International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Research Article on Data Mining in Addition to Process Mining: Similarities and Dissimilarities

S. Sowjanya Chintalapati[1], Ch.G.V.N.Prasad[2], J. Sowjanya[3], R.Vineela[4]

[1, 3, 4] Assistant Professor, Department of CSE, Sri Indu College of Engineering and Technology, Ibrahimpatnam, Hyderabad, India

[2] Professor, Head of the Department (CSE), Sri Indu College of Engineering and Technology, Ibrahimpatnam, Hyderabad, India

Abstract: ***Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. On the other hand Process mining is a process management technique that allows for the analysis of business processes based on event logs. The basic idea is to extract knowledge from event logs recorded by an information system. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs. This paper mainly supports a "missing link" between data mining and traditional model-driven BPM. Its primary objective is the discovery of process models based on available event log data. The discovered process models can be used for a variety of analysis purposes. This article provides an introduction to process mining. It addresses basic concepts necessary to understand and apply process mining.***

Keywords: Data mining, Analyzing data, Knowledge discovery, Business intelligence (BI), process mining, event log data.

1. Introduction

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in databases. [1]

In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. [4].

Primarily, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed.

The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data. [3]

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process. [3]

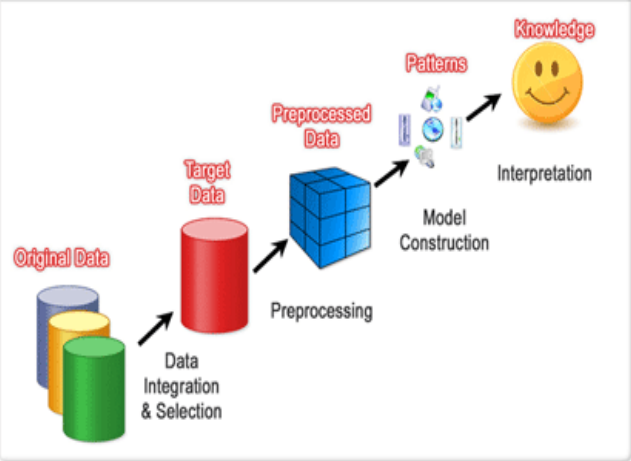

Figure 1: An iterative knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps: [15]

1) **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2) **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3) **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4) **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5) **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6) **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
7) **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. [2]

Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. [16] The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.[3]

Unlike Data mining, process mining focuses on the process perspective: It includes the temporal aspect and looks at a single process execution as a sequence of activities that have been performed. Most data mining techniques extract abstract patterns in the form of, for example, rules or decision trees. In contrast, process mining creates complete process models, and then uses them to precisely highlight where the bottlenecks are.[13] Both Data Mining and Process Mining goes under the concept, called Business Intelligence.

Business intelligence refers to techniques and tools that are used to analyze large amounts of digital data and retrieve valuable business knowledge out of them. And that is true for data mining techniques as well as process mining techniques - albeit with different perspective on the analysis and the results they produce. [14]

2. Similarities and Differences between Data Mining and Process Mining

2.1 Similarities

1. Both techniques are used to analyze large amounts of data that it would be impossible to analyze manually.
2. Both techniques produce information that can be used for making business decisions.
3. Both techniques use the "mining" techniques where algorithms traverse through large volumes of data, looking for patterns and relationships.

Of course there are some similarities, as both techniques can be categorized as Business Intelligence. But, as mentioned before, the two techniques have different perspectives and goals.[17]

2.2 Differences between Data mining and process mining

Data mining techniques are using multi-dimensional views (cubes) on data which can be drilled up and down (in different aggregated levels). For example, a sale of a product could have the related dimensions: region, month, measure, product, version and so on and it is then possible to slice and look at the cube of data and aggregated data in various ways .A multi-dimensional cube is diagrammatically represented as shown in the figure 2 below:

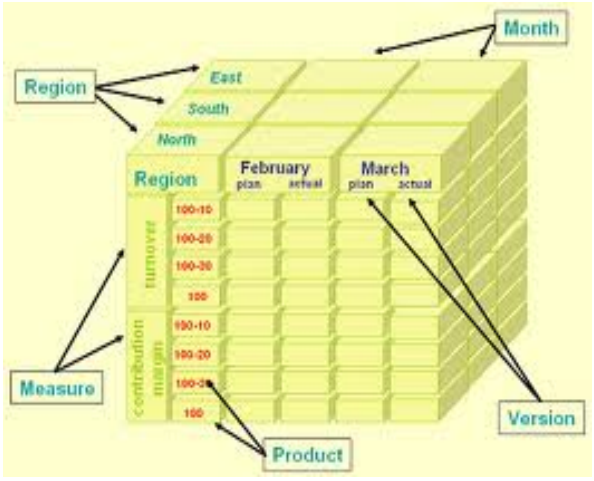

Figure 2: An Example for sale of a product with its related dimensions

1) Data mining techniques are primarily used to find patterns in large data sets. With data mining techniques it may be possible to find that certain categories of customers demand a certain product, or to find that the customers who most frequently buy product A are also the ones most often buying product B, or that the products placed on a specific location in the shop while running an advertising campaign, are also the ones that sell the best. [17]
   For Example a store which, through data mining techniques, found out that the customers who shopped the most was also that most often buying at special Italian cheese that otherwise was not often sold. Traditionally retailers would try to remove products with very low turnover rates and replace them with products [5] with better sales - the problem is that the removal of goods according to the principle could lead to the best customers having to look somewhere else (for the special Italian cheese).
2) The inputs to data mining are tables with data.
3) Process mining is not used to find relationship data patterns, but rather to find process relationships in the data. Finding process relationships that provide an overview of processes and activities in the process, and deviations and process performance such as throughput, bottlenecks and discrepancies refer figure 3.

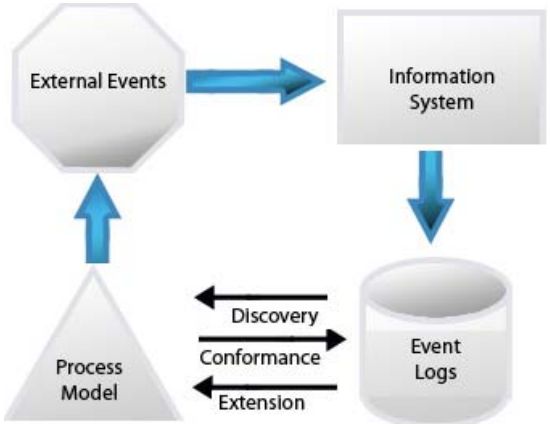

Figure 3: Process Mining

4) Process mining is the use of data mining tools for the purpose of information extraction from event logs. e.g. the audit trails of a workflow management system or the transaction logs of an enterprise resource planning system. These can be used to identify process models, products, informal organizations of business units, and even monitor deviations. [10]
5) Process mining's perspective is not on patterns in the data but in the processes the data represents.
6) The goal of process mining is to find information about the business processes.
7) The inputs to the process mining analysis are event logs, audit trails, and data and events stamped in the IT systems.

3. Process Mining Technology

There are lots of data mining tools that are used to support business decisions in specific areas (for example: which products should be placed together in the supermarket, or: where you should send your marketing flyer), but they do not work well for processes.[12]

At the same time, organizations spend lots of money on modeling processes. Because the process modeling is done manually, these models are quickly becoming outdated and out of touch with reality — and so they often they end up as dead piles of paper that have no value.[6]

Process mining technology combines the strengths of both data mining and process modeling: By automatically creating process models based on existing IT log data, process mining yields live models that are connected to the business and can be updated easily at any point in time. Refer figure 4 Process mining has more in common with data mining than just the "mining" part: Just like data mining, process mining takes on the challenge to process large volumes of data that simply [5] cannot be evaluated by hand anymore. Enterprise IT systems collect more and more data about the business processes they support. These data usually reflect very closely what happened in "the real world" and can be a great source of insight for understanding and improving the business.

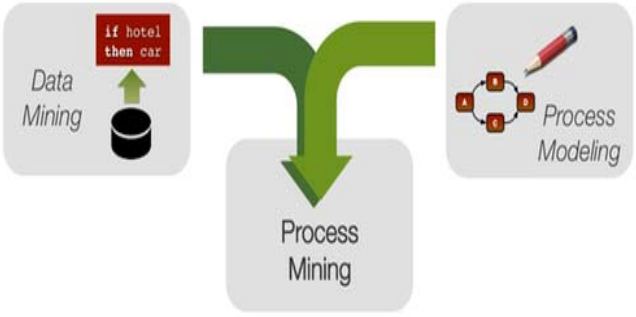

Figure 4: Process Mining Technology

Unlike data mining, process mining focuses on the process perspective: It includes the temporal aspect and looks at a single process execution as a sequence of activities that have been performed. Most data mining techniques extract abstract patterns in the form of, for example, rules or decision trees. In contrast, process mining creates complete process models, and then uses them to precisely highlight where the bottlenecks are.[8]

In data mining, generalization is very important to avoid what is called "over fitting the data". [9,11]This means that one wants to strip away all the examples that do not match the general rule. In process mining, generalization is also necessary to deal with complex processes and understand the main process flows. However, understanding the exceptions is often important to discover inefficiencies and points of improvement. In data mining, models are often trained to make predictions about future similar instances in the same space. Quite a few data mining and machine learning methods operate as a "black box" that spills out predictions without the possibility to trace back the "why".[7]

Because today's business processes are so complex, accurate predictions are often unrealistic. The gained knowledge and deeper insights from the discovered patterns and processes

help to deal with the complexity, which is where the true value is.[5]

4. Process Mining bridging data mining and big data (Event data) and business process management

Process mining builds the bridge between data mining as a business intelligence approach and business process management.

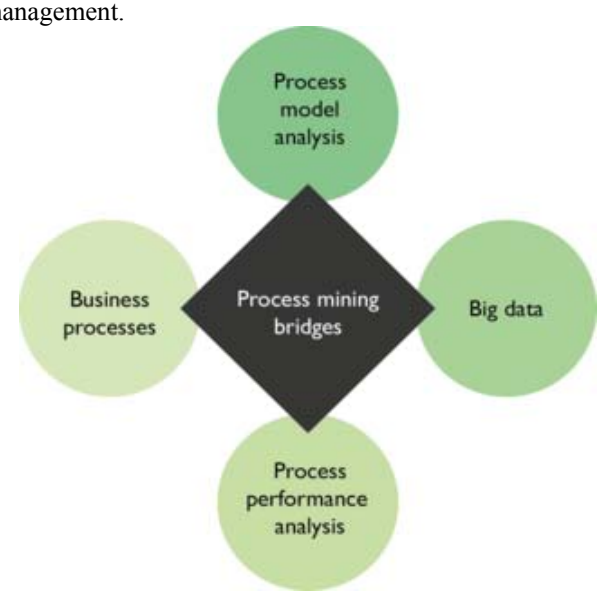

Figure 5: Process mining as a link between data mining and traditional BPM.

Process mining is the "missing link" between data mining and traditional BPM (Business Process Management). Refer figure 5. Data mining provides valuable insights through analysis of data, but is generally not concerned about processes. This is where process mining comes into the picture and gives the opportunity to get the same benefits of data mining, when working with processes and process improvements. [17]. Process mapping can be done with mining techniques instead of brown -paper workshops and interviews. And the process performance analysis can be made on existing data mining techniques without first collecting data through work studies.

5. Conclusion

Now a days Business processes become more and more complex and information systems support or even automate the execution of business transactions in modern companies. Business intelligence aims to support and improve decision making processes by providing methods and tools for analyzing data. Business Intelligence (BI) and Process Mining (PM), is presented a framework for improving the decision-making processes in organizations. Process mining builds the bridge between data mining as a business intelligence approach and business process management. Its primary objective is the discovery of process models based on available event log data.

6. Future Scope

Process mining offers auditors a new and powerful tool to perform analytic procedures in many Business applications. Researchers have explored many different analytic procedure tools, ranging from simple ratio analysis to continuity equations (as a means of modeling business processes and cluster analysis. Process mining does not replace these techniques, but rather, provides way of refining their results [18].

References

[1] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012, Effective Pattern Discovery for Text Mining Ning Zhong, Yuefeng Li, and Sheng-Tang Wu.
[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.
[3] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/
[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
[5] http://fluxicon.com/blog/2011/02/how-process-mining-compares-to-data-mining/
[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word-Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.
[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazione dell'Informazione, 2000.
[8] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
[10] http://www.infotech.monash.edu.au/research/about/centres/ccsl/solutions/process-mining. html.
[11] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
[12] Van der Aalst, W. M. P. 2011. Process Mining: Discovery, Conformance and Enhancement of Business Processes, (1st Edition) Berlin; Heidelberg: Springer.
[13] Van der Aalst, W. M. P., Andriansyah, A., Alves de Medeiros, A. K., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., Van den Brand, P., Brandtjen, R., and Buijs, J. 2012. "Process mining manifesto," In BPM 2011 Workshops Proceedings, pp. 169–194.
[14] Van der Aalst, W. M. P., Weijters, A. J. M. M., and Maruster, L. 2002. "Workflow Mining: Which Processes can be Rediscovered," (Vol. 2480) Presented at the Proc. Int'l Conf. Eng. and Deployment of Cooperative Information Systems (EDCIS 2002), pp. 45–63.
[15] Adriansyah, A., Van Dongen, B. F., and Van der Aalst, W. M. P. 2011. "Conformance checking using cost-based fitness analysis," In Enterprise Distributed Object

Computing Conference (EDOC), 2011 15th IEEE International, pp. 55–64.
[16] C.W. Gunther and W.M.P. van der Aalst. Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics. In G.Alonso, P. Dadam, and M. Rosemann, editors, International Conference on Business Process Management (BPM 2007), volume 4714 of Lecture Notes in Computer Science, pages 328-343.Springer-Verlag, Berlin, 2007.
[17] http://www.allaboutrequirements.com/
[18] PROCESS MINING: RESEARCH OPPORTUNITIES IN AIS by Michael Alles Rutgers Business School Newark, NJ, USA alles@business.rutgers.edu.Mieke Jans, PhD, Hasselt University, Belgium Mieke.jans@uhasselt.be.Miklos VasarhelyiRutgers Business School Newark, NJ, USA,miklosv@andromeda.rutgers.edu,Sep 12, 2010.

Author Profile

Sarada Sowjanya .C completed her Mtech from JNTU Kakinada with Distinction. She did her BE from RTM NAGPUR UNIVERSITY, Currently working as Asst. Professor in CSE dept, Sri Indu College of Engg & Technology her research interests include: Data Mining and Data Base Management Systems.

Dr. Ch G.V.N. Prasad currently working as a Professor and HOD of CSE department, Sri Indu College of Engg & Technology. He gained 12 years of experience in IT industry (8 years in National Informatics Centre, Govt. of India, as Scientist and Software Analyst in AT&T in US) and 11 years of experience in Teaching (As a Professor & HOD of CSE Dept).

J. Sowjanya completed her Mtech from SIT, JNTUH with Distinction. She did her BE from Muffakham-Jah College of engineering and Technology, OU. Currently working as a Asst. Professor in CSE dept, Sri Indu College of Engg & Technology. Her areas of interest are: Data Mining, Web Technologies

R. Vineela Completed her Mtech from JNTUH with Distinction. She completed her MCA from St.John's PG College OU. Currently working as a Asst. Professor in CSE dept, Sri Indu College of Engg & Technology. Her areas of Interest are: Data Mining, Web Designing and Database Management System.