Detecting Group Anomalies in Tera-Scale Multi-Aspect Data via Dense-Subtensor Mining

Aug-13-2020

Kijung Shin (kijungs@kaist.ac.kr)

1 General Information

- Version: 2.0
- Date: Aug 13, 2020
- Authors: Kijung Shin, Bryan Hooi, Jisu Kim, and Christos Faloutsos

2 Introduction

D-Cube (**D**isk-based **D**ense-block **D**etection) is an algorithm for detecting dense subtensors in web-scale tensors. **D-Cube** has the following properties:

- Scalable: D-Cube can handle large data not fitting in memory or even on a disk.
- Fast: Even when data fit in memory, D-Cube outperforms its competitors in terms of speed.
- Accurate: D-Cube detects dense subtensors in real-world tensors accurately, providing theoretical accuracy guarantees.

Detailed information about the method is explained in the following papers:

- Kijung Shin, Bryan Hooi, Jisu Kim, and Christos Faloutsos, *D-Cube: Dense-Block Detection in Terabyte-Scale Tensors*, ACM International Conference on Web Search and Data Mining (WSDM) 2017, Cambridge, UK
- Kijung Shin, Bryan Hooi, Jisu Kim, and Christos Faloutsos, *Detecting Group Anomalies in Tera-Scale Multi-Aspect Data via Dense-Subtensor Mining,* Frontiers in Big Data 2021.

3 Installation

- This package requires the following software to be installed in the system and set in PATH.
  - Hadoop 1.x.x. from http://hadoop.apache.org
  - Java 1.6.x. or higher, preferably from sun

- For compilation (optional), type *./compile.sh*
- For packaging (optional), type *./package.sh*
- For demo (optional), type *make*

4 Input File Format

The input file lists all tuples in a relation. Each line corresponds to a tuple and consists of dimension attributes values and a measure attribute value, which are separated by a comma. Additionally, we assume the followings:

- Dimension attributes values are integers between 0 and (cardinality -1).
- Measure attribute values are in the last column of each row
- Measure attribute values are integers

example_data.txt is an example input file.

5 Output Files Format

For each found block, two files are created. For example, for the n-th found block, the following two files are created:

- *block_n.tuples*: this file lists tuples included in the n-th block. This file has the same format with the input file.
- *block_n.attributes*: this file lists attribute values included in the n-th block. Each line consists of the order of an attribute and a value of the attribute.

output directory contains the examples of the output files. Statistics, including the volumes, masses, and densities of found blocks, are printed in the console.

6 Running D-Cube Serial Version

- How to Run

```
./run_single.sh input_path output_path dimension density_measure policy mass_threshold num_of_blocks
```

- Parameters
    - *input_path*: path of the input file. See 4 for the detailed format of the input file
    - *output_path*: path of the local directory for output files. See 5 for the detailed format of the output files

- *dimension*: number of dimension attributes
- *density_measure*: density measure to use. This parameter should be one among [ari, geo, susp]
- *policy*: policy to use for selecting attribute from which values are removed: This parameter should be one among [density, cardinality]
- *mass_threshold*: mass-threshold parameter, which should be greater than or equal to one
- *num_of_*blocks: number of blocks to find

7 Running D-Cube Hadoop Version

- How to Run

```
./run_hadoop.sh input_path output_path dimension density_measure policy mass_threshold
num_of_blocks num_of_reducers log_path
```

- Parameters
  - *input_path*: path of the input file in HDFS. See 4 for the detailed format of the input file
  - *output_path*: path of the HDFS directory for output files. See 5 for the detailed format of the output files
  - *dimension*: number of dimension attributes
  - *density_measure*: density measure to use. This parameter should be one among [ari, geo, susp]
  - *policy*: policy to use for selecting attribute from which values are removed: This parameter should be one among [density, cardinality]
  - *mass_threshold*: mass-threshold parameter, which should be greater than or equal to one
  - *num_of_*blocks: number of blocks to find
  - *num_of_redcuers: number of reducers to use*
  - *log_path*: path of the local directory for logs.