INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ETL TOOLS IN DATA MINING A REVIEW

THIRUMAGAL R[1], SUGANTHY R[2], MAHIMA S[3], KESAVARAJ G[4]

[1]*M.Phil(Department Of Computer Science)Vivekanandha College of Arts and Sciences for Women,Namakkal, thirumagalmca@gmail.com*
[2] *M.Phil(Department Of Computer Science)Vivekanandha College of Arts and Sciences for Women,Namakkal,, rsuganthy@live.com*
[3] *Assistant Professor,Vivekanandha College of Arts and Sciences for Women,Namakkal,, blessie_john@yahoo.com*

[4]*Assistant Professor,Vivekanandha College of Arts and Sciences for Women,Namakkal,, kesavaraj2020@gmail.com*

Abstract

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. 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 tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. This paper presents an overview of the data mining tools like WEKA, ETL, Spatial ETL.

Keywords: Data mining, WEKA, ETL, Spatial ETL, Dashboards, Text Mining

1. Introduction

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, **data mining is actually part of the knowledge discovery process.** The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process. In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

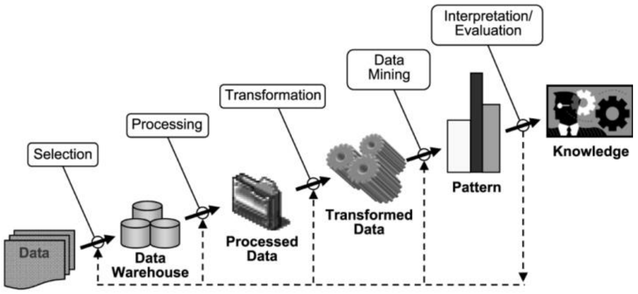

Figure 1: Data mining is the core of Knowledge discovery process

Organizations that wish to use data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products and systems as they are brought online, or they can build their own custom mining solution. For instance, feeding the output of a data mining exercise into another Computer system, such as a neural network, is quite common and can give the mined data more value. This is because the data mining tool gathers the data, while the second program (e.g., the neural network) makes decisions based on the data collected. Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. This paper presents an overview of the data mining tools available. For example – WEKA, ETL, Spatial ETL,and it briefly discuss about the nature of ETL tools developed by various concerns.

2. Basic data mining tools

Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and Text-mining tools. Below is a description of each.

2.1 Traditional Data Mining Tools

Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight Trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

2.2. Dashboards

Installed in computers to monitor information in a database, dashboards reflect data changes and updates Onscreen —often in the form of a chart or table —enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

2.3. Text-mining Tools

The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from

Different kinds of text —from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes. When evaluating data mining strategies, companies may Decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although Acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a Dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its Marketing and sales arms, and an interrogation tool so auditors can identify fraud activity.

4. ETL Tool

ETL (**Extract Transform Load**) tools are designed to save time and money by eliminating the need of 'hand-coding' when a new data warehouse is developed. Now a days ETL tool becomes popular. They are also used to facilitate the work of the database administrators who connect different branches of databases as well as integrate or change the existing databases.

The main purpose of the ETL tool is:

- Extraction of the data from legacy sources (usually heterogeneous)
- Data Transformation (data optimized for transaction --> data optimized for analysis)
- Synchronization and Cleansing of the data
- Loading the data into data warehouse.

There are several requirements that must be had by ETL tools in order to deliver an optimal value to users, supporting a full range of possible scenarios.

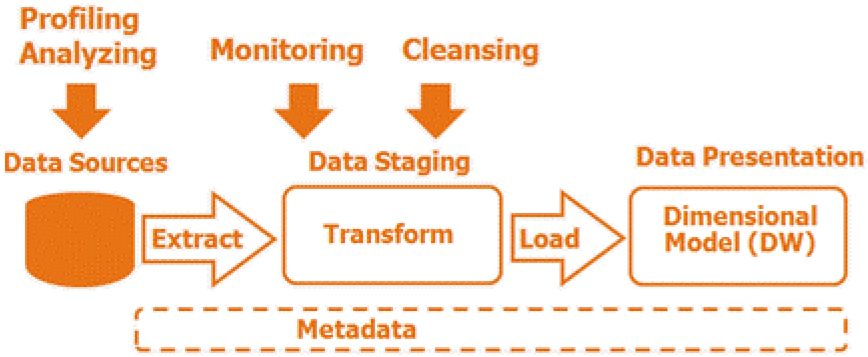

Figure 3:ETL

4.1 Extract

The first part of an ETL process involves extracting the data from the source systems. In many cases this is the most challenging aspect of ETL, since extracting data correctly sets the stage for how subsequent processes go further. In general, the goal of the extraction phase is to convert the data into a single format appropriate for transformation processing. An intrinsic part of the extraction involves the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data may be rejected entirely or in part.

4.2 Transform

The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources require very little or even no manipulation of data.

4.3 Load

The load phase loads the data into the end target, usually the data warehouse (DW). Depending on the requirements of the organization, this process varies widely. As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

- For example, a financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all this data and consolidate it into a uniform presentation, such as for storing in a database or data warehouse.

4.4 Real-life ETL cycle

The typical real-life ETL cycle consists of the following execution steps:

1. Cycle initiation
2. Build reference data
3. Extract (from sources)
4. Validate
5. Transform (clean, apply business rules, check for data integrity, create aggregates or disaggregates)
6. Stage (load into staging tables, if used)
7. Audit reports (for example, on compliance with business rules. Also, in case of failure, helps to diagnose/repair)
8. Publish (to target tables)
9. Archive
10. Clean up

5. ETL tools comparison criteria

The information provided below lists major strengths and weaknesses of the most popular ETL vendors.

5.1 IBM (Information Server Infosphere platform)

Advantages:

- strongest vision on the market, flexibility
- progress towards common metadata platform
- high level of satisfaction from clients and a variety of initiatives

Disadvantages:

- difficult learning curve
- long implementation cycles
- became very heavy (lots of GBs) with version 8.x and requires a lot of processing power

5.2 Informatica Power Center

Advantages:

- most substantial size and resources on the market of data integration tools vendors
- consistent track record, solid technology, straightforward learning curve, ability to address real-time data integration schemes
- Informatica is highly specialized in ETL and Data Integration and focuses on those topics, not on BI as a whole
- focus on B2B data exchange

Disadvantages:

- several partnerships diminishing the value of technologies
- limited experience in the field.

5.3 Microsoft (SQL Server Integration Services)

Advantages:

- broad documentation and support, best practices to data warehouses
- ease and speed of implementation
- standardized data integration
- real-time, message-based capabilities
- relatively low cost - excellent support and distribution model

Disadvantages:

- Problems in non-Windows environments. Takes over all Microsoft Windows limitations.
- unclear vision and strategy

5.4 Oracle (OWB and ODI)

Advantages:

- based on Oracle Warehouse Builder and Oracle Data Integrator – two very powerful tools;
- tight connection to all Oracle data warehousing applications;
- Tendency to integrate all tools into one application and one environment.

Disadvantages:

- focus on ETL solutions, rather than in an open context of data management;
- tools are used mostly for batch-oriented work, transformation rather than real-time processes or federation data delivery;
- long-awaited bond between OWB and ODI brought only promises - customers confused in the functionality area and the future is uncertain

5.5 SAP Business Objects (Data Integrator / Data Services)

Advantages:

- integration with SAP
- SAP Business Objects created a firm company determined to stir the market;
- Good data modeling and data-management support;
- SAP Business Objects provides tools for data mining and quality; profiling due to many acquisitions of other companies.
- Quick learning curve and ease of use

Disadvantages:

- SAP Business Objects is seen as two different companies
- Uncertain future. Controversy over deciding which method of delivering data integration to use (SAP BW or BODI).
- Business Objects Data Integrator (Data Services) may not be seen as a stand-alone capable application to some organizations.
- Costly

5.6 Sun Microsystems

Advantages:

- Data integration tools are a part of huge Java Composite Application Platform Suite - very flexible with ongoing development of the products
- 'Single-view' services draw together data from variety of sources; small set of vendors with a strong vision

Disadvantages:

- relative weakness in bulk data movement
- limited mindshare in the market
- support and services rated below adequate

6. ETL Software implementation in Parallel processing

A recent development in ETL software is the implementation of parallel processing. This has enabled a number of methods to improve overall performance of ETL processes when dealing with large volumes of data.

ETL applications implement three main types of parallelism:

- **Data**: By splitting a single sequential file into smaller data files to provide parallel access.
- **Pipeline**: Allowing the simultaneous running of several components on the same data stream. For example: looking up a value on record 1 at the same time as adding two fields on record 2.
- **Component**: The simultaneous running of multiple processes on different data streams in the same job, for example, sorting one input file while removing duplicates on another file.

All three types of parallelism usually operate combined in a single job.

An additional difficulty comes with making sure that the data being uploaded is relatively consistent. Because multiple source databases may have different update cycles (some may be updated every few minutes, while others may take days or weeks), an ETL system may be required to hold back certain data until all sources are synchronized. Likewise, where a warehouse may have to be reconciled to the contents in a source system or with the general ledger, establishing synchronization and reconciliation points becomes necessary.

7. Spatial ETL Tool

Spatial ETL tools provide the data processing functionality of traditional Extract, Transform, Load (ETL) software, but with a primary focus on the ability to manage spatial data (which may also be called *geographic*, *map* or *location* data). A Spatial ETL system may translate data directly from one format to another, or via an intermediate format; the latter being more common when transformation of the data is to be carried out. The following figure 4 presents the Spatial ETL.

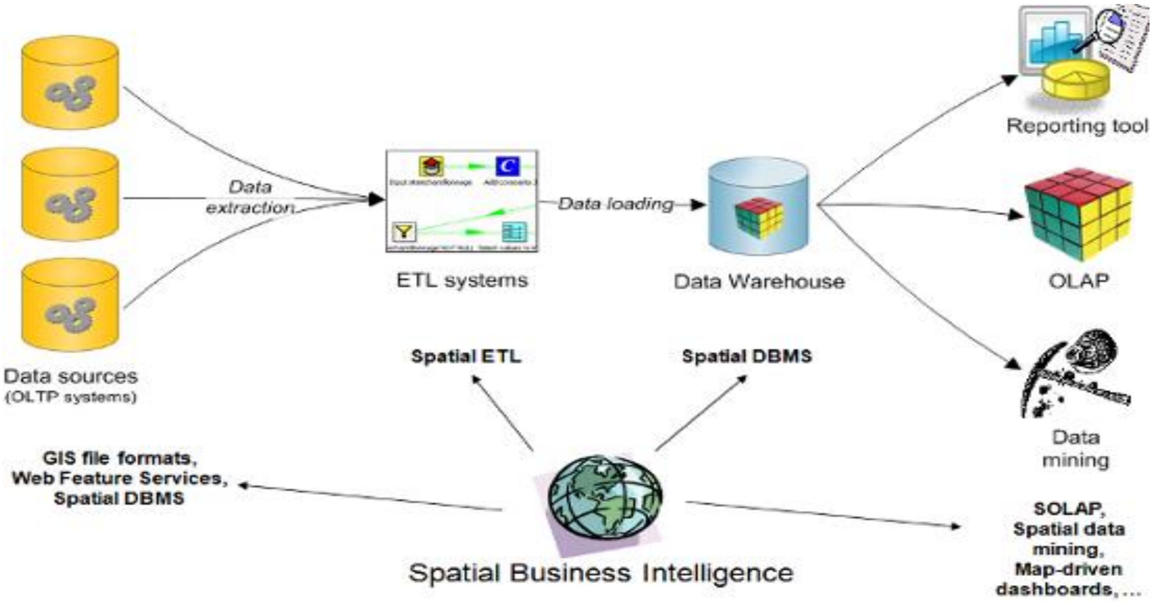

Figure 4:Spatial ETL

Spatial ETL Uses

Spatial ETL has a number of distinct uses to which it is put.

- Data cleanup: The removal of errors within a dataset
- Data Merging: The bringing together of multiple datasets into a common framework - Conflation is a good example of this
- Data verification: The comparison of multiple datasets for verification and quality assurance purposes
- Data translation: Conversation between different data formats.

8. Conclusion

Data mining is popular in the science and mathematical fields but also is utilized increasingly by marketers trying to distill useful consumer data from Web sites. Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, and the web. ETL systems are commonly used to integrate data from multiple applications, typically developed and supported by different vendors or hosted on separate computer hardware. ETL tools Pull large volumes of data from different sources, in different formats, restructure them and load into a warehouse.

REFERENCES

1.Nisbet, R, J. Elder, and G. Miner. 2009. The Handbook of Statistical Analysis & Data Mining Applications. Academic Press (Elsevier). Burlington, MA.

2.R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang,and C.-J. Lin. LIBLINEAR: A library for large linearclassification.Journal of Machine Learning. Research,9:1871–1874, 2008.

3.E. Frank and S. Kramer. Ensembles of nested di-chotomies for multi-class problems. InProc 21st International Conference on Machine Learning, Banff,Canada, pages 305–312. ACM Press, 2004.

4. Olga Fedotova ; Gladys Castillo ; Leonor Teixeira ; Helena Alvelos,International Journal of Engineering Science and Technology, 2011, Vol.3(7), p.6064.

5.S. Celis and D. R. Musicant. Weka-parallel: machinelearning in parallel. Technical report, Carleton College, CS TR, 2002.

6. C. Shearer. The CRISP-DM model: The new blueprint for data mining.Journal of Data Warehousing, 5(4), 2000.

7.Topsy. N.p., n.d. Web. 20 June 2013.www.//topsy.com/s/cloveretl

8. Roy, Krishna. "Javlin Elucidates CloverETL Strategy as It Continues to Take Aim at Data Integration." MIS Impact Report (2013): 1–4.

9. "GoodData Selects CloverETL to Enrich Data Integration – GoodData." GoodData. N.p., 6 December 2012. Web. 20 June 2013. www.gooddata.com/in-the-news/gooddata-selects-cloveretl-to-enrich-data-integration.

10. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

11. Fu, P., and J. Sun. 2010. *Web GIS: Principles and Applications*. ESRI Press. Redlands, CA. ISBN 1-58948-245-X.