Abstract

Exchange of enormous data and information securely and frequently via Internet is very common and demanded in today's fast track scenario of world. The idea behind the proposed Sensitive Information Security Model Based on Term Clustering (SIS-TC) is to provide the security to a large volume of text documents which contain very important and sensitive information or data or both. These documents are first broken into its constituent parts, called terms, by using knowledge repository and then term clusters are made by finding out the similar terms of each category. These clusters represent the categories of Noun, Pronoun, Numeral, Punctuation etc. Only one instance of a cluster is kept and become the cluster representative. Firstly, the term frequency of each different occurred term (or word) is calculated and then all the duplicate copies of each term are removed, so that to transform it into the low dimensional data. Such reduced data set drastically decreases the total size of the complete data and space as well, and increases the performance of the system by the ratio of 65%-70%. Next, this reduced data is divided into High Risk Data (HRD) and Low Risk Data (LRD) to provide different level of security to each type. Therefore, HRD is symmetrically encrypted whereas LRD is encrypted non-symmetrically. This paper also includes the analytical experimental
results based on the test data set of 8 text documents of varying sizes.

References

- Jiawei Han, and Michele Line Kamber, Data Mining: Concepts and Techniques, 2nd ed. , Elsevier, 2006.
- Shihab A. Hameed, Habib Yuchoh, and Wajdi F. Al-khateeb, "A Model For

Index Terms

Computer Science

Information Security

Keywords

Term Clustering; Knowledge Repository; Sensitive Information; High Dimensional And Low Dimensional Data; Symmetric And Non-symmetric Encryption