IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 2, April 2014.
www.ijiset.com
ISSN 2348 - 7968

Review on Data Mining Tools

Heena Agrawal[1], Pratik Agrawal[2]

**[1] Lecture/ComputerTechnology Department, RTMNU/Rajiv Gandhi College of Engineering and Research,
Nagpur, Maharashtra, India**

**[2] Assistant Professor/Computer Science and Engineering Department, PRMIT&R,
Badnera, Maharashtra, India**

Abstract

Data mining is one of the most important steps of the knowledge discovery in databases process and is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization over coming decades. This review paper explores the data mining tools which have been developed to support knowledge management process.Data mining has become an essential factor in various fields including business, education, health care, finance, scientific etc because of the large amount of the data. To analyse this vast amount of data and depict the fruitful conclusions and inferences, it needs specific data mining tools. This paper discusses the knowledge discovery process, data mining, various open source tools in the field of data mining from past to the present and explores the future trends.

Keywords: *Data mining; Data mining tools; Knowledge management*

1. Introduction

The Data mining has a long history, with strong roots in statistics, artificial intelligence, machine learning, and database research. Data mining is a step in the knowledge discovery from databases (KDD) process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) across the data.

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data.

The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge.

KDD process consists of iterative sequence methods as follows:
1. Selection: Selecting data relevant to the analysis task from the database
2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources
3. Transformation: Transforming data into appropriate forms to perform data mining
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns
5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating the useful patterns into terms that human understandable.

1.1 CATEGORIES OF DATA MINING TOOLS

Most of the data mining tools can be classified into three categories: Traditional data mining tools, dash boards and text-mining tools. Description of each is as follows [11]:

1.1.1. Traditional Data Mining Tools

Traditional mining programs help the companies to establish data patterns and trends by using various complex algorithms and techniques. Some of these tools are installed on the desktop computers to monitor the data and emphasize trends and others capture information residing outside a data base. Majority of these programs are supported by windows and UNIX versions. However, some software specializes in one operating system only. In addition to that some may work in only one database type. But, Most of the software will be able to handle any data using online analytical processing or a similar technology.

1.1.2. Dashboards

Dashboards reflect data changed and update on screen. Dashboards is normally installed in computers to monitor information in a database and it reflects data changes and updates the data in the form of a chart or table on the screen. It enables the user to see how the business is performing. Historical data can be referenced and checks against the current status in order to see the changes in the business. By this way, dashboards is very easy to use and helps the manager a lot with great appeal to have an overview of the company's performance.

1.1.3. Text-Mining Tools

The third type of data mining tools is called as a text-mining tool because of its ability to mine data from different kind of text starting from Microsoft Word, Acrobat PDF documents to simple text files. This provides facility of scanning the content and converts the selected into a format that is compatible with the tools database without opening different applications.

The paper is organised as follows I) first part contains the introduction about the data mining domain ii) second part contains the literature review about data mining tools that are available and their comparison III) third part describes the conclusion derived from the overall literature reviews about the tools.

2. Literature review

Authors Ralf Mikut and Markus Reischl In their paper "Data mining tools" proposed nine types of tools:

• Data mining suites (DMS) focus largely on data mining and include numerous methods. They support feature tables and time series, while additional tools for text mining are sometime available. The application focus is wide and not restricted to a special application field, such as business applications; however, coupling to business solutions, import and export of models, reporting, and a variety of different platforms are nonetheless supported. In addition, the producers provide services for adaptation of the tools to the workflows and data structures of the customer. DMS is mostly commercial and rather expensive, but some open-source tools such as RapidMiner exist. Typical examples include IBM SPSS Modeler, SAS Enterprise Miner, Alice d'Isoft, DataEngine, DataDetective, GhostMiner, Knowledge Studio, KXEN, NAG Data Mining Components, Partek Discovery Suite, STATISTICA, and TIBCO
Spotfire.

• Business intelligence packages (BIs) have no special focus to data mining, but include basic data mining functionality, especially for statistical methods in business applications. BIs are often restricted to feature tables and time series, large feature tables are supported. They have a highly developed reporting functionality and good support for education, handling, and adaptation to the workflows of the customer. They are characterized by a strong focus on database coupling, and are implemented via a client/server architecture. Most
BI softwares are commercial (IBM Cognos 8 BI,Oracle DataMining, SAPNetweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM, and PolyVista),
but a few open-source solutions exist (Pentaho).

• Mathematical packages (MATs) have no special focus on data mining, but provide a large and extendable set of algorithms and visualization routines. They support feature tables, time series, and have at least import formats for images. The user interaction often requires programming skills in a scripting language. MATs are attractive to users in algorithm development and applied research because data mining algorithms can be rapidly implemented, mostly in the form of extensions (EXT) and research prototypes (RES). MAT packages exist as commercial (MATLAB and R-PLUS) or open-source tools (R, Kepler). In principle, table calculation software such as Excel may also be categorized here, but it is not included in this paper. Most tools are available for different platforms but have weaknesses in database coupling.

• Integration packages (INTs) are extendable bundles of many different open-source algorithms, either as stand-alone software (mostly based on Java; as KNIME, the GUI-version of WEKA, KEEL, and TANAGRA) or as a kind of larger extension package for tools from the MAT type (such as Gait-CAD, PRTools for MATLAB, and RWEKA for R). Import and export support standard formats, but database support is quite weak. Most tools are available for different platforms and include a GUI. Mixtures of license models occur if open-source integration packages are based on commercial tools from the MAT type. With these characteristics, the tools are attractive to algorithm developers and users in applied research due to expandability and rapid comparison with alternative tools, and due to easy integration of application-specific methods and import options.

• EXT are smaller add-ons for other tools such as Excel, Matlab, R, and so forth, with limited but quite useful functionality. Here, only a few data mining algorithms are implemented such as artificial neural networks for Excel (Forecaster XL and XLMiner) or MATLAB (Matlab Neural Networks Toolbox). There are commercial or open-source versions, but licenses for the basic tools must also be available. The user interaction is the same as for the basic tool, for example, by using a programming language (MATLAB) or by embedding the extension in the menu (Excel).

• Data mining libraries (LIBs) implement data mining methods as a bundle of functions. These functions can be embedded in other software tools using an Application Programming Interface (API) for the interaction between the software tool and the data mining functions. A graphical user interface is missing, but some functions can support the integration of specific visualization tools. They are often written in JAVA or C++ and the solutions are platform independent. Open source examples are WEKA (Java-based), MLC++ (C++ based), JAVA Data Mining Package, and LibSVM (C++ and JAVAbased) for support vector machines. A commercial example is Neurofusion for C++, whereas XELOPES (Java, C++, and C_) uses different license models. LIB tools are mainly attractive to users in algorithm development and applied research, for embedding data mining software into larger data mining software tools or specific solutions for narrow applications.

• Specialties (SPECs) are similar to DMS tools, but implement only one special family of methods such as artificial neural networks. They contain many elaborate visualization techniques for such methods. SPECs are rather simple to handle as compared with other tools, which eases the use of such tools in education. Examples are CART for decision trees, Bayesia Lab for Bayesian networks, C5.0, WizRule, Rule Discovery System for rule-based systems, MagnumOpus for association analysis, and JavaNNS, Neuroshell,

• RES are usually the first—and not always stable—implementations of new and innovative algorithms. They contain only one or a few algorithms with restricted graphical support and without automation support. Import and export functionality is rather restricted and database coupling is missing or weak. RES tools are mostly opensource. They are mainly attractive to users in algorithm development and applied research, specifically in very innovative fields. Examples are GIFT for content-based image retrieval, Himalaya for mining maximal frequent item sets, sequential pattern mining and scalable linear regression trees, Rseslibs for rough sets, and Pegasus for graph mining. Early versions of today’s popular tools such as WEKA and RapidMiner started in this category and shifted later to other categories as DMS.

• Solutions (SOLs) describe a group of tools that are customized to narrow application fields such as text mining (GATE), image WIREs Data Mining and Knowledge Discovery Data mining tools processing (ITK, ImageJ), drug discovery (Molegro Data Modeler), image analysis in microscopy (CellProfilerAnalyst), or mining gene expression profiles (Partek Genomics Suite, MEGA). The advantage of these solutions is the excellent support of domainspecific feature extraction techniques, evaluation measures, visualizations, and import formats. The level of data mining methods ranges from rather weak support (particularly in image processing) to highly developed algorithms. In some cases, more general tools from types DMS or INT also support specific domains (KNIME, Gait-CAD for peptide chemoinformatics). There are many commercial and open-source solutions. A large variety of tools actually requires a fuzzy categorization with gradual memberships to different types. Examples are tools including a set of different algorithms (LIB) with an additional GUI acting as an INT, DMS, including special methods for narrow application fields and others. In these cases, a main type was assigned and the other fuzzy memberships are discussed in the Excel table in the additional material section.

Many advanced tools for data mining are available either as open-source or commercial software. They cover a wide range of software products, from comfortable problem-independent data mining suites, to business-centered data warehouses with integrated data mining capabilities, to early research prototypes for newly developed methods. In this paper, nine different types of tools are presented: DMS, BIs, MATs, INT, EXT, SPECs, RES, LIBs, and SOLs. They vary in many different characteristics, such as intended user groups, possible data structures, implemented tasks and methods, interaction styles, import and export capabilities, platforms and license policies are variable. Recent tools are able to handle large datasets with single features, time series, and even unstructured data-like texts; however, there is a lack of powerful and generalized mining tools for multidimensional datasets such as images and videos.

Authors Agathe MERCERON* and Kalina YACEF in their paper "Educational Data Mining: a Case Study" used a range of tools. Initially the author worked with Excel and Access to perform simple SQL queries and visualisation. Then they used Clementine for clustering and our own data mining platform for teachers, Tada-Ed for clustering, classification and association rule (Clementine is very versatile and powerful but Tada-Ed has preprocessing facilities and visualisation of results more tailored to our needs). The author used SODAS to perform symbolic data analysis. Consider an example of student data. In SODAS, the population is partitioned into sets called symbolic objects. Our symbolic objects were defined by the number of attempted exercises and were characterized by the values taken for these newly calculated variables: the number of successfully completed exercises, the average number of correct steps per attempted exercise, the average number of mistakes per attempted exercise. The author obtained a number of tables to compare all these objects.

In paper Data Cleaning: Problems and Current Approaches, Erhard Rahm and _ Hong Hai DoA explained that a large variety of tools is available on the market to support data transformation and data cleaning tasks, in particular for data warehousing. Some tools concentrate on a specific domain, such as cleaning name and address data, or a specific cleaning phase, such as data analysis or duplicate elimination. Due to their restricted domain, specialized tools typically perform very well but must be complemented by other tools to address the broad spectrum of transformation and cleaning problems. Other tools, e.g., ETL tools, provide comprehensive transformation and workflow capabilities to cover a large part of the data transformation and cleaning process.

A general problem of ETL tools is their limited interoperability due to proprietary application programming interfaces (API) and proprietary metadata formats making it difficult to combine the functionality of several tools [8].

The author discusses tools for data analysis and data reengineering which process instance data to identify data errors and inconsistencies, and to derive corresponding cleaning transformations.

Data analysis and reengineering tools

MIGRATIONARCHITECT (EvokeSoftware) is one of the few commercial data profiling tools. For each attribute, it determines the following real metadata: data type, length, cardinality, discrete values and their percentage, minimum and maximum values, missing values, and uniqueness. MIGRATIONARCHITECT also assists in developing the target schema for data migration.

Data mining tools, such as WIZRULE (WizSoft) and DATAMININGSUITE (InformationDiscovery), infer relationships among attributes and their values and compute a confidence rate indicating the number of qualifying rows. In particular, WIZRULE can reveal three kinds of rules: mathematical formula, if-then rules, and spelling-based rules indicating misspelled names, e.g., "value Edinburgh appears 52 times in field Customer; 2 case(s) contain similar value(s)". WIZRULE also automatically points to the deviations from the set of the discovered rules as suspected errors.

Data reengineering tools, e.g., INTEGRITY (Vality), utilize discovered patterns and rules to specify and perform cleaning transformations, i.e., they reengineer legacy data. In INTEGRITY, data instances undergo several analysis steps, such as parsing, data typing, pattern and frequency analysis. The result of these steps is a tabular representation of field contents, their patterns and frequencies, based on which the pattern for standardizing data can be selected. For specifying cleaning transformations, INTEGRITY provides a language including a set of operators for column transformations (e.g., move, split, delete) and row transformation (e.g., merge, split). INTEGRITY identifies and consolidates records using a statistical matching technique. Automated weighting factors are used to compute scores for ranking matches based on which the user can select the real duplicates.

4. Conclusions

In this paper the overall literature survey related to different data mining tool for different applications are mentioned. It is observed data mining tools are very helpful for the efficient working. Different applications used various methods and algorithms for implementing mining tool and they have been proved as efficient in their domain of work. This review would help the researchers to focus on the various issues of data mining. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing.

References

[1] Ralf Miku and Markus Reischl, " Data mining tools", wires.wiley.com/widm, Volume 00, Januar y / Februar y 2011.

[2] Agathe Merceron And Kalina Yacefa,"Educational Data Mining: a Case Study".

[3] Jiawei Han and Jing Gao, " Research Challenges for Data Mining in Science and Engineering".

[4] Erhard Rahm_ Hong Hai Do, "Data Cleaning: Problems and Current Approaches".

[5] Wei Fan And Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations Volume 14, Issue 2.