Duo-Mining Techniques in Knowledge Discovery Process in Data Base

Ch Narsimha Chary*, Kancharla Bharath Reddy, Pillalamari Vinay Bhuahan

Research Scholar, Sri Satya Sai University of Technology & Medical Sciences, Bhopal, Madhya Pradesh, India

ABSTRACT

Duo mining is used frequently in a mixture of industries and its enduring to gain in both popularity and acceptance. Duo Mining is a blend of data and text mining. This paper suggests Data mining architecture in addition with Knowledge discovery process. It also presents the comparison between data mining and text mining. As Data mining handles various processes like text Mining, Multi-media, Web mining etc.Text Classification, Clustering, Keyword based Association are the terms that are used to describe the process of Text Mining

Keywords : Association, Clustering, Data Mining knowledge, Pattern mining, Regression, Data Mining, classification, Duo Mining, Pattern Recognition, Text Mining

I.   INTRODUCTION

Duo Mining is the variation of data and text mining. It has demonstrated especially well for the banking and credit card companies in order to take better decisions. As separate capabilities, of the pattern finding technologies of data mining and text mining have been around for years. However, it is only recently that enterprises have been started to use the two in acycle - and have discovered that it is a combination that is worth more than the sum of its parts.

They are similar because they both "mine" large Amounts of data, and looking for significant patterns. However, what they evaluate is quite different. Instead of only being able to analyse the structured data they collect from transactions, they can add call logs from customer services and further analyse customers.

In addition, spending patterns from the text mining side. These new developments in text mining technology that go beyond simple searching methods are the key to information discovery, which is generally work on the unstructured data.

There are several methods of data mining which handle the following or application of mining:

· Spatial mining
· Multimedia mining
· Text mining
· Web mining

Spatial mining:

Spatial is a three–dimensional object, and mining is extraction of patterns. Non-trivial searches "robotic" as possible to diminish human effort. It refers to the extraction of knowledge, spatial relationship, or other fascinating patterns not explicitly stored in spatial databases. Such mining demands an incorporation of data mining with spatial database technology.

Multimedia mining:

Its stores and manages a large collection of multimedia data, such as audio, video, image, image, hypertext data, which contain text, textmarkups and linkage. Multimedia database system is gradually more common due to the trendy use of audio video equipment, digital cameras and the internet.
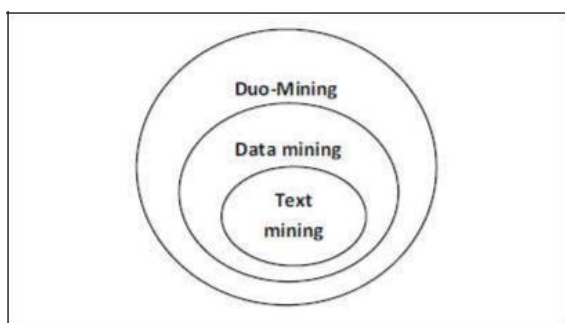
Text mining:

In text mining, the goal is to discover unidentified information, something that no one yet knows and so could not have yet written down. Text mining is a distinction on a field called data mining that tries to find motivating patterns from bulky databases.

Web mining:

The application of data mining techniques to find out patterns from the Web. According to analysis targets, web mining can be divided into three are Web usage mining, Web content mining and Web Structure mining.

We can use data mining in so many different kinds of places in the world are as follows:

- ✓ Analysts and managers who deals with strategic and tactical decision-making.
- ✓ Managers responsible for revenue and cost
- ✓ Detection.
- ✓ Risk managers in insurance, to minimize the
- ✓ Risk of claim and to maximize the profit.
- ✓ Educators to improve educational processes to conduct researchers, provide analysis of education effectiveness and
- ✓ Institutional decision.
- ✓ Scientist to provide new knowledge for researchers in various fields.
- ✓ Use knowledge discovery for various opportunities like sales, forecasting, market researches etc.

They are similar because they both "mine" large amounts of data, and looking for significant patterns. However, what they evaluate is quite different. Instead of only being able to analyse the structured data they collect from transactions, they can add call logs from customer services and further analyse customers and spending patterns from the text mining side. These new developments in text mining technology that go beyond simple searching methods are the key to information discovery, which is generally work on the unstructured data.

II. DATA MINING AND ITS TASKS

Data mining is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is to design and work efficiently with large data sets. Data mining is the databases. In data mining the data can be mined by component of wider process called knowledge discovery from database. The definition of data mining is closely related to another commonly used term knowledge discovery. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. Many areas of theory and technology in current era are databases, artificial intelligence, data mining and statistics is a study of three strong large technology pillars. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data-mining algorithm, analyzing results and taking appropriate action. The data, which is accessed, can be stored in one or more operational databases. In data, mining the data can be mined by various processing.

Supervised Learning

In supervised learning (often also called directed data mining), the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in

regression analysis. To proceed with directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

Unsupervised Learning:

In unsupervised learning, all the variables are treated in same way; there is no distinction between dependent and explanatory variables. However, in contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning requires, target variable should be well defined and that a sufficient number of its values are given. Unsupervised teach typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown.

Data mining as a term used for the specific classes of six activities or tasks as follows:

1. Classification
2. Estimation
3. Prediction
4. Affinity grouping or association rules
5. Clustering
6. Description and visualization

The first three tasks - classification, estimation and prediction rules are examples of directed data mining or supervised learning. In directed data mining, the goal is to use the available data to build a model that describes one or more particular attribute(s) of interest (target attributes or class attributes) in terms of the rest of the available attributes. The next three tasks – association rules, clustering and description are examples of undirected data mining i.e. no attribute is singled out as the target, the main goal is to establish some relationship among all attributes.

Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The well-defined classes and a training set consisting of reclassified examples characterize the classification task. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

✓ Classification of credit applicants as low, medium or high risk
✓ Classification of mushrooms as edible or poisonous
✓ Determination of which home telephone lines are used for internet access

Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. Some examples of estimation tasks include:

✓ Estimating the number of children in a family from the input data of mothers' education
✓ Estimating total household income of a family from the data of vehicles in the family
✓ Estimating the value of a piece of a real estate from the data on proximity of that land from a major business center of the city.

Prediction

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the efforts, it is possible to check. Predictive tasks feel different because the records are classified according to some predicted future behavior or way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include:

✓ Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer
✓ Predicting which customers will leave within next six months

✓ Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

Association Rules

An association rule is a rule, which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: "30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

Clustering

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a pre-processing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods .Further data set can be numeric or categorical. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are

no predefined classes. The records are grouped together based on self-similarity. Clustering is often done as a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster [6]. Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data "Explosive growth of data from terabytes to pet bytes

✓ Data gathering and data accessibility.
✓ Traditional techniques are infeasible for unprocessed data.

[2]Data mining is the process of extracting patterns from large data sets by combining of methods from figures and artificial intelligence with the database management system. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of the profiling practices, such as marketing, surveillance, fraud recognition and methodical discovery. Data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.
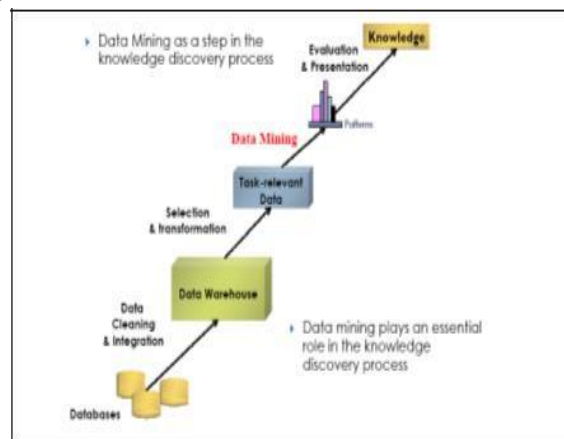
III. Knowledge discovery from data (KDD) process in the data mining:

- ✓ **Data cleaning**: Remove clatter and conflicting data.
- ✓ **Data integration**: Combine multiple data sources.
- ✓ **Data selection**: Data appropriate to analysis tasks are retrieved from the data.
- ✓ **Data transformation**: Transform data into appropriate form for mining.

Data mining: Haul out data patterns
- ✓ **Knowledge** representation: Use visualization and knowledge demonstration tools to present the mined data to the client
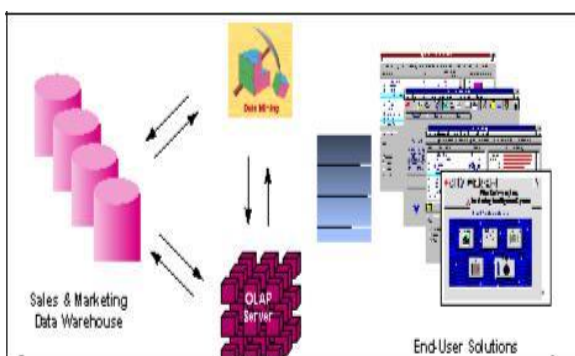
We can use data mining in so many different kinds of places in the world are as follows:
- ✓ Analysts and managers who deals with strategic and tactical decision-making.
- ✓ Managers responsible for revenue and cost detection.
- ✓ Risk managers in insurance, to minimize the risk of claim and to maximize the profit.
- ✓ Educators to improve educational processes to conduct researchers, provide analysis of education effectiveness and institutional decision.
- ✓ Scientist to provide new knowledge for researchers in various fields.

Use knowledge discovery for various opportunities like sales, forecasting, market researches etc.

IV. TEXT MINING AND ITS TYPES WITH TRADITIONAL TECHNIQUE:

It is the process of extracting fascinating and non-trivial information and knowledge from unstructured text. Text mining has been defined as [6] "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources"

Text mining is similar to data mining, except that data mining gear are designed to handle structured data from databases or XML files, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents, HTML files, etc. As a result, text mining is a much better solution for companies, where large volumes of diverse types of information must be multipart and managed.
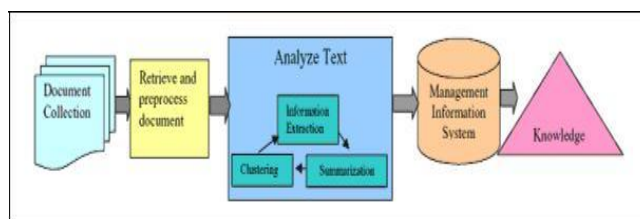
Process of text mining:

- **i.** Text pre-processing
  - o Semantic text analysis
- • Features generation
  - o Bag of words
- • Features selection
  - o Simple counting
  - o Statistics
- • Text mining
  - o Classification
  - o Clustering
  - o Association
  - o Analysis results

We can also knob text data in following ways:

- ✓ Modelling semi-structured data.

- Information retrieval (IR) from unstructured documents.
- Text mining

Information retrieval: [7] Information retrieval trouble locating appropriate documents based on user input, such as keywords or example documents.

Emblematic IR system:

- Online collection catalogs.
- Online document board systems.

Common difference between Information retrieval vs. database systems:

- Some DB troubles are not at hand in IR, e.g..., update, transaction management, complex objects.
- Some IR problems are not addressed well in DBMS e.g... Unstructured documents, approximate search using keywords and relevance.

Several techniques used in reclamation of information:

Credentials can be described by a set of diplomat keywords called **index terms**.

Unusual index terms have unstable relevance when used to describe document contents.

This effect is captured through the handing over of numerical weights to each index terms of a document. (E.g. frequency)

**Index terms** ->Attributes

Weights ->Attributes values

Text mining:
- Text classification.
- Text clustering.
- Keyword based association rule.

Text classification: Involuntary classification for the bulky number of online text documents (web pages, e-mails, corporate intranets etc.). Text document classification is differs from the classification of relational statistics. Documents databases are not prearranged according to attribute value pairs.
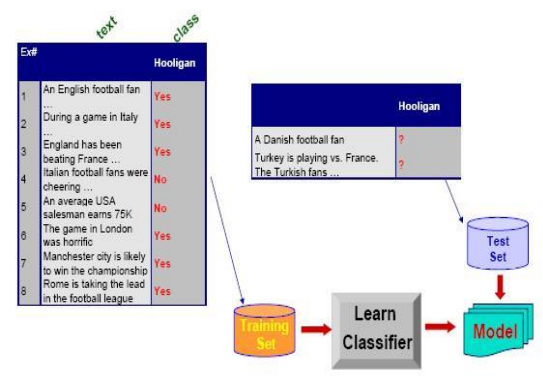
Following steps are taken in the process of Classification

- Data preprocessing.
- Description of training set and test sets.
- Creation of the classification model using the preferred classification algorithm.
- Classification model substantiation.
- Classification of new/unknown text documents.

Classification schema:
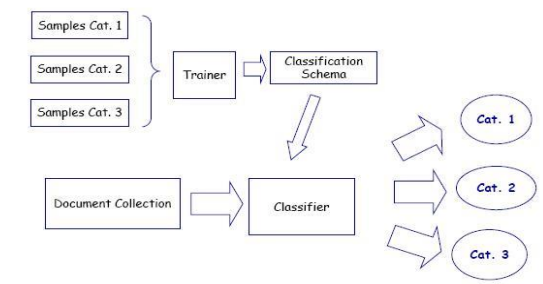
## Example of text classification

Text clustering: The process of isolating a dataset into reciprocally elite groups such that the members of each group are as "close" as possible to one another, and unlike groups are as "far" as possible from one another, where distance is considered with esteem to all available variables.

Clustering schema:
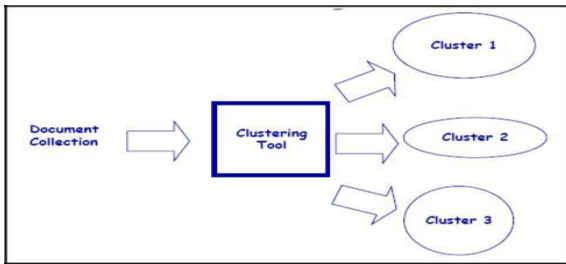
Example of text clustering:
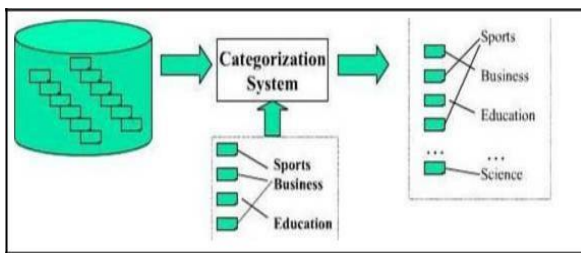
Keyword based association rule: Collect sets of keywords or terms that take placerecurrentlycollectively and then find the association or parallel relationships among them.

Though doing keyword based association rule, we include following some variety of steps to analysis it:

✓ [7]Pre-process the text data by parsing, stemming, removing stop words, etc.

✓ Evoke association mining algorithms

   o Consider each document as a transaction.

   o View a set of keywords in the document as a set of items in the transition

Terms level association mining

   o No need for human effort in tagging documents.

   p The number of meaningless result and the execution time is greatly reduced.

V. CONCLUSION

In this paper, we have illustrated the various methods of Data mining. The proper use of Duo

Mining is a combination of text and data mining have been illustrated. It also concludes the proper analysis of text and data mining in correlation with Techniques, methods, process and architecture. The use of text mining and its various types has been correlated with data mining.

VI. REFERENCES

[1]. Donald Michie, Data Mining Discovering Interesting Relationships in Large Data Sets, Retrieved from http://www.aaai.org /aitopics/ pmwiki/ pmwiki.php/AITopics/DataMining

[2]. Wikipedia, free encyclopaedia, Data mining, Retrieved from http://en.wikipedia.org/ wiki/ Data mining

[3]. Discovering hidden value in your data warehouse, Retrieved from http:// www.thearling.com /text/ dmwhite/ dmwhite.htm

[4]. Mining Object, Spatial, Multimedia, Text, and Web Data, Retrieved from http:// www.dataminingtools.net/ wiki/ applications_ of_data_mining.php

[5]. Wikipedia, free encyclopedia, web mining, Retrieved from http://en.wikipedia.org/ wiki/ Web_mining

[6]. Biomarker information extraction tool (BIET) development using natural language processing In addition, machine, retrieved from http:// portal.acm.org/citation.cfm?id=1741927

[7]. Mining Text and Web Data retrieved fromhttp://www.slideshare.net/pierluca.lanzi/ machine-learning-and-data-mining-19- mining-text-and-web-data