CHAPTER-7

Discretization and Concept Hierarchy Generation

7.Discretization and Concept Hierarchy Generation

7.1 Introduction

Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value labels can be used to replace actual data values. These methods are typically recursive, where a large amount of time is spent on sorting the data at each step. The smaller the number of distinct values to sort, the faster these methods should be. Many discretization techniques can be applied recursively in order to provide a hierarchical or multiresolution partitioning of the attribute values known as concept hierarchy.

A concept hierarchy for a given numeric attribute attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data y collecting and replacing low-level concepts (such as numeric value for the attribute age) by higher level concepts (such as young, middle-aged, or senior). Although detail is lost by such generalization, it becomes meaningful and it is easier to interpret.

Manual definition of concept hierarchies can be tedious and time-consuming task for the user or domain expert. Fortunately, many hierarchies are implicit within the database schema and can be defined at schema definition level. Concept hierarchies often can be generated automatically or dynamically refined based on statistical analysis of the data distribution.

7.2 Discretization and Concept Hierarchy Generation for Numeric Data:

It is difficult and laborious for to specify concept hierarchies for numeric attributes due to the wide diversity of possible data ranges and the frequent updates if data values. Manual specification also could be arbitrary.

Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below- binning histogram analysis entropy-based discretization and data segmentation by “natural partitioning”.

Binning:

Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies.

Histogram Analysis:

Histograms can also be used for discretization. Partitioning rules can be applied to define range of values. The histogram analyses algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre-specified number of concept levels have been reached. A minimum interval size can be used per level to control the recursive procedure. this specifies the minimum width of the partition, or the minimum member of partitions at each level.

Cluster Analysis:

A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower kevel in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy.

Segmentation by natural partitioning:

Breaking up annual salaries in the range of into ranges like ($50,000-$100,000) are often more desirable than ranges like ($51, 263, 89-$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform "natural" intervals. In general the rule partitions a give range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit. The rule can be recursively applied to each interval creating a concept hierarchy for the given numeric attribute.

7.3 Discretization and Concept Hierarchy Generation for Categorical Data:

Categorical data are discrete data. Categorical attributes have finite number of distinct values, with no ordering among the values, examples include geographic location, item type and job category. There are several methods for generation of concept hierarchies for categorical data.

Specification of a partial ordering of attributes explicitly at the schema level by experts:

Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level. A hierarchy can be defined at the schema level such as street < city < province <state < country.

Specification of a portion of a hierarchy by explicit data grouping:

This is identically a manual definition of a portion of a concept hierarchy. In a large database, is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate-level data.

Specification of a set of attributes but not their partial ordering:

A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

Specification of only of partial set of attributes:

Sometimes a user can be sloppy when defining a hierarchy, or may have only a vague idea about what should be included in a hierarchy. Consequently the user may have included only a small subset of the relevant attributes for the location, the user may have only specified street and city. To handle such partially specified hierarchies, it is important to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

7.4 Review Question

1 Explain about Introduction to Concept Hierarchy?

2 Explian Discretization and Concept Hierarchy Generation for Numeric Data:

3 Explain Discretization and Concept Hierarchy Generation for Categorical Data:

7.5 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