Efficient Recommendation of De-identification Policies using MapReduce

ABSTRACT

Many data owners are required to release the data in a variety of real world application, since it is of vital importance to discovery valuable information stay behind the data. For example, all registered hospitals in California of US are required to submit specific demographic data on some patients which have been in good condition [1]. However, publishing those data containing sensitive information could violate individual's privacy.

EXISTING SYSTEM

In existing System re-identification attacks on the AOL and ADULTS datasets have shown that publish such data directly may cause tremendous threads to the individual privacy. Thus, it is urgent to resolve all kinds of re-identification risks by recommending effective de-identification policies to guarantee both privacy and utility of the data. De-identification policies is one of the models that can be used to achieve such requirements, however, the number of de-identification policies is exponentially large due to the broad domain of quasi-identifier attributes. To better control the trade off between data utility and data privacy, skyline computation can be used to select such policies, but it is yet challenging for efficient skyline processing over large number of policies.

DRAWBACKS

- Datasets have shown that publish such data directly may cause tremendous threads to the individual privacy.
- The number of de-identification policies is exponentially large due to the broad domain of quasi-identifier attributes.

PROPOSED SYSTEM

In this proposed system one parallel algorithm called SKY-FILTER-MR, which is based on MapReduce to overcome this challenge by computing skylines over large scale de-identification policies that is represented by bit-strings. To further improve the performance, a novel approximate skyline computation scheme was proposed to prune unqualified policies using the approximately domination relationship. With approximate skyline, the power of filtering in the policy space generation stage was greatly strengthened to effectively decrease the cost of skyline computation over alternative policies. Extensive experiments over both real life and synthetic datasets demonstrate that our proposed SKY-FILTER-MR algorithm substantially outperforms the baseline approach by up to four times faster in the optimal case, which indicates good scalability over large policy sets.

ADVANTAGES

- SKY-FILTER-MR, Algorithm is based on MapReduce to overcome this challenge by computing skylines over large scale de-identification policies.

- To improve the performance, a novel approximate skyline computation scheme was proposed

SYSTEM REQUIREMENTS

- **H/W System Configuration:-**

- Processor - Pentium –IV
- RAM - 4 GB (min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

- **S/W System Configuration:-**

- Operating System : Windows 7 or 8 32 bit
- Application Server : Tomcat5.0/6.X
- Backend coding : Java
- Tool : Virtual Box
- Environment : Ubuntu
- Technology : Hadoop