On Privacy Preserving of Text and Sparse Binary Data with Sketches - Review

Shumin Guo

Problem Description

Text and binary data has the property of sparsity, which means that only very few attributes of a data tuple have non-null values, and high dimensionality, which renders existing privacy preservation methods ineffective. These properties make existing algorithm about privacy preservation not be directly used into these types of data. So, how to effectively preserve privacy for text and binary sparse data becomes a research problem.

Contribution of this paper

This paper proposes a sketch based method for privacy protection of text and sparse binary data. This method reduces the high dimension to a lower one by aggregating(sketching) data tuple using a binary random matrix. By controlling the dimension of the sketch matrix, we can achive an expected anonymous privacy level. Experiments with the proposed method show that we can get a good balance between the utility over classification and clustering algorithms and the privacy.

Weaknesses of this paper

As the title of this paper says, text and sparse binary data are the supposed data type for the sketch method, but this is not very clearly modeled in the paper.

For the experiment part, it will be better if the author makes a comparison between the performance of the existing methods, such as k-anonymity, and the sketch based method.

Although this paper has discussion about controlling the dimension of the sketched data set, and prove that the level of anonymity can be controlled by adjusting the dimension of the sketch matrix, there is no discussion about how the dimension and the level of sparsity can affect the performance of the sketch algorithm.

Still, this paper fails to point out the property of data mining algorithms that can be used over the sketched data.

Are there any possible attacks to the sketched data set?