Data Mining
Big Data's Increasing Challenge and Payoff

For Teachers

Introduction

This video explores the processes of data mining. Increases in the processing power of computers available to business and government have enabled the detailed statistical analysis of huge data stores. The video presents local experts explaining the processes involved in data mining including the collection, warehousing, purification and analysis of useful data. Doug Campbell from De Loitte Analytics, Anthony Ugoni Head of Analytics from the National Australia Bank and Peter O’Hanlon from the Institute of Analytics Professionals, and others explain the processes and social implications of this technology using examples from business, border protection, traffic signalling, queuing processes, fraud detection, law enforcement and others.

Timeline

00:00:00 Data Mining: An Overview
00:05:05 Data
00:09:53 The Uses of Data mining
00:14:40 The Process of Data Mining
00:19:16 Tips for Successful Data Mining
00:21:27 Ethical Concerns
00:23:54 Credits
00:24:49 End Program

Related Titles

Creating Apps 1: Research and Planning
Creating Apps 2: Development and Deployment
Data: Ethical Use and Storage

Recommended Resources

http://www.officemuseum.com/data_processing_machines.htm
http://en.wikipedia.org/wiki/Tree_(data_structure)
http://en.wikipedia.org/wiki/Binary_search_tree
http://www.knime.org/
**Student Worksheet**

**Initiate Prior Learning**

1. How much information about you is stored in computer databases? Try to make a list of as many as you can think of. Compare lists with other students and make a single list. Discuss the implications of your findings with the class.

2. Where and how is data stored and accessed? Find definitions of the following terms and share your results with the class
   
a) Data mining

b) iCloud

c) Cloud storage

d) Data noise

e) Data warehousing
3. Research the history of punched card database sorting methods. Design a punched card database, which can be manually sorted using holes in the cards. Store information about 20 different motorcars on your cards, with at least 8 pieces of data on each card. For example use make, model, fuel consumption, price, accessories, and transmission. You will need 20 cards. Demonstrate your punched card sorting system to solve searches such as all the cars with automatic transmission. What are the obvious disadvantages of a punched card system compared to a computer-based database system?
Active Viewing Guide

1. View the introduction to the video and write a description of the many database systems you may be interacting with during a day.

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

2. What types of data would be collected at the supermarket checkout?

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

3. What types of information might be hidden in the data collected at the supermarket checkout?

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

4. What use could businesses make of the data collected at the supermarket?

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

5. What items would normally be purchased along with disposable nappies? How can this information help a retailer?

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
6. What types of data could be collected by traffic control systems and how could it be used to improve traffic flow?

_________________________________________________________________________________

_________________________________________________________________________________

7. The following quotation comes from The Age newspaper.

“Today, the underground economy is worth up to $20.7 billion of potential tax revenue, and some experts say a big crackdown by the Australian Taxation Office has failed to stem the cash flow.”

How could data mining help to detect tax fraud?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

8. Data mining requires computers to process data using large numbers of simple repeated calculations. Why does this require powerful computers?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

9. What are the three main processes used to create a data warehouse.

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

10. Normally businesses don’t keep data for very long, preferring to keep it up to date so that accounting processes such as billing are accurate. What is the advantage of keeping historical data?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________
11. Why do many data mining projects fail?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

12. What does the term "big data" mean? What are some of the challenges to the data miners in dealing with this data?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

13. How are decision trees used to make decisions?

_________________________________________________________________________________

_________________________________________________________________________________

_________________________________________________________________________________

14. Text is difficult to extract meaning from because computers are not very good at understanding semantics or the meaning of words and phrases in their context. Consider the following sentence: "She rose in the morning and picked the rose". Why would it be difficult to write computer programs that can determine the meaning of the word "rose" in this sentence?

_________________________________________________________________________________

_________________________________________________________________________________
Extension Activities

1. Holidays are often times when people have accidents. Often this is because they do things they don't normally do such as swim, bushwalk or household repairs. Create a survey to find out what types of accident students have had during their lifetime and when they occurred. Enter the information into a relational database such as Microsoft Access with a separate table for each entity, e.g. person, accident and a linking table called person_accident. You will need a many to many relationship between the person and accident because a person can have many accidents and many different persons may have the same accident. What other tables do you think you may need to make useful enquiries such as sorting by the severity of the accident. Do males have more accidents? Which times of the year do accidents tend to occur? What is the most common accident? Report your findings.

2. Research "data purification". What is it? Why is it important? The road transport authority of NSW, Australia, issues 5-year licences. What problems could be caused if someone dies after one year and they are not removed from the database?

3. KNIME is an open source (free to use) data-mining tool. Research "KNIME" on the web. If possible, install it and experiment with its tools. Describe three of its more important features.

4. Data integration structures data into a common format and then organises it into groupings such as indexed tables, hierarchies and tree structures which make searching the data easier. Research relational database tables and tree structures. What is the advantage of using a tree data storage structure? How is a tree structure searched?
Suggested Student Responses

Initiate Prior Learning

1. How much information about you is stored in computer databases? Try to make a list of as many as you can think of. Compare lists with other students and make a single list. Discuss the implications of your findings with the class.
   Students should be able to create fairly extensive combined lists and should be able to see the possible implications in terms of privacy and security and risks of fraud.

2. Where and how is data stored and accessed? Find definitions of the following terms and share your results with the class
   a) Data mining
   b) iCloud
   c) Cloud storage
   d) Data noise
   e) Data warehousing
   f) Data Analysis
   g) Data transactions
   h) Web analytics
   i) Google Analytics
   Students should be able to find working definitions of the terms and may already be using some of the cloud storage solutions such as iCloud without much thought about the implications.

3. Research the history of punched card database sorting methods. Design a punched card database, which can be manually sorted using holes in the cards. Store information about 20 different motorcars on your cards, with at least 8 pieces of data on each card. For example use make, model, fuel consumption, price, accessories, and transmission. You will need 20 cards. Demonstrate your punched card sorting system to solve searches such as all the cars with automatic transmission. What are the obvious disadvantages of a punched card system compared to a computer-based database system?
   Students will need file cards, scissors and hole punching machines. A paperclip can be used for selecting card by passing it through holes along the edges of the cards associated with data types.
Active Viewing Guide

1. View the introduction to the video and write a description of the many database systems you may be interacting with during a day.  
   **Students may list systems such as a transport system, a mobile phone system, traffic control systems, and school role marking system, Retail sales systems and many others.**

2. What types of data would be collected at the supermarket checkout?  
   **Names and product codes of items purchased, quantities of items of each type purchased, credit card details, rewards card ID numbers (which are tied to a database of the information in the customer profile filled out when the customer applied for the cards).**

3. What types of information might be hidden in the data collected at the supermarket checkout?  
   **Rewards cards allow a user's purchasing history and personal details such as address and email address. By participating in these schemes you give permission for the retailer to mine this data and even sell some of the non-personal data on to other businesses such as travel marketing and airline companies.**

4. What use could businesses make of the data collected at the supermarket?  
   **Customers often buy related products. For example customers purchasing airline tickets may be in the market to buy travel luggage or holiday accommodation. They could then be targeted more effectively by email promotions or offers. People who buy luxury items or live in areas with high land or home prices may have higher disposable incomes and be targeted for more the marketing of expensive items.**

5. What items would normally be purchased along with disposable nappies? How can this information help a retailer?  
   **Baby food, baby-cleaning products, wipes and skin protection products. Customers could be offered bundles at reduced prices or the products could be located close to each other on the shelves making customers more likely to purchase products on impulse.**

6. What types of data could be collected by traffic control systems and how could it be used to improve traffic flow?  
   **Number and size of vehicles passing a point, average speeds at any location and time could all be used to see the rate of change of traffic lights or on-ramp flow control lights. Accident rates could be used to detect dangerous sections of road or make sure sufficient tow trucks are available at peak accident times.**

7. The following quotation comes from The Age newspaper.  
   “Today, the underground economy is worth up to $20.7 billion of potential tax revenue, and some experts say a big crackdown by the Australian Taxation Office has failed to stem the cash flow.”
   How could data mining help to detect tax fraud?  
   **By combining data from different government databases it is possible to detect welfare and cash-in-hand tax cheats. See the AUSTRAC website for examples of government attempts to deal with tax avoidance.**

8. Data mining requires computers to process data using large numbers of simple repeated calculations. Why does this require powerful computers?  
   **Each calculation requires a small amount of time but the number of calculations may be in the millions and so requires a fast computer to complete the processes in a practical amount of time.**
9. What are the three main processes used to create a data warehouse.
   Staging
   Integration
   Accessing

10. Normally businesses don't keep data for very long, preferring to keep it up to date so that accounting processes such as billing are accurate. What is the advantage of keeping historical data?
    Historical data can be mined for information about the spending habits of consumers over time and so identify trends such as increasing popularity of products or changes in preferences over time or by age group.

11. Why do many data mining projects fail?
    Lots of irrelevant data described as "noise" can make it difficult to extract useful information at the access stage. A lot of thought needs to go into the type of information that needs to be extracted from the data before it is stored and structured.

12. What does the term "big data" mean? What are some of the challenges to the data miners in dealing with this data?
    Big data refers to the way in which interacting systems produce huge volumes of data when computers exchange data between the different systems. For example an airline check-in will talk to luggage transporting systems, in-flight food catering systems, systems which allocate seating and luggage location on the airline, customs information and security systems, departure tax systems, travel agencies and many other systems which each generate their own data. All of this data can be captured and mined but it requires very complex data mining tools.

13. How are decision trees used to make decisions?
    Decision trees are networks of binary decisions (questions that can be answered yes or no) that can assist (for example) managers of banking accounts. By following the pathway through the tree of decisions that has led to successful sales in the past, they can identify customers who are likely to buy certain financial products, such as home loans.

14. Text is difficult to extract meaning from because computers are not very good at understanding semantics or the meaning of words and phrases in their context. Consider the following sentence: "She rose in the morning and picked the rose". Why would it be difficult to write computer programs that can determine the meaning of the word "rose" in this sentence?
    The meaning of word depends on the structure or grammar of the sentence and on an understanding of the context or previous exposure to the use of the words, which is something we learn as children but is difficult to teach a computer. Searching for meanings, in textual data, which are important to business, is difficult for computers to execute.
Extension Activities

1. Holidays are often times when people have accidents. Often this is because they do things they don't normally do such as swim, bushwalk or household repairs. Create a survey to find out what types of accident students have had during their lifetime and when they occurred. Enter the information into a relational database such as Microsoft Access with a separate table for each entity, e.g. person, accident and a linking table called person_accident. You will need a many to many relationship between the person and accident because a person can have many accidents and many different persons may have the same accident. What other tables do you think you may need to make useful enquires such as sorting by the severity of the accident. Do males have more accidents? Which times of the year do accidents tend to occur? What is the most common accident? Report your findings.

   With help, student should be able to design and construct a simple multi-table database to answer the queries for the data set they have collected.

2. Research "data purification". What is it? Why is it important? The road transport authority of NSW, Australia, issues 5-year licences. What problems could be caused if someone dies after one year and they are not removed from the database?

   Data purification is the process of removing redundant, irrelevant or inaccurate data from a database. It is often a mandatory requirement under government regulation. Failure to keep data up-to-date could lead to prosecution or embarrassment for the business.

3. KNIME is an open source (free to use) data-mining tool. Research "KNIME" on the web. If possible, install it and experiment with its tools. Describe three of its more important features.

   Answers will vary but essentially KNIME is a very sophisticated data-mining tool that provides a graphic user interface to create drag and drop workflows which process and report complex queries and explore relationships between data.

4. Data integration structures data into a common format and then organises it into groupings such as indexed tables, hierarchies and tree structures which make searching the data easier. Research relational database tables and tree structures using the links below. What is the advantage of using a tree data storage structure? How is a tree structure searched?

   Tree structures are used to represent hierarchical data and store data in a way that makes it easily searchable (see binary search trees and tree traversal). Tree structures are searched using special search algorithms that systematically follow each branch of the tree to find data at the "leaves".