International Journal of Computer Science and Informatics

Volume 2 | Issue 4 Article 11

April 2013

Impact of Duo-Mining in Knowledge Discovery Process

Aditi Chawla
School of Eng. & Tech , Noida International University, U.P, aditichawla@ymail.com

Deepti Sachdeva
Sanlok Institute of Management and Information Technology, Gurgaon, deepti.sachdeva@ymail.com

Impact of Duo-Mining in Knowledge Discovery Process

Aditi Chawla & DeeptiSachdeva

School of Eng. & Tech , Noida International University, U.P
Sanlok Institute of Management and Information Technology, Gurgaon
E-mail : aditichawla@ymail.com & deepti.sachdeva@ymail.com

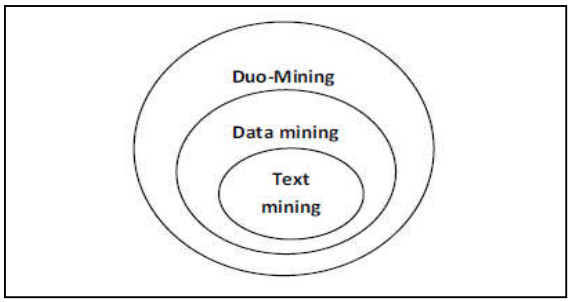
Abstract **-** Duo mining is used frequently in a mixture of industries and its enduring to gain in both popularity and acceptance. Duo-Mining is basically a blend of data and text mining. This paper suggests Data mining architecture in addition with Knowledge discovery process. It also presents the 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.

I. DUO- MINING

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 patternfinding 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. [1]

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 analyze the structured data they collect from transactions, they can add call logs from customer services and further analyze 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.

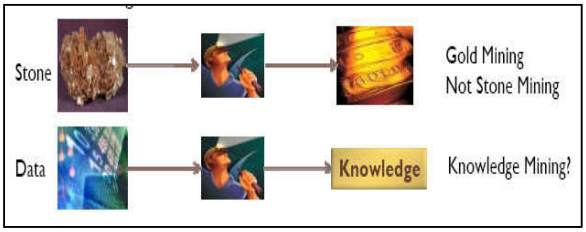


II. DATA MINING AND ITS TASKS

Today Data Mining is used everywhere in a collection of data and its availability and accessibility. But why We use Data Mining.

- 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 bycombining ofmethods from figures and artificialintelligence withthe 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 theprofiling practices, such as marketing, surveillance, fraud recognition and methodicaldiscovery.



The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population

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.

Data mining generally have four classes of responsibilities engage in it are as follow:

Clustering:Is the undertaking of discovering groups and structures in the data that are in some way or another "comparable", without using known structures in the data.

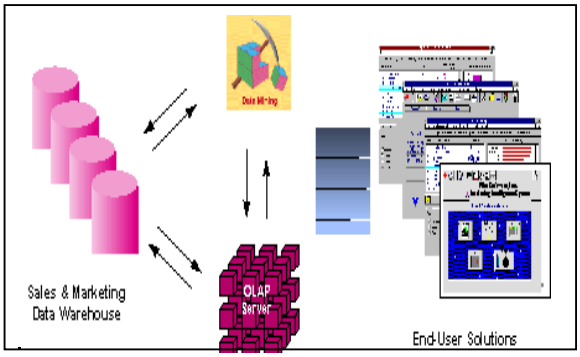
Classification: Is the task of generalize known structure to apply to new data. For example, an email program might attempt to classify an email as justifiable or spam.

- Regression:Attempts to hit upon a function which models the data with the smallest amountof error.
- Association rule learning:Searches for associations between variables. For example a shop might gather data on shopper purchasing habits. Using association rule learning, the shop can conclude which products are recurrently bought jointly and use this information for marketing purposes. This is sometimes referred to as market basket analysis
- chronological pattern mining:A chronological rule: A→ B, says that event A will be immediately followed by event B with a certain assurance.
- Digressionrecognition: Discovering the most considerable changes in data.
- Data visualization: using graphical methods to show patterns in data.

Architecture of data mining:

The ideal starting point is the data warehouse which contains tracking of all customers' data which means domestic data and outdoor data. This warehouse can be implemented in various RDBMS like oracle, Sybase which optimized the elasticity and prompt data access.

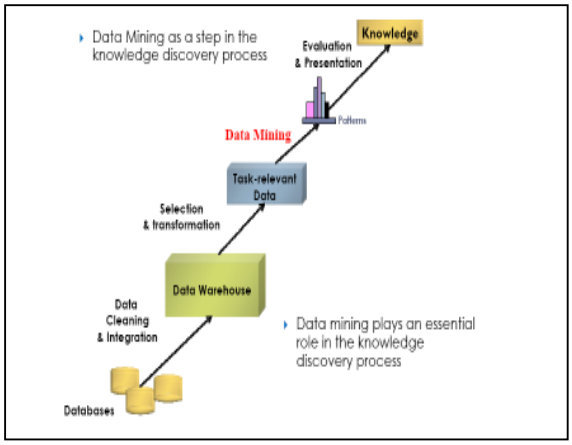
[3] Online analytical processing (OLAP) this types of server unable the end users business model warehouse. Example the multi-dimensional structure allows the user to investigation the data to recap the data in view of region, line etc.to be applied when, navigating through the data



The data mining server integrates with the data warehouse and OLAP server to drive in ROI application directly into the infrastructure and advanced process centric metadata templatedefine the data mining objectives for specific business issue, prospective optimization. Integration with the data warehouse unable the operational decisions to be implemented directly as a warehouse grows with the new decision and results the organization can make the best use of taking effective decisions.

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



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

- Spatial mining
- Multimedia mining
- Text mining
- Web mining

Spatial mining: Spatial is basically a three–dimensional object, and mining is extraction of patterns**.**Non-trivial searchas "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:[4] its stores and manages a large collection of multimedia data, such as audio, video, image, image, hypertext data, which contain text, text markups and linkage. Multimedia database system isgradually 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.

[5]According to analysis targets, web mining can be divided into three are Web usage mining, Web content mining and Web Structure mining.

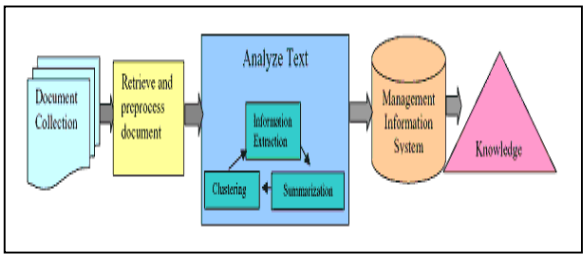
We can use data mining in so many different kinds of places in the world are as follows:

- 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.

III. TEXT MINING AND ITS TYPES WITH TRADITIONAL TECHNIQUE:

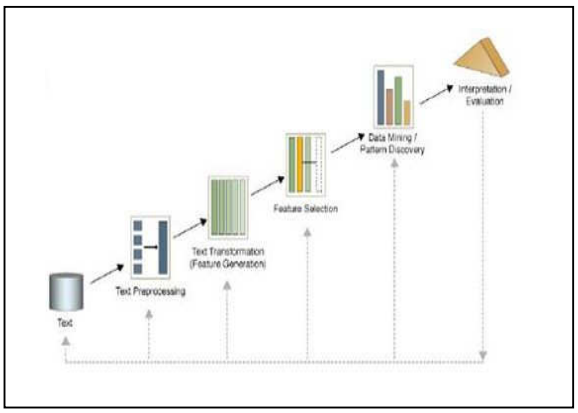
It is the process of extracting fascinating and non-trivial information and knowledge from unstructured text. Text mining has been defined as [6]"the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources"

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:

- Text preprocessing
  - Semantic text analysis
- Features generation
  - Bag of words
- Features selection
  - Simple counting
  - Statistics
- Text mining
  - Classification
  - Clustering
  - Association
  - Analysis results



We can also knob text data in following ways:

- Modeling 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

Types of Text mining:

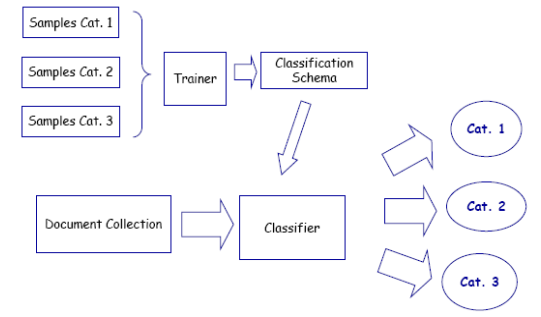
- Text classification.
- Text clustering.
- Keyword based association rule.

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.

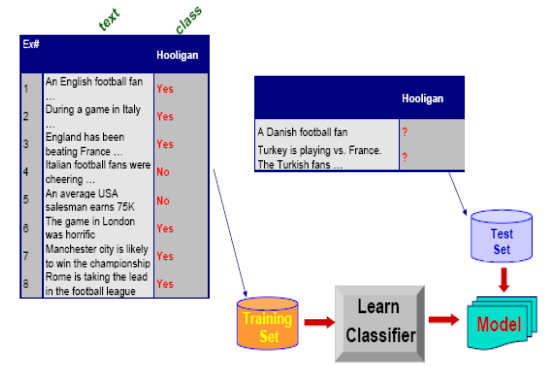
Following steps are taken in the process of Classification

- 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:

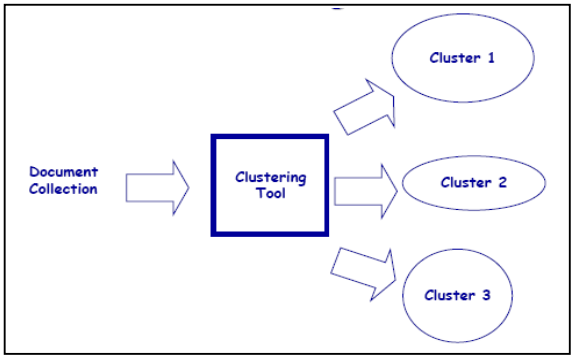


Example of text classification

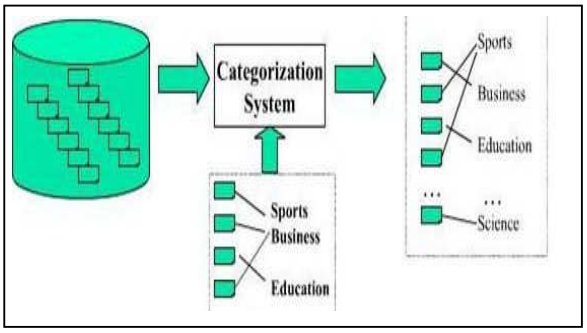


Text clustering: The process of isolating a dataset into reciprocallyelite groups such that the members of each group are as "close" as possible to one another, and unlike groups are as "far" as possible from one another, where distance is considered with esteem to all available variables.

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 to follow some variety of steps to analysis it:

- [7]Preprocess the text data by parsing, stemming, removing stop words, etc.
- Evoke association mining algorithms
    - Consider each document as a transaction.
    - View a set of keywords in the document as a set of items in the transation

Terms level association mining

- No need for human effort in tagging documents.
- The number of meaningless result and the execution time is greatly reduced.

IV CORRELATION BETWEEN DATA AND TEXT MINING

Points	Data Mining	Text Mining
Structure Wise	Structure	Unstructured and semi-structured
Representation in terms of Data	instantly forward	multipart
Space required	tens of thousands	tens of thousands
Complex object used	arithmetical& categorical data	Textual data
Several methods are used to conclude them	Data analysis, mechanism learning statistic, neural networks	Data mining, information retrieval, NLP
Implementation	large implementation since 1994	large implementation stating 2000
Market analysis	10^5 analysts at large and mid sizecompany	10^8 analysts' corporate staff and person users.

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.

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 encyclopedia, Data mining, Retrieved from http://en.wikipedia.org/ wiki/ Data_mining

[3]. Discovering hidden value in your data warehouse, Retrieved fromhttp:// 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 and machine, retrieved from http:// portal.acm.org/citation.cfm?id=1741927

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

◈◈◈