Knowledge Discovery Process in data mining

A Research Paper Writing Proposal

Submitted by

Samita Prashar (11412661)

TO

School of Computer Applications

In Partial Fulfillment of the Requirement For the

Award of the Degree of

Master of Computer Applications

Under the guidance of

Mr. Kumar vishal

April, 2015

CERTIFICATE

This is to certify that Samita prashar (11412661) have completed their MCA Research Paper Writing Proposal titled "**Knowledge Discovery Process in data mining**" under my guidance and supervision. To the best of my knowledge, the present work is the result of their original investigation and study. No part of the dissertation proposal has ever been submitted to any other degree or diploma.
The proposal is fit for the submission and the partial fulfillment of the conditions for the award of the degree of Master in Computer Applications.

Date: ----------

Signature of the Advisor

DECLARATION

We hereby declare that the research paper writing proposal entitled, "**Knowledge Discovery Process in data mining** ", submitted for the MCA Degree is entirely our original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Investigator Name:

Registration Number:11412661

Introduction

Data mining is a process of secreted discover pattern and information from the obtainable data. Data mining is a software that a one of number logical tools for study data.
Data mining is called as information finding in large data, firms and organization to make calculated decisions by assemble study build up commercial access data. It is different multiplicity of tools like query logical processing tools, reporting tools, and resolution Support System (DSS) tools etc.
Data mining is allows to users study data from lots of different scope or angle classify and go over the main points the associations recognized. Data taking out is the procedure of decision correlation or pattern along with dozens of field in huge relational database.
The person are use in the singular tools toward sufficient in the culture. both daytime the individual being are use the infinite information and these information be in the singular field. during the structure of credentials, graphical format record . Not just in the direction of investigate these facts but get a high-quality resolution and continue the facts . The client will mandatory the facts must be retrieve by the folder and create the improved resolution .This method is essentially we call as a data mining or information focus or just KDD(Knowledge Discovery Process).

Literure Review

Dr. Pragnyaban Mishra [1] we have alert a variety of technique, approach and similar area of the study which are useful and marked , significant field of data mining technology. We are alert that big organization are operate in dissimilar places of the dissimilar country. all position of process might produce as big volume of information. trade end result makers require way in as of every source and take planned decision. These types of vast quantity of data are offered in the form bytes which have considerably unchanged in the area of knowledge and business. toward evaluate direct in addition to create result of such kind of vast total of information we need technique called the information removal which self-control convert inside lots of field. This document impart additional amount of purpose of the information removal and also focus scale of the information removal which resolve useful in the additional investigate.

Richard A. Huebner[2](EDM) principles for Educational data mining is an rising regulation that focus on the apply data mining tools and technique educationally related to tha data. If focuses on analyze educational data to build up the models for humanizing knowledge experience and humanizing institutional efficiency. Future research can study how common the acceptance of learning data mining.

Padhraic Smyth [3] data mining to date has mainly focused on computational and algorithmic issue rather than the additional conventional arithmetical aspect of data study. This paper provide the feature appraisal of the origin of data mining and discuss some of the most important theme in present do research in data mining, and including scalable algorithms for considerable data

sets, discover work of fiction pattern in data, study of text, net, and associated to the multi-media data set.

M.S.B. PhridviRaja[4] Information Stream Mining is one of the part fast a lot of useful consequence and is making progress at a quick speed with new methods, methodologies and result in the a variety of application and interrelated to medicine, computer knowledge, bio-information and supply market calculation text, audio and video handing out to name a few. With the huge online data generate from the more than a few sensors, Internet chat, Twitter, Face book, Online Banking communication or ATM etc, the thought of with excitement altering data. In this paper, we provide the algorithm for the result regular patterns from data stream and with a case study and recognize the research issues in handling data streams. The problem of handling the streams for cluster, cataloging and matter discovery is still a confront In this paper we find common pattern from the data stream and definite the which use common pattern generate hierarchy.

Rajashree Shettar1 [5] In order Pattern Mining involve apply data mining method to huge net data repositories to take out usages patterns. The rising reputation of the (www) standard for World Wide Web, many websites is usually thousands of guests skill in every day. the progression hierarchy algorithm is implement for pattern mining . The net log information which is careful as resulting data of the net has been careful for the finding of everyday in order patterns. The consequences have been exposed that the succession algorithm hierarchy perform better than the well-known Generalized Sequential Pattern (GSP) algorithm. Comparative study :-The research show that the successively time of succession tree algorithm is earlier than the average GSP algorithm and also progression hierarchy algorithm discover further number of patterns than the average GSP algorithm.

Nikita Jain[6] the conception of data mining was abbreviation and importance towards its methodologies . The data mining base on the Neural system and Genetically Algorithm. The key knowledge and customs to accomplish the data mining on Neural system and Genetic Algorithm are also survey. Data mining is a dealing out investigation and study. Data mining is regarding processing data and identify pattern.

Chun-Nan Hsu[7]Many application of information finding and data mining such as the regulation discovery for semantic question mark optimization, database combination and decision support (SS), have need of the information to be inconsistent data. databases usually transform in the end and make mechanism exposed information incompatible. Useful information should be strong adjacent to database change so that it is doubtful to become conflicting later than database change. This paper define this idea of toughness and relational database that hold many relations and describe how strength of rst-order Horn-clause rules can be predictable and useful in information finding. Our experiment explain that the opinion move toward can correctly calculate the strength of a rule.hope work mostly focus on apply the advance to singular selection of KDD application in database managing.

Pedro Domingos [8]This paper we alert on the assess VFDT, a result-tree knowledge structure base on Hoeffding trees. A lot of organization today have too much very large total of databases, they include databases that produce exclusive of limit at a rate of some million records

per day. This term paper we describe and evaluate the VFDT,an anytime system that builds conclusion trees using regular memory and even time per example. VFDT can in-business tens of thousands of examples per second by means of off-the-shelf hardware. In data flow mining the incoming data comes in streams, which potentially can amount to time without end Upcoming work We arrangement to compare VFDT with SPRINT/SLIQ.

Peng Peng[9] The large purpose of trade cleverness within commercial require used for information removal software growing in on a daily basis .To get better the good organization and feature of the reuse information removal software with decrease the phase and price on the increase information removal application system. This document propose is latest part documentation structure of information removal. All the way from side to side componention of information removal algorithm, this structure tools different center algorithms of information removal inside the structure of method. during this approach, the effectiveness and class of increasing information removal software be better considerably to gather a variety of function load.

CHENGQI ZHANG[10]In this document present an well-prepared process in support of removal both optimistic ,unenthusiastic relationship policy in catalog.The way extend conventional relations toward incorporate relationship policy. The way have be evaluate use together artificial with actual-earth database, with original consequences are show or efficiency . decision making purpose like as manufactured goods placement ,venture and study untried results have been established and proposed come within reach of is useful well-organized and capable.

Objective

The Main objective of is work is:
Develop a information and drive data mining assistant to support the researchers in data-intensive, information loaded domains.

Research Methodology

Data mining is a procedure representation that describe the usually use proceed and removal expert apply to undertake harms information removal process force be profit as of the experience. like as commerce understanding ,information understanding ,information preparation, model and appraisal exploitation
1. commerce understanding: This step focus on the thoughtful the objectives and requirements from a commerce point of view and also translate data mining difficulty description with design first round project graph to realize the objectives target.

2. information understanding: This step start from the first data collection and getting used to with the data.they are including the particular aims of the classification of data feature problems and first insights into the data,and finding of appealing data subsets.

3. Information preparation: This step cover for all actions needed to make the final dataset,which constitute the data that will be feed into DM tool(s) in the after that step. It include Table, record, and feature group data clear out, structure of new attribute and conversion of data.

4. Modeling: At this position different model techniques are chosen and apply. Modeling usually involves for the use of more than a few method for the same DM difficulty type and the calibration of their parameter to most select values.

5. Evaluation: After one or more model have been build that have high class from a data Study point of view the model is evaluate of the consequences and then method review, and Purpose of the next step.

6. Deployment: Now the exposed information must be prepared and on hand in a way that the purchaser can use. they can employment the plan and plan monitor and repairs or generating of ending report and analysis of the procedure sub steps.

Data mining not just to study these data get a good quality decision and continue the data. They compulsory the information be supposed to retrieve from the folder and prepare the improved conclusion. Information removal concept or method be capable of apply for a variety of field like as medication ,marketing, engineering, web mining, customer relationship management system etc.

learning information removal is a fresh performance of information removal that be able to be helpful of the information connected to the ground of instruction like student record system.

Knowledge Discovery Process - Data Mining issue & challenge

Input problem as regards Knowledge discovery process information Mining are regarding unfinished in sequence, loud & lost information, stage of improbability with fervor & quick-shifting information situation. Information removal application resting on database toward contribute a unrefined information used for contribution. The issue within database / information (for example instability incomplete, sound, quantity enhance issue via the point in moment it reach information removal mission .extra troubles start while a end result of the competence or import of the data store.

Limited Information

A catalog is frequently calculated purpose similar as of information removal from time to time the property or attribute that would make simpler the knowledge duty be not there they can be real request as of the actual earth. uncertain information cause troubles since but various attribute critical toward information regarding the function field be not at hand inside the information it might be there impractical to find out large information regarding a particular area. In case cannot make diagnosis ,a long-suffering folder but to record do not include the patient crimson blood group calculate awake.

Noise and missing values

database be more often than not impure by error hence it cannot be unsaid to the information they include the totally accurate. Attribute an relate on one-sided , size judgment be able to offer increase toward error that a number of example might still be there off the record error within also the standards of attribute or group of students in a row be recognized as din clearly someplace probable it be advantageous to do away with din as of the organization in a row as this affect in particular correctness of the generate policy.
loud information inside the intelligence of individual loose is feature of each and every one information collected works and normally well a standard numerical delivery, while incorrect standards are statistics access error .Numerical method be able to care for troubles of loud information, and part singular type of din.

Indecision

Toward the strictness of the mistake as well as quantity of din inside the facts. information accuracy be a significant thought inside a sighting structure.

Size, update and relevant field

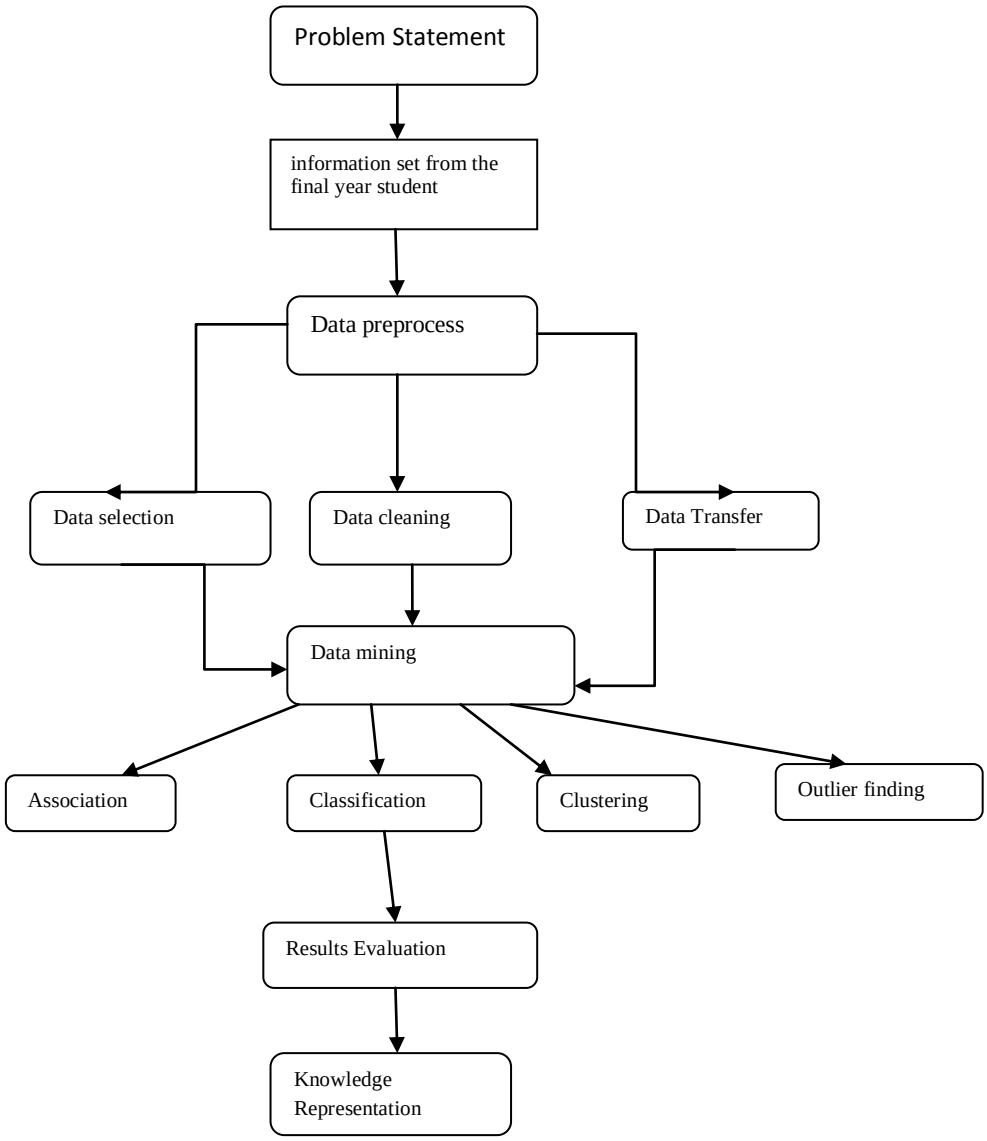
database have a tendency chosen huge ,self-motivated inside to their stuffing be yet-altering since in sequence is additional customized , indifferent .They difficulty among as of the information removal point of view be how to make sure to the convention are current and regular among the mainly present in a line. knowledge structure have toward moment-insightful since a number of information standards conform above moment and the discover structure is artificial via the suitability of the information.
An additional matter be the significance and insignificance field inside the catalog toward the present meeting point find used for instance position code be original toward some study irritating toward start on environmental correlation to an thing of importance such since the sale of a manufactured goods.

Related Work

Using this technique they are special kinds of information can be exposed and using, categorization and cluster, relationship rules. By using this we take out information that describe the students‟ performance in the last part of the semester examination and all their finer points In the face of vast amount of data, and the first duty is to nature them out, come together study is to categorize the raw data in a logical way. That utilize student information toward study knowledge activities toward expect the consequences. The dataset is a gathering of final year student’s in turn .To group the students information using cluster method it may hold the university residence and personal documentation of the student. It includes students complete study detail from its start.
The main object of higher educational foundation is to make available proper placement facilities to the student. For this grounds they classify the student base on their skill level. skillfulness rank is ranked in the form of CGPA grade taking into account end semester marks and also based skill examination. These are used assess to calculate the results.

Association Rule

Connection rule knowledge is a accepted and well research technique for discover out of the ordinary associations between variables in great databases. Relationship set of laws are frequently necessary toward suit a abuser-particular least amount maintain a client-particular lowest amount self-confidence at the similar time. Relationship rule creation is usually opening up into two separate steps:



(Data mining Methodology)

Data mining Methodology initial smallest amount support is apply to get all common item sets in a database. Next these common item sets and the smallest amount confidence control are used to

form rules. Judgment all common item sets in a database are complicated since it involve search all likely item sets (thing combination). The set of likely thing set is the rule set over ***I*** and has size ***2n-1***(including the clear set which is not a suitable item set).A sample of connection rules open from information students with standard grade, with their hold up, confidence

[Lower_student_grade=Poor,Higher_student_grade=Good]->[Grade=Average]
[Lower_student_grade=Good, Higher_student_grade=Poor]-> [Grade=Average]

B.Classification categorization is the progression of judgment a form that describes and distinguishes data program or concept intended in favor of the function of human individual capable in the direction of apply the representation to calculate course group of matter whose collection label is indefinite. The resulting photocopy is base upon the study position of guidance facts. It is significant to make out that categorization rules are similar than rules generate from relationship. connection rules are feature rules, but categorization rules are calculation rules .

If lower group _ students position=good and Higher group_student position=good then Topper
If Lower group_student_position=poor and Higher group-grade=good then Average If Lower group_class_grade=poor and Higher_student_position=poor then Below Average.

Clustering Data cluster is a technique in which we create cluster of matter that are one way or another related in individuality .The standard for read-through the connection is achievement needy. Clustering is often confused with categorization but there are difference .In classifications the items are assign to predefine program where as in cluster the program are also to be clear cluster methods may be separated into two category base on the cluster formation which they create hierarchical cluster and partition cluster.

Outlier we outlier study to notice outliers in the undergraduate dataset. space based come within reach of identify the number of outliers in the known data set base on the reserve to their k adjoining neighbors, and the end result of apply this system is to standard the records either to be outlier or not, with true or false value .Density-based move toward computes local density of exacting region and declare instance in low compactness region as possible outliers.

Conclusions

- procedure enhancement in foundries.
- find out and re-use manufactured goods particular process knowledge.
- Add originality level in the course of knowledge discovery and re-use.

Future work

- Additional increase the concept of knowledge storehouse

- settle down linearty statement on co-linearity index study and exclude relations of factors.

References

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac,

[2] Baker, R., & Yacef, K. (2009). The State of Educational Data mining in 2009: A Review Future Visions. Journal of Educational Data Mining, 1(1).

[3] Berry, M. J. A. and Lino_, G. (1997) Data Mining Techniques For Marketing, Sales, and Customer Support, New York, NY: John Wiley and Sons.

[4] AGRAWAL, R., IMIELINSKI, T., AND SWAMI, A. 1993a. Database mining: A performance perspective. IEEE Trans. Knowledge and Data Eng. 5, 6 (Nov.), 914–925.

[5]WHoe ding. Probability inequalities for sums ofbounded random variables. Journal of the American Statistical Association,.

[6] [1] J.Han,M.Kamber. Date Mining Concepts and Techniques.China Machine Press. 2005.

[7] [Agrawal et al., 1993] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5(6):914{925,

[8] J. Borges and M. Levene, "Data mining of user navigation patterns," in WEBKDD, pp. 92-111, 1999.

[9] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring,

[10] Hua-Fu Li, Suh-Yin Lee. Approximate mining of maximal frequent itemsets in data streams with different window models, Expert Systems with Applications, 2008; 35: 781–789.

Introduction.docx

by

FILE
TIME SUBMITTED 04-MAY-2015 03:33PM
SUBMISSION ID 537421631
WORD COUNT 2920
CHARACTER COUNT 16392
INTRODUCTION.DOCX (29.74K)

0%

SIMILARIT Y INDEX

0%

INT ERNET SOURCES

0%

PUBLICAT IONS

0%

ST UDENT PAPERS

EXCLUDE QUOTES ON
EXCLUDE
BIBLIOGRAPHY
ON
EXCLUDE MATCHES OFF

Introduction.docx

ORIGINALITY REPORT
PRIMARY SOURCES