PIVOT: Privacy-preserving Outsourcing of Text Data for Word Embedding

Against Frequency Analysis Attack

(Poster submission)

Yanying Li and Wendy Hui Wang
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ
{yli158, Hui.Wang}@stevens.edu

Boxiang Dong
Department of Computer Science
Montclair State University
Montclair, NJ
dongb@montclair.edu

Abstract

In this paper, we design PIVOT, a new privacy-preserving method that supports outsourcing of text data for word embedding. PIVOT includes a 1-to-many mapping function for text documents that can defend against the frequency analysis attack with provable guarantee, while preserving the word context during transformation.

Introduction

There has been a significant growth in the volume and variety of unstructured text data. Technologies such as text analytics and Natural Language Processing (NLP) are developed to process and analyze those massively collected data. However, the high complexity of these techniques hinders common users to perform sophisticated text analysis for their own business and research use.

In this paper, we consider an outsourcing paradigm. Alice, a data owner who has a private text document \( D \), intends to outsource \( D \) to a third-party service provider for NLP analysis without leaking sensitive information in \( D \). Word embedding has been used commonly in many NLP tasks. It enables machine learning techniques to process raw text data. In this paper, we assume that the server performs a prediction-based word embedding method (e.g., word2vec (Mikolov et al. 2013)) to generate the vectors of text data. In this paper, we consider an outsourcing paradigm. Alice, a data owner who has a private text document \( D \), intends to outsource \( D \) to a third-party service provider for NLP analysis without leaking sensitive information in \( D \). Word embedding has been used commonly in many NLP tasks. It enables machine learning techniques to process raw text data. In this paper, we assume that the server performs a prediction-based word embedding method (e.g., word2vec (Mikolov et al. 2013)) to generate the vectors of text data.

PIVOT is a privacy-preserving outsourcing of text data for word embedding. To our best knowledge, this is the first work on privacy-preserving outsourcing of text data that considers the accuracy of word embedding as the utility goal.

Preliminaries

In this paper, we use the notations shown in Table 1.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D \rightarrow D' )</td>
<td>document before/after transformation</td>
</tr>
<tr>
<td>( w_i )</td>
<td>a word in ( D )</td>
</tr>
<tr>
<td>( S_j )</td>
<td>replacement candidate words of ( w_i )</td>
</tr>
<tr>
<td>( s_{i, j} )</td>
<td>the ( j )-th replacement word of word ( w_i )</td>
</tr>
</tbody>
</table>

Data Utility. In this paper, we consider the quality of word embedding as the utility of the outsourced data. We measure the accuracy loss of word embedding as the change of distance between the vectors that are generated from \( D \) and \( D' \). Formally, given two vectors \( v \) and \( v' \), the distance of \( v \) and \( v' \) is measured as following, where \( \cos(v, v') \) measures the cosine similarity of \( v \) and \( v' \):

\[
\text{dist}(v, v') = 1 - \cos(v, v') = 1 - \frac{v \cdot v'}{|v| \cdot |v'|}.
\]

Let \( V \) and \( V'' \) be the set of vectors generated from \( D \) and \( D' \). We measure the distance between \( V \) and \( V'' \) as the absolute difference between the sum of the cosine distance of
all vector pairs in $V$ and $V'$.

$$\text{dist}(D, D') = \sum_{w_i, w_j \in W, s_i^p \in S_i, s_j^p \in S_j} |\text{dist}(w_i, w_j) - \text{dist}(s_i^p, s_j^p)|$$

Our utility goal is to minimize the distance between the original document $D$ and its transformed document $D'$. We design a grouping-based 1-to-many word mapping (GOMM) method that maps multiple instances of a single word to different words. To defend against the frequency analysis attack, we assume the attacker may obtain the prior knowledge of word frequency distribution of $D$. Then by collecting the word frequency information in the transformed document $D'$, the attacker can perform the frequency analysis attack, a well-known attack, to map the words in the transformed document $D'$ back to the original words in $D$ using their frequency distribution (Naveed, Kamara, and Wright 2015).

**Privacy Model.** To quantify the privacy guarantee against the frequency analysis attack, we define the notion of $\ell$-privacy, which is adapted from a well-known notion $\ell$-diversity (Machanavajjhala et al. 2006). Formally, given $D$ and $D'$, $D'$ satisfies $\ell$-privacy if for each replacement word $s \in D'$, the attacker’s attack probability on $s$ must satisfy:

$$P(s \rightarrow w) \leq \frac{1}{\ell},$$

where $w \in D$ is the original word of $s$, and $\ell > 1$ is a user-specified integer. Intuitively, the higher $\ell$ is, the stronger the transformation is against the frequency analysis attack.

**Grouping-based One-to-Many Word Mapping**

We design a grouping-based 1-to-many word mapping (GOMM) method that maps multiple instances of a single word to different words. To defend against the frequency analysis attack, the frequency distribution of the transformed document is always (almost) uniform. Figure 1 shows an example of before and after mapping.

![Frequency](attachment://before.png)  ![Frequency](attachment://after.png)

(a) Freq. of orig. words (before mapping)  (b) Freq. of transformed words (after mapping)

**References**

