CCEA
Rewarding Learning

FACTFILE: GCE DIGITAL TECHNOLOGY

UNIT A21: INFORMATION SYSTEMS

Data Mining

Learning Outcomes

Students should be able to:

- Explain what is meant by data mining;
- Describe how digital technology can be used in data mining to gather, store, process and analyse large volumes of data; and
- Explain the importance of big data in the operation and competitiveness of organisations in health, finance and retail sectors.

Content in Data Mining

- What is data mining?
- What is big data?
- Digital technology in data mining.
- The importance of big data to organisations.
- Questions.

What is data mining?

Data mining refers to the process of analysing large data sets (Big Data) with a view to discovering patterns and trends that go beyond simple analysis. It combines the application of artificial intelligence, statistics and database systems in the analysis of groups of structured and unstructured data sets which prove difficult to analyse using traditional means. The main aim of data mining is to extract information from a data set and transform it into an appropriate format for future use.

Patterns such as those identified above are then presented as a summary of the input data and can be used for further analysis, e.g. they may be input into decision support systems to support predictions for the future. The data mining process stops at the process of pattern extraction and all other activities carried out after this point are not considered part of the data mining process.

What is big data?

Big data is a term associated with data sets that are so complex that traditional database (such as RDBMS's) and other processing applications are unable to capture, curate (the process of organising data from a range of data sources), manage and process them within an acceptable time frame.

Big data challenges can be defined as the 3V's:

- Volume – the amount of data to be processed
- Variety – the number of types of data to be analysed
- Velocity – the speed of data processing

As such, big data can be defined as “high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making”, Gartner 2001.

The use of digital technology in data mining

Big Data Gathering
Everyday consumers make around 11.5 million payments using PayPal, 7000 tweets are made on Twitter every second, 136000 updates are posted on Facebook every minute! Social media is one of the biggest sources of Big Data. Consumer good companies actively scan social media websites to decipher user preferences, choices and perceptions towards their brands. Retailers, healthcare organisations, financial services companies all now utilise social media for brand building.

We have more and more data available to use today and data sets grow rapidly, mainly due to the increase in the use of mobile phones, sensing devices, software logs, cameras, RFID readers and other similar technologies As such, what is considered to be big data will depend more on the capabilities of the user and the applications at their disposal. Digital technology allows us to collect data for further analysis using methods such as online forms, mobile phone data transmissions, email data, stock market data, market research, personal digital assistants, smart phones, tablets, net-books and many other such technologies.

Data sources can be categorised into internal and external. The internal data includes sources such as customer details, product details, sales data etc. External sources include data collected from business partners, data suppliers, internet, government and market research companies. In essence the commonly used data sources are:

- social media;
- machine data – data generated from devices such as RFID chip readers, GPS results; and
- transactional data – data generated from companies such as eBay, Amazon and large stores such as Tesco’s.

Big Data Storage
While Big Data cannot be measured in terms of the amount of data, for many organisations the term big data may relate to the need to store data sets in the cumulative range of terabytes to many petabytes of data. The key requirements of big data storage therefore are that it can handle very large amounts of data and keep scaling to keep up with the growth of data sets, in addition to being able to provide high speed Input/Output operations necessary to support the delivery of data analytics as they are carried out.

Big data practitioners such as Google, Facebook etc all run what are known as hyperscale computing environments which consist of a vast number of servers with Direct Attached Storage (DAS). Each unit will generally have PCIe flash storage devices to support data storage and high speed access to data sets.

Smaller organisations can support the storage of big data through the use of clustered Network Attached Storage (NAS) devices. This is file access shared storage which can easily be scaled out to meet the increased capacity or computing requirements required for big data analysis. As NAS systems scale outwards they can be difficult to manage as they tend to span out in a hierarchical manner (many devices with many folders within folders).

Object-based storage systems offer an alternative to NAS devices and the problems it can lead to. Each file stored in a object-based storage system will be given its own unique identifier and index to support high speed access to a particular data file or data set.

Big Data Processing and Analysis
Big data processing techniques analyse data sets at terabyte or even petabyte scale. Here we will look briefly at some big data processing techniques applied during data mining.

Some of the methods of processing applied to big data include:

- Cluster analysis – where groups of data records are identified;
- Classification – where the data mining process is used to determine an appropriate structure to new data, in the way for example an email application may classify some emails as spam;
- Anomaly detection – where unusual records are identified. Such anomalies may merit further investigation as a point of interest to the organisation or they may be representative of data errors;
- Association rule mining and sequential pattern

mining – where dependencies between data items can be identified, for example the use of data sets by a supermarket to determine which patterns of products are purchased together;

- Regression – where relationships between data variables are investigated to help how a change in an independent variable can impact upon a dependant data variable; and

- Summarisation – where data is summarised in a visual format .

The importance of big data to organisations

Big data can help companies gain insight into potential revenue increases or help companies determine how the can gain or retain customers and improve operations.

How the financial services sector uses big data

Data is collected continually about each of us by members of the financial sector; when we become customers of a bank, when we telephone a call centre or make financial transactions of any kind for example. Customers no longer need to visit branches to withdraw or deposit money or even make investments, and purchases are made with debit or credit cards and even through mobile devices. Each interaction generates data and since customer profiles and transaction patters change rapidly, financial organisations need a reliable means of analysing data. Many banks now even make use of social media interactions to gauge customer feedback on new products or TV advertisements.

The financial sector includes banking services and even life and general insurance services and has been actively using big data platforms over the past few years with some of their key objectives being:

- Ensuring they are complying with regulations – using traditional data processing platforms to support this objective has become increasingly expensive and unsustainable. For example rather than using traditional means to check customer names and aliases against a customer blacklist, complex fuzzy matching can be applied on name matching and contact information across much larger data sets at a much lower cost.

- Improving risk analysis – complex algorithms can be run on large volumes of transaction data to help identify fraudulent activity or to perform risk analysis . Live data streams can be processed to support trading decisions and to support trending and forecasting on the stock market for example.

- Understanding customer behaviour and transaction patterns – customer data can be consolidated from a variety of sources and analysed to predict customer spending, mortgage defaults.

- Improving Services – customer data can be analysed to help identify patterns which can lead to customer dissatisfaction. Email content, telephone call recordings and social media comments can be analysed to determine positive / negative feelings of customers towards products on offer by the organisation.

How the health sector uses big data

Big data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. More and more data is being collected about patients and can be used to help pick up early warning signs of serious illness at stage when treatment may be simpler and cheaper than if it had not been spotted until later.

Smart phones which enable us to measure steps walked, diet and sleep patterns can now also allow us to upload our data for compilation and tracking against other user. In the near future you would anticipate being able to share this data with your GP who can use it as part of their diagnostic toolbox to support diagnosis. Some health organisations take data from various sources (health and insurance records, sensors worn by the individual, genetic data and even social media content) and use it to tailor health care programs for their clients. Such a high volume of data can also be analysed alongside other clients to help identify health patterns and threats using sophisticated modelling processes while on an individual level the treatment programmes can be developed based on reliable and real-time data collected with regards the patients genetic makeup and lifestyle.

The way we visit and interact with doctors is also likely to change in the future. Telemedicine allows patients to receive treatment remotely using internet connections. All of these interactions leave a data trail which can be analysed to provide insight into trends in health and healthcare provision. Big data can also be used to support clinical trials to allow researchers to select the best subjects. Data sharing arrangements following medical trials can even be used to support breakthroughs in cures for illness while mobile phone location data was used to track population movement and predict the spread of the Ebola virus in Africa a number of years ago, thus providing medical care providers with insight into the best areas to locate treatment centres and to support.

How the retail sector uses big data
Retail organisations are using big data to help understand customers' behaviour in an attempt to match customer needs to products made available by the retail store. Big data is applied at every stage of the retail process and can be used to predict trends, forecast demand, optimising price and identifying those customers most likely to be interested in new products.

- Predicting trends and forecasting demand - retailers today make user of a wide range of data to help identify trends in products. Trend forecasting algorithms comb social media posts and ad-buying data is analysed to help determine what products will be pushed forward by marketing organisations.

- Price optimisation - Once there is an understanding of the products people are interested in buying, retailers are able to use big data to determine where the demand will be. Through the analysis of demographic data and economic indicators, spending habits of customers can be identified while algorithms which track millions of transactions every day can be used to track demand against inventory and competitor activity to ensure a retailer can respond quickly to real time changes in market activity.

- Identifying potential customer - Data collected through transactional records and loyalty programs allows demand to be forecast on the basis of geographic areas.

Questions

1. Explain what is the meant by the term big data? [6]

2. Define the term data mining. [3]

3. Describe the problems associated with the gathering and storing big data for further analysis by any organisation. [9]

4. Data analytics can be used to support the processing of big data. Describe four methods of processing applied to the analysis of big data. [8]

5. Research the topic of big data in health care. Describe in detail how big data analysis has improved health care across the world. [6]

6. Identify three methods used to store data appropriately to support big data analysis. [9]