SPECIALIST READING

A Find the answers to these questions in the following text.

1 What tool is often used in data mining?
2 What AI method is used for the following processes?
   a Separate data into subsets and then analyse the subsets to divide them into further subsets for a number of levels.
   b Continually analyse and compare data until patterns emerge.
   c Divide data into groups based on similar features or limited data ranges.
3 What term is used for the patterns found by neural networks?
4 When are clusters used in data mining?
5 What types of data storage can be used in data mining?
6 What can an analyst do to improve the data mining results?
7 Name some of the ways in which data mining is currently used.

DATAMINING

Data mining is simply filtering through large amounts of raw data for useful information that gives businesses a competitive edge. This information is made up of meaningful patterns and trends that are already in the data but were previously unseen.

The most popular tool used when mining is artificial intelligence (AI). AI technologies try to work the way the human brain works, by making intelligent guesses, learning by example, and using deductive reasoning. Some of the more popular AI methods used in data mining include neural networks, clustering, and decision trees.

Neural networks look at the rules of using data, which are based on the connections found or on a sample set of data. As a result, the software continually analyses value and compares it to the other factors, and it compares these factors repeatedly until it finds patterns emerging. These patterns are known as rules. The software then looks for other patterns based on these rules or sends out an alarm when a trigger value is hit.

Clustering divides data into groups based on similar features or limited data ranges. Clusters are used when data isn't labelled in a way that is favourable to mining. For instance, an insurance company that wants to find instances of fraud wouldn't have its records labelled as fraudulent or not fraudulent. But after analysing patterns within clusters, the mining software can start to figure out the rules that point to which claims are likely to be false.

Decision trees, like clusters, separate the data into subsets and then analyse the subsets to divide them into further subsets, and so on (for a few more levels). The final subsets are then small enough that the mining process can find interesting patterns and relationships within the data.

Once the data to be mined is identified, it should be cleansed. Cleansing data frees it from duplicate information and erroneous data. Next, the data should be stored in a uniform format within relevant categories or fields. Mining tools can work with all types of data storage, from large data warehouses to smaller desktop databases to flat files. Data warehouses and data

Data stores
You must first have data to mine. Data stores include one or several databases or data warehouses.

Cleanse data
Data must be stored in a consistent format and free from errors and redundancies.

Search data
Actual mining occurs when data is combed for patterns and trends. Rules for patterns are noted.

Analyse reports
Someone must analyse mining results for validity and relevance.

Report findings
The mining results can then be reviewed and interpreted, and a plan of action determined.

marts are storage methods that involve archiving large amounts of data in a way that makes it easy to access when necessary.

When the process is complete, the mining software generates a report. An analyst goes over the report to see if further work needs to be done, such as refining parameters, using other data analysis tools to examine the data, or even scrapping the data if it's unusable. If no further work is required, the report proceeds to the decision makers for appropriate action.

The power of data mining is being used for many purposes, such as analysing Supreme Court decisions, discovering patterns in health care, pulling stories about competitors from newswires, resolving bottlenecks in production processes, and analysing sequences in the human genetic makeup. There really is no limit to the type of business or area of study where data mining can be beneficial.

[Adapted from 'Data Mining for Golden Opportunities', Smart Computing Guide Series Volume 8 Issue 1, January 2000]

B Re-read the text to find the answers to these questions.

1 Match the terms in Table A with the statements in Table B.

Table A

a	Data mining	c	Cleansed data
b	AI	d	Data warehouse

Table B

i Storage method of archiving large amounts of data to make it easy to access
ii Data free from duplicate and erroneous information
iii A process of filtering through large amounts of raw data for useful information
iv A computing tool that tries to operate in a way similar to the human brain

2 Mark the following as True or False:

a Data mining is a process of analysing known patterns in data.
b Artificial intelligence is commonly used in data mining.
c In data mining, patterns found while analysing data are used for further analysing the data.
d Data mining is used to detect false insurance claims.
e Data mining is only useful for a limited range of problems.

3 Complete the following description of the data mining process using words from the text:

Large amounts of data stored in data ..... are often used for data ..... . The data is first ..... to remove ..... information and errors. The ..... is then analysed using a tool such as .... An analysis report is then analysed by an ..... who decides if the ..... need to be refined, other data ..... tools need to be used, or if the results need to be discarded because they are ..... . The analyst passes the final results to the ..... makers who decide on the ..... action.