RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-31

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

- BASIC PRINCIPLES OF ATTRIBUTE-ORIENTED INDUCTION
- BASIC PRINCIPLES OF ATTRIBUTE ATTRIBUTE-ORIENTED ORIENTED INDUCTION
- BASIC ALGORITHM FOR ATTRIBUTE ORIENTED INDUCTION
- BASIC ALGORITHM FOR ATTRIBUTE ORIENTED INDUCTION
- EXAMPLE
- MCQ
- REFERENCES

Basic Principles of Attribute-Oriented Induction

Data focusing

– task-relevant relevant data, including including dimensions dimensions, and the result is the initial initial relation.

• Attribute-removal

– remove attribute A if there is a large set of distinct values for A but

– (1) there is no generalization operator on A, or

– (2) A's hi hg er l l eve concepts are expressed i t f d in terms of other att ib t ttributes.

Basic Principles of Attribute Attribute-Oriented Oriented Induction

Attribute-generalization

– If there is a large set of distinct values for A, and

there exists a set of generalization operators on of generalization operators on A,

then select an operator and generalize A.

• Attribute-th h ld reshold control

• Generalized relation threshold control

– control the final relation/rule size.

Basic Algorithm for Attribute Oriented Induction

Initial Relation

– Query processing of task-relevant data, deriving the initial

relation.

• Pre Generalization

– Based on Based on the analysis analysis of the number of distinct distinct values in each

attribute, determine generalization plan for each attribute:

removal? or how high to generalize?

Basic Algorithm Algorithm for Attribute Attribute-Oriented Oriented Induction

Prime Generalization

– Based on the PreGen plan, perform generalization to the

right level to derive a “p g rime eneralized relation”,

accumulating the counts.

• Presentation

– User interaction: (1) adjust levels by drilling, (2) pivoting,

(3) mapping into rules, cross tabs, visualization

presentations.

Example

DMQL: Describe DMQL: Describe general general characteristics characteristics of graduate graduate students students

in the Big-University database

use Big_University_DB

mine characteristics as “Science_Students”

in relevance to name, gender, major, birth_place,

birth date residence birth

date, residence, phone#, gpa

from student

where status in “graduate”

• Corresponding SQL statement:

Select name, gender, major, birth place _ , birth date _ ,

residence, phone#, gpa

from student

where status in {“Msc”, “MBA”, “PhD” }

Multiple Choice Question

1. Various visualization techniques are used in _____ step of KDD.
a) selection
b) transformaion
c) data mining.
d) interpretation.

2. Extreme values that occur infrequently are called as _____.
a) outliers
b) rare values.
c) dimensionality reduction.
d) All of the above.

3. Box plot and scatter diagram techniques are _____.
a) Graphical
b) Geometric
c) Icon-based.
d) Pixel-based.

4. _____ is used to proceed from very specific knowledge to more general information.
a) Induction
b) Compression.
c) Approximation.
d) Substitution.

5. Describing some characteristics of a set of data by a general model is viewed as _____
a) Induction
b) Compression
c) Approximation
d) Summarization

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm
- https://www.geeksforgeeks.org/
- http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf DATA MINING BOOK WRITTEN BY Micheline Kamber
- https://www.javatpoint.com/three-tier-data-warehouse-architecture
- M.H. Dunham, " Data Mining: Introductory & Advanced Topics" Pearson Education
- Jiawei Han, Micheline Kamber, " Data Mining Concepts & Techniques" Elsevier
- Sam Anahory, Denniss Murray," data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, " Pearson Education
- Mallach," Data Warehousing System", TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.