Fast, Accurate and Flexible Algorithms for Dense Subtensor Mining

Nov-08-2016

Kijung Shin (kijungs@cs.cmu.edu)

1 General Information

- Version: 2.0
- Date: Nov-08-2016
- Authors: Kijung Shin (kijungs@cs.cmu.edu)

2 Introduction

M-Zoom (Multidimensional Zoom) and **M-Biz (Multidimensional Bi-directional Zoom)** are algorithms for detecting dense blocks in tensors. They have the following properties:

- *Scalable*: scales almost linearly with all input factors
- *Provably accurate*: provides high accuracy in real data as well as theoretical guarantees
- *Flexible*: supports high-order tensors, various density measures, multi-block detection, and size bounds

Detailed information about the methods is explained in the following papers

- Kijung Shin, Bryan Hooi, and Christos Faloutsos. "*Fast, Accurate and Flexible Algorithms for Dense Subtensor Mining."* ACM Transactions on Knowledge Discovery from Data (TKDD) (Accepted)
- Kijung Shin, Bryan Hooi, and Christos Faloutsos. "*M-Zoom: Fast Dense Block Detection in Tensors with Quality Guarantees*", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) 2016, Riva del Garda, Italy

3 Installation

- This package requires that java 1.7 or greater be installed in the system and set in PATH.
- 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:

- Measure attribute values are in the last column of each row
- Measure attribute values are integers

example_data.txt is an example of the 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 M-Zoom

6.1 How to Run

```
./run_mzoom.sh input_path output_path dimension density_measure num_of_blocks lower_bound upper_bound
```

6.2 Parameters

- *input_path*: path of the input file. See 4 for the detailed format of the input file
- *output_path*: path of the 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, es_alpha], where alpha should be a number greater than zero.
- *num_of_*blocks: number of blocks to find
- *lower_bound* (optional): minimum size of blocks to find
- *upper_bound* (optional): maximum size of blocks to find

7 Running M-Biz

7.1 How to Run

```
./run_mbiz.sh input_path output_path dimension density_measure num_of_blocks lower_bound upper_bound
```

7.2 Parameters

- *input_path*: path of the input file. See 4 for the detailed format of the input file
- *output_path*: path of the directory for output files. See 5 for the detailed format of the output files
- *dimension*: number of dimension attributes
- *density_measure*: *density_measure*: density measure to use. This parameter should be one among [ari, geo, susp, es_alpha], where alpha should be a number greater than zero.
- *num_of_*blocks: number of blocks to find
- *lower_bound* (optional): minimum size of blocks to find
- *upper_bound* (optional): maximum size of blocks to find