SOFT SET THEORY FORDATA REDACTION

Student Name

AMIN LUDDIN BINADNAN

ID:

CB09I14

THESIS SUBMITED IN FULFILMENT OF THE DEGREE OF

COMPWER SCIENCE

FACULTY OF COMPUTER SYSTEM AND SOFTWARE

ENGINEERNG

2012

ABSTRACT

The recent changes in utility structureso development in renewable technologies and increased. There are many data exist all stored data stored in the computer using intemet, everyday data was stored. This data poses a problem when we need to use data" but the data are too numerous and scattered on the internet blur of data. Therefore, there are techniques required and are introduced to overcome this problem. Discussion discussed is Knowledge Discovery in Databases and techniques used are multi-soft set of techniques. Dataset is a set of multi-value data. By using Multi soft sets irq can rcduce the data based on the theory of soft sets

ABSTRAK

Dengan adanya teknologi informasi sekarangini, jumlah data-data semakin banyak. Kesemua data disimpan disimpan di dalam computer dengan adanya internet, semakin hari semakin banyak data yang disimpan.Perkara ini menimbulkan masalah apabila kita memerlukan data untuk kegunaan, tetapi data yang ada terlalu banyak dan berselerak di intemet. Oleh sebab itu, terdapat teknik-teknik yang diperlukan dan diperkenalkan untuk mengatasi masalah ini perbincangan yang dibincang adalah Knowledge Discovery in Databases dan teknik yang digunakan ialah teknik multi soft set. Dataset yang digunakan ialah set data multi value. Dengan menggunakan Multi soft set in, dapat mengurangkan data berdasarkan teori soft set

Table of Contents

		Page
DECLARATION		II
ACKNOWLEDGMENTS		III
ABSTRACT		IV
ABSTRAK		V
CONTENTS		VI
LIST OF TABLES		IX
LIST OF FIGURES		X
LIST OF ABBREVIATIONS		XI
LIST OF SYMBOLS		XII
CHAPTER I		1
1.2	PROBLEM STATEMENT	2
1.4	Scope	3
1.5	Contribution	3
1.6	Thesis Organization	3
CHAPTER II		4
2.1	Knowledge Discovery from Databases	4
2.1.1	Definition	5
2.1.2	KDD Processes	5
2.1.3	KDD Application	9
2.1.4	Data Reduction	9
2.2	Soft set theory	11
2.2.3	Soft set for data reduction	13

CHAPTER III 15

METHODOLOGY 15

3.1 Information Systems and Set Approximations 15

3.2 Soft Set Theory 19

3.3 Reduct in Information Systems using Soft Set Theory 22

3.4 Multi-soft sets construction from multi-information systems 22

3.5 AND and OR operations in multi-soft sets 26

3.6 Attribute reduction 27

CHAPTER IV 29

RESULTS AND DISCUSSION 29

4.1 Software Design 29

4.1.1 Interface design 30

4.2 Datasets 32

4.3.1 Experimental Results 33

4.3.2 AND and OR operations in multi-soft sets 35

4.3.3 Attribute reduction 36

4.4 AND and OR operations in multi-soft sets 40

4.4.1 Attribute reduction 41

CHAPTER V 44

CONCLUSION AND FUTURE WORK 44

REFERENCES 45

LIST OF TABLE

Table Number		Page
3.1	An Information system	17
3.2	An information system from small dataset	20
3.3	Tabular representation of soft set in the above example	22
3.4	The multi Boolean information systems	26
4.1	Dataset	34
4.2	dataset	34
4.3	decomposition of a multi-valued information system into multi-valued information systems	35
4.4	A decomposition of a multi-valued information system	41

LIST OF FIGURES

Figure Number		Page
2	Overview of the steps that compose the KDD process	8
3	A decomposeition of a multi valued information System	25
4.1	Start interface	32
4.1	Creator interface	32
4.3	About Interface	33
4.4	Calculation interface	33

LIST OF ABBREVIATIONS

KDD: Knowledge Discovery in Database

DM: Data Mining

IT:Information Technology

LIST OF SYMBOLS

U: Universal set

x: Variable

f (x): function of x

(a,b): open interval

{}: braces

$\in$: element of

$\notin$: not element of

<<: much less than

>>: much greater than

$\cap$: intersection

|A|: cardinality

supp(u): support

CHAPTER I

INTRODUCTION

In this chapter, thesis will coming out of the overview from this research, this thesis consist six parts. The first thesis part is introduction with whole of thesis. Follow by the problems statements, thesis did show what happen real world problems using large data .Continue part of research with objective is coming out solve where the project is determined. Next are scopes of the system, follow up contribution and lastly the thesis organization which describe of the thesis

1.1 RESEARCH BACKGROUND

The real world, what human consist is a large of data to be analysis, information and save of the information to represent in an information table, Where set of attribute describe of the set of object. Some particular property will face by particular problem, entire of attribute set necessary to preserve attribute property [1]. To describe of entire data make of suffer to describe when time-consuming and hard to understand, apply or verify when consist rules. Problems make solver to deal with this problems so that required attribute reduction. Objective of this thesis to solve problems is reduce of the number attributes and make it at same time. Molodstov has proposed the theory soft set at 1999, it is for handling some information has uncertain. Binary , basic , and elementary that called soft set [3]. The theory of soft set is parameter to mapping while crisp part of universe. The structure of soft set , we cal classified the object into binary (yes/1or no/0).the Boolean-valued has deal soft set standard . Data analysis and decision support can deal by using theory of soft set standard. Concept reduction is a application to supporting fundamental application. The theory of soft set using dimensionality reduction have been proposed and compared. Dimensionality of

reduction only can use at Boolean-valued application Normal peoples still blur what the theory soft set. So that the theory of soft set just popularity to researcher and every year paper published make a good position. According Maji et al soft set theory has been introduce some relative operation concept .Some application of soft set was begin by Aktas and Cagman that the concept in algebra. Some application also, soft set theory BCK/BCI algebra eas introduce by JUN and Park. Three theories to distinct to deal the vagueness, first theory membership is decided by adequate parameter. Second theory rough set to employee equivalence classes and grade of membership using fuzzy set theory. Here we try to establish link between soft sets and fuzzy soft sets and soft rough sets [2]

1.2 PROBLEM STATEMENT

However, the thesis research of soft sets, hard to represent data multi value of dataset to multi soft set of dataset. In real situation application, depending data we present the set of parameter, a result of parameter will have value like contain, grade and multiple. For examples, grade of degree mathematical student can classified into three value like a high,, medium and low. The real situation, each parameter will determine a partition of the universe, since is contain two or more disjoint subset. Multi value information cannot directly convert into multi soft set like a multi value system.

1.3 OBJECTIVES OF RESEARCH AND SCOPE OF WORKS

This research focuses on the development of new techniques for

a. To present the idea of multi-soft sets to deal multi-valued information systems.
b. To describe the idea of dimensionality reduction for categorical (multi-valued) information systems (dataset) under soft set theory.
c. To develop a system for data reduction using Visual Basic.

1.4 Scope

The scopes of this project are described as follow:

a. The data used in the project is based on categorical dataset.
b. The technique used is based on soft set theory (multi-soft sets).
c. The software developed later using Visual Basic.

1.5 Contribution

This contribution consist three main, the first main contribution will coming out multi soft set. Second main contribution applicability of data reduction using soft set theory under multi- value information using multi-set and AND operation. Lastly the operation reduct at multi value information can obtain using soft set theory. Although result will presented as some result, main of this part paper reveal interconnection multi-value information and reduction using rough and soft set theory

1.6 Thesis Organization

To organize of this project, we divide by section. Section 2 explanations about dimensionality reduction. Section 3 explanations about information system follow up approximation. Section 4 explains about fundamental soft set theory. Section 5 explains about reduction information system using theory called soft set theory. Finally, at last section we conclude our work at section 6.

CHAPTER II

LITERATURE REVIEW

This chapter briefly discusses on existing literature related with the proposed project. There are two main sections in this chapter. The first section introduces on Knowledge Discovery from Databases. The second section describes some brief Soft set theory.

2.1 Knowledge Discovery from Databases

This section presents a definition, processes, and applications of Knowledge Discovery in Databases (KDD), following by data reduction process.

2.1.1 Definition

Nowadays database has been rapidly growing , so that make a database need technology to explain and summaries the information maybe contain some important . Now database growing double every 36 month, it more important to analyzing to gain information. Process data mining is part knowledge discovery, which is useful pattern and out meaning-full in large of data. Knowledge discovery give more definition, different people different definition what actually knowledge discovery. First what definition is knowledge discovery as “a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from collections of data” it’s especially to reveal some new and to get useful information from data [1] . Second definition of knowledge discovery is non-trivial process of identifying valid , able to understand pattern of data , since for implementing a KDD case study[2]. Other researcher give definition is multi step of data to be manipulated and uncover useful knowledge from data mining in held database. [2]

.

2.1.2 KDD Processes

KDD is a non-trivial process of identifying valid, novel, potential, useful, and understand able data patterns. KDD process has been to apply at managing valuable Taiwanese airline passenger such as apply the KDDD process and

data mining approach to explore information on demographic, travel behavior and perception of service quality, to identify valuable passenger using database air line. At here process KDD process depend a researcher to explain [2]. According Jehn-Yih Wong KDD process as a below:

a. Selecting application domain.

During step selecting specific high-value-added area and keynote of problems and knowledge needs are the early jobs through this step it is important to make sure the availability of sufficient information related to the problem.

b. Selecting target data.

This regard as application domain, state, problems, and the KDD and DM goals to determine data types.

c. Pre-processing data.

This step cleans or transforms data to ensure research validity.

d. Extracting knowledge.

DM is widely employed during the knowledge-discovery stage in the KDD processes. It mainly seeks meaningful rules or knowledge using automatic or semi -automatic algorithms A series of steps explore s hidden knowledge by selecting the DM task, mining techniques and algorithms, and implementing DM The tasks are categorized into six types: classification; estimation; pre-diction; affinity grouping; clustering ; and description. The classification task involves examining features of a newly presented object and assigning it to a predefined class set.

e. Interpretation and evaluation.

Redundant or irrelevant patterns are removing d by examining graphic s, logic, and other information. Results are translated into terms that are easy for users to understand[2].

While according Brachman KDD process so very interactive, including numerous stage with many decision made by the user. Here we generally outline some of basic stage

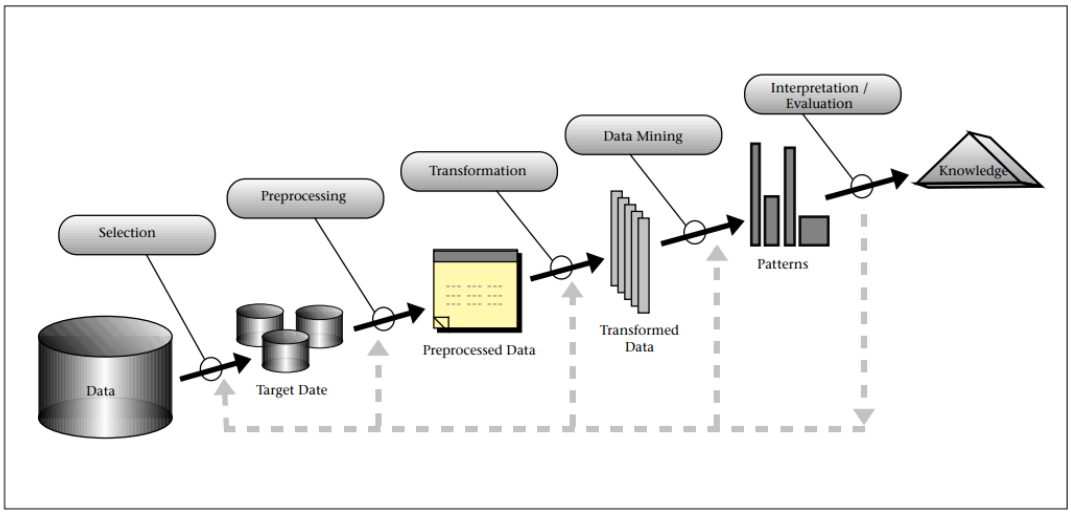

Figure 2: Overview of the steps that compose the KDD process [7]

The process of KDD as described in Figure 2 consists of the following steps [7]:

a. Developing an understanding of the application domain, the relevant prior knowledge, and the goal(s) of the end user.
b. Creating or selecting a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
c. Data cleaning and preprocessing: this step includes, removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes.
d. Data reduction and projection: finding useful features to represent the data depending in the goal of the task. This may include dimensionality reduction or transformation to reduce the effective number of variables under consideration or to find the invariant representations of the data.

e. Matching the goals to a particular data mining method such as summarization, classification, regression, clustering etc. Model and hypothesis selection, choosing the data mining algorithm(s) and methods to be used for searching for data patterns.

f. Exploratory analysis and model and hypothesis selection: choosing the data mining algorithms(s) and selecting method(s) to be used for searching for the data patterns. This process includes deciding which models and parameters might be appropriate and matching particular data mining method with the overall criteria of the KDD process.

g. Data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data mining method by correctly performing the preceding steps.

h. Interpreting mined patterns: possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.

8

i. Acting on discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

2.1.3 KDD Application

In today's society, the Information Technology (IT) is an increasingly part of all economic, technological, educational and even cultural sectors. Through applications such as e-commerce, networking, and digital administration, the IT evolution has become one of the most important factors in shaping the future of our social system [20]. For the example operational stage s of applying the KDD procedure to a large Taiwanese airline includes three sub-procedures involving eight stage s within the operation procedure he data consists of personal information, and passenger opinion surveys on consumption trends, and airline services from November to December.[2]. To performance analyzing and order to explore the factor having impact on the success of university student, to able explore system has been developing called MUSKUP to test on student data. Using this software all task it can keep information together by knowledge discovery together. What coming out from this system students' family easy to be associated w ith student success application.

.

2.1.4 Data Reduction

Reduction make a many definition like a subset of attributes that jointly sufficient and individually necessary for preserving a particular property of a given information table. Reduct generally definition of attribute first there variety of property that can be observe in a information table. Second of definition is preservation of certain property by attribute set can be evaluated by different measure, since that can be define as different function. Third of definition can be define as monotocity property of a particular fitness function. Reduct of knowledge is an important step at knowledge discovery and method of general rules. We can see many researcher researches about the reduct but half of them just using static. As we know rough set is part knowledge discovery, the method reduct using standard rough set method are effective to some extent but to solve problems practice has some problems . using standard rough set are not always sufficient to solve decision system. The reason why still problems is a not taking into account the fact that part of reduct is chaotic , we can say it not stable . Dynamic data, incremental data and noise data make the analysis

results instable and uncertain. All of these limit the application of rough set theory [14]. Commonly rough set theory is used to extract rule from and reduce attribute in database , which attribute are characterized by partitions [15]. Data Reduction is the transformation of numerical or alphabetical digital information derived empirical or experimentally into a corrected, ordered, and simplified form. Columns and rows are moved around until a diagonal pattern appears, thereby making it easy to see patterns in the data.When the information is derived from instrument readings then there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, coding, sorting, collating, and producing tabular summaries, and when the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are likely to be needed. Often the data reduction is undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined. Coding of Data Reduction: Coding involves three stages: Open Coding – Data is broken down and examined. The aim is to identify all the key statements in the interviews that relate to the aims of your research and your research problem. After identifying the key statements you can then put the key points that relate to each other into categories giving a suitable heading for each category. Axial Coding – After the open coding stage, this stage is to put the data back together and part of this process means re-reading the data you've collected so you can make precise explanations about the area of interest. During this stage new categories may be developed and used. Questions like this are asked usually in the axial stage – Can I put certain codes together under a more general code than keeping them separate in two? Selective Coding - This is the final stage of coding, this involves aiming to make the finishing touches to your categories and finish so you can group them together. When grouped together, you will then have to produce diagrams to show how your categories link together. The key part of this is to select a main category, which will form the main focal point of your diagram. Also you will need to look for contradictive data on previous research rather than data which supports it.[5]. An important knowledge discovery problem is to establish a reasonable upper bound on the size of a data set needed for an accurate and efficient analysis. For example, for many applications increasing the data set size 10 times for a possible accuracy gain of 1% cannot justify huge additional computational costs. Also, overly large training data sets can result in increasingly complex models that do not generalize well [8].

Traditionally, the concept of Data Reduction has received several names, e.g. editing, condensing, filtering, thinning, etc, depending on the objective of the reduction task. Data reduction techniques can be applied to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintains the integrity of the original data. That s, mining on the reduced dataset should be more efficient yet produce the same analytical results. There has been a lot of research into different techniques for the data reduction ask which has leaded to two different approaches depending on the overall objectives. The first one is to reduce the quantity of instances, while the second is to select a subset of features from the available ones. The later, known broadly as dimensionality reduction can be done in two ways, namely, feature selection and feature extraction. Feature selection refers to reducing the dimensionality of the space by discarding redundant, dominated or least information carrying features [11].

2.1.5 Data Reduction Process

The multiple-valued datasets will be transferred as multi-soft sets [13]. Further, the AND operation in the sets [14] is used for data reduction.

2.2 Soft set theory

This section presents a history of soft set, applications of soft set and Soft set for data reduction.

2.2.1 History

Theory of soft sets is introduced by Molodtsov. This theory is a relatively new approach to discuss vagueness. It is getting popularity among the researchers and a good number of papers is being published every year. In Maji et al discussed theoretical aspect of soft sets and they introduced several operations for soft sets. Some applications of soft sets are discussed in. In concept of fuzzy soft sets is introduced [10]. Soft set theory is getting popularity among the researchers working in diverse areas due to its applications in these fields. It is a new tool to deal with uncertainty, alongside with fuzzy sets and rough sets [9].In 1999, Molodtsov

introduced soft sets and established the fundamental results of the new theory. It is a general mathematical tool for dealing with objects which have been defined using a very loose and hence very general set of characteristics. A soft set is a collection of approximate descriptions of an object. Each approximate description has two parts: a predicate and an approximate value set. In classical mathematics, we construct a mathematical model of an object and define the notion of the exact solution of this model. Usually the mathematical model is too complicated and we cannot find the exact solution. So, in the second step, we introduce the notion of approximate solution and calculate that solution. In the Soft Set Theory (SST), we have the opposite approach to this problem. The initial description of the object has an approximate nature, and we do not need to introduce the notion of exact solution. The absence of any restrictions on the approximate description in SST makes this theory very convenient and easily applicable in practice. We can use any parameterization we prefer with the help of words and sentences, real numbers, functions, mappings, and so on. It means that the problem of setting the membership function or any similar problem does not arise in SST [12].

2.2.2 Application of soft set

Most of our traditional tools for formal modeling, reasoning, and computing are crisp, deterministic, and precise in character. But many complicated problems in economics, engineering, environment, social science, medical science, etc., involve data which are not always all crisp[3]. Applications of soft sets in decision-making problems have been studied by many authors in different contexts . In present paper we discuss the concept of reduction of parameters in soft sets. Majiet al. initiated the concept of application of soft sets in decision-making. Unfortunately errors were pointed out in this initial level work in by Chen et al. They rejected the point of view presented in pointed out some odd situations which may occur when method of reduction of parameters in case of soft sets given in [8] is applied. So, they introduced the concept of reduction of normal parameters [9]. It has beenseen that there is a very close relationship between soft sets and rough sets. So it is natural to ask, “Can we develop a method of reduction of parameters for soft sets as we do in case of rough sets for attribute reduction without losing important information?” In this paper,

we try to find answer to this question. Applications of Soft Set Theory in other disciplines and real life problems are now catching momentum. Molodtsov successfully applied the soft theory into several directions, such as smoothness of functions, game theory, operations research, Riemann integration, Perron integration, theory of probability, theory of measurement, and so on. Majietal.gave first practical application of soft sets in decision making problems. It is based on the notion of knowledge reduction in rough set theory. Maji et al defined and studied several basic notions of soft set theory [12]

2.2.3 Soft set for data reduction

The idea of reduct and decision making using soft set theory was firstly proposed by Majiet al.[4]. In [4], the application of soft set theory to a decision making problem with the help of Pawlak's rough mathematics was presented. The reduction approach presented is using Pawlak's rough reduction and a decision can be selected based on the maximal weighted value among objects related to the parameters. Chen et al. [5-6] presented the parameterization reduction of soft sets and its applications. They pointed out that the results of reduction proposed by Maji is incorrect and observed that the algorithms used to compute the soft set reduction and then to compute the choice value to select the optimal objects for the decision problem proposed by Maji are unreasonable. They also pointed out that the idea of reduct under rough set theory generally cannot be applied directly in reduct under soft set theory. The idea of Chen et al. for soft set reduction is only based on the optimal choice related to each object. However, the idea proposed by Chen is not error free, since the problems of the sub-optimal choice is not addressed. To this, Kong et al.[7] analyzed the problem of suboptimal choice and added parameter set of soft set. Then, they introduced the definition of normal parameter reduction in soft set theory to overcome the problems in Chen's model and described two new definitions, i.e. parameter important degree and soft decision partition and use them to analyze the algorithm of normal parameter reduction. With this approach, the optimal and sub-optimal choices are still preserved. Zou[8] proposed a new technique fordecision making of soft set theory under incomplete information systems. The idea is based on the calculation of weighted-average of all possible choice values of object and the weight of each possible choice value is decided by the distribution of other objects. For fuzzy soft sets, incomplete

data will be predicted based on the method of average probability. All those techniques are still based on Boolean information systems. As to this date, no researches have been done on dimensionality reduction in multi-valued information systems under soft set theory. Since every rough set [9] can be considered as soft set as presented in [10], thus, an alternative approach with potential for finding reduct in multi-valued information systems is using soft set theory. Still, it provides the same results for rough reduction [11,12]

CHAPTER III

METHODOLOGY

In this chapter, the proposed dimensionality reduction is proposed.

3.1 Information Systems and Set Approximations

An information system is a 4-tuple (quadruple) $S=(U,A,V,f)$, where $U=\{u_1,u_2,u_3,\cdots,u_{|U|}\}$ is a non-empty finite set of objects, $A=\{a_1,a_2,a_3,\cdots,a_{|A|}\}$ is a non-empty finite set of attributes, $V=\bigcup_{a\in A}V_a$, V_a is the domain (value set) of attribute a, $f:U\times A\to V$ is an information function such that $f(u,a)\in V_a$, for every $(u,a)\in U\times A$, called information (knowledge) function.

An information system is also called a knowledge representation systems or an attribute-valued system. An information system can be intuitively expressed in terms of an information table (see Table 1).

Table 3.1. An information system

U	a_1	a_2	...	a_k	...	$a_{\|A\|}$
u_1	$f(u_1,a_1)$	$f(u_1,a_2)$	...	$f(u_1,a_k)$	...	$f(u_1,a_{\|A\|})$
u_2	$f(u_2,a_1)$	$f(u_2,a_2)$	...	$f(u_2,a_k)$	...	$f(u_2,a_{\|A\|})$
u_3	$f(u_3,a_1)$	$f(u_3,a_2)$	...	$f(u_3,a_k)$	...	$f(u_3,a_{\|A\|})$