ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2012,
Vol. 5, No. (1):
Pgs. 131-134

Knowledge Penetration Process: A Step by Step Approach

ANUPAM BHATIA[1] and R.K. CHAUHAN[2]

Kurukshetra University P.G. Regional Centre, Jind (India).
Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, (India).

(Received: February 12, 2012; Accepted: June 04, 2012)

ABSTRACT

In this paper we have presented the detailed model of Knowledge Penetration Process. Knowledge Penetration Process (KPP) is improved and modified KDD. The major modification in KPP over KDD is the usage of intermediate results, auxiliary data, auxiliary statistics and auxiliary tuples.

Keywords : Knowledge Penetration Process, Knowledge Discovery in Databases, Preprocessing, Pre Mining, Mining, Descriptive Data Mining, Predictive Data Mining.

INTRODUCTION

Knowledge Penetration Process

In KDD, an improving runtime is possible by tolerating the results having low quality and vice versa. The major issues involved with KDD which leads to the evolution of KDD are

(i) Redundant Action
Different instances of KDD require similar task because of isolated view of some instances.

(ii) Ignored Outcomes of Previous KDD Instances
Most of the algorithms in Data Mining consider that nothing is known about the data under consideration, whereas results of previous instances increase the knowledge and hence can improve the quality of forthcoming processes.

(iii) Ignored Intermediate Results
Data Mining algorithms rarely go for computations of intermediate results and hence cannot be used for future analysis.

To overcome the above said considerations we have defined the Knowledge Penetration Process as an improved KDD. Knowledge Penetration Process is qualitative and quicker than KDD. KPP facilitates to bring forward the computations of all items such as intermediate results and auxiliary data. The same item may be interpreted as intermediate results as well auxiliary data. Both are helpful to improve performance and quality of an analysis. These are considered separately because they have different effects in KPP. While pre-computing intermediate results save time in its succeeding instance of KPP, computing auxiliary data can improve the quality of succeeding instance of KPP. Both intermediate data

and auxiliary results are tuples fulfilling some special condition or some statistics of the data. All kind of statistical parameter are potential characteristics of the data. Therefore auxiliary statistics and auxiliary tuples are essential in the KPP process.

We have designed the five step model of Knowledge Penetration Process as shown in Fig. 1.

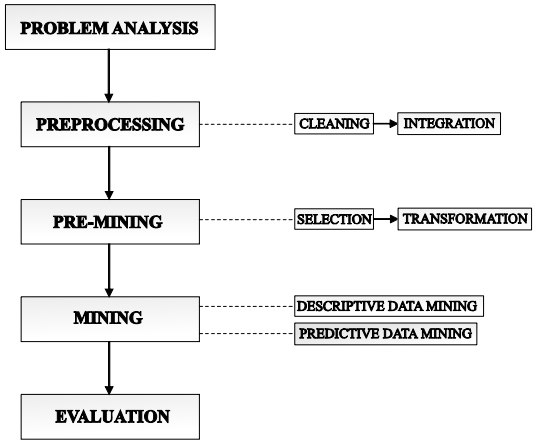

Fig. 1. Knowledge Penetration Process

Problem Analysis

It is a initial phase of Knowledge Penetration Process which does not need any computation. In this phase the problem description is studied and understood by the analyst. The analyst thoroughly study about the problem and plan the implementation of forthcoming phases of Knowledge Penetration Process e.g. what attributes are to be select, which function is to be used during the Data Mining phase etc. This step is the root of the success of the results being generated by Knowledge Penetration Process.

Preprocessing

It includes two steps also available in KDD, which are parameter independent i.e. *Cleaning* and *Integration*. Preprocessing step is directly inherited from KDD.

Pre Mining

Pre Mining phase of KPP includes Selection and Transformation both of which are parameter dependent. Selection and Transformation steps of KPP are quite different from traditional KDD.

Selection

The preprocessing phase has already processed intermediate results for all kind of potential analysis. These intermediate results are computed using all the data from fact tables. It is a critical task to join the set of tables looking at requirement of the query under consideration as well as constraints of query language. If only a fraction of data from the table is relevant then intermediate result may be irrelevant. In KPP, selection process has a commutative approach to process the intermediate results after the selection of the data. Selection of auxiliary statistics and auxiliary tuples must be taken care during the process. The basic idea of selection of auxiliary statistics is to split the data set into multiple smaller similar parts. Selection of auxiliary statistics is a complex task and depends on the selection predicate. It is essential to limit the degree of freedom in KPP queries on pre-computed intermediate results and auxiliary data. On Line Analytical Processing (OLAP) operations like Slice, Dice, Drill Down and Roll Up may be used for the purpose. If the splitted part of the data set fulfills a predicate partially or completely, then most probable value of statistics may be calculated. If a query cannot satisfy the OLAP operations then pre-computed auxiliary statistics may not be used. If a selection predicate does not contain any aggregate function and an intermediate result or auxiliary data need an analysis has the type of auxiliary tuple. The selection method of KPP tests whether an auxiliary tuple satisfies a selection predicate effectively as per constraints of query language.

Transformation

Transformation consolidates the data into appropriate forms for KPP. It becomes essential when specific scale attributes are required. Transformation in KPP involves some or all of the following.

1. Aggregation

This step may be used for construction of data cube for analysis of data at multiple

granularities, where aggregation or summary operations are applied to the data.

2. Generalization

Generalization of the data is required when low level data is needed to be replaced by higher level data.

3. Normalization

Several analysis use normalized attributes for comparison especially when a range of attributes differ in size. During the selection, when selection predicate selects subset of the data set, in majority of the cases it does not remain normalized, hence renormalization is required to receive normalized intermediate results and auxiliary data.

Mining

Mining is the core processing step of KPP. It refers to penetrate knowledge from large amount of data stored in databases, data warehouses and other data repositories. The interesting patterns are presented to the user and may be stored as new component of knowledge base. Concept of Mining step is inherited from the Data Mining step of KDD. Data Mining can be broadly categorized in two parts **i.e.** Descriptive Data Mining and Predictive Data Mining. Same is true in the case of KPP, however, the approach of implementation is quite different. Descriptive Data Mining characterizes the general properties of database e.g. Classification. Predictive Data Mining performs the inference on the current data in order to make predictions e.g. Prediction. The construction of a classifier requires some parameters for each pair of attribute value where one attribute is the class attribute and another attribute is selected by the analyst. These parameters may be used as intermediate result for constructing the classifier. Yet, the class attribute and rest all attributes that analyst considers as relevant attributes must be the attributes of the tables that might be used for analysis in future. Hence, attribute values of class attribute are always frequent. When pre-computing the frequencies of pairs of frequent attribute values, the set of computed frequencies should also include the frequencies that a potential application needs as values of the class attribute and relevant attribute are typically frequent. Splitting a fact table into several parts having similar data is necessary to select relevant data in KPP. Thereby, a selection method applies a selection predicate to each junk of data to receive the statistics of those data that fulfill the predicate. In order to use auxiliary statistics and / or auxiliary tuples in analysis of future instances of KPP, these statistics and tuples must either be applicable in any analysis or there exist an appropriate statistics or tuple for each potential analysis. Potential analysis which uses the same data set might differ in the subset of data the analyst is interested in. An analyst can perform operations on data which transform it such as aggregating tuple. Aggregating data is a potential operation of an analyst on a data set which may need special handling. Furthermore, using a selection of pre-selected auxiliary tuples as training set for a classifier is beneficial because there is no need to access the persistent storage to randomly retrieve tuples as auxiliary tuples can be stored in small cache which is fast to read and hence can reduce runtime.

i. Descriptive Data Mining

It is the step of Mining phase of KPP that characterize the general properties of the data in the databases. In our research, this step is referred as *Classification*. Databases contain hidden information that can be used for intelligent decision making. Because of the diversity of discipline contributing to the knowledge penetration, it is essential to provide a clear classification which may help potential users. Moreover DM systems can be distinguished on the granularity or levels of abstraction of the penetrated knowledge. In our research, we propose application based classification model because different applications often require the integration of application specific methods.

ii. Predictive Data Mining

It is the step of Mining phase of KPP that performs inference on the current data in order to make predictions. In our research, this step is referred as *Prediction*.

Evaluation

In our research to evaluate and test the quality of results, two major tools are used.

i) Confusion Matrix

ii) ROC Curve

REFERENCES

1. Ester, Martin. Sander, Joerg. "Knowledge Discovery in Databases" .*Springer Verlag*, Heidelberg. (2000).
2. Micheline Kamber Jiawei Han. "Data Mining: Concepts and Techniques" .*Morgan Kaufmann* (2006).
3. Piatetsky, Gregory. "Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics". *Data Min Knowl Disc (2007)* **15:** 99–105 (2007).
4. Orfanidis, Paraskevas. Russomanno, David J. "Preprocessing enhancements to improve data mining algorithms". (2008).
5. Chauhan, R K. Bhatia, Anupam. " KPP : A Step Ahead to KDD". *RIMT Journal of Strategic Management and Information Technology,* **5**(1-4): (2008).
6. Bhatia, Anupam. Chauhan RK. "Knowledge Penetration Process : A Splitted KDD", *Global Journal of Computer Science and Technology*; **11**, (6): (2011).
7. Bhatia, Anupam. Chauhan RK. "Preprocessing in Knowledge Penetration Process". *Emerging Trends in I.T.*, Vol. 1. (2012).