Chapter 4

DATA, TEXT, AND WEB MINING

Learning Objectives

- Define data mining and list its objectives and benefits
- Understand different purposes and applications of data mining
- Understand different methods of data mining, especially clustering and decision tree models
- Build expertise in use of some data mining software

Learning Objectives

- Learn the process of data mining projects
- Understand data mining pitfalls and myths
- Define text mining and its objectives and benefits
- Appreciate use of text mining in business applications
- Define Web mining and its objectives and benefits

Data Mining Concepts and Applications

- **Data mining (DM)**

  A process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases

Data Mining Concepts and Applications

- Major characteristics and objectives of data mining
  - Data are often buried deep within very large databases, which sometimes contain data from several years; sometimes the data are cleansed and consolidated in a data warehouse
  - The data mining environment is usually client/server architecture or a Web-based architecture

Data Mining Concepts and Applications

- Major characteristics and objectives of data mining
  - Sophisticated new tools help to remove the information are buried in corporate files or archival public records; finding it involves massaging and synchronizing the data to get the right results.
  - The miner is often an end user, empowered by data drills and other power query tools to ask ad hoc questions and obtain answers quickly, with little or no programming skill

Data Mining Concepts and Applications

- Major characteristics and objectives of data mining
  - Striking it rich often involves finding an unexpected result and requires end users to think creatively
  - Data mining tools are readily combined with spreadsheets and other software development tools; the mined data can be analyzed and processed quickly and easily
  - Parallel processing is sometimes used because of the large amounts of data and massive search efforts

Data Mining Concepts and Applications

- **How data mining works**
  - Data mining tools find patterns in data and may even infer rules from them
  - Three methods are used to identify patterns in data:
    1. Simple models
    2. Intermediate models
    3. Complex models

Data Mining Concepts and Applications

- **Classification**

  Supervised induction used to analyze the historical data stored in a database and to automatically generate a model that can predict future behavior
- Common tools used for classification are:
  - Neural networks
  - Decision trees
  - If-then-else rules

Data Mining Concepts and Applications

- **Clustering**

  Partitioning a database into segments in which the members of a segment share similar qualities

- **Association**

  A category of data mining algorithm that establishes relationships about items that occur together in a given record

Data Mining Concepts and Applications

- **Regression** is a well-known statistical technique that is used to map data to a prediction value
- **Forecasting** estimates future values based on patterns within large sets of data

Data Mining Techniques and Tools

- Data mining tools and techniques can be classified based on the structure of the data and the algorithms used:
  - Statistical methods
  - **Decision trees**

    Defined as a root followed by internal nodes. Each node (including root) is labeled with a question and arcs associated with each node cover all possible responses

Data Mining Techniques and Tools

- Data mining tools and techniques can be classified based on the structure of the data and the algorithms used:
  - Case-based reasoning
  - Neural computing
  - Intelligent agents
  - Genetic algorithms
  - Other tools
    - Rule induction
    - Data visualization

Data Mining Techniques and Tools

- A general algorithm for building a decision tree:
  1. Create a root node and select a splitting attribute.
  2. Add a branch to the root node for each split candidate value and label
  3. Take the following iterative steps:
     a. Classify data by applying the split value.
     b. If a stopping point is reached, then create leaf node and label it. Otherwise, build another subtree

Data Mining Techniques and Tools

- **Gini index**

  Used in economics to measure the diversity of the population. The same concept can be used to determine the 'purity' of a specific class as a result of a decision to branch along a particular attribute/variable

Data Mining Techniques and Tools

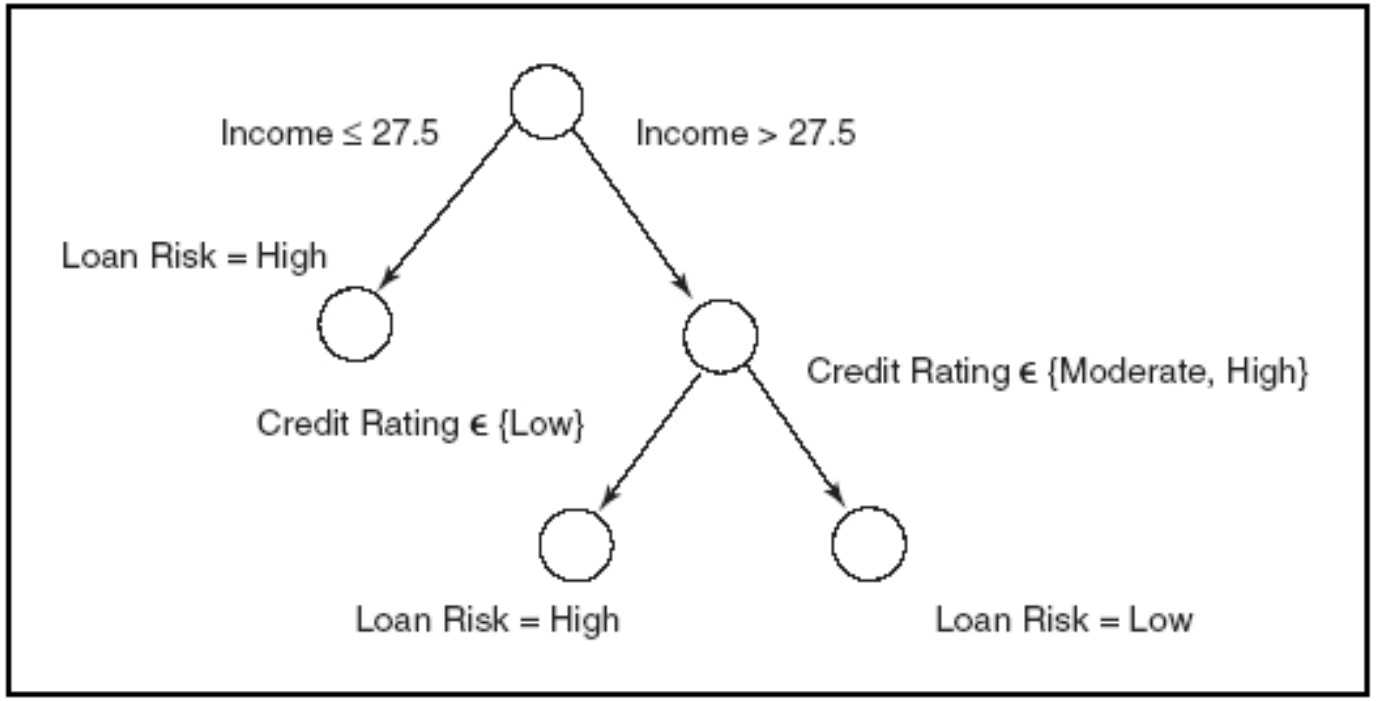


FIGURE 7.1 Decision Tree Using Gini Index for Split Criteria

Data Mining Techniques and Tools

- **Cluster analysis for data mining**
  - Cluster analysis is an exploratory data analysis tool for solving classification problems
  - The object is to sort cases into groups so that the degree of association is strong between members of the same cluster and weak between members of different clusters

Data Mining Techniques and Tools

- Cluster analysis results may be used to:
  - Help identify a classification scheme
  - Suggest statistical models to describe populations
  - Indicate rules for assigning new cases to classes for identification, targeting, and diagnostic purposes
  - Provide measures of definition, size, and change in what were previously broad concepts
  - Find typical cases to represent classes

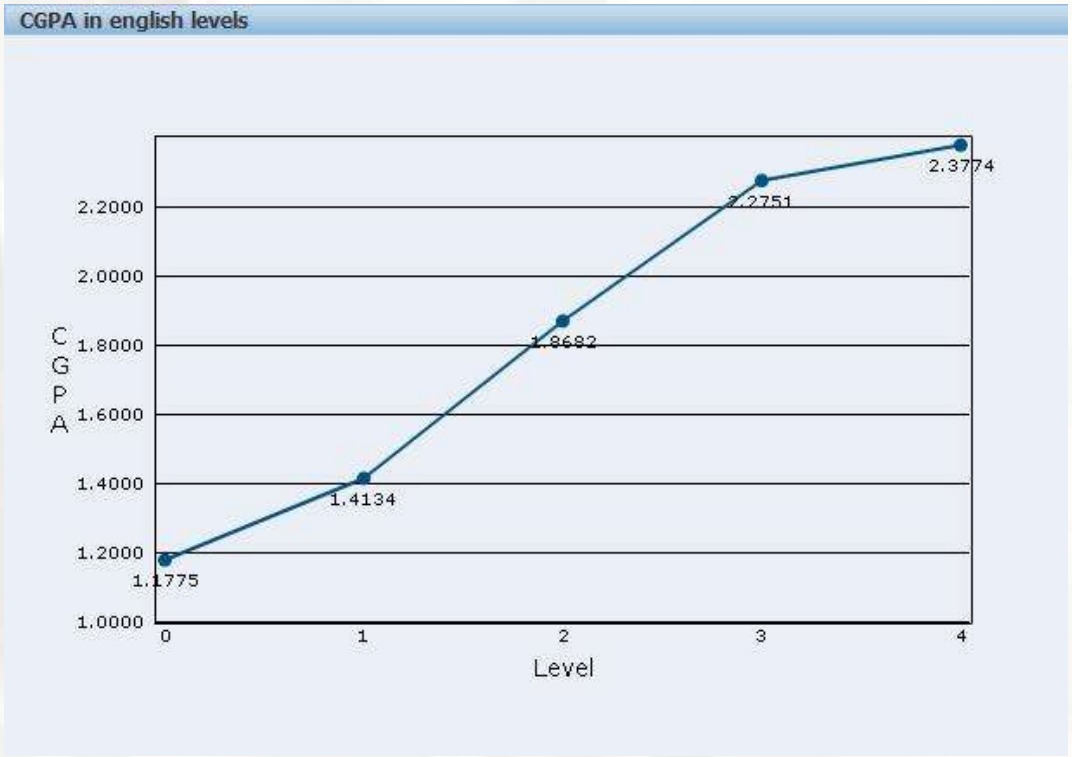
Data Mining Techniques and Tools

- **Cluster analysis methods**
  - Statistical methods
  - Optimal methods
  - Neural networks
  - Fuzzy logic
  - Genetic algorithms
- Each of these methods generally works with one of two general method classes:
  - Divisive
  - Agglomerative

Data Mining Techniques and Tools

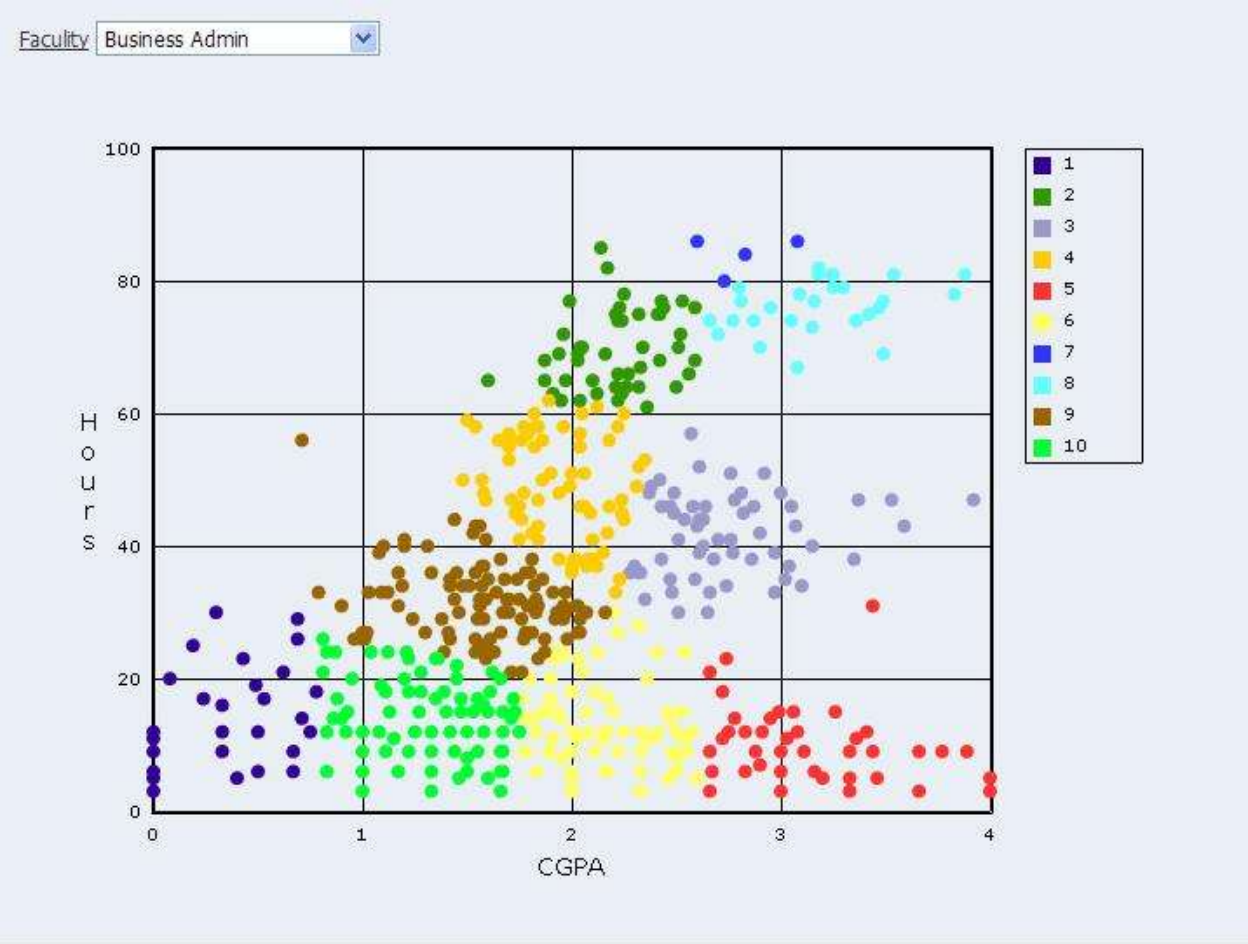
- **Hierarchical clustering method and example**
  1. Decide which data to record from the items
  2. Calculate the distances between all initial clusters. Store the results in a distance matrix
  3. Search through the distance matrix and find the two most similar clusters
  4. Fuse those two clusters together to produce a cluster that has at least two items
  5. Calculate the distances between this new cluster and all the other clusters
  6. Repeat steps 3 to 5 until you have reached the prespecified maximum number of clusters

Data Mining Concepts – Ex.

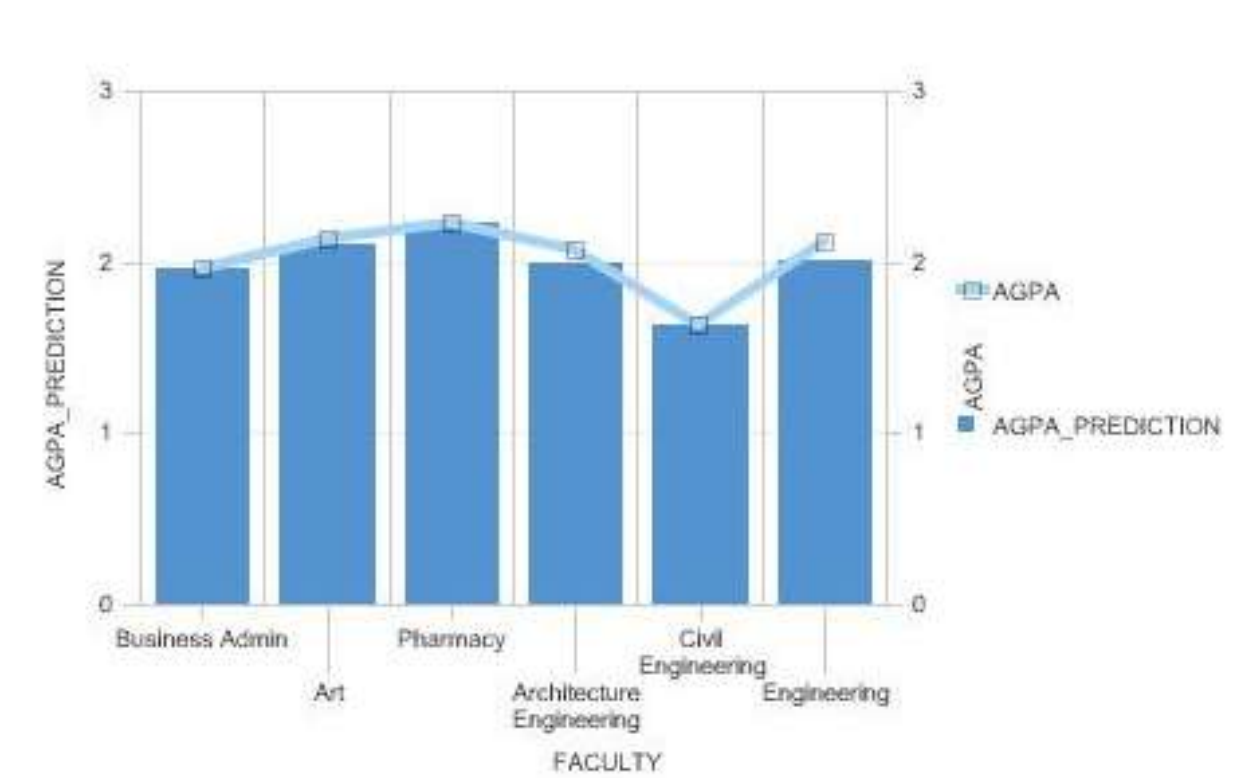


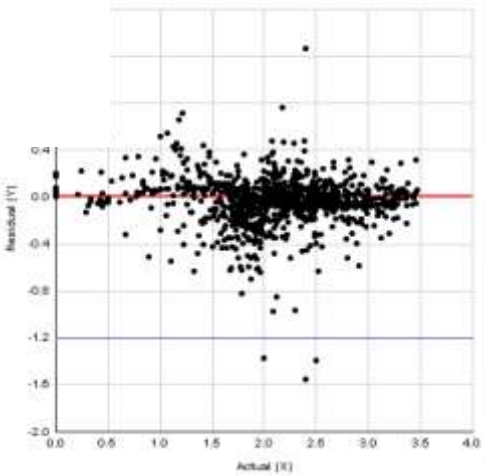
Data Mining Concepts – Ex.

Data Mining Concepts – Ex.



Data Mining Concepts - Ex





Data Mining Concepts and Applications

Data mining applications

- Marketing
- Banking
- Retailing and sales
- Manufacturing and production
- Brokerage and securities trading
- Insurance
- Computer hardware and software
- Government and defense
- Airlines
- Health care
- Broadcasting
- Police
- Homeland security

Text Mining

- **Text mining**

  Application of data mining to nonstructured or less structured text files. It entails the generation of meaningful numerical indices from the unstructured text and then processing these indices using various data mining algorithms

Text Mining

- Text mining helps organizations:
  - Find the “hidden” content of documents, including additional useful relationships
  - Relate documents across previous unnoticed divisions
  - Group documents by common themes

Text Mining

- Applications of text mining
  - Automatic detection of e-mail spam or phishing through analysis of the document content
  - Automatic processing of messages or e-mails to route a message to the most appropriate party to process that message
  - Analysis of warranty claims, help desk calls/reports, and so on to identify the most common problems and relevant responses

Text Mining

- Applications of text mining
  - Analysis of related scientific publications in journals to create an automated summary view of a particular discipline
  - Creation of a “relationship view” of a document collection
  - Qualitative analysis of documents to detect deception

Text Mining

- **How to mine text**
  1. Eliminate commonly used words (stop-words)
  2. Replace words with their stems or roots (stemming algorithms)
  3. Consider synonyms and phrases
  4. Calculate the weights of the remaining terms

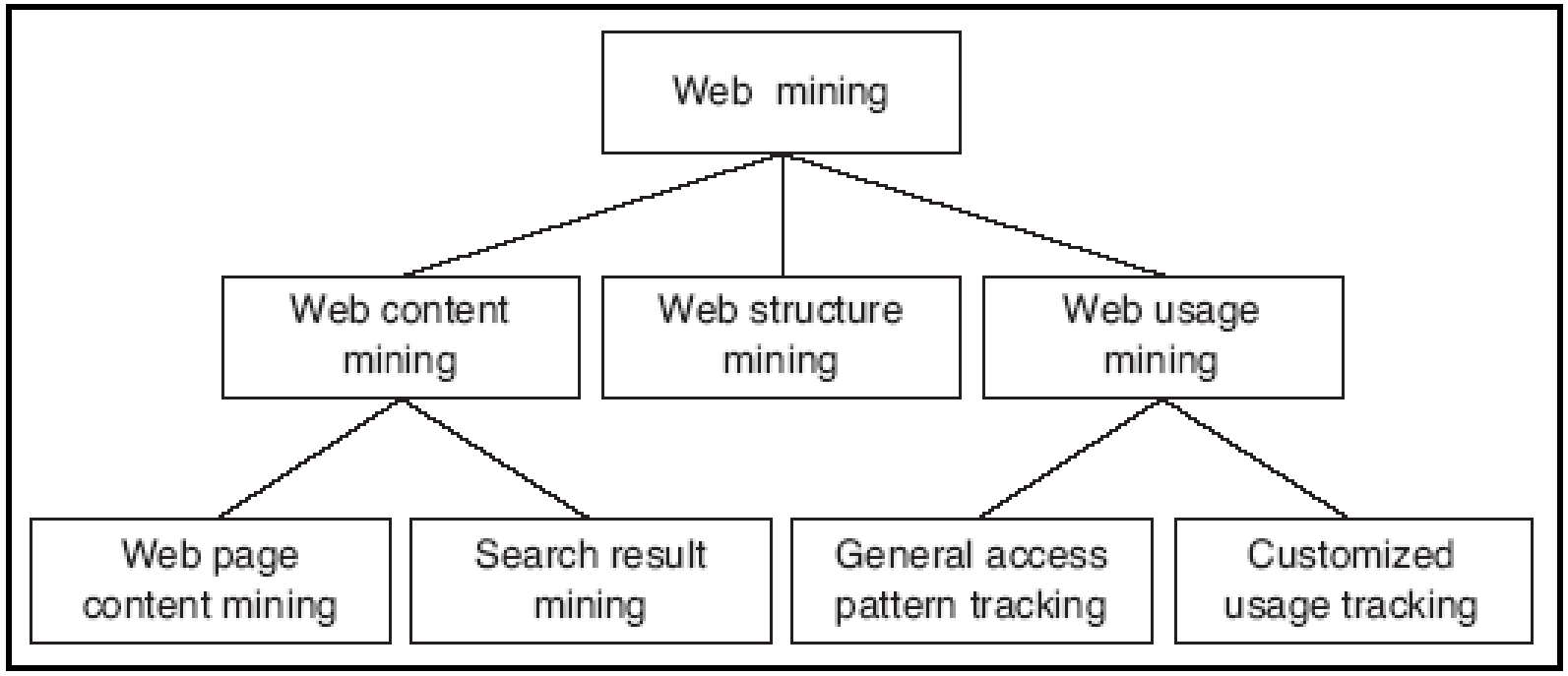
Web Mining

- **Web mining**

  The discovery and analysis of interesting and useful information from the Web, about the Web, and usually through Web-based tools

Data Mining Project Processes

FIGURE 7.5 Types of Web Mining



Web Mining

- **Web content mining**

  The extraction of useful information from Web pages
- **Web structure mining**

  The development of useful information from the links included in the Web documents
- **Web usage mining**

  The extraction of useful information from the data being generated through webpage visits, transaction, etc.

Web Mining

- Uses for Web mining:
  - Determine the lifetime value of clients
  - Design cross-marketing strategies across products
  - Evaluate promotional campaigns
  - Target electronic ads and coupons at user groups
  - Predict user behavior
  - Present dynamic information to users

Data Mining Project Processes

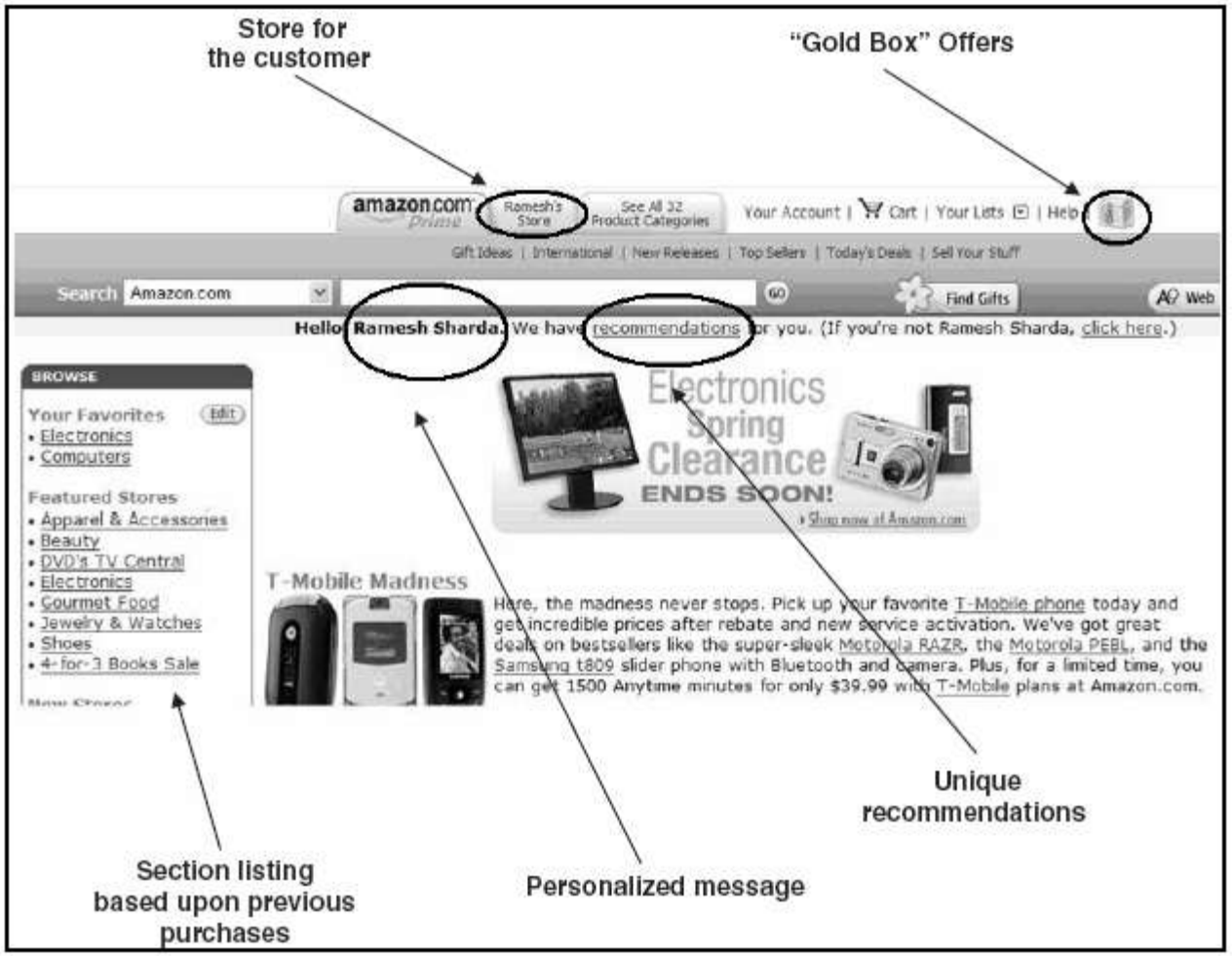


FIGURE 7.6 Example of Customization Using Web Usage Mining