Project: Extracting Top-K Insights from Multi-dimensional Data

1. Project Description
OLAP tools have been extensively used by enterprises to make better and faster decisions. Nevertheless, they require users to specify group-by attributes and know precisely what they are looking for. This paper takes the first attempt towards automatically extracting top-k insights from multi-dimensional data.

2. Reading Material
Please read the paper from the beginning up to Section 5.2 to understand the algorithms. Sections 6 and 7 are optional to read, which describe the ways to optimize the runtime. Please read Section 8 to understand the evaluation methodology.

3. Algorithms to Implement
Please implement the following techniques:
Algorithms 1 and Algorithm 2 in Section 5.

Note: There is no need to implement the systems configuration layer as described in Section 4.1. You may just specify some default insight types to extract and the null hypothesis. To compute the null hypothesis, it is a good practice to scan the data to understand what distribution would be best to start with as the null hypothesis. A trivial null hypothesis may not lead to valuable insights.

There is also no need to implement the user interface.

4. Results to Produce
Dataset: the DBLP dataset we used in our homework assignments

Results to produce:
- Top-10 insights from DBLP: Try different values of tau = 1, 2, 3, 4.
- Top-10 insights on collaborations using a composite extractor. Example: which authors have collaborated the most over time.
- Top-10 insights for authors whose articles contain the word "graph" in the title.
- Top-10 insights for words used in the titles.
5. Ideas for Extension (20%)
This project also requires an extension. We offer a few ideas below. Please feel free to choose one idea to explore in your project. Students are also welcome to propose their own idea.

1) Insights on keywords or key phrases extracted from titles. For example, the keyword "cloud" started to appear more frequently from 2005, or the key phrase “deep learning” started to increase rapidly in 2013.

2) Optimization using the EKIO algorithm discussed in Section 6. Also show performance evaluation results using EKI and EKIO algorithms on the TPC-H dataset. Reproduce Figure 16 for these algorithms.