DATA GENERALIZATION AND SUMMARIZATION

A technical report for

CSc 564 Data Warehousing and Data Mining

Submitted by

Brihat Ratna Bajracharya

19/075

Submitted to

Mr. Bal Krishna Subedi

Central Department of Computer Science and Information Technology

16 December 2019

TABLE OF CONTENT

1. INTRODUCTION....3
1.1 Concept Description....3
2. APPROACHES OF DATA GENERALIZATION....3
2.1. Data Cube Approach....3
2.1.1. OLAP Operations....4
2.1.1.1. Roll Up Operation....4
2.1.1.2. Drill Down Operation....4
2.1.1.3. Slice and dice....4
2.1.1.4. Pivot....4
2.1.2. Limitation of Data cube approaches....6
2.2. Attribute Oriented Induction (AOI) Approach....6
2.2.1. AOI Techniques....6
2.2.1.1. Data focusing....6
2.2.1.2. Data generalization by attribute removal or attribute generalization....7
2.2.1.3. Count and aggregate value accumulation....7
2.2.1.4. Attribute generalization control....7
2.2.1.5. Generalization data visualization....8
3. CONCLUSION....8
REFERENCES....8

1. INTRODUCTION

Data Generalization is a process which abstracts a large set of task-relevant data in a database from a low level conceptual levels to higher ones. From a data analysis point of view, data generalization is a form of descriptive data mining, which describes data in a concise and summarative manner and presents interesting general properties of the data. Data generalization summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute age) with higher-level concepts (e.g., young, middle-aged, and senior), or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions (e.g., removing birth date and telephone number when summarizing the behavior of a group of students).

1.1 Concept Description

A concept refers to a data collection such as frequent buyers, graduate students, and so on. Concept description is a form of data generalization that generates descriptions for data characterization and comparison. This is also known as class description when the concept to be described refers to a class of objects. Concept description generates descriptions for data characterization and comparison. Characterization provides a concise and succinct summarization of the given data collection, while Comparison (or Discrimination) provides descriptions comparing two or more data.

2. APPROACHES OF DATA GENERALIZATION

2.1. Data Cube Approach

In data cube approach, (also known as OLAP approach) the data for analysis are stored in a multi-dimensional database (called data cube). In general, the data cube approach materializes data cubes by first identifying expensive computations required for frequently-processed queries. These operations typically involve aggregate functions, such as count(), sum(), average(), and max(). These computations are performed and their results are stored in data cubes. Such computations may be performed for various levels of data abstraction. These materialized views can then be used for decision support, knowledge discovery, and many other applications. A set of attributes may form a hierarchy or a lattice structure, defining a data cube dimension. For example, date may consist of the attributes day, week, month, quarter, and year which form a lattice structure, and a data cube dimension for time. A data

cube can store pre-computed aggregate functions for all or some of its dimensions. The pre-computed aggregates correspond to specified group-by's of different sets or subsets of attributes. Generalization and specialization can be performed on a multidimensional data cube by roll-up or drill-down operations. Some of the operations that can be done in this approach is given below.

2.1.1. OLAP Operations

2.1.1.1. Roll Up Operation

The roll-up operation (also called the drill-up operation) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. It reduces the number of dimensions in a data cube, or generalizes attribute values to higher level concepts.

2.1.1.2. Drill Down Operation

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions. Since many aggregate functions need to be computed repeatedly in data analysis, the storage of pre-computed results in a multidimensional data cube may ensure fast response time and offer flexible views of data from different angles and at different levels of abstraction.

2.1.1.3. Slice and dice

The slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube. Figure 1 shows a slice operation where the sales data are selected from the central cube for the dimension time using the criterion time = “Q1.”

The dice operation defines a sub-cube by performing a selection on two or more dimensions. Figure 1 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions:

(location = “Toronto” or “Vancouver”) and (time = “Q1” or “Q2”) and (item = “home entertainment” or “computer”).

2.1.1.4. Pivot

Pivot (also called rotate) is a visualization operation that rotates the data axes in view to provide an alternative data presentation. Figure 1 shows a pivot operation where the item and

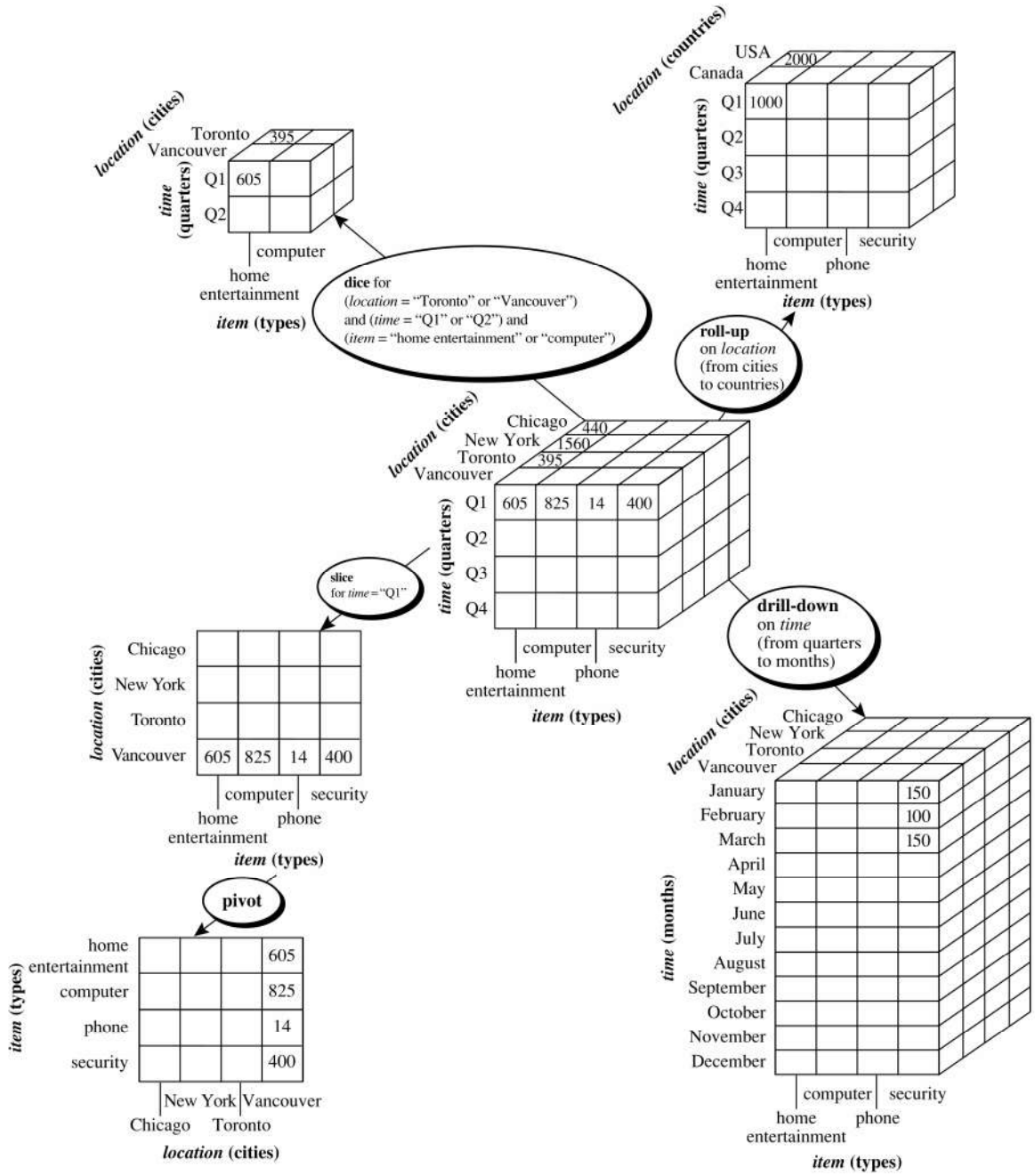

Figure 1: Different OLAP Operations

(source: Han, J. et al. Data Mining Concepts and Techniques 3rd ed (page 147))

location axes in a 2-D slice are rotated. Other examples include rotating the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes.

2.1.2. Limitation of Data cube approaches

Data cube approaches cannot incorporate non-numeric data but in reality the database can include various data types like spatial, text or image which should be included in the concept description. In addition to this, the aggregation of attributes in the database may include sophisticated data types such as the collection of non-numeric data, the merging of spatial regions, composition of images, integration of texts and grouping of object pointers. Furthermore, the data cube approach in data generalization is user control process i.e. the selection of dimensions and application of OLAP operations are directed and controlled by users. This require users to have a good understanding of the role of each dimension. But it is more productive if we can make this process automated that helps users to determine which attributes to include in the analysis so that the given data set can be generalized to produce effective summarization of the data.

2.2. Attribute Oriented Induction (AOI) Approach

This approach was first proposed in 1989. While the data cube approach is based on materialized views of the data and performs offline aggregation before OLAP or data mining query is submitted for processing. The attribute oriented induction on the other hand is a query-oriented, generalized-based, online data analysis technique.

The general idea of attribute-oriented induction is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set.

The generalization is performed by either attribute removal or attribute generalization. Aggregation is performed by merging identical generalized tuples and accumulating their respective counts. This reduces the size of the generalized data set. The resulting generalized relation can be mapped into different forms (e.g., charts or rules) for presentation to the user.

2.2.1. AOI Techniques

2.2.1.1. Data focusing

Data focusing corresponds to the specification of the task-relevant data (i.e., data for analysis). The data are collected based on the information provided in the data mining query. Because a data mining query is usually relevant to only a portion of the database, selecting the relevant data set not only makes mining more efficient, but also derives more meaningful results than mining the entire database. Next operation of attribute-oriented induction is data

generalization, which can be performed in either of two ways: attribute removal and attribute generalization.

2.2.1.2. Data generalization by attribute removal or attribute generalization

Attribute removal is based on the rule: *If there is a large set of distinct values for an attribute of the initial working relation, but either*

(case 1) there is no generalization operator on the attribute (e.g., there is no concept hierarchy defined for the attribute), or

(case 2) its higher-level concepts are expressed in terms of other attributes, then the attribute should be removed from the working relation.

Attribute generalization is based on the following rule: *If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute.*

2.2.1.3. Count and aggregate value accumulation

The aggregate function, count(), is associated with each database tuple. Its value for each tuple in the initial working relation is initialized to 1. Through attribute removal and attribute generalization, tuples within the initial working relation may be generalized, resulting in groups of identical tuples. In this case, all of the identical tuples forming a group should be merged into one tuple. The count of this new, generalized tuple is set to the total number of tuples from the initial working relation that are represented by (i.e., merged into) the new generalized tuple. For example, say by attribute-oriented induction, 52 data tuples from the initial working relation are all generalized to the same tuple, T. These 52 identical tuples are merged to form one instance of T, with a count of 52.

Other popular aggregate functions are sum() and avg().. For a given generalized tuple, sum() contains the sum of the values of a given numeric attribute for the initial working relation tuples making up the generalized tuple. Taking above example, the sum value for tuple T would then be set to the total number of units sold for each of the 52 tuples. Also, the aggregate function avg() is computed according to the formula avg() = sum() / count().

2.2.1.4. Attribute generalization control

Attribute removal and Attribute generalization claim that if there is large set of distinct values for an attribute, further generalization should be applied. This raise a question of how large

set of distinct values to be considered. This depends upon the user. Some may prefer this value to remain at rather low abstraction level while others go for higher levels. The control of this process is called attribute generalization control. If the attribute is generalized "too high," it may lead to over generalization making resulting rules not very informative. And if the attribute is not generalized to a "sufficiently high level", under generalization may result which also makes the rules obtained not informative. Thus, a balance should be attained in attribute-oriented generalization.

2.2.1.5. Generalization data visualization.

The generalized relations obtained from above techniques can be presented in the form of cross-tabulation forms, and various kinds of graphic presentation such as pie charts and bar charts and quantitative characteristics rules like showing how different value combinations are distributed in the generalized relation. These falls under generalization data visualization.

3. CONCLUSION

To conclude, 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. Data generalization approaches include data cube-based data aggregation and attribute-oriented induction. Concept description is the most basic form of descriptive data mining. It describes a given set of task-relevant data in a concise and summarative manner, presenting interesting general properties of the data. Data cube approach is based on materialized views of the data and performs offline aggregation while attribute oriented induction approach is a query-oriented, generalized-based, online data analysis technique. Both of them are used to extract relevant information from vast amount of data and hence summarize large data in concept space involving fewer attributes.

REFERENCES

1. Han, J., Kambar, M., Pei, J. (2012). '*Data Generalization by Attribute-Oriented Induction*', in Data Mining Concepts and Techniques, 3rd ed. (pp. 166-171). Waltham, MA: Elsevier.
2. Han, J., Kambar, M., Pei, J. (2012). '*Typical OLAP Operations*', in Data Mining Concepts and Techniques, 3rd ed. (pp. 146-148). Waltham, MA: Elsevier.