Data Mining And Visualization of Large Databases

AbdulRahman R. Alazmi raphthorne@yahoo.com
College of Petroleum and Engineering
Kuwait University
Kuwait

AbdulAziz R. Alazmi fortinbras222@hotmail.com
College of Petroleum and Engineering
Kuwait University
Kuwait

Abstract

Data Mining and Visualization are tools that are used in databases to further analyse and understand the stored data. Data mining and visualization are knowledge discovery tools used to find hidden patterns and to visualize the data distribution. In the paper, we shall illustrate how data mining and visualization are used in large databases to find patterns and traits hidden within. In large databases where data is both large and seemingly random, mining and visualization help to find the trends found in such large sets. We shall look at the developments of data mining and visualization and what kind of application fields usage of such tools. Finally, we shall touch upon the future developments and newer trends in data mining and visualization being experimented for future use.

Keywords: *Applications of Data Mining, Business Intelligence, Data Mining, Data Visualization, Database Systems.*

1. INTRODUCTION

Since the inception of information storage, the ability to sift through and analyze huge amounts of information was a dream sought out for in many ways and through different ways. With the advent of electronic and magnetic data storage, rational databases emerged as one of the efficient and widely used method to store data. Data stored in such large databases are not always comprehendible by humans, it needed to be filtered and analyzed first. Stored records are raw amounts of data poor in information, not only is it large and seamlessly irrelevant but also continuously increasing, updating and changing [1].Here is where data mining and visualization comes into the picture. Data mining and visualizations are knowledge discovery tools [2] used for autonomous analysis of data stored in large sets in many different ways. Large data sets of data cannot possibly be analyzed manually; mining tools and visualization provide automated means to comprehend such data sets. Data mining is defined as the automated process of finding patterns, relationships, and trends in the data set. On the other hand, data visualization is the process of visually representing the data set in a meaningful and comprehendible manner [3]. In fig. 1, the figure shows what Data Mining is and is not.

Data mining is a knowledge discovery process; it is the analysis step of knowledge discovery in databases or KDD for short. As an interdisciplinary field of computer science, it involves techniques from fields such as artificial intelligence AI, machine learning, probability and statistics theory, and business intelligence. As in actual mining, where useful substance is mined out of large deposits hidden deep with mine. Data mining mines meaningful and hidden patterns, and it’s highly related to mathematical statistics. Though utilizing pattern recognition techniques, AI techniques, and even socio-economic aspects are taken into consideration. Data mining is used in today’s ever-growing databases to achieve business superiority, finding genome sequences, automated decision making, monitoring and diagnosing engineering processes, and for drug discovery and diagnosis in medical and health care [4]. Data Mining, as with other Business Intelligence tools, efficiency is affected by the Data Warehousing solution used [5] [6].

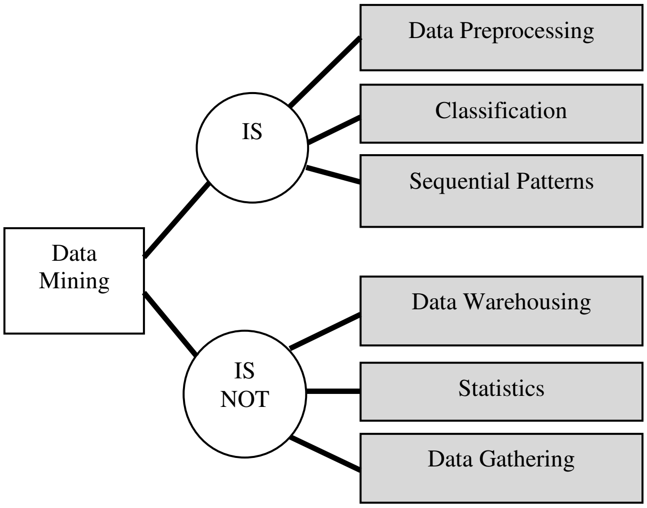

FIGURE 1: What is Data Mining?

Data Visualization is a data mining application considered as an information-modeling paradigm, in which seamlessly random data may be represented in an appealing graphical manner. Visualization of collected data can be found as early as the middle of the 19th century, where Dr. John Snow made a map of central London, pinpointing the locations of the possible sources of the cholera and its victims. Thus allowed for the detection of the hidden relation of the alleged sources of cholera (the water pumps) and its victims, and helped in ridding of the disease [7], other examples are also given in [8]. Visualization can be divided in to seven main subfield according to Frits [9], visualization algorithms, volume visualization, information visualization, multi-resolution methods, modeling techniques, and architecture and interaction techniques. Human beings understand and comprehend graphics more easily than numbers and letters. Human brains can interpret graphs, charts, icons and models quicker than numbers in tables; this is in contrast to computers, where numerical representation is perceived more efficiently. For example, a pie chart showing the classification of a university student will be understood quicker than the same data represented in a table, as Fig. 2 shows. Visualization of such data helps the human brain figure out and perceives such knowledge hidden in the data. The goal of data visualization is to not only summarize the large dataset, but also provide a better way of exploring the knowledge hidden and waiting to be found there automatically and autonomously. Visualization of datasets helps in explicitly showing proximity, enclosure, similarity, connection, and continuity. Analysis of data through visualization is further divided into two main categories, the Exploratory Data Analysis EDA, and Qualitative Data Analysis QDA; we shall see both analysis types in the paper.

Data Visualization and Data Mining can be used together, or in sequence, whether Data Visualization first or Data Mining first [10] [11]. It's worth noting that data mining and visualization today are available on most platforms. Cost for data mining applications range from high-end DBMS, and huge processing power costing in millions to business class systems costing in several hundreds of dollars. In addition, it is worth mentioning that data mining and visualization efficiency depends on the type of DBMS, and the processing power available for the application.

In this paper, we shall review data mining and visualization in light of their usage in large databases. We shall see the current trends, main tools and application of such technologies and have a look at the latest and possible future uses for data mining and visualization. In the next section we shall demonstrate and give a background on the developments that led to the modern data mining and visualization technologies we came to know today. In section III, we will talk about the tasks and techniques used today and available in the market for data mining and visualization. In sections IV and V, we shall see the applications were data mining and data visualization is used. In section VI, we examine some of the tools used; in section VII we shall see latest developments and future uses of such technologies. In section VIII, we see some of the challenges. Finally, we conclude the paper in the conclusion and references.

2. BACKGROUND

The notion of automated discovery tool in a large data set has prevailed in the development of data storage technologies. Tracing the roots of data mining to the early days of mathematical regression and probability theories in the eighteenth century, we can see that mathematical models such as regression and Bayesian theories provided means of analyzing large data sets effectively. With electronic computers taking the exclusive position for data storage in the twentieth century, early commercial computers quickly over took manual and other means of data storage. By the 1950s, early high level languages were developed; this development dramatically changed how humans interact with computers. Computerized data storage was not only used for storage but also for querying. Further on after the advancements in both hardware and software, rational database systems RDBS were developed. Structured Query Languages SQLs were used for semi-automatic acquisition of knowledge through querying the data storage, although tedious programming and substantial efforts have to be done.

By the late 1970s and through the 1980s, developments in computer networking, data storage and software had led to the break-through developments of the famous *online-analytical processing* OLAP techniques. Further developments in databases such as multidimensional and spatial database, and the dramatic cost reduction in data storage have lead the way for complex databases with sizes ranging in terabytes to petabytes ever growing and 24 hours online connected. Such developments led to the development of more complex algorithms derived from both AI and neural networks to efficiently search the database to automatically and autonomously acquire knowledge, more efficient and complex than OLAP. In the early 1990s, we can safely say that early data mining and visualization tools were developed.

As modern RDBMS dominate the market, data mining tools are developed to search such solutions. RDBMS usually store the data in the form of bytes or *Binary Data Objects* blobs; this makes the data mining of such datasets even more elusive and harder. Data mining was known by many names including knowledge extraction, information discovery, data archaeology, and data pattern processing. The term data mining however is the most popular term in the database field, since then it was also incorporated both the AI and machine-learning fields as well [12]. With further developments in data mining tools happening today, and the huge increase in processing power and decrease of hardware and network costs, accessible and efficient data mining can be achieved at a moderate costs. The business sector, scientific research, and health care are the dominant users of such data mining and visualization tools. Standards such as the European Cross Industry Standard Process for Data Mining CRISP-DM were developed to create a cross platform compatible data mining interface to cope with the increasing demand for data mining across many different applications and fields. Today, data mining software packages imply complex algorithms and techniques for searching, pattern recognition, and forecasting complex global stock exchange markets changes. Oracle, IBM and Microsoft are the most prominent providers of commercial data mining software. Such advanced and available intelligence software have influenced and played a major role in reshaping the security practices and techniques applied by international intelligence agencies such as FBI and CIA.

Data visualization is an emerging field, developed to counter the ever-increasing growth of databases in both size and complexity. Developed from the statistical, probabilistic and data representation fields to make sense of large quantitative data sets found in databases. As with data mining, data visualization techniques began as mathematical tools that summarize large datasets into a single representation or values. Mathematical models included time-series graphs, cartography, and fitting equation [13]. Even before computers were available, visualization operation on data was done [14], such as Francis Galton's weather maps fating back to 1980's [15]. Today complex techniques are used in visualization. Visualizations are mainly used for business and scientific research applications. Usually data visualizations, unlike data mining, work on raw data such as numbers or letters as in names [16]; this makes the visualization process consume both time and energy. Such a problem is frequently faced with large DBMS.

According to D. Keim in [11], Visualization techniques can achieve several ends that include visual bases for data hypotheses, evidence for or against a trend in the data, and/or data models for demonstration purposes. Data visualization is used to visualize and present the data set, test hypothesis, or explore dataset freely. Visualization of data also helps in communication and easily emphasizes trends otherwise buried within the dataset. As stated by Friedman [17] "*main goal of data visualization is to communicate information clearly and effectively through graphical means*".

Data Visualization techniques can be used to *pre-process* data before Data mining techniques is used. These include segmenting, sub setting, and aggregation techniques. Visualization techniques of Data can be categorized into three categories, which are *Data Visualization, Distortion, and Interactive Techniques*. The first type include among others Geometric, Graph-Based, and Icon-Based. This type is seen in the form of Histograms, Scatter plots, and Shape Coding. Distortion types include the use of Perspective Wall, and Hyper-box, the latter being techniques used for multivariate datasets [18]. The latter type, the Interactive techniques can be in the form of Projections, Zooming, and Detail on Demand. There still exist among researchers of Data Mining and machine learning fields the need to incorporate and embed Data Visualization tools into Data Mining tools.

3. RELATED WORK

In [19] the author reviews several interactive visualization techniques that are used in the context of data mining. The paper also retrospectively defines visualization techniques in the world of data mining; these can be defined as expressing data sets to discover trends, for *exploration*, or can visualize the workings of complex data mining processes, for *comprehension*. The paper focuses on data visualization, while in our survey we shall review both data mining and data visualization and their integration as one field.

Authors C. Romero and S. Ventura of [20] give a survey data mining techniques in the field of education. Not just in e-learning but also in traditional class rooms. Data mining can help in improving educational courses through knowledge discovery of facts in the past history of a specific course. These include: feedback for the educators such as effectiveness of content, students' classifications, and mistakes in the teaching process, feedback for the students such as suggesting helpful educational content available for them. The paper surveys data mining and a few data visualization techniques used in education such as classification, text mining, sequential patterns and visualization. In our survey, data mining and visualization techniques, trends, and application will be discussed not only for education but for a wider range of fields.

In [21], the paper reviews the history of Knowledge Discovery and Data Mining KDDM process, its definitions, models, and standards. The survey suggests a need for the standardization of the KDDM methodology rather than its somewhat haphazard usage in industry. While effective models are used, they are however separate in form and methodology. This can affect the field of KDDM as it matures by making ambiguous and redundant set of models and techniques. Data mining being a step in the KDDM process helps in understanding processes and gives input for decision support systems. Both KDD and KDDM are related, while the latter is not only concerned with databases, but other sources of data. KDDM models range from industrial to academic, each having several different steps. Important steps are data extraction, preparation, mining, and evaluation. The survey compares several KDDM models, while in our survey we do not take the whole KDD or KDDM process in the picture; we focus on data mining and visualization alone. Data mining is mostly a step in the middle of any KDD or KDDM model, and it is a pivotal one with many dimensions and factors.

Other body of work usually surveys a specific field in the data mining and data visualization techniques, such as [22] for web mining, [23] for data visualization in bioinformatics, and [24] for data mining in e-commerce. In this survey, we take a broader view in many fields and trends.

4. TASKS AND TECHNIQUES

Data mining and data visualization were developed from mathematical methods of pattern recognition and probabilistic theories to deal with unstructured, time varying, and fuzzy data in huge amounts. Such techniques allowed for finding correlations, relations and assertions. We shall touch upon some of the main tasks associated with data mining and visualization and the techniques to achieve such tasks that are popular in the field of data mining and visualization in the following paragraphs.

Data mining tasks include *Classification, Association Rules, Clustering, Anomaly detection, Summarization, Regression, and Sequential Patterns,* M. Sousa et al [3]. Visualization tasks include *statistical modeling, regression modeling, information abstraction*, *mindmaps*, and usually *data presentation* in other forms like graphs, maps, and histograms. All data mining and visualization techniques and algorithms relay on three main steps, model representation, then evaluation of the model, finally model search to identify patterns [25]. Model representation depends on the dataset itself, most datasets require certain models, clustering usually is effective with demographical datasets. Second is model evaluation, where the model used is evaluated to make sure it matches the nature of the dataset. Finally the model search, it's done after evaluation of the model is verified, it extracts the knowledge we need from the dataset.

Classification is the process classifying sets of data based on common attributes. Classification is considered a classical data mining techniques as it's highly related to statistics method used before data mining was conceived. This classification help divide the large dataset into further smaller and correlated datasets. Such classification is the basis for further analysis, as classifications divide the datasets into smaller correlated groups; this is called consolidation of data. Then we have *association rules*, it's the process of testing or when implying a set of hypotheses are made against a certain data set. These hypotheses are called *rules.* After the verification of the plausibility of such rules, or associations, then we subject the dataset against such association rules. Associations rules can find hidden links between otherwise unrelated data; the *beer-diaper* links used in *market baskets* is an example for such associated rules. Market baskets are defined as items usually bought together; such an analysis is used heavily in *marker research*.

Clustering is the technique of grouping of several objects unto groups of similar attributes in order to simplify large, complex sets. Clustering is a learning technique and therefore it has no correct answer. Clustering can be hierarchical and non-hierarchical. Hierarchical clustering clusters groups of data in size (can be from small to large or vice versa), and it comes in two flavors, Agglomerative and Divisive. The first clusters each record alone, and then merges clusters together. The second, does the opposite, it starts with one full cluster and then subdivides the cluster. The non-hierarchical clustering has two flavors as well; the difference here is that no hierarchic clusters are used. The first type is the single pass methods, where the database is scanned once to create the cluster. The second type is the relocation method, where records are relocated from one cluster to another for optimization. Several passes against the database may be used, as opposed to the single pass methods [26].

Sequential Patterns the use of sequential pattern algorithms on sets of sequential data (e.g. bills made on the same month). The goal is to find a trend or pattern that happen in sequence. Rule induction task is used to find hidden if-then rules in the dataset. These rules are based on statics analysis and probabilistic models. Derived if-then rules are further used in analyzing the dataset in the future.

Data mining techniques are varied and interdisciplinary, since they come from varied fields. Neural Networks are techniques frequently used in data mining. These techniques are from the field of Artificial Intelligence AI. Neural Networks link different attributes through vectors intelligently; it has considerable training time when compared to other techniques, and has little confidence intervals that depend on the number of neighbors. Also AI derived techniques tend to be more sophisticated and show human like-intelligence in finding hidden correlations.

Nearest neighbor technique is another classical technique to classify records of data based on their resemblance or closeness to a specific record. This technique is used to compare newer or updated records to a pivotal or historical important record; it tries to mimic the human comparison process. Decision trees are techniques from *machine learning* field. When compared to neural networks, *decision trees* are much faster in performance, due to less computational overhead. Decision trees algorithms are greedy algorithms that divide the cases or classified groups in the training set of data until no more cases in the dataset can be logically or ontologically divided. Their drawback is their need of large datasets to provide efficient results. Different kinds of decision trees exist, we mention two kinds for example. Classification and Regression Trees CARTs, these trees split the data set into 2 way splits for decision making. Other type of decision trees is Chi Square Automatic Interaction Detection CHAID. CHAID trees on the other hand create splits in the dataset using the Chi square tests, creating multi-way splits in the dataset.

Moving on to data visualization, techniques for visualization vary depending on the type, usually they are classified as query independent techniques, and query dependent techniques. Query independent techniques directly visualize data set without any assertions. On the other hand query dependent techniques will visualize depending on a query specified prior. We shall look at techniques of both classes.

In D. Keim's work in [27] the authors presented a novel technique called pixel-oriented visualization techniques. Pixel oriented techniques are mappings of the data values into a 2D or 3D map of colored pixels depending on the value of the data. The colored maps give immediate and precise information on the trend or the average values of the dataset [28]. Pixel oriented techniques are further divided into query dependent and independent pixel techniques. The query dependent pixel oriented techniques tend to form a map of the current trend of the data. Usually this is not very useful as most of the time the data values or the colored map is not very easy to read out. On the other hand, the query dependent pixel oriented techniques are more effective, as the finished colored graphs indicates how the data is scattered or varied around the queried data set or target values.

Other visualization techniques are the geometric projection techniques. These techniques are sometimes summed under the projection techniques. These techniques find efferent or convenient 'projections' of the data in multiple dimensions. In addition, these techniques are used chiefly in EDA, as most of the time multiple projections are done for further exploring the dataset. Since the search space is very large in terms of multi-dimensional datasets, exploration can prove to be very difficult. Systems developed specifically for geometric projection pursuits found in [29], automatically find such convenient and interesting projections more easily and effectively.

5. VISUALIZATION OF LARGE DATABASES

As discussed previously in this paper, data visualization is an important application that helps to convey knowledge mined *graphically*. As human beings, we are more familiar with drawings, icons, and graphs then we are with numbers and tables. Raw numerical data or even alphanumerical data can be represented in a map; chart, bars, pie chart, or even a histogram to visually identified and convey important trends and correlations visually. Data mining in large databases is still a difficult task; due to the fact of the huge amount of raw data need to be processed. A turnaround is to partition the data into sets, and tackle them individually. This makes supporting tools such as *Visualization Tools* needed. Data visualization is considered with two kind of analysis, first, Exploratory Data analysis EDA and model visualization [30], second is the Qualitative Data Analysis QDA.

By EDA, it is meant the careful exploration of the data set graphically to identify a pattern, a recurring trend or behavior that connects different views or visualization. EDA helps to identify patterns without preconceived knowledge, hypothesis, or suggested models used on the data set. Model visualization is the use of predefined models, such as XY charts, 3D plots, or box plots to model the data. Usually visualization of data plays on the key idea that human beings are more capable in analyzing and understanding graphs than digits and letters. Figure 2 include a very simple, yet effective example of tabular versus visualized data sets. Visualizations such as Venn diagrams and clustering help the observers see grouping and partitions in a dataset more easily than rows of alphanumerical records. QDA on the other hand, is the analysis of non-numerical data. QDA is considered with database containing images, text, links, or other kinds of data that is not numerical or alphanumerical.

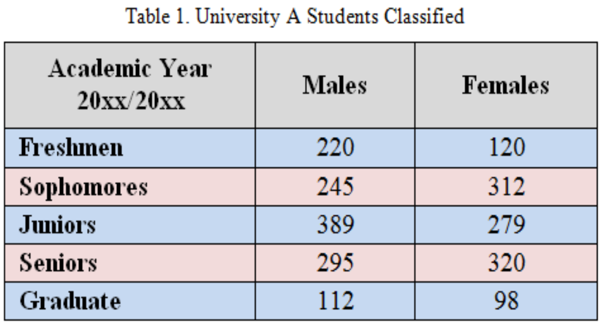

Table 1. University A Students Classified

Academic Year 20xx/20xx	Males	Females
Freshmen	220	120
Sophomores	245	312
Juniors	389	279
Seniors	295	320
Graduate	112	98

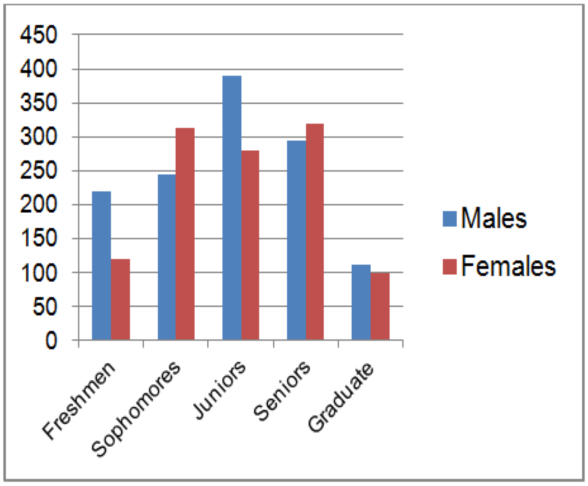

FIGURE 2 (a): Histogram Representation

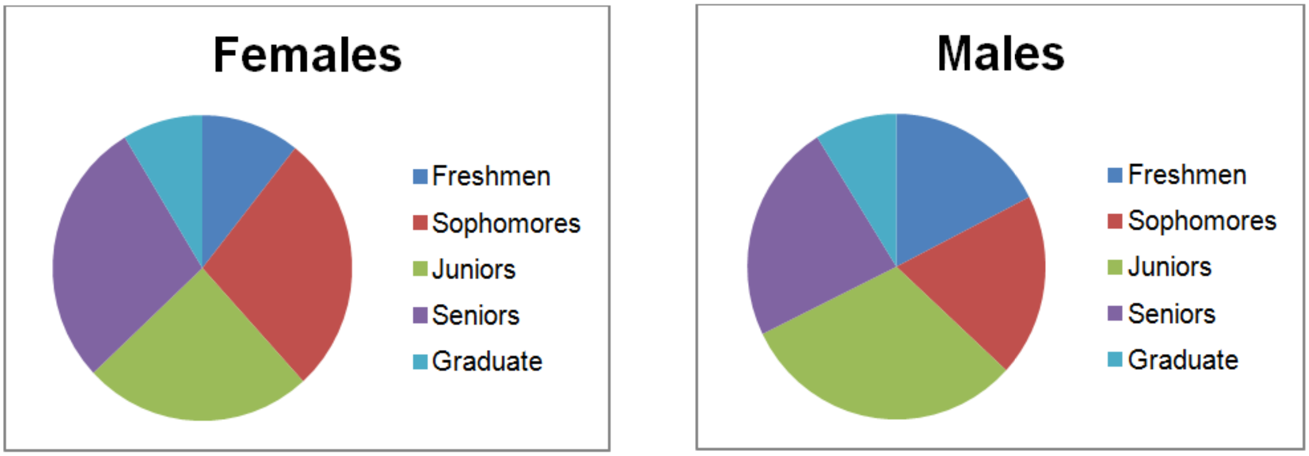

FIGURE 2 (b): Pie-Chart Representation

FIGURE 2: Different Data Representation

Prominent visualization techniques used to visualize large datasets is charting [31]. Charts or namely pie charts are the most common form of data visualization. Pie charts are both easy to understand and an elegant and fast delivery method. As most people are familiar with pie charts, they convey information relatively in a fast and direct way, see Fig2. Large database consists of millions of data records that are updated frequently, an example of such databases are Geographical Information Systems GISs. Visualization is used in GISs to visualize dataset.

Making understanding more visually and less tedious of the gathered satellite data over the course of time, weather maps and contour maps with colored regions depending on different fauna and flora. This is usually the case with real-time DBMS, where the visualization must support some kind of animation over time [32]. The time intervals of such sequenced data play a role in making the visualization more realistic and effective. A well-known example of GIS application is Google Earth and Google Maps. Google Earth and Google Maps have both changed how we deal with GISs; it is a free, online, fully visualized GIS. It is provided to the masses an interactive and visually appealing GIS system. Most people do not know that Google Map is just another GIS system visualizing a large real-time GIS system. Data visualization is used widely in computer networking as in data network traffic plotting, as we shall see later in this paper [33], Market segmentations, Anomaly detection [34], and Manufacturing [35] are among the best domains were data visualization provides tangible results.

6. APPLICATIONS OF DATA MINING

Applications of data mining vary, depending on the nature of the data to be mined. Since its inception data mining was used in various other fields. The classical application of data mining encompasses statistical and probabilistic applications. These classical applications included for example, population census studies, biosphere analysis, and marine life and oceanography. As a prominent use for data mining is data visualization, we have touched upon this application in the previous section; we focused on visualization since its importance as an application for data mining. We have selected other important applications of data mining used widely today. Many of these applications have branched out to become separate but related fields.

6.1 Spatial Data Mining

Spatial databases are databases that have unique data; this data is about space and geometry, such as the coordinates of earth, maps, and satellite data. This data is in the form of geological or geographical data. Such databases are extremely large and the data seems for the most part unrelated, and without any signs or correlations. Data mining is a natural candidate to find logic and make sense of such data. Data visualization, which was discussed earlier, is another tool of data mining heavily used in spatial databases.

Spatial databases are also used in geographical for marketing, traffic control and analysis, and GIS systems [36]. Economic geographers use such spatial data to acquire global market information such as customers' demography, manage inventory, and have logistic advantages. Another interesting feature of data visualization of large databases is that the visualization also finds relation among the non-spatial data in the database, such as local maxima and minima.

Algorithms used in spatial databases data mining and visualizations fall into neighborhood graph algorithms [37]. Beside algorithms, machine-learning techniques are also used for geometric clustering, since we have so much data, largely in 3D or topological in 2D [38]. Applications of data mining in spatial databases are mostly statistical such as the Global autocorrelation. Global autocorrelation is, basically, the calculations of the average, variance and mean of the special data [39]. On the other hand, Density analysis is an EDA process in which visualization can show at a glance we the data values are mostly concentrated, such as plotting on a map or the globe of the earth. Famous applications that rely heavily on spatial data mining include Microsoft Bing Maps and Google Earth and Google Maps. Such applications offer up-to-date information, with search capabilities allowed for end users, such as finding names, streets, and locations.

6.2 Business Intelligence

Data mining help business intelligence in many ways and for that, it is one of the fundamental tools of business intelligence. Business intelligence's (BI) goals are to gain a competitive advantage over competitors, increase productivity and effectiveness of current business operations, and to maintain a balance and control of risk management. Business intelligence is a usual task of any Enterprise Resource Planning ERP solution [40]. Business-intelligence mine habits and trends of customers' data stored as records through internet cookies and sales profiles. This mining helps in discovering the customers' segmentations, and demographics. Data mining provides market basket analysis; items purchased together are identified and in turn bundled and advertised together. Anomalies can be also caught using intelligent mining tools; such tools mine the transactions and try to extract anomalies. Anomalies may be deliberate, such as fraudulent transactions or they could be unintentional, a glitch or bug in the program or just an odd transaction that may never be presented again in the entire database. Fraudulent transactions are caught due to their recurring characteristics, such as credit card theft, identity thefts or account hackings.

Global economies today around the world are information driven, known as knowledge-based economies [41]. Business intelligence is one of the top proponents and drivers for the development of technologies of data mining. Most data mining tools in the market today are integrated in enterprises tools such as Enterprise Resource Planning tools, and Customer Relationship Management tools. Top applications of business intelligence include market research, risk management, Market baskets, and fraud and anomaly detection. Automated business intelligence through data mining support is being used by modern enterprises in decision-making, and drive knowledge based decision rather than human imitation based decisions.

Business intelligence achieves what is known as a competitive advantage. A competitive advantage is defined as the advantage that other rivals lack, the specialty or secret skill others lack. One of the main reasons to acquire such an advantage as a competitive advantage is the competitive pressure. According to [42], competitive pressure is degree of pressure that companies feel from rivals in the market and possible new entrants. For gaining a competitive advantage, enterprises develop market research groups that analyze through data mining large datasets. Market research finds what products dominate the market, and the hidden elements that set such products from others in the market. As an example, media networks use data mining in their market research to set the common factors between audience and the program's scheduled slot. Large media groups used to hire human experts to schedule their programs slots, now the use of fully automated data mining tools for scheduling is the common trend. Results were equivalent or better than the human manual scheduling [43]. Data mining tools also discover the market's baskets, as mentioned earlier, market basket are associations of certain products that are highly likely bought together. The retail industry is dominated by market baskets predictions, giant retailers such as Wal-Mart, Costco, and K-Mart, are among the main adopters of such business intelligence achieved by data mining.

6.3 Text Mining

Another widely used application of data mining is text mining. Text mining deals with textual data rather than records stored in a regular database. It is defined as an automated discovery of hidden patterns, knowledge, or unknown information from textual data [44]. Most of data found on the World Wide Web WWW is text, after distilling the multimedia elements; most of knowledge out there is text. Text mining utilizes different techniques and methodologies, mostly linguistic and grammatical techniques, such as the Natural Language Processing NLP. Techniques of text mining originated from computational linguistics, statistics, and machine learning, such techniques were developed to make machines, specifically speaking computers, understand human language.

Text mining mines large sums of documents and articles stored in a database or even fetched from the web. The way that text mining works is very complex, were NLP algorithms try to parse sentences and matching verbs with nouns to make logical connections between all of the elements in a single sentence. Computers today are not able to understand human languages directly, not without complex AI and NLP algorithms, so mining text is still considered a daunting task, which consumes resources. Text mining to be effective, it involves a training period for the text-mining tool to comprehend the hidden and recurring patterns and relations. The process of textually mining documents involve both steps, first the linguistically analysis then the semantically analysis of the plain text. After scrutinizing plain text, mining then finally can relate nouns and verbs, mining out some hidden traits found in the text, traits such as the frequency of use of some verbs, entity extractions such as the main characters, and possible summarizations of long documents. Text mining is used in business applications, scientific research, and in medical and biological research [45]. TM is very useful in finding and matching proteins' names and acronyms, and finding hidden relations between millions of documents.

6.4 Web Mining

With the revolution of the Internet that have changed how databases are used, this revolution brought the term of web mining. Web mining is considered as a subfield of data mining, it's regarded as the vital web technology that is used heavily to monitor and regulate web traffic. Web mining is further divided into three main sub groups, web content mining, web structure mining, and web usage mining [46]. Web content mining is the mining of content found on the web, this include metadata, multimedia, hyperlinks and text. Web structure mining is considered with the semantics and hyperlinks making up a website or a network. Web structure mining are usually is used by search engines to 'crawl' the web and find all possible links forming a network. Web usage mining is considered with the traffic patterns in the World Wide Web WWW. Most of the data is mined from the web servers and web proxies. Web servers log most traffic, such logs are the data needed to construct an overview map of the traffic coming and going to that web site.

Web mining is used in Information Retrieval IR systems, such as search engines. Web mining is also used in web trafficking measures, were traffic is traced and monitored. But for the most times, web mining is used for business intelligence [47], as it can search the web with all its fuzziness to retrieve business oriented information from the web.

7. TOOLS

Data mining tools are basically software packages, whether integrated packages or individual packages. These sophisticated software tools often require special data analysts. Such analysts are trained to use such tools, as data mining itself is not a straightforward process. It is worth mentioning that data mining tools need a substantial investment in hardware and software, as well as human resources. Deployment of data mining tools and packages is also an overwhelming task, in size and management, as it needs careful planning and management. In the next paragraphs, we shall look into some of the used data mining tools and data visualization tools.

7.1 Data Mining Tools

Data mining tools are also called *siftware*, for the sole reason that they 'sift' through the dataset. Data mining tools varies depending on level of their sophistication and projected level of accuracy. In 2008, the global market for business intelligence software, data mining centric software, reached over 7.8 billion USD, a vast amount. IBM *SPSS* is an example of business intelligence software package [48]; it is integrated data mining software with diverse business intelligence capabilities. IBM also provides online services for web mining, these services are called *Surfaid* Analytics; they provide sophisticated tools for web mining [49]. Other data mining with business intelligence capabilities is *Oracle Data Mining* [50], a part of the company's flagship RDBMS software suite. SAS also offers its *SAS Enterprise Miner* [51], as a part of its enterprise solutions. SAP, a world-renowned business solution provider, offers world known ERP solutions along with providing other mining tools software that can be integrated into their ERP solutions. Other software companies include Microsoft; it offers *SQL Server Analysis Services*, a platform dependent solution integrated in Microsoft SQL platform for Microsoft Windows Server. Microsoft also offers a less sophisticated product, namely the *PowerPivot*, a mining tool for small and middle size enterprises, with limitations and ease of use to match with its nature of use. Open source mining tools exist; they include the *Waikato Environment for Knowledge Analysis* or WEKA [52].

With the huge decline of the costs of both storing and acquiring data, through utilizing mining tools to mine web, documents, or the use of data acquisition tools such as RFID tag readers and imaging devices, data mining tools are being adapted more rapidly and incorporated into almost every business tools in the market today.

7.2 Data Visualization tools

For Data Visualization tools, we have checked IBM's *Parallel Visual Explorer*. This software package is used for market analysis, oil exploration, engineering and aerospace applications, and agriculture to name a few.

For medical fields, Parallel Visual Explorer is used to analyze various effects of treatments on the immune system. It helps in visualizing many different diverse effects on the patients' immune system [53]. For manufacturing, this tool helps in monitoring the processing parameters. Process parameters are vital for effective streamlined production. For agricultural usages, this tool helps in determining which seed to plant by analyzing the soil parameters with taking in consideration the weather conditions. Finally, Parallel Visual Explorer is also used for market research such as providing visual aids to help market analyst find customers trends, habits, and buying sprees.

An interesting visualization tool is Cave5D [54]. It's a data visualization tool developed by the university of Wisconsin-Madison. The inventors of this tool are Glen Wheless, Cathy Lascara, from the center for Pacific Oceanography, with Bill Hibbard and Brian Paul back in 1994. This software ran as a package for the Vis5D software. Cave5 provides interactive 3D, time variable visualization of dataset in a virtual environment. Cave5D integrates Vis5D's libraries and framework; it uses its graphical libraries to model the dataset. An image showing Cave5D in actual usage is shown in Fig. 3.

Vis5D [55] is a visualization system used for 3D animated simulation of weather and geological data. Vis5D uses 5D arrays that contain the time sequences of the 3D spatial datasets. Vis5D was incorporated into Cave5D through its extendable PLI libraries. Cave5D and Vis5D have their limitations as only relatively medium to small datasets can be visualized and animated at the same time.

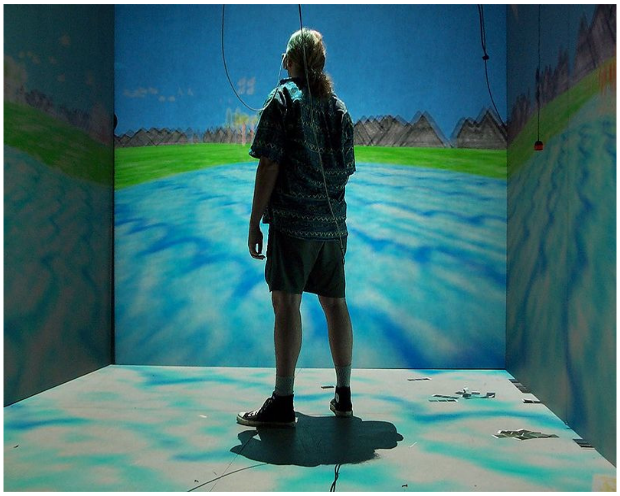

FIGURE 3: Cave5D

In [56] the paper offer a system that offers a simple interface that overcomes the difficulties faced by other visualization tools. The system utilizes tree structures to visualize the data. Its interface allows the users to zoom in on data set as well as dynamic branching. Navigation controls are also given, to allow for smooth switching in and out of the dataset trees. Visualizations can be in pie charts, scatterplots, and histograms among others. This proposed system was compared to *Polaris* of [57].

FlowScan is network traffic flow reporting and visualizing tool proposed by Dave Plonka in [33]. FlowScan is a collection of software that includes flow collection engine, a database, and a visualization tool. At 2000, FlowScan is an early indication of the need for visualization for data, especially for prolific network data.

Tools such as *Spotfire* and *XGobi* provide the user with predefined query visualization tools. These tools also have interactive functionalities such as zooming and brushing, which enable for finer graining the results. Academic tools such as Visage and Polaris offer similar functionality, but with custom block building query tools. Polaris, which is a visual query declarative language, has been extended to the Tableau software. Polaris offers Gant charts, scatter plots, maps, and tables.

Visualization tools are abundant. These tools range from internet network visualization, music information network, social network tagging, and web feeds visualization tools. Internet visualization tools are abundant over the internet, these include Mapping the Blogsphere, Websites as Graphs, and Opte Project. These tools offer to visualize the network from a single computer as neural networks. Music information tools such as Tuneglue, MusicMap allow the user to have a visual map of the artist of their choice and the other related artists, bands, and musical movements that influence the target of the search. Fidg't, TwittEarth, and Flickr Related Tag Browser all offer visualized social networking information. The first, offer you to Flickr and Last.FM tags to compare them to your network tagging activities. The second tool correlates a map of the world and the tweets made from twitter arising from their geographical locations. The third of this kind, offers the search through a visualized map of tags and their related tags. Other visualization tools such as Visualizing Information Flow in Science allows for a visualized view of citations used throughout scientific journals and are used to evaluate them [58].

8. FUTURE TRENDS

Future trends for data mining lie in the hands of innovation and scientific breakthrough. As data mining is both a difficult problem, and a relatively new problem that incorporates many interdisciplinary fields. We shall see some new trends that will shape the way that data mining will be used in the upcoming future. Visualization tools are also witnessing a rise, credited to the newer technologies in human-computer interactions.

8.1 Cloud Computing Based Data Mining

A relatively new trend in utilizing and benefiting from data mining tools for middle-sized and small enterprises, incapable of supporting a full-fledged data mining solution, is *cloud computing* based data mining tools [59]. Because small and middle-sized enterprises usually lack the infrastructure and budget available for large enterprises, they tend to try this new cost effective trend. Cloud computing promises to provide data mining tools benefits at relatively lower costs form such small or middle sized enterprises. Cloud computing provides web data warehousing facilities, were the actual data warehouse [60] [61] application is outsourced and accessed entirely through the World Wide Web. Cloud based data mining also provides sophisticated mining analysis of the dataset, comparable to actual data mining software, as the enterprise specifies and demands.

Aside from lowering the costs of the data mining software tools infrastructure, cloud based mining also provides expertise that is not available in such middle-sized and small enterprises. Most cloud based data mining providers tend to have data experts, data analysis, and a broader experience with data mining then their clientele. Usually start-ups or entrepreneur level enterprises lack not only the financial resources but also the human resources and expertise in the Information Technology IT field, not to mention in the data analysis field.

The *Infrastructure-As-Service IAS* helps middle-sized and start-up enterprises to be rid from the burden of software, hardware, and human resources management costs. It also helps in reducing the already limited budget. The main downfalls of cloud computing based data mining are the dependency and privacy issues that occur from the fact that another party that the enterprise have to agree to store its data on its machines and data warehouses facilities. Such issues are the main reason that limit and turn off large capable enterprises from going with cloud based data mining solution. These enterprises, large enough and have huge IT resources, can set up their own data mining solutions instead of taking the much less needed risks. *Dependency* is another problem, it means that the whole service depends on the other party, not the enterprise itself, meaning that the enterprise is pretty much tied up with what the service provider has to offer, huge switching costs. The privacy concerns arise from the fact that the enterprise's data is technically not under its control or even possession, the other party has it, it utilize its resources to give results and analysis. The privacy concern entails the misuse of the data, mostly causing confidentiality risks.

8.2 Data Conditioning Tools

Data conditioning is currently a technique that is not only meant for data mining. It is used for intelligent routing, privacy and protection as well as for data mining. As data grows today in unprecedented rate, the need to clean up the huge piles of data is necessary. Reports suggest that more than 80% of enterprises data are unstructured and fuzzy data [62]. The other goal of data conditioning is to elevate or at least minimize the interference of IT people. This would quicken the BI step, and in turn make it ubiquitous for the end-users, whether business or science users.

The key technique used for data conditioning for data mining is data warehousing. Data warehousing is used for organizing such unstructured data, it's the middleware that transfer data from the transactional database into a structured, aggregated warehouse [63]. Data warehousing is tasked with data extractions transformations, and load, this is known as the ETL process were the data is modified to be stored in the warehouse. Data in the data warehouse is not like its previous form were it was in the original database, it's an aggregated more cleaned version.

Usually data mining processes are done on the data stored in the data warehouse as it has already cleaned and formatted for the analysis tool, which will mine useful knowledge. The quality of the mined knowledge depends heavily on the data warehouse design and model used. Finally, we can deduce that data mining efficiency as well as quality is highly affected by the level of structures and aggregation found in the data warehouse [64].

8.3 Human like Intelligence

The goal of today's data mining tools is to reach human experts level, in terms of accuracy and innovation. The promise of such intelligence lies in incorporating more AI techniques into data mining tools. This newfound intelligence will help incorporate data mining into fields that was not usual for such mining to occur. Technically the data mining is one of the main uses of AI algorithms commercially available today among other data-mining related fields [65].

Such intelligence incorporation has led to fraud detection mining tools, summarization, predictive analysis, and information retrieval tasks to name a few. IBM's SPSS, statistical modeling software, usages many AI techniques, incorporating machine learning also. Data mining seems to be the most prominent frontier were AI is currently thriving. In addition, a new technique rising in the field of AI in data mining is soft computing. *Soft computing* is considered with computing techniques that tolerate and exploit imprecision, uncertainty, approximation and reasoning [66]. This new and promising technique allows for traceability, robustness, and close resemblance, forming the new term of Machine IQ. *Fuzzy logic* also is a contributor to the advancement of newer more intelligent data mining techniques.

8.4 Interactive Visualization

The trend for the visualization tools is being more and more interactive with the user [67]. This is due to the advances in User Interfaces (UI) designs, from graphical interfaces, voice recognition, to touch sensitive displays. This trend of visualization graphs is called *advanced* visualization as opposed to the olden types of *static* graphs such as pie charts, histograms and scatter plots. While the interactive -advanced- visualization tools do have limits such as the need of a multimedia medium such a monitor of a computer, laptop, or a tablet, they are still have the edge of being able to show more complex structures through zooming in and out, 3D rotation, and/or changes in datasets by enabling user input. These types of interactive tools can also be embedded into systems and websites, due to their nature of being targeted toward end users and able to have multiple outputs.

Interactive visualization must also keep their level of details to a tolerable degree, because some tools might go as far as to require programming of languages or structures, to evaluate datasets. While this may be acceptable for scientists and researchers, however, among business users it is unwelcomed. On the other hand, performance is another parameter that will appreciate among the latter, but might not be a key aspect to the former group. All groups of users welcome the level of accessibility of such interactive tools, such as changing the colors, font sizes, and font types among other features and configurations that allow any user with any level of visual media perception to use such charts and figures.

Live data feed is another factor contributing to the popularity of data visualization, especially interactive data visualizations. Hot in the data feed categories are the customers’ reaction to the business, decision makers would highly appreciate the visualization of their large data sets of their customers’ reaction, live and interactive. This is true especially in the case companies that have electronic data, such as websites.

9. CHALLENGES

Data Mining and Data Visualization is usually more effective if the data on which to be mined are conditioned beforehand. Future directions show the usage of visualizations output as inputs for Data Mining through the tight integration of implementing visual and pattern recognition algorithms in Data Mining functions themselves. Selecting a data mining algorithm can also be challenging. The user must select an algorithm that would represent the set of data accurately; a method to evaluate the representation; and a search criterion [68].

9.1 Challenges in Data Mining

Currently data mining, in the form we know it today, has not really achieved the potential of what was expected, envisioned in the late 1980’s or early 1990’s. The vision of becoming a mainstream application, it’s widely used but to a degree still limited, data mining hasn’t reached that vision. Challenges come in many forms, mainly in three categories, technical, legal, and ethical challenges, all of which they hinder adaption of data mining as a common practice. We shall examine some of these challenges that hinder the further development of data mining in the next few paragraphs.

Technically, data mining is not an application widely adapted by enterprises. It did not reach the level of a common desktop application, still. Although this was the intended goal envision for data mining. It was intended to grow until it reaches the desktop level. The technical issues, such as the huge and elaborate hardware and software infrastructure, are a usual suspect, because data mining requires substantial resources to be deployed and careful thinking and planning, to be effective. Usually the cost of a typical data mining tool in millions, as such is evident in integrating these tools into full ERP systems.

Other than costs, technical issues reside in the human resources as well; data mining require expert data analysts. These analysts will design and perform tasks on the data mining tool. Finally, the technical challenges can also manifest themselves in the limitations we have today in the current tools. Most data mining tools are not extendable, or easily upgradable or adaptable to other applications. These tools are hardwired into using a set of models based on certain best known methodologies. For example, a data measurement for business intelligence is hardly ever useful for a medical application. In addition, the limitations of today’s tools are such that they cannot really replace, although we have good progress in this direction, the human element.

Ethical challenges plaguing data mining originate from the public concerns about their personal data found on the Internet. The privacy issues stems from the fact that mining can link, find, and relates the public profiles, personal preferences and possibly private data such as emails and photos. The initial goal of data mining tools is not to identify such individuals; to counter act this, most dominated data are anonymized before it is published into the public internet. However, still huge concerns are raised around how enterprises may want to exploit the individuals' data for such mining purposes.

Other than privacy concerns around data mining application for business, governments are utilizing data mining tools for its own security and national security purposes. Such governmental security agencies are sifting through public data to locate certain wanted individuals, possible terrorists or other convicts. These uses, along with the business uses of data mining tools; made public awareness of their legal and privacy implications more evident in programs like *Total Information Awareness* program [69]. The Total Information Awareness Program was a secret program sponsored by the Pentagon; it was aimed at national security and the possible identification of terrorists. It used mining tools to mine private individuals' records on a massive scale. Public awareness against the exploitation of individuals' privacy and private data forced the congress in 2001 to stop the funding for this program. Legal and regulatory acts were issued to address these ethical concerns, acts like the *Health Insurance Portability and Accountability Acts* HIPAA, in the United States stated by the congress. The HIPAA act requires a prior consent from individuals regarding the use of their information and the notification of the purpose will their information will be used. Another ethical issue in raised by data mining tools is that they made *Globalization* far easier. Globalization has dire consequences on emerging economies, emerging businesses that can never compete with international top companies utilizing data mining.

Legislatively data mining has resulted in new levels of transparency in the free market globally. This is due to the vast data decimation across the internet willingly or unwillingly. *Wikileaks* for example, have had a hand in decimating much documentation otherwise thought to be secrets. The term *data quality* [70] is a relatively recent term, refers to authentic, complete, and accurate data and that the source of this data is legally liable for its authenticity of its quality. Governments, international legislative and professional organizations have made standardizations and regulations regarding the quality of published data from companies and other agencies. According to Will Hedfield case study in [59], more than 25% of critical data in leading organizations' databases are inaccurate and incomplete.

9.2 Challenges in Data Visualization

Challenges in visualizing large databases arise from the fact that an intuitive interface is as important and critical as any part of the visualization tool. Allowing the user to have a full view of the database, and then allow the user to zoom in on a piece of information, as more zooming is allowed, navigation tools to allow the user to change the path of zooming-in different directions, which can greatly complicate the querying and negatively decrease its performance. Suggested solution include the use of cubes to visualize the data hierarchy levels, however data cubes grow more complex far more quickly with the growth of the data set, and making the visualization huge, complex, and unintuitive. Another problem in visualizing databases is when the database itself is large in size, the form, tool, or graph used to visualize the data mined cannot be anything but cluttered or difficult to grasp.

Challenges facing Visualization tools are the limitations of the current human-computer interaction frontiers. For example, in the past decade, touch sensitive screen, although available, were limited in functionality, whether because of the software, or hardware itself. This had led to limitations of their use, but on the turn of its end, the last decade witnessed a rise in the touch sensitive screens with integrated touch sensitive functionalities such as the slates, pads, and tablets available in the market, ranging from several developers. This allows for more and better interaction levels with the visualization graphs such as the zooming, especially dynamic branching in zooming in on 3D objects [71]. Other potential of these human-computer interaction see the use of holograms, and 3D screens [72] [73] [74]. Each of which has its set of challenges as well, such as availability, cost, and reliability since most of these technologies are offered by limited vendors.

Other hazards still exist in the form of designing such interfaces to cope with these levels of interactions [75], [76]. For example, double clicking and dragging objects should still be in use instead of demanding the user to delve into levels and levels of menus and submenus, or the use of difficult finger swipes or error prone voice commands.

Data feeds to such tools are also a challenge; can a visualization tool offer interaction and real time modeling of data feeds? Most visualizations offer interaction but they are fed with linear sets of data, not online data which is refreshed at rates that may reach each second.

Unstructured data, which can be more than 70% - 80% of an organization data [77], also offer difficulties for visualization tools as well. Unstructured data comes in many forms; one form is text in a document. A document can be in any form, data and information are interleaved together, and must be extracted in order to infer facts. Data Mining tools and computational intelligence tools can be used to structure these data, in the form of an index. Even though the Data Mining tools have formatted the data, it is still difficult to visualize. For example it may be represented as a neural network, which is still hard for the user to navigate through, in and out, because of its branching paths, which are scattered in multiple directions.

10. CONCLUSION

Data mining is a vast, yet an emerging, computer science field. Widely vast and encompassing many other subfields such as *web mining* and *text mining*, and overlaps with fields such as such as *text mining, machine learning, fuzzy logic, probabilistic reasoning,* and *computational intelligence*. Data mining and Visualization have developed a lot since its inception from hundreds of years ago. Many application of data mining have gotten a huge adaption and user base. Google, the internet giant is one of the main adaptors of data mining. Data visualization today is helping to solve many engineering and scientific problems in ways that were unimaginable before, such as Map-Reduce algorithms.

Future trends in data mining and visualization are becoming more apparent with every new introduction of newer data mining solutions and data visualizations tools. Most of such advancements are based on AI, such tools aims at human like intelligence. The aim is of is to replace the human factor in decision-making. Data conditioning is a promising solution to the amounts of data that is on the rise, which is of need to be mined. For visualization tools, interactivity is the new wave, allowing users to touch, rotate, and select how to view data sets on its wake.

Future work includes more investigation on the current challenges facing the development and wide spread use of data mining and visualization techniques. The current emerging automated data conditioning tools that provide a more effective dataset to be processed by the data mining tools had an enormous impact on how data mining and visualization tools are designed. With those emerging tools in mind, data mining and visualization tools can achieve more than accurate results than before. Also more work should be done in terms of the current ethical issues associated with data mining and visualization techniques, namely the anonymization problem associated with the privacy concerns of the general public. How to ethically sift through data records without harming or breaching others privacy.

Today, most leading enterprises and organizations depend heavily on such tools for decision support, and business intelligence. Finally, it is clear now how data mining and visualization tools are essential in the knowledge discovery process, and they have an enormous impact on businesses and the research facilities. While both are separate and have their respectful principals and methods, their integration is eminent for the benefit of both fields.

11. REFERENCES

[1] D. Alexander "Data Mining" Internet: http://www.laits.utexas.edu/~norman/BUS.FOR/course.mat/Alex/, [Mar. 11, 2012].

[2] B. Palace, “Data Mining,” Internet: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm, spring 1996 [Feb. 25, 2012].

[3] M. Sousa, M. Mattoso, and N. Ebecken “Data mining: a database perspective,” In Proc. *International Conference on Data Mining*, 1998, pp.413-431.

[4] G. Dennis Jr, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki “DAVID: Database for Annotation, Visualization, and Integrated Discovery,” *Genome Biology*, vol. 4, pp.3-14, August 2003.

[5] V. Friedman, “Data Visualization: Modern Approaches,” Internet: http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches, Aug. 2, 2007 [Mar. 12, 2012].

[6] R. Mikut, and M. Reischl “Data mining tools” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 1, pp.431-443 , September/October 2011.

[7] D. Tegarden, “Business Information Visualization,” *Communications of AIS,* vol. 1, January 1999.

[8] S. Few, “Human Perception,” Internet: http://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html, Sept. 16, 2010 [Mar. 16, 2012].

[9] F. Post, G. Nielson, and G. Bonneau “Data Visualization: the State of the Art,” United States of America: Springer, 2002, pp.464.

[10] G. Grinstein, and B. Thuraisingham, “Data Mining and Data Visualization” in Proc. of *the IEEE Visualization '95 Workshop on Database Issues for Data Visualization*, October 1995, pp.54-56.

[11] D. Keim “Visual Techniques for Exploring Databases,” *International Conference on Knowledge Discovery in Databases (KDD '97)*, California, USA, August 1997.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” in AI Magazine, *American Association for Artificial Intelligence AAAI*, vol. 17, pp. 37-54, Fall 1996.

[13] M. Friendly “A Brief History of Data Visualization,” *Handbook of Computational Statistics: Data Visualization*, vol.2, pp. 15-56, 2008.

[14] E. Tufte, “The Visual Display of Quantitative Information,” *Cheshire, CT: Graphics Press*, 1986, pp.200.

[15] S. Allen, “The Value of Many Eyes,” Internet: www.interactiondesign.sva.edu/classes/datavisualization/updates, Jul. 29, 2010 [Apr. 1, 2012].

[16] P. Kochevar, “Database Management for Data Visualization,” *Database Issues for Data Visualization*, vol.871, pp.107-117, 1994.

[17] V. Friedman, “Data Visualization and Infographics,”Internet: http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics, Jan. 14, 2008 [Jan. 14, 2008].

[18] B. Alpern, and L. Carter “Hyperbox,” in Proc. of IEEE Conference on Visualization ‘91, October 1991, pp. 133-139.

[19] M. Ferreira de Oliveira, “From visual data exploration to visual data mining: a survey,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 378-394, July-September 2003.

[20] C. Romero, and S. Ventura “Educational data mining: A survey from 1995 to 2005,” *Expert Systems with Applications*, vol.33 (2007) pp. 135–146, 2007.

[21] L. Kurgan and P. Musilek, “A survey of Knowledge Discovery and Data Mining process models,” *The Knowledge Engineering Review,* vol. 21, pp. 1- 24, March 2006.

[22] Q. Zhang and R. Segall, “Web Mining: A Survey of Current Research, Techniques, and Software,” *International Journal of Information Technology & Decision Making,* vol.7, pp.683-720, December 2008.

[23] G. Pavlopoulos, A. Wegener, and R. Schneider, “A survey of visualization tools for biological network analysis,” *BioData Mining,* vol.1, pp.12, November 2008.

[24] N. Raghavan, “Data mining in e-commerce: A survey,” *SADHANA Academy Proceedings in Engineering Sciences*, vol.30, pp.275-289, April-June 2005.

[25] B. Gaddam, D. Ghosh, N. Ahmed, S. Donepudi, and V. Khadilkar, “Computational Intelligence in Data Mining,” Internet: http://www.cs.lamar.edu/faculty/disrael/COSC5100/ComputationalIntelligenceInDataMining.pdf, [Apr. 1, 2012].

[26] A. Berson, S. Smith, and K. Thearling, “An Overview of Data Mining Techniques,” Excerpted from the book *Building Data Mining Applications for CRM*, McGraw Hill: USA, 1999, pp.488.

[27] D. Keim “Pixel-oriented Visualization Techniques for Exploring Very Large Databases,” *Journal of Computational and Graphical Statistics*, vol. 5, pp. 58-77, March 1996.

[28] D. Keim, and H. Kriegel “Visualization Techniques for Mining Large Databases: A Comparison” *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, pp.923-938, December 1996.

[29] D. Asimov, “The Grand Tour: A Tool for Viewing Multidimensional Data,” *SIAM Journal of Science & Statistical Computing,* vol. 6, pp. 128-143, 1985.

[30] M. Oliveira, and H. Levkowitz “From Visual Data Exploration to Visual Data Mining: A Survey,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 378 - 394, July-September 2003.

[31] Information Management, “Charting information management how your business works,” Internet: www.information-management.com/media/ui/mk2010.pdf, 2010 [Apr. 1, 2012].

[32] S. Casner “A Task-Analytic Approach to the Automated Design of Graphic Presentations,” *ACM Transactions on Graphics*, vol.10, pp.111–151, April 1991.

[33] D. Plonka, “FlowScan - Network Traffic Flow Visualization and Reporting Tool,” *14th Systems Administration Conference (LISA 2000),* New Orleans, Louisiana, USA, December 3– 8, 2000, pp. 305-317.

[34] M. Marwah, R. Sharma, R. Shih, C. Patel, V. Bhatia, M. Mekanapurath, R. Velumani, and S. Velayudhan, “Visualization and Knowledge Discovery in Sustainable Data Centers,” Compute 2009 ACM Bangalore Chapter Compute, Bangalore, India, January 2009.

[35] MAIA Intelligence, “Business Intelligence in Manufacturing”, 2009, Internet: www.maia-intelligence.com, 2008 [Apr. 1, 2012].

[36] M. Ester, H. Kriegel, and J. Sander “Spatial Data Mining: A Database Approach” Advances in Spatial Databases, vol. 1262, pp47-66, 1997.

[37] M. Erwig, and R. Gueting, “Explicit Graphs in a Functional Model for Spatial Databases,” *IEEE Transactions on Knowledge and Data Engineering*, vol.6, pp.787-803, October 1994.

[38] K. Zeitouni, “A Survey of Spatial Data Mining Methods Databases and Statistics Point of Views,” *Information Resources Management Association International Conference IRMA 2000, Data Warehousing and Mining, Anchorage*, Alaska. pp. 229-242

[39] R. Geary, “The Contiguity Ratio and Statistical Mapping,” *Incorporated Statistician*, vol. 5, pp. 115-145, 1954.

[40] S. Chaudhuri, and V. Narasayya, “New Frontiers in Business Intelligence” *The 37th International Conference on Very Large Data Bases*, Seattle, Washington, pp.1502-1503.

[41] A. Mocanu, D. Litan, S. Olaru, and A. Munteanu “Information Systems in the Knowledge Based Economy” *WSEAS Transactions on Business and Economics*, vol. 7, pp.11-21, January 2010.

[42] T. Ramakrishnan, M. Jones, and A. Sidorova, “Factors influencing business intelligence (bi) data collection strategies: An empirical investigation,” *Decision Support Systems*, vol. 52, pp. 486–496, January 2012.

[43] M. Fitzsimons, T. Khabaza, and C. Shearer, "The Application of Rule Induction and Neural Networks for Television Audience Prediction," *In Proceedings of ESOMAR/EMAC/AFM Symposium on Information Based Decision Making in Marketing*, Paris, November 1993, pp. 69-82.

[44] M. Hearst, "What Is Text Mining?" Internet: http://people.ischool.berkeley.edu/~hearst/text-mining.html, Oct. 17, 2003 Oct. 17, 2003 [May 2, 2012].

[45] K. Cohen KB, L. Hunter, "Getting Started in Text Mining," *Public Library of Science* PLOS, vol. 4, pp.20-22, January 2008.

[46] F. Facca, and P. Lanzi "Mining interesting knowledge from weblogs: a survey," *Data & Knowledge Engineering,* vol.53, pp. 225–241, 2005.

[47] A. Abraham, "Business Intelligence from Web Usage Mining," *Journal of Information & Knowledge Management*, vol. 2, pp. 375-390, 2003.

[48] IBM, "SPSS", Internet: http://www-01.ibm.com/support/docview.wss?uid=swg21506855, [Apr. 1, 2012].

[49] IBM, "SurfAid Analytics", Internet: http://surfaid.dfw.ibm.com, [Apr. 1, 2012].

[50] Oracle, "Oracle Data Miner 11g Release 2," Internet: http://www.oracle.com/technetwork/database/options/odm/dataminerworkflow-168677.html, Jan. 2012 [Apr. 1, 2012].

[51] SAS, "SAS Enterprise Miner," Internet: http://www.sas.com/technologies/analytics/datamining/mine, Sept. 2, 2010 [Apr. 15, 2012].

[52] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *Special Interest Group on Knowledge Discovery and Data Mining SIGKDD Explorer News*, vol. 11, pp. 10-18, June 2009.

[53] IBM, "IBM Parallel Visualizer," Internet: www.pdc.kth.se/training/Talks/SMP/.../ProgEnvCourse.htm, Sept. 22, 1998 [Apr. 15, 2012].

[54] Cave5D, "Cave5D Release 2.0," Internet: www.mcs.anl.gov/~mickelso/CAVE2.0.html, Aug. 5, 2011 [Apr. 17, 2012].

[55] W. Hibbard and D. Santek, "the Vis5D System for Easy Interactive Visualization", *Proceedings of IEEE Visualization,* pp 28-35, 1990.

[56] C. Stolte , D. Tang , and P. Hanrahan, "Multiscale Visualization Using Data Cubes," in Proc.*of the IEEE Symposium on Information Visualization (InfoVis'02)*, October 2002, pp. 28-29.

[57] C. Stolte , D. Tang , P. Hanrahan, "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp.52-65, January 2002.

[58] C. Chapman, "50 Great Examples of Data Visualization," Internet: http://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/, 2012 [Mar. 22, 2012].

[59] W. Hedfield "Case study: Jaeger uses data mining to reduce losses from crime and waste," Internet: www.computerweekly.com, 2009 [Apr. 1, 2012].

[60] Inmon W.H., "Building the Data Warehouse," Indiana, USA: J. Wiley&Sons, 1994. pp.576.

[61] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim, and A. Valencic, "Data Modeling Techniques for Data Warehousing," Internet: www.redbooks.ibm.com/redbooks/pdfs/sg242238.pdf, Feb. 1998 [Nov. 16, 2011].

[62] K. Lynn, "Search Fuels Business Intelligence for Decision Making," *TNR Global*, Available at http://www.tnrglobal.com/blog/tag/business-intelligence/, 2004-2012 [Mar. 20, 2012].

[63] L. Greenfield, "The Data Warehousing Information Center," Internet: http://www.dwinfocenter.org/against.html, 1995 [Mar. 8, 2012].

[64] J. Lawyer, and S. Chowdhury, "Best Practices in Data Warehousing to Support Business Initiatives and Needs," In Proc. *37th Annual Hawaii International Conference on System Sciences*, January 2004, pp.9.

[65] S. Badawi, "AI Computer Vision Blog," Internet: blog.samibadawi.com, Mar. 26, 2012 [Apr. 23, 2012].

[66] S.K. Pal, "Soft Data Mining, Computational Theory of Perceptions, and Rough-Fuzzy Approach", Information *Sciences (Special Issue on Soft Computing Data Mining)*, vol. 163, pp.5-12, 2004.

[67] The Global Community of Information Professionals, "What is Information management?" Internet: www.aiim.org/what-is-information-management , 2012 [Apr. 1, 2012]

[68] B. Gaddam, and S. Donepudi, "Computational intelligence," Internet: http://cs.lamar.edu/faculty/disrael/COSC5100/ComputationalIntelligenceInDataMining.pdf, 2005 [Mar. 20, 2012].

[69] K.A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *Columbia Science and Technology Law Review*, vol. 5, December 15, 2003, Available at: http://www.stlr.org/cite.cgi?volume=5&article=2
Retrieved on 1st of April 2012

[70] Carlos Rodríguez, Florian Daniel, Fabio Casati, Cinzia Cappiello "Toward Uncertain Business Intelligence: The Case of Key Indicators" *IEEE Internet Computing*, vol.14, pp.32-40, July-Aug. 2010.

[71] D. A. Keim, C. Panse, and M. Sips "Information Visualization: Scope, Techniques and Opportunities for Geovisualization" *Exploring Geovisualization*, pp.23-52, June 27, 2005. Available at http://bib.dbvis.de/uploadedFiles/124.pdf

[72] T. M. Lehtima□ki, K. Sa□a□skilahti, M. Kowiel, and T. J. Naughton, "Displaying Digital Holograms of Real-World Objects on a Mobile Device using Tilt-Based interaction" *9th Euro-American workshop on Information Optics (WIO)*, pp.1-3, July 2010.

[73] Zebra Imaging, "Motion Displays," Internet: http://www.zebraimaging.com/products/motion-displays/, 2010 [Apr. 9, 2012].

[74] Z Space, "The ZSpace Experience," Internet: http://zspace.com/about-zspace/, 2012 [Apr. 9, 2012].

[75] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough, "SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny," *Nucleic Acids Research*, vol. 37, pp.380-386, December 2009.

[76] Van den Berg, J. P. "A literature survey on planning and control of warehousing systems" *IIE Transactions*, vol. 31, pp.751–762, 1999.

[77] O. Grabova, J. Darmont, J. Chauchat, and I. Zolotaryova; "Business Intelligence for Small and Middle-Sized Enterprises," in the *Special Interest Group on Management of Data SIGMOD Record*, vol. 39, pp. 39-50, December 2010.