Applying the Knowledge Discovery in Databases (KDD) Process to Fermilab Accelerator Machine Data

K. Yacoben and L. Carmichael
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia Illinois 60510, USA

Abstract

During day to day operations Fermilab collects a substantial amount of accelerator machine data. This repository of data provides an ideal basis for applying the knowledge discovery in databases (KDD) process. Knowledge discovery in databases is a new and emerging field that defines a set of techniques and tools for decision support and data analysis [1]. This paper describes our approach for applying the KDD process to accelerator machine data in order to improve machine operation, performance and understanding. In the initial sections of this paper we will describe our motivations and goals, along with a brief description of what the KDD process is. In the last section we will describe our steps in the application of the KDD process. This description will be in the context of a real accelerator controls application shot data analysis (SDA).

1 Introduction

A unique characteristic of the Fermilab accelerator control system is its ability to capture accelerator data of various types in vast amounts. During the 1997-98 accelerator shutdown an effort is planned to move a majority of these data capture facilities from operations specific indexed data storage to commercial databases. This will further enhance Fermilab's control system by adding new dimensions to the potential use of all stored data. These conditions provide an ideal basis for a new field of decision support and data analysis called knowledge discovery in databases (KDD). The KDD process is a new generation of computational methods and techniques designed to facilitate the extraction of useful knowledge from large repositories of data [1].

The motivation for applying the KDD process is found in the following issues. First, to take advantage of the wealth of information in captured accelerator data advanced analysis tools need to be developed. Also through the use of these tools' users will be able to improve their understanding of machine data and thus lead to more in-depth analysis techniques. The second motivational issue deals with helping operations staff meet the demands of the next collider run. With retirement of the "Main Ring" and the addition of the "Main Injector" the operations staff will have to overcome multiple obstacles. Not only are they operating a new accelerator but also working with increased performance requirements and new operational scenarios. Lastly we are motivated by the observations of certain deficiencies in the accelerator controls system that were highlighted in the previous collider run. Our goal in this work is to provide a facility that implement the core features of the KDD process and demonstrates future potential. Initially the targeted users for this facility are the operator and application developer community.

2 Knowledge discovery in databases

It has been shown that the amount of information stored in databases is dramatically increasing everyday [2]. On that same token the tools and facilities used to understand and analyze this data has not increased at the same rate [2]. This is further compounded by the realization that efficient and intelligent analysis of this data can be a valuable asset to any process. To address these issues the computer science community has created a new field called Knowledge discovery in databases.

A formal definition of Knowledge discovery in databases is the non-trivial process of identifying valid, previously unknown and potentially useful information from large datasets [3]. Knowledge discovery in databases uses methods and techniques that are derived from the areas of statistical and data analysis, decision support and machine learning. It differs from the traditional analytical methods since it overcomes some of the obstacles posed by large datasets. In traditional data analysis knowledge extraction is performed by multiple analysts who are very familiar with a given dataset. Through the use of statistical techniques these analysts manually probe the data searching for useful and interesting facts. However, as the dataset increases in size and dimensional complexity the amount of effort required by these traditional methods exceeds human capabilities. It should be noted that KDD is not a replacement for traditional methods, but is an augmentation to them.

2.1 KDD process overview

The KDD process is an iterative multi-step process that is highly interactive and closely tied to human interactions. Figure 1 represents the basic components of the KDD process. The input for the KDD process draws data from three major datasets: raw data, meta and aggregate data and domain knowledge. At the heart of the process is the knowledge discovery support environment (KDSE). This is the environment through which the knowledge discovery will happen. The last section of the KDD process is the knowledge discovery application (KDA). The KDA plays a key role in the KDD process since it is through these applications that the output of KDD uses, validates and periodically feeds

knowledge back into the KDD process. Connecting each step in the KDD process is a bi-directional transition. These transition points allow processing to move forward or turn back and return to an earlier stage. This dynamic nature allows the process to perform self tuning. Throughout the KDD process the user is intimately involved and it is this involvement that forms the backbone of the process.

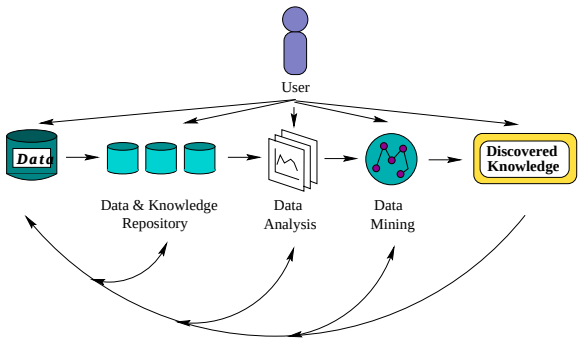

Figure. 1. KDD Process

2.2 Data & knowledge repository

The basic input to the KDD process is raw data. However, by itself the raw data has very limited potential and most often is useless. The problems with the raw data used in KDD stem from the fact that the dataset is large, dynamic and some times incomplete. To be of use the raw data must be processed and placed within a framework that supports KDD. The processing performed on the raw data includes selection, cleaning, transformation and data reduction.

A typical framework used with KDD is the data warehouse. Data warehouses are databases designed to meet the needs of decision support and online data analysis.

Another input to the KDD process is domain knowledge. Domain knowledge is knowledge about the task domain. It is composed of meta-data (facts) and a set of rules (or some other representation) that use the data as basis for decision making. Meta-data are derived values from various steps in the KDD process. In the early steps of the KDD process this information is used to help drive and guide the raw data processing.

2.3 Knowledge discovery support environment (KDSE)

At the core of the KDD process is the Knowledge Discovery Support Environment (KDSE). The KDSE is an integrated environment of analysis tools that empower to perform KDD. It provides a framework to support the dynamic nature of the KDD process. The KDSE enables the user to enter/leave any step in the KDD process, refine datasets, propose complex queries. The base functionality of the KDSE includes integrated query tools, query reuse tools, graphical display tools, shared objects and a knowledge representation scheme. The KDSE empowers the user to perform data analysis in their own way. Further more it captures the analysis process which enables reuse and refinement.

A critical function within the KDD process is the ability to capture and represent domain knowledge. How domain knowledge is stored has a dramatic effect on its usability. The KDSE supports knowledge capture and representation through a set of tools that are tailored toward the knowledge representation schema. By using these tools the analyst can add, review, and modify the domain knowledge. The types of representation scheme used in KDD include production rules, decision trees and neural nets.

The techniques provided by the KDSE for knowledge extraction is characterized as Data mining. The term Data mining is a popular topic and according to some authors is equivalent to KDD. However, data mining only incorporates the set of algorithms and techniques needed to explore large datasets. It does not include higher level algorithms, like algorithms which reason on the merits of other data mining algorithms and determine which functions to use. The algorithms used in data mining include clustering, association, classification, sequence analysis (trending) and prediction.

2.4 Knowledge discovery application (KDA)

At the end of the KDD process is the knowledge discovery application (KDA). The KDAs are applications that work with the results of the KDD process. There are two classes of KDAs, ones which work with discovered information and ones who incorporate some of the functionality of the KDD process. Overall KDAs are an integral part of the KDD process by providing validation and insight into the process.

3 Application for the KDD process

In this section we present the application of the KDD process to Fermilab machine data. The initial target application of KDD is the shot data analysis(SDA) to be done during Collider Run II (CRII). Shot data analysis is the process performed by a run coordinator after a shot to determine the optimal accelerator configuration for the next shot. It requires the run coordinator to extract meaningful and relevant information from various sources, which can include very large datasets. Further more, all of this activity must be performed in an efficient and timely manner.

Following the stepwise and iterative nature of the KDD process we are applying KDD in multiple phases, with each phase building on the work of previous ones. In the first phase we are focusing on the Data component of the KDD process. Initially all work is concentrated on the creation of a data warehouse and warehouse support environment.

3.1 Data warehouse

The initial phase of the KDD process involves the creation of a data warehouse as the repository for SDA machine data. The design of the initial data warehouse is based on three objectives, it must support the current shot data analysis facility, incorporate only SDA data and the design is flexible enough for later expansion.

Fermilab's machine data generally falls into one of two categories: physical hardware measurements made at Front-

Ends (F/E) and computed data generated by a set of tasks termed Open-Access Frontends (OAFs). The warehouse currently resides in a set of Sybase tables located on a PC. Population of these table is performed by a set of central tasks, Shot Data Acquisition, Save-Restore (SR) and LUMBERJACK, which define the logging facilities at Fermilab. These tasks issue a set of at a requests to DPMs (Data Pool Managers) which merge similar requests into a list of unique requests which are then sent to the appropriate F/Es and OAFs. The path that this data takes to reach the warehouse is illustrated in figure 2. Also shown in the diagram are the OAM attributes which will be discussed in the next section.

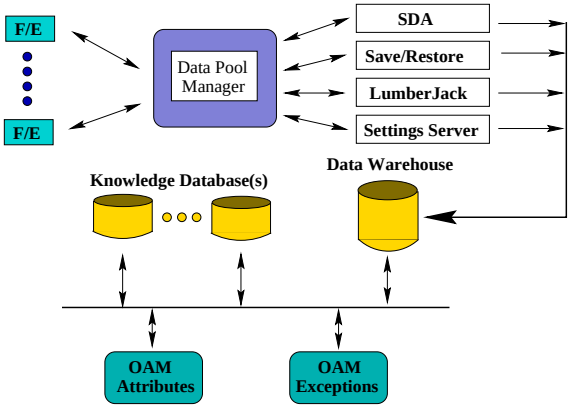

Figure. 2. Data flow

3.2 Data attributes

One of the crucial issues facing many scientific facilities is not the gathering of data, but how to efficiently use the acquired data. This issue becomes even more critical as data growth rapidly outpaces the growth of sophisticated data analysis tools. Fermilab currently provides several tools that are used to analyze data. Unfortunately, these tools are generally restricted to data viewing and are strongly tied to non database storage mediums, such as filesharing. Part of the focus in implementing this stage of the KDD process is in modifying the available viewing facilities to access warehouse data. Additionally, it would be desirable to have more sophisticated methods of analyzing the data. One such method, currently under development, involves the design of a facility that allows for the creating and viewing of data attributes. Data attributes can be defined as tasks embedded in a Finite-State Machine formulation where each state is represented by a set of transitions and a set of arbitrary functions of warehouse data. Figure 3. illustrates the GUI, currently under development, that is used to create and visualize the data attributes.

These data attributes actually construe an abstraction of warehouse data and can be visualized as a layer on top of the warehouse data. This layer is variable, as opposed to fixed nature of the warehouse data, and represents more of a Knowledge-Base. This designation stems from the realization that data attributes can be continuously modified as knowledge about your system and data grows.

Figure. 3. Data attributes definition interface

4 Conclusion

Future iterations of the KDD process will include the expansion of the data warehouse to include additional sources of data and the development of additional tools for analysis and extraction of pertinent information from the data. Ultimately this will lead to the construction of a mechanism for the automation of knowledge discovery.

References

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications Of The ACM* 39(11): 27-28.

[2] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining to Knowledge Discovery. In *Advances In Knowledge Discovery and Data Mining*, ed. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P, and Uthurusamy, R. cambridge, Mass: AAAI/MIT Press, 1-31.

[3] Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. 1992. Knowledge Discovery in Databases: An Overview. *AI Magazine* :57-70, Fall.

[4] Brachman, R., and Anand, T. The Process of Knowledge Discovery in Databases. In *Advances In Knowledge Discovery and Data Mining*, ed. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P, and Uthurusamy, R. cambridge, Mass: AAAI/MIT Press, 37-58.

[5] Han, J,, and Fu, Y. Attribute-Oriented Induction in Data Mining. In *Advances In Knowledge Discovery and Data Mining*, ed. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P, and Uthurusamy, R. cambridge, Mass: AAAI/MIT Press, 339-421.