CHAPTER-10

CONCEPT DESCRIPTION:

CHARACTERIZATION AND COMPARISION

10.1 Introduction

10.2 What is concept description?

10.3 Complex data types and aggregation:

10.4 User-control versus automation:

10.5 Data Generalization and Summarization-Based Characterization

10.6 Attribute-Oriented Induction

10.7 Review Question

10.8 References

10.CONCEPT DESCRIPTION:

CHARACTERIZATION AND COMPARISION

10.1 Introduction

Data mining can be classified into two categories: descriptive data mining and predictive data mining. Descriptive data mining describes the data set in a concise and summative manner and presents interesting general properties of the data. Predictive data mining analyzes the data in order to construct one or a set of models, and attempts to predict the behavior of new data sets.

Data base is usually storing the large amounts of data in great detail. However users often like to view sets of summarized data in concise, descriptive terms. Such data descriptions may provide an overall picture of a class of data or distinguish it from a set of comparative classes. Moreover, users like the ease and flexibility of having data sets described at different levels of granularity and from different angles. Such descriptive data mining is called concept description and forms an important component of data mining.

10.2 What is concept description?

The simplest kind of descriptive data mining is concept description. A concept usually refers to a collection of data such as frequent_buyers, graduate_students, and so on. As a data mining task, concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data. It is some times called class description, when the concept to be described refers to a class of objects. Characterization provides a concise and succinct summarization of the given collection of the data, while concept or class comparison (also known as discrimination) provides discriminations comparing two or more collections of data. Since concept description involves both characterization and comparison, techniques for accomplishing each of these tasks will study.

Concept description has close ties with the data generalization. Given the large amount of data stored in database, it is useful to be describe concepts in concise and succinct terms at generalized at multiple levels of abstraction facilities users in examining the general behavior of the data. Given the

ABCompany database, for example, instead of examining individual customer transactions, sales managers may prefer to view the data generalized to higher levels, such as summarized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income. Such multiple dimensional, multilevel data generalization is similar to multidimensional data analysis in data warehouses. The fundamental differences between concept description in large databases and online analytical processing involve the following.

10.3 Complex data types and aggregation:

Data warehouses and OLAP tools are based on a multidimensional data model that views data in the form of a data cube , consisting of dimensions (or attributes) and measures(aggregate functions). However, the possible data types of the dimensions and measures for most commercial versions of these systems are restricted. Many current OLAP systems confine dimensions to non-numeric data, similarly, measures (such as count (), sum (), average ()) in current OLAP systems apply only to numeric data. In contrast, for concept formation, the database attributes can be of various data types, including numeric, nonnumeric, spatial, text, or image. Furthermore, the aggregation of attributes in a database may include sophisticated data types, such as the collection of nonnumeric data, the merging of spatial region, the composition of images, the integration of texts, and the grouping of object pointers. Therefore, OLAP, with its restrictions on the possible dimension and measure types, represents a simplified model for data analyses. Concept description in databases can handle complex data types of the attributes and their aggregations, as necessary.

10.4 User-control versus automation:

On-line analytical processing in data warehouses is a purely user-controlled process. the selection of dimensions and the application of OLAP operations, such as drill-down, roll-up, slicing, and dicing, are directed and controlled by the users, although the control in most OLAP systems is quite user-friendly, users do require a good understanding of the role of each dimension. Furthermore, in order to find a satisfactory description of the data, users may need to specify a long sequence of OLAP operations. In contrast, concept description in data mining strives for a more automated process that helps determine which dimensions (or attributes) should be included in the analyses, and the degree to which the giver data set should be generalized in order to produce an interesting summarization of the data.

Recently, data warehousing and OLAP technology has been evolving towards handling more complex types of data and embedding more knowledge discovery mechanisms. As this technology continues to develop , it is expected that additional descriptive data mining features will be integrated into future OLAP systems.

Methods for concept description, including multilevel generalization, summarization, characterization, and comparison are outlined below. Such methods set the foundation for implementation of two major functional modules in data mining: multiple-level characterization and comparison. In addition, you will also examine techniques for the presentation of concept a description in multiple forms, including tables, charts, graphs, and rules.

6.5 Data Generalization and Summarization-Based Characterization

Data and objects in databases often contain detailed information at primitive concept levels. .For example, the item relation in sales database may contain attributes describing low-level item information such s item _ID , name , brand, category, supplier, place_made, and price. It is useful to be able to summarize a large set or data and present it at a high conceptual level.. For example, summarizing a large set of items relating to Christmas season sales provides a general description of such data , which can be very helpful for sales and marketing managers. This requires an important functionality in data mining: data generalization.

Data generalization is a process that abstracts a large set of task-relevant data in a database from a relatively low conceptual level to higher conceptual levels. Methods for the efficient and flexible generalization of large data sets can be categorized according to two approaches :(1) the data cube (or OLAP) approach and (2) the attribute –oriented induction approach .In this section, we describe the attribute-oriented induction approach.

10.6 Attribute-Oriented Induction

The attribute-oriented induction (AOI)) approach to data generalization and summarization-based characterization was first proposed in 1989,a few years prior to the introduction of the data cube approach. The data cube approach can be considered as a data warehouse-based, pre-computation-oriented, materialized-view approach. It performs off-line aggregation before an OLAP or data mining query is submitted for processing. On the other hand, the attribute-oriented induction approach, at least in its initial proposal, is a relational database query –oriented, generalization –based, on-line data analysis technique. However, there is no inherent barrier distinguishing the two approaches based on on-line aggregation versus off-line pre computation. Some aggregations in the data cube can be computed on-line, while off-line while off-line pre -computation of multidimensional space can speed up attribute –oriented induction as well.

10.7 Review Question

1. What is concept description?

2 Explain Complex data types and aggregation:

3 Discus User-control versus automation:

4 Defferentiate Data Generalization and Summarization-Based Characterization

5 Explain about Attribute-Oriented Induction

10.8 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition