the problems related to symbolic methods. This algorithm extracts simple association rules and this extraction produce redundant association rules. Therefore we suggest alternative algorithm to solve these problems by extracting complex association rules are shown in figure 20. This could be achieved based on recall and precision measures. In order to be able to extract different number of rules and to control this number. We use different generalization levels in MultiSOM. Let c being a cluster, P_c^* being the set of peculiar features associated to c , $t_i \in P_c^* \quad i = 1.....N$. Let N is the number of peculiar features.

1. For all cluster $c \in C$
2. Find P_c^* the set of peculiar features associated to c
3. Create A (set of features) such that:

$$A = \{ t_i \mid t_i \in P_c^*, Prec(t_i) = 1, Rec(t_i) = 1 \}$$
4. Create B (set of features) such that:

$$B = \{ t_i \mid t_i \in P_c^*, Prec(t_i) \neq 1, Rec(t_i) = 1 \}$$
5. Create E (all possible subsets of B) such that:

$$E = \{ b_k \mid \exists d \in D_c, b_k \subseteq d \}$$

 b_k being a subset of B
 d being a document from D_c
 D_c being the set of documents associated to a cluster

$$D_c = \{ d \mid d \in c \}$$
6. if $|A| \geq 2$ then
7. For all $t_i \in A$
8. $t_i \rightarrow A \setminus \{t_i\}$ (*informative rule*)
9. End for
10. End if
11. if $A \neq \emptyset$ and $E \neq \emptyset$ Then
12. For all $b_k \in E$
13. $b_k \rightarrow A$
14. End for
15. End if
16. End for

Figure 20 Algorithm for Extracting Complex Association Rules from Clustering Model

5.4. Experimental Results

Our test database consists of 1000 patents (see Appendix B). Four different subfields have been extracted to represent four different viewpoints: Use, Advantages, Titles and Patentees. We use only USE viewpoint in our experiment. The size of the description space of this database is 234. Our experiment about extracting complex association rules is based on an optimal number of clusters generated by SOM method which is 100 clusters (neurons) [4]. Then, from this optimal clustering we generalize it by applying the generalization mechanism. So we have the following specified number of each generalization level:

- 81, 64, 49, 36, 25, 16, 9 and 4 neurons.

	Total Number of Generated Association Rules	Total Number of New Generated Association Rules	Confidence
Original level (100 clusters)	66	66	100%
1st generation level (81 clusters)	127	61	100%
2nd generation level (64 clusters)	153	27	100%
3rd generation level (49 clusters)	173	20	100%
4th generation level (36 clusters)	186	13	100%
5th generation level (25 clusters)	193	7	100%
6th generation level (16 clusters)	197	4	100%
7th generation level (9 clusters)	197	0	100%
8th generation level (4 clusters)	197	0	100%
Total Number of Association Rules		106	

Table 4 Number of Complex Association Rules Generated using Algorithm

Our experiment consists in extracting association rules from the single Use viewpoint. Both the original MultiSOM and its generalizations are used for extracting the association rules. We apply our proposed algorithm (see figure 20).

The results of knowledge discovery (as numeric association rules) that presented in table 4. Figure 21 shows some of extracted the important complex numerical association rules from the several levels of MultiSOM.

Extract of association rules from symbolic model. The global rule count defined for the symbolic model includes the count total rules ($conf = 100\%$ $minsup = 0.004$), ($conf = 100\%$ $minsup = 0.009$) and ($conf = 100\%$ $minsup = 0.013$). We notice from the results that number of extract to association rules of symbolic methods as table 5. The following table displays extract association rules in symbolic methods by using $conf$ and $minsup$. The number of extract association rules of numeric methods are 106 as in figure 21. Extracted complex association rules are important rules from information nature side that have been presented. So we conclude activity of association rules discovered algorithm from numeric methods comparing with symbolic methods which do not give large information to data analysts in selection of important rules from huge group of generated association rules. According to the extracted results, we conclude that numrical methods is better than the symbolic methods.

	<i>Minsup</i>	Confidence	Number of association rules
1	0.004 (1/220)	100%	1322155
2	0.009 (2/220)	100%	775
3	0.0136 (3/220)	100%	242
4	0.018 (4/220)	100%	89
5	0.0227 (5/220)	100%	28
6	0.0272 (6/220)	100%	13

Table 5 Extracted Symbolic Association Rules from USE Database

```

56 ==> 60
60 ==> 56
38 ==> 161
161 ==> 38
74 ==> 121, 190
121 ==> 74, 190
190 ==> 74, 121
210 ==> 87, 167
157 ==> 198, 201, 204
198 ==> 157, 201, 204
204 ==> 157, 198, 201
1 ==> 37, 40, 62, 192
37 ==> 1, 40, 62, 192
40 ==> 1, 37, 62, 192
192 ==> 1, 37, 40, 62
24 ==> 47, 51, 83, 116, 120
47 ==> 24, 51, 83, 116, 120
51 ==> 24, 47, 83, 116, 120
120 ==> 24, 47, 51, 83, 116
25 ==> 39, 81, 92, 115, 131, 160, 184, 188, 212, 215, 222
39 ==> 25, 81, 92, 115, 131, 160, 184, 188, 212, 215, 222
222 ==> 25, 39, 81, 92, 115, 131, 160, 184, 188, 212, 215
8 ==> 35, 67, 80, 86, 146, 150, 159, 162, 164, 168, 178, 193, 226, 227
80 ==> 8, 35, 67, 86, 146, 150, 159, 162, 164, 168, 178, 193, 226, 227
227 ==> 8, 35, 67, 80, 86, 146, 150, 159, 162, 164, 168, 178, 193, 226

```

Figure 21 Some Extracted Complex Numerical Association Rules from USE Database

5.5. Conclusion

The concept of discover knowledge has been exposed in this chapter from huge amounts of data throughout association analysis between its features. Also we displayed the common methods to complete this process, which is the symbolic method that depends on analysis of data rule and studying most important features during through repetition amount and generate of several kinds of association rules. Apriori method generates huge number of association rules in spite of small size of database, this leads to generate a lot of redundant rules. In order to solve these problems Close method has come to rely closing concepts to generate important

association rules. In spite of that it produces huge amounts of association rules that is difficult to extract the important rules from them through depending on some measures as support and confidence, this leads to delete a lot of important rules. So the numeric method have been recently adopted, especially neural clustering methods to discover knowledge such as MultiSOM method so this method has solved problems that related with symbolic methods. Numeric methods characterized by great ability to summing up data, and let the weak associations, in addition to low cost in knowledge discover comparing with symbolic methods. So, we proposed to work on unsupervised neural network such as (MultiSOM) which is one of the clustering methods could solve the related problems with symbolic methods which are shown in figure 19. The algorithm of extract simple association rules from clustering model has partially solved of the related problems with symbolic methods. This algorithm of extracts simple association rules and this extraction produce redundant association rules. Therefore we have suggested alternative algorithm to solve these problems by some extracting complex association rules which are shown in figure 20.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

Presently databases have become huge and complex, especially for text data, which have their representation in descriptive space with high dimensional space, so the investment and working on it become very difficult, and need a time from data analysts to extract the important knowledge. So it is important to use the data mining process or knowledge discovery in databases to generate association rules. It depends at the beginning on the methods known as symbolic methods. But using these methods in knowledge discovery is expensive and inaccurate process because it generates a huge number of association rules, therefore they generate huge amount of redundant association rules which complicate the important association rules selection. So numeric methods come to solve problems and gain reasonable number of important association rules. This study has enabled to know the concept of knowledge discovery in databases, these data could be from different kinds such as text data. Also we determined the steps of knowledge discovery databases, especially the step of data mining. This step contains some techniques such as classification, clustering and association rules. Our interest in this thesis was limited in text data mining for extracting the important information from it. We have also clarify the steps of text data processing through their indexation. These documents have been presented with different models the vector model is the most important one because it shows the strength of different information that exist in every document, additionally their strength that exist in text database. In order to do knowledge discovery processes, the vector model allows to classify or clustering data correctly. So we apply clustering methods for dropping data in to different clusters. Therefore, we exposed the important clustering methods to knowledge discovery processes. We

concentrated on neural networks types that followed unsupervised learning approach such as MultiSOM. Also, its generalization mechanism has been invested to control the number of extracted numerical complex association rules. So, we proposed to work on unsupervised neural network such as (MultiSOM) which is one of the clustering methods could solve the related problems with symbolic methods which are shown in figure 19. The algorithm of extract simple association rules from clustering model has partially solved of the related problems with symbolic methods. This algorithm of extracts simple association rules and this extraction produce redundant association rules. Therefore we have suggested alternative algorithm to solve these problems by some extracting complex association rules which are shown in figure 20.

6.2. Recommendations

We are looking forward to continue our studying in this field throughout suggest some futurity recommendations which reflect the following:

- 1- Estimation of extract association rules of any types of data.
- 2- Developing the suggested algorithm to extract more association rules.
- 3- Working on suggest new and accurate types of data features, and then definite new types to clustering which contain these data.
- 4- Working on suggest new algorithm to generate more complex numeric association rules in every space which could adopt the previous algorithm as in figure 20 to some extract of association rules through utilize generalization mechanism to cluster MultiSOM.
- 5- Determination of association rules during generalization process to contain complex association rules which the rule terminals is group of features.
- 6- Working on suggest additional types of association rules, help in tasks of explanation knowledge.
- 7- Working on suggest additional standards to estimate extraction to complex numerical association rules, which has been suggested in this thesis.

REFERENCE

1. **FAYYAD U. M., SHAPIRO G. P., SMYTH P., UTHURUSAMY R., (1996)**, *“Advances in Knowledge Discovery and Data Mining”*, AAAI/MIT Press, California, pp.560.
2. **LAROSE D. T., (2006)**, *“Data Mining Methods and Models”*, Wiley-IEEE Press, New Jersey, USA, pp.344.
3. **AGRAWAL R., SRIKANT R., (1994)**, *“Fast Algorithms for Mining Association Rules”*, In Proceedings of the International Conference on Very Large Databases, Santiago, USA, pp. 487-499.
4. **LAMIREL J. C., AL-SHEHABI S., (2005)**, *“Efficient Knowledge Extraction from Unsupervised Multi-Topographic Neural Network Models”*, 5th Workshop on Self-Organizing Maps, Paris, France, pp.291-298.
5. **HAMMER B., RECHTIEN A., STRICKERT M., VILLMANN T., (2002)**, *“Rule Extraction from Self-Organizing Networks”*, International Conference on Artificial Neural Networks, Osaka, Japan, pp.877-883.
6. **ATTIK M., AL-SHEHABI S., LAMIREL J. C., (2006)**, *“Clustering Quality Measures for Data Samples with Multiple Labels”*, IASTED International Conference on Databases and Applications, Austria, pp.58-65.
7. **LAMIREL J. C., AL-SHEHABI S., FRANCOIS C., HOFFMANN M., (2004)**, *“New Classification Quality Estimators for Analysis of Documentary Information: Application to Patent Analysis and Web Mapping”*, François, vol. 60, no. 3, pp.445-462.
8. **Lent B., Swami A., Wisdom J., (1997)**, *“Clustering Association Rules”*, in the Proc of 13th Intl Conference on Data Engineering, pp. 220.

9. **Agrawal R., Srikant R., (1994),** “*Fast Algorithms for Mining Association Rules in Large Databases*”, 20th International Conference on Very Large Data Bases, pp. 487-499.

10. **Jadhav S. R., and Kumbargoudar P., (2007),** “*Multimedia Data Mining in Digital Libraries: Standards and Features*”, Proceedings of Conference Recent Advances in Information Science and Technology, India, pp 54-59.

11. **Sotiris G. and Dimitris D., (2006),** “*Association Rules Mining: A Recent Overview*”, International Transactions on Computer Science and Engineering, vol. 32, no. 1, pp. 71-82.

12. **Yuan, Y., Huang T., (2005),** “*Matrix Algorithm for Mining Association Rules*”, Taiwan, vol. 3644, pp. 370-379.

13. **Mohammed J. Z. and Ching J. H., (2002),** “*An Efficient Algorithm for Closed Itemset Mining*”, In Proceedings of the 2nd SIAM International Conference on Data Mining, pp. 457–473.

14. **Kohonen T., (1990),** “*The self-organizing maps*”, Proceedings of the IEEE, vol. 78, pp. 1464–480.

15. **ZAÏANE O. R., (1999),** “*Principle of Knowledge Discovery in Databases*”, CMPUT, Canada, pp. 581–583.

16. **BRACHMAN R., ANAND T., (1996),** “*The Process of Knowledge Discovery in Databases: A Human-Centered Approach*”, In Advances in Knowledge Discovery and Data Mining, California, USA, pp. 40-45.

17. **FAYYAD U. M., SHAPIRO G. P., SMYTH P., (1996),** “*From Data Mining to Knowledge Discovery in Databases*”, California, pp.37-51.

18. **SALTON G., (1989)**, “*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*”, Addison Wesley, Amsterdam, North Holland, pp. 115–123.
19. **FAYYAD U. M., DJORGOVSKI S. G., WEIR N., (1996)**, “*From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey*”, AI Magazine, California, vol. 17, no. 2, pp.51–66.
20. **BERRY M. J. A., LINOFF G. S., (2004)**, “*Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*”, 2nd edition Wiley-IEEE Press, USA, New Jersey, pp.643.
21. **HEARST M. A., (1997)**, “*Text Data Mining: Issues, Techniques, and the Relationship to Information Access*”, Presentation notes for UW/MS Workshop on Data Mining, Lawrence, Erlbaum, pp. 257–274.
22. **FELDMAN R. DAGAN I., (1995)**, “*Knowledge Discovery in Textual Databases (KDT)*”, In proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, pp.112-117.
23. **SIMOUDIS E., (1996)**, “*Reality check for data mining*”. IEEE Expert, New York, vol. 11, no. 5, pp. 26-33.
24. **Dixon M., (1997)**, “*An Overview of Document Mining Technology*”, http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps. (Data Download Date: 01 April, 2015).
25. **MANNING C. D., RAGHAVAN P., SCHÜTZE H., (2008)**, “*Introduction to Information Retrieval*”, USA, pp. 40-105.
26. **Salton G., (1983)**, “*Introduction to Modern Information Retrieval*”, McGraw-Hill, USA, pp. 46-194.

27. **TAN P. N., STEINBACH M., KUMAR V., (2006)**, *“Introduction to Data Mining”*, Addison-Wesley, New York, USA, pp.769.

28. **Bradley P.S., Fayyad U.M., (1998)**, *“Refining Initial Points for K-Means Clustering”*, in Proceed of the 1st malison, Kaufmann, Morgan, pp. 91-99.

29. **ROBERTO J., BAYARDO J., AGRAWAL R., (1999)**, *“Mining the Most Interesting Rules”*, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, pp.145-154.

30. **François C., Polanco X., (2000)**, *“Information visualization and Analysis for Knowledge Discovery: using a Multi Self-Organizing Mapping”*, 4th European Conference of Principles and practice of Knowledge Discovery in Databases (PKDD), Lyon, France, pp.12-16.

31. **LAMIREL J. C., TOUSSAINT Y., FRANCOIS C., POLANCO X., (2001)**, *“Using a MultiSOM Approach for Mapping of Science and Technology”*, "In ISSI, Australia, vol. 1, pp. 339-351.

32. **RIJSBERGEN C. J. V., (1979)**, *“Information Retrieval”*, Butterworth, London, UK, pp. 577–597.

33. **GUHA S., RASTOGI R., SHIM K., (1998)**, *“An Efficient Clustering Algorithm for Large Databases”*, ACM SIGMOD, New York, USA, pp. 73–84.

34. **KARYPIS G., HAN E. H., KUMAR V., (1999)**, *“Chameleon: Hierarchical Clustering using Dynamic Modeling”*, IEEE Computer, vol. 32, no. 8, pp.68-75.

35. **ZHANG T., RAMAKRISHNAN R., LIVNY M. (1996)**, *“Birch: An Efficient Clustering Method for Very Large Databases”*, In ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, pp. 103-114.

36. **KOHONEN T., (2001)**, “*Self-Organizing Map*” Springer-Verlag, New York, vol. 30, pp. 945-947.

37. **SZATHMARY L., NAPOLI A., KUZNETSOV S. O., (2007)**, “*A Multifunctional Itemset Mining Algorithm*”, Proceedings of the 5th International Conference on Concept Lattices and their Applications, pp.22-33.

38. **Han J., Kamber M., (2001)**, “*Data Mining Concepts and Techniques*”, Morgan Kaufmann Publishers, London and New York, pp. 33–50.

39. **ENGLAND R., BEYNON D., (1981)**, “*Statical Algorithms: Remark AS R39: A Remark on AS 136: A K-Means Clustering Algorithm*”, Canada, vol. 30, no. 3, pp.355-356.

40. **LAMIREL J. C., AL SHEHABI S., (2005)**, “*Efficient Knowledge Extraction Using Unsupervised Neural Network Models*”, 5th Workshop on Self-Organizing Maps, Paris, France, pp.291-298.

41. **Hamerly G., Elkan C., (2002)**, “*Alternatives to the K-Means Algorithms that Find Better Clustering*”, in Procceed of the 11th Intl Conference on Information and Knowledge Management, Virginia, pp. 600-607.

42. **Singh L, Chen B., Haight R., Scheuermann R., (1999)**, “ *An Algorithm for Constrained Association Rule Mining in Semi-Structured Data*”, In proc. Of the 3rd pacific-Asia conf., China, pp. 148-158.

43. **Tan P. N., Kumar V., Srivastava J., (2002)**, “*Selecting the Right Interestingness Measure for Association Pattern*”, in Proc. of the 8th Intl. Conf. On Knowledge Discover and Data Mining, Canada, pp. 32-41.

44. **Robertson S. E. and Sparck J. K. (1976)**, “*Relevance Weighting of Search Terms*”, Journal of the American Society for Information Science, vol. 27, pp. 129–146.

APPENDIX A

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: AL-JIBOURI, Ali

Date and Place of Birth: 14 January 1980 Diyala, Iraq

Marital Status: Married

Phone: +9647710019083

Email: alialjubury96@yahoo.com



EDUCATION

Degree	Institution	Year of Graduation
M.Sc.	Çankaya University, Information Technology	2015
B.Sc.	Al-Mustansirya University \ College Of Computer Science	2003
High School	ABE HANEFA AL NOMAN	1999

FOREIN LANGUAGES

English, Beginner Turkish.

HOBBIES

Football , Reading, Travel, Swimming.

APPENDIX B

MULTIPLE INDEXATION FOR PATENTEES

Our example is a set of 1000 patents about oil engineering technology recorded during the year 1999.

B1. The Analysis Phase

The MultiSOM application role has been firstly planed by the control expert in purpose of get answers to such different types of questions on the patents that:

- 1: “Which are the relationships between the patentees?”
- 2: “Which are the advantages of the different oils?”,
- 3: “Does a patentee works on a limit engineering technology, for which advantage and for which use?”,
- 4: “Which is the technology that is utilized by a given patentee without being used by another one?”,
- 5: “Which are the major advantages of a specific oil component and do this advantages have been refered in all the patents using this unit?”.

An analysis has been achieved on all the possible types of question led the expert to define various viewpoints on the patents that could be related to different closed semantic domains appearing in these questions. One of the major goal of the expert was to be able to use each viewpoints in apart in purpose to get answers to domain closed questions (like questions 1,2) while maintaining the possibility of a multi viewpoint communication in order to get answers to multi-domain questions (like questions 3,4,5) that might also includes negation (like question 4). The specific viewpoints which have been focus by the expert from the set of possible questions are:

- 1: Patentees,
- 2: Title (often contains information on the specific components used in the patent),
- 3: Use,
- 4: Advantages.

B2. The Technical Realization

The role of this phase includes in mapping the four specific viewpoints highlighted by the domain expert in the preceding phase in four different maps. A preliminary task contains in obtaining the index set (i.e. the vocabulary set) related to each viewpoint from the full text of the patents. This task has been itself divided into three elementary steps. At the step 1, the structure of the patent abstracts is parsed in order to extract the subfields corresponding to the **Use** and to the **Advantages** viewpoints¹. At the step 2, the rough index set of each subfield is constructed by the use of a basic computer-based indexing tool [4]. This instrument extracts terms and noun phrases from the subfield content according to a normalized terminology and its syntactical variations. It eliminates as well usual language templates. At the step 3, the normalization of the rough index set related to each viewpoint is performed by the domain expert in order to obtain the last index sets. The normalization of the **Title**, **Use** and **Advantages** subfields consists in selecting a single representative among the terms or noun phrases which represent the same concept (for ex., “oil fabrication” and “oil engineering” noun phrases will be both assimilated to the single “oil engineering” noun phrase). The normalization of the **Patentees** viewpoint is operated in the same way regarding that the same firm can appear with different names in the set of published patents. After the construction of the final index sets, the patents are re indexed apart for each viewpoint thanks to these sets. Figure 22 presents a patent abstract consist its generated multi-index. A classical *IDF-Normalization* step [44] is applied to the index vectors associated to the patents in order to reduce the influence of the most widespread terms of the indexes. The table 6 summarizes the results of the patent indexation and the map building.

1. The index count of the **Title** field is significantly higher than the other ones. An analysis of the indexes explain that the information contained in the patent titles is both sparser, of higher diversity, and more precise than the ones contained in the **Use** and **Advantages** fields. Thanks to the expert opinion, the high level of generality of the **Use** and **Advantages** fields, which consequently led to poorer generated indexes, could be explained as an obvious strategy of the Patentees for indirectly protecting their patents.

2. The final patentees number (i.e. 32) has been significantly decrease by the expert as contrasted to the one primary generated by the computer-based indexing tool. The main part of this reduction is not due to variations in patentee names. It is related to the fact that the prior goal of the study was to consider the main companies and their relationships. Thus, the patentees corresponding to small companies have been grouped into a same general index: “Divers”.
3. On the Patentees map, the class number is close to the last number of retained patentees. Most of these patentees will then be associated to separate classes on the Patentees map.
4. Only 62% of the patents have an Advantages field and 75% a Use field. Consequently, some of the patents will not be indexed for the all the expected viewpoints. The role of the mechanism of communication between viewpoints (see next section) will then be to generate indirect evaluation of the contents of these patents on their missing viewpoints through their associations with other patents.

Title:	Lubricating oil composition - includes a base oil, sulphur-containing organic molybdenum compound, one organic acid salt compound,
Patentee(s):	TONEN CORP
Abstract:	Lubricating oil composition includes a base oil, (a) a sulphur-containing organic molybdenum compound, and (b) at least one organic acid salt compound selected from (b-1) and (b-2), wherein (b-1) is an organic acid metal salt compound excluding copper carboxylate and being a metal salt of an organic acid selected from a carboxylic acid, an aliphatic sulphonic acid, an aromatic sulphonic acid, an alkyl salicylic acid and an alkyl phenolic compound in which the metal is a metal selected from 1A, 3A-7A, 8 and 1B-6B groups of the Periodic Table, and (b-2) is an ammonium salt and an amine salt compound the organic acid selected from the carboxylic acid, the aliphatic sulphonic acid, the aromatic sulphonic acid, the alkyl salicylic acid and the alkyl phenolic compound
Use:	The lubricating oil composition is used for internal combustion engines, automatic gear changers, gears,
Advantages:	The lubricating oil composition exhibits excellent friction-reducing effect over a wide range from low temperatures to high temperatures
Final indexation:	<u>adv.friction coefficient stability on a wide range of temperature</u> ; <u>adv.friction reduction</u> ; <u>adv.low temperature</u> ; <u>soc.TONEN CORP.</u> ; <u>titre.base oil</u> ; <u>titre.lubricant composition</u> ; <u>titre.lubricating</u> ; <u>titre.organic acid salt</u> ; <u>titre.sulfur-containing organo molybdenum</u> ; <u>use.automatic transmission</u> ; <u>use.engine oil</u> ;

Figure 22 Example of a Patent Abstract with Its Generated Multi-Index

The multi-index that has been generated for the above patent abstract corresponds to the “**Final indexation**” field. The terms of the generated multi-index are prefixed by the name of the viewpoint to which they are associated: “adv.” for the **Advantages**

viewpoint, “titre.” For the **Title** viewpoint, “use.” for the **Use** viewpoint, “soc.” for the **Patentees** viewpoint.

	Patentees	Titles	Uses	Advantages	GlobMin (WEBSOM)	Glob Max (WEBSOM)
Number of indexed documents (NID)	1000	1000	745	624	1000	1000
Number of rough indexes generated (NRI)	73	605	252	231	1395	1395
Number of final indexes (NFI)	32	589	234	207	1075	1075
Number of map classes with members (/100)	28	55	57	61	89	238

Table 6 Summary of the Results of Patent Indexation and Map Building

Note that the NRI (resp. NFI) of the “global viewpoint” are less than the sum of the NRIs (resp. NFIs) of all the specific viewpoints (i.e. 1089) because there are similar indexes occurring in different viewpoints.