ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

# A Framework Based On Weighted Indexing For the Prevention of the Data Leakage in Cloud Computing Transaction

C.Sureshkumar[1], Dr.K.Iyakutti[2], Dr.C.Rekha[3]

Associate Professor, Department of Computer science, MKU College, Madurai, Tamil Nadu, India[1]

Professor, Department of Physics and Nanotechnology, SRM University, Chennai, India[2]

Assistant Professor, Department of Computer Science, R.D.Govt Arts College, Sivaganga, Tamil Nadu, India[3]

**Abstract*:*** Cloud Computing is the talk of the today's technology for the deployment of service to the consumers at their door steps. The consumers are free of the mechanism of process of deployment of the solution, but enjoying the fruit of the technology. In this research the data security issues are highlighted. Data leakage in the cloud computing transactions is a biggest security threat estimated today. A framework for the prevention of the data leakage is devised and discussed. The uninvestigated area of data leakage prevention mechanism based on the semantic relations is discussed in this research. This research proposes a frame work based on the weighted indexing. The weight estimation is done through the back propagation network. It is followed by the relatedness estimation based on the various properties of the text.

***Keywords***: Data Security, Data leakage. weighted indexing, back propagation network`

## I. INTRODUCTION

The leading global provider is the International Data Corporation (IDC) for market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC's view on cloud computing is that the cloud software market reached $22.9 billion in revenue in 2011, a 30.9% growth rate. IDC expects cloud software will grow to $67.3 billion by 2016 at a compound annual growth rate (CAGR) of 24%. SaaS delivery will significantly outpace traditional software product delivery, growing nearly five times faster than the software market as a whole and becoming the significant growth driver to all functional software markets. By 2016, the cloud software model will account for $1 of every $5 spent on software [1].

SaaS is an application hosted on a remote server and accessed through the Internet. A simple and concrete example of SaaS is the "free" email (also called web-based e-mail) systems offered on the Internet such as Microsoft Hotmail, G-mail and Yahoo mail. Software as a service (or SaaS) is a way of delivering applications over the Internet as a service. Without taking the pain of installing and looking for the maintenance of software, one can simply access it via the internet and thereby avoid the need for complex software and hardware management. The utilization of SaaS system is steadily rising. A typical characteristic of SaaS, is storage of client's data location accessible from the Internet. This means the data is no longer stored on the client's personal computer, but in a data center operated by the SaaS provider and hence the data is not completely in control of the user. The important feature which makes the security as the vital factor to be considered is because the deployment is done over the internet. Cloud vendor must look further than the expected security measures like restricting user access, password protection etc. Data encryption is the usual method the vendors follow to secure their clients data. Many vendors use private or public key encryption to guarantee data security.

## II. Scope of Research

Computing and IT services has become the article of trade in the last decade. Organizations view the data and the transactions related to their business formulates the core competency. The strategic decisions are based on the information from the data and the transactions. The data thus plays a vital role in the organizations data today activity and the development. Data and the services related to the data manipulations are given a strong care from the organizations point of view.

When the organization moves towards the SaaS for obtaining and utilizing the benefits of SaaS, the organization must take the risk of storing their data in the service provider's area. If it is in the case of public cloud, the data will be stored with other like data from other organizations. The service provider might also duplicate the data in several locations in order to provide the high availability of the data for accessing.

Most enterprises are familiar with the traditional on-premise model, where the data continues to reside within the enterprise boundary, subject to their policies. Consequently, there is a great deal of discomfort with the lack of control and knowledge of how their data is stored and secured in the SaaS model. There are strong concerns about data breaches, application vulnerabilities and availability that can lead to financial and legal liabilities [2].

There are numerous security issues for cloud computing as it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Therefore, security issues for many of these systems and technologies are applicable to cloud computing [3]. With the increasing popularity of enterprise cloud computing and its public connectivity via the internet it is the next frontier for viruses, worms, hackers and cyber-terrorists to start probing and attacking [4].

Data loss and leak prevention is a serious security issue for the cloud, as the number of incidents continues to increase. Whether it's a spiteful attempt, or an unintentional fault, data loss can reduce a Cloud's brand, diminish value, and damage the goodwill and reputation. Data loss can compromise intellectual property or cause an organization to violate compliance regulations [5].

## III. Problem Formulation

The distributed nature of the cloud model necessarily involves more transits of data in networks, thus opens new taxing security risks [6]. To handle the massive amount of data present in cloud and the popularity that gains cloud computing over the times of yore, security invites major concentration for all who are using it and also those who want to utilize it but would not able to do so because no one can assure them in terms of security of their data on the cloud [7].

Most of the security solutions secure data at rest by restricting access to it and encrypting it, the state of art relies on robust policies and pattern matching algorithms [8]. Data leak prevention focused on building policies [9], developing watermarking schemes [10] and identifying the forensic evidence for post-mortem analysis [11].

The current state of the art in data leak prevention focuses on pattern matching, which suffers from the general shortcoming of misuse detection techniques: an expert needs to define the signatures. Given the elusive definition of data leaks, signatures should be defined per corporation basis, making the widespread deployment of current data leak prevention tools a challenge. Data leaks can occur by accident between individuals who are completely legitimate. The detection of such data leak requires the understanding of semantics [8].

## IV. Proposed Methodology

The following framework gives the outline of the proposed approach. In this proposed framework the Data in motion is concentrated. Any time data is set into motion accessed in an unconventional way, forwarded to a co-worker, sent to a printer, etc. its security is put at risk [12].

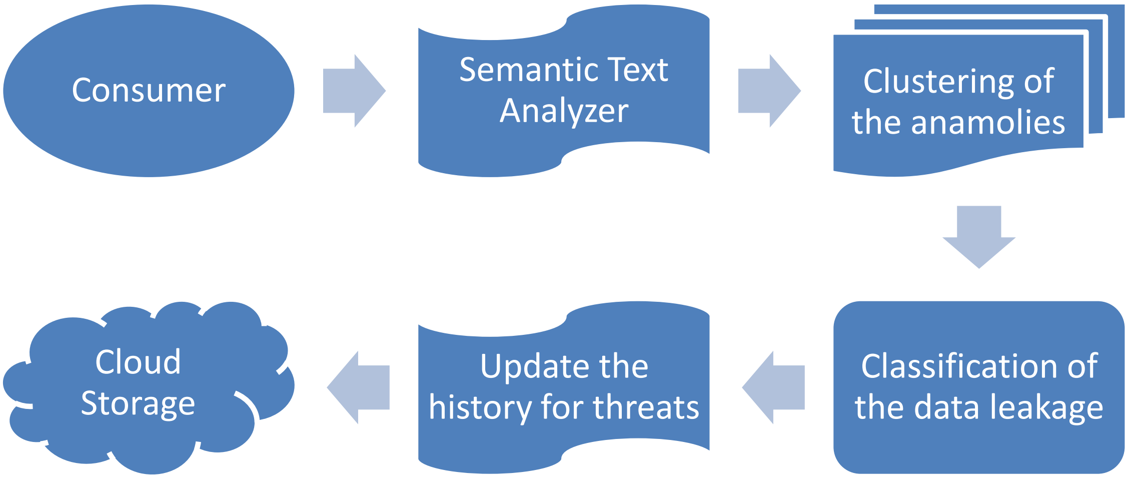

Consumer
Semantic Text Analyzer
Clustering of the anamolies
Classification of the data leakage
Update the history for threats
Cloud Storage

FIG 1 : PROPOSED MODEL FOR THE DATA LEAKAGE PREVENTION

The above framework employs in four stages. The first stage the communication from the consumer is being tapped and the semantic analysis of the communication is done. In the second stage the clustering of the dataset is done based on the similarities and the anomaly is detected for the leakage. The next step is to classify the data leakage threat. After classification the history of the threats is being update. This history of the threats serves as the basic repository for the avoidance of the data leakage in the future.

### 1. *Pseudo code*

The proposed approach in the first stage is to develop the semantic text analyzer for the first stage in the data leakage prevention mechanism. This process is developed in two parts. The first part is involved in the fixing up of the weight for the indexes, passing the information to the semantic interpreter and the second part is used to estimate the relatedness.

### 2. *STEP 1 (a): Semantic text analyzer*

The first part is concerned with the fixing up of the weights for the tokens of the text to analyze. The weight fixation of the proposed work is done by the back propagation network. Back propagation is a form of supervised learning for multi-layer nets, which is called as generalized delta rule. The pseudo code of the algorithm is given as follows

### 3. *Back Propagation phase*

```
        Input: Text to be processed
Step 1: Select a pattern $X_k$ from the training set T and present it to the network
Step 2: Compute activations and signals for the input neurons, hidden neurons
and output neurons
Step 3: Compute the error over the outputs with the desired outputs
Step 4: Use the error calculated in step 3 to compute the change in the hidden to output layer weights and the change in input to hidden layer weights such that a global error measure gets reduced
Step 5: Update all the weights hidden to output layer weights
        $w_{hj}^{K+1} = w_{hj}^{K} + \Delta w_{hj}^{K}$
        Input to hidden layer weights
```

$$\mathrm{w}_{\mathrm{ih}}^{\mathrm{K+1}} = \mathrm{w}_{\mathrm{ih}}^{\mathrm{K}} + \Delta\, \mathrm{w}_{\mathrm{ih}}^{\mathrm{K}}$$

Step 6: Repeat steps 1 through 5 until the global error falls below a predefined threshold
Output: weights for the intended tokens and pass to the semantic interpreter.

*4. STEP 1(b):*

The second part of the algorithm deals with the estimation of the relations based on the semantic processing. The relatedness is slatted based on the following properties

- Verbs
- Connectors
- Modalities
- Adjectives
- Pronouns

*5. STEP 2: Semantic clustering for the anomalous detection*

```
INPUT: Semantically analyzed dataset, database for the threat patterns
BEGIN
     Mapping of Semantic component with the dataset
    Grouping and ranking based on the dataset
END
OUTPUT: Semantic clusters with top ranks and the irrelevant clusters with lower rank
```

*6. Step 3: Semi supervised classification*

```
INPUT: Semantic ranked clusters
BEGIN
    Training labels construction based on the semantic scoring
    Constructing the classifier with the semi supervised learning
END
OUTPUT: Anomalous behaviour patterns
```

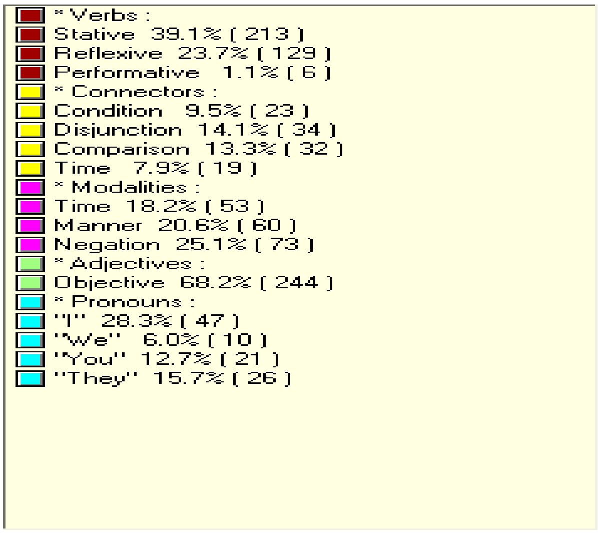
*7. Step 4: Updating of the history of threats*

```
INPUT: Anomalous behaviour patterns
BEGIN
    Database update with the anomalous behaviour patterns
END
OUTPUT: Updated Database for threats
```

## V. RESULTS AND DISCUSSIONS

To measure the similarity measure the similarity is measured for 30 sentence pairs used by [13]. The result obtained has been compared with the approached used in [13] and [14]. The table shows the comparison of the results.

The relations are plotted as the graph between various properties. The relatedness of the semantic analysis is processed for the identification of the various Equivalent classes and Relations between equivalents. The first stage of the algorithm is tested in with the text files and the results based on the semantic analysis and the building of relationship is done. The results obtained based on the text file given is



🟥 * Verbs :
🟥 Stative 39.1% ( 213 )
🟥 Reflexive 23.7% ( 129 )
🟥 Performative 1.1% ( 6 )
🟨 * Connectors :
🟨 Condition 9.5% ( 23 )
🟨 Disjunction 14.1% ( 34 )
🟨 Comparison 13.3% ( 32 )
🟨 Time 7.9% ( 19 )
🟪 * Modalities :
🟪 Time 18.2% ( 53 )
🟪 Manner 20.6% ( 60 )
🟪 Negation 25.1% ( 73 )
🟩 * Adjectives :
🟩 Objective 68.2% ( 244 )
🟦 * Pronouns :
🟦 "I" 28.3% ( 47 )
🟦 "We" 6.0% ( 10 )
🟦 "You" 12.7% ( 21 )
🟦 "They" 15.7% ( 26 )

Fig 2 : Relatedness based on the measures

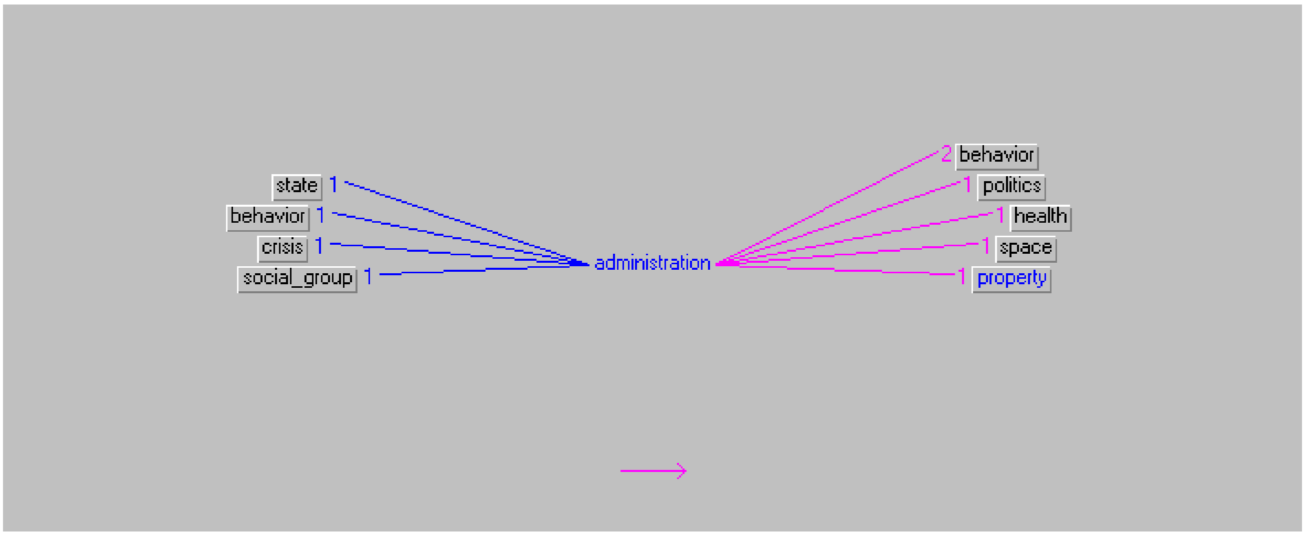


FIG 3:STAR GRAPH FOR THE SEMANTIC RELATION WITH REFERENCE FIELD <ADMINISTRATION>

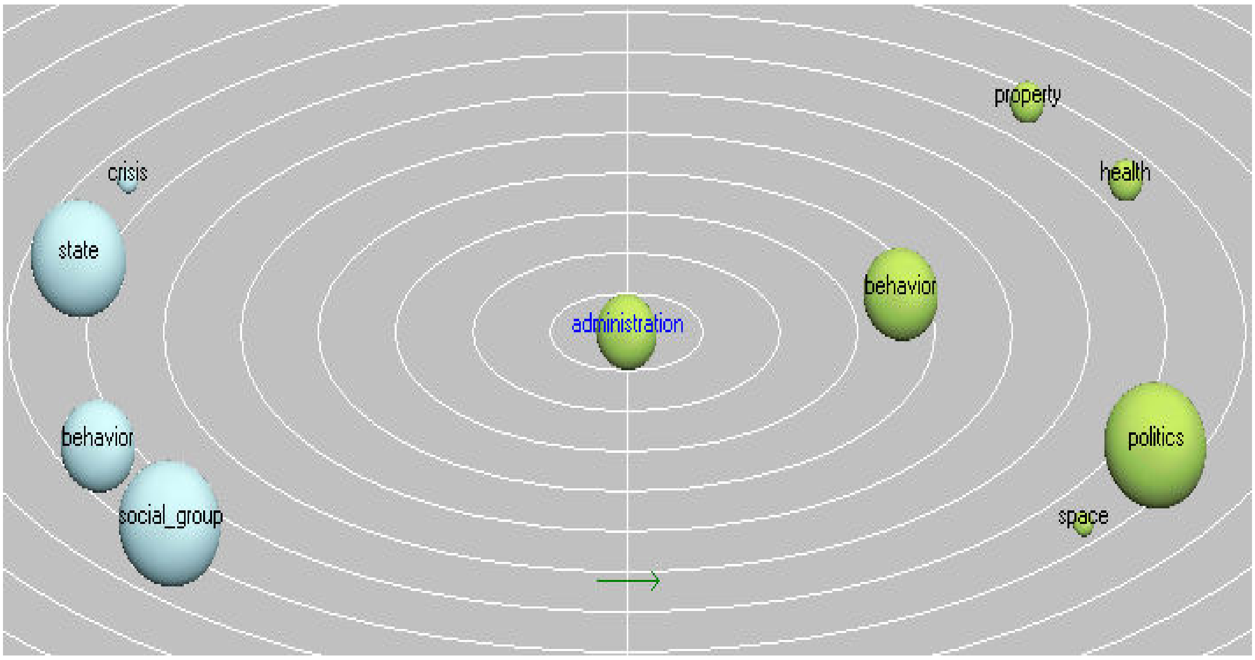


FIG 4: THE AREA GRAPH TO INDICATE THE DISTANCE RELATION

The similarity is measured as rating the similarity of meaning of the sentence pairs on the scale from 0.0 to 4.0 where the 0.0 represents the lower range - minimum similarity and the 4.0 represents the higher range - maximum similarity, the results are presented as the following table. The comparison shows the clear improvement of the proposed approach. The improvement is due to the proposed approach uses the weight fixation through the back propagation network. The similarity measure is fixed by the continuous training in the network.

TABLE 1

SENTENCE DATA SET RESULTS COMPARISON OF THE PROPOSED APPROACH WITH THE EXISTING METHODS

| R & G No. | R&G Word Pair | Human Similarity (Mean) | Li et al similarity method | Semantic text similarity method | Proposed Method |
|---|---|---|---|---|---|
| 1 | Cord Smile | 0.01 | 0.33 | 0.06 | 0.2925 |
| 5 | Autograph Shore | 0.01 | 0.29 | 0.11 | 0.090833 |
| 9 | Asylum Fruit | 0.01 | 0.21 | 0.07 | 0.1825 |
| 13 | Boy Rooster | 0.11 | 0.53 | 0.16 | 0.475833 |
| 17 | Coast Forest | 0.13 | 0.36 | 0.26 | 0.32 |
| 21 | Boy Sage | 0.04 | 0.51 | 0.16 | 0.136667 |
| 25 | Forest Graveyard | 0.07 | 0.55 | 0.33 | 0.2925 |
| 29 | Bird Woodland | 0.01 | 0.33 | 0.12 | 0.13 |
| 33 | Hill Woodland | 0.15 | 0.59 | 0.29 | 0.314167 |
| 37 | Magician Oracle | 0.13 | 0.44 | 0.2 | 0.216667 |
| 41 | Oracle Sage | 0.28 | 0.43 | 0.09 | 0.0975 |
| 47 | Furnace Stove | 0.35 | 0.72 | 0.3 | 0.325 |
| 48 | Magician Wizard | 0.36 | 0.65 | 0.34 | 0.368333 |
| 49 | Hill Mound | 0.29 | 0.74 | 0.15 | 0.1625 |
| 50 | Cord String | 0.47 | 0.68 | 0.49 | 0.530833 |
| 51 | Glass Tumbler | 0.14 | 0.65 | 0.28 | 0.303333 |
| 52 | Grin Smile | 0.49 | 0.49 | 0.32 | 0.346667 |
| 53 | Serf Slave | 0.48 | 0.39 | 0.44 | 0.476667 |
| 54 | Journey Voyage | 0.36 | 0.52 | 0.41 | 0.444167 |
| 55 | Autograph Signature | 0.41 | 0.55 | 0.19 | 0.205833 |
| 56 | Coast Shore | 0.59 | 0.76 | 0.47 | 0.509167 |
| 57 | Forest Woodland | 0.63 | 0.7 | 0.26 | 0.281667 |
| 58 | Implement Tool | 0.59 | 0.75 | 0.51 | 0.5525 |
| 59 | Cock Rooster | 0.86 | 1 | 0.94 | 1.018333 |

| | | | | | |
|---|---|---|---|---|---|
| 60 | Boy Lad | 0.58 | 0.66 | 0.6 | 0.65 |
| 61 | Cushion Pillow | 0.52 | 0.66 | 0.29 | 0.314167 |
| 62 | Cemetery Graveyard | 0.77 | 0.73 | 0.51 | 0.5525 |
| 63 | Automobile Car | 0.56 | 0.64 | 0.52 | 0.563333 |
| 64 | Midday Noon | 0.96 | 1 | 0.93 | 1.0075 |
| 65 | Gem Jewel | 0.65 | 0.83 | 0.65 | 0.704167 |

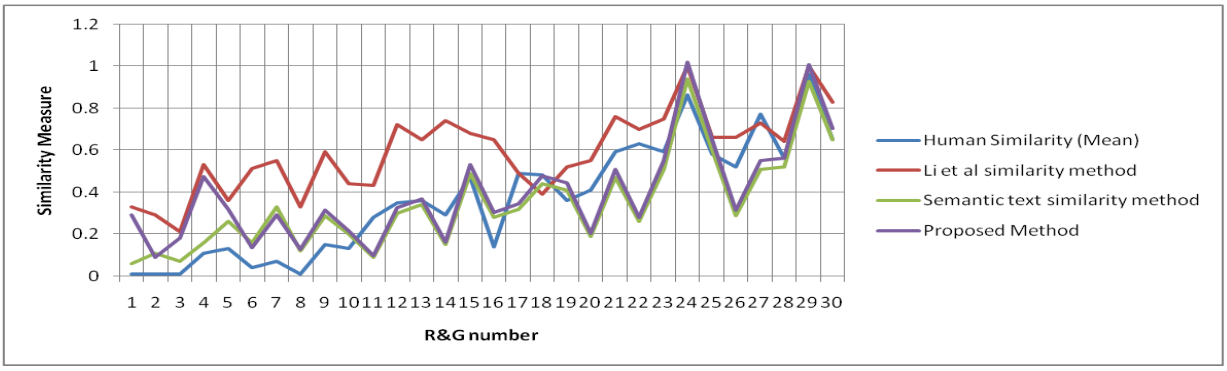


FIG 5: COMPARISON OF THE PROPOSED APPROACH WITH THE EXISTING APPROACHES

The stage two of the algorithm is simulated and the following screenshots shows the semantic clustering and the anomaly is shown in the different color

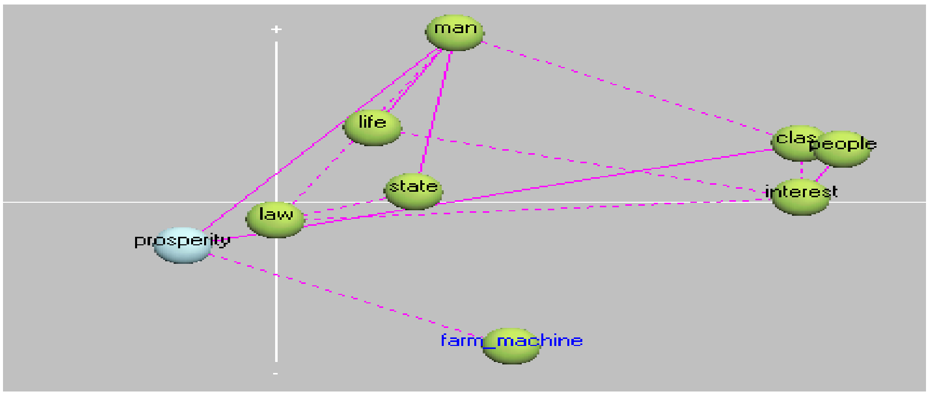


FIG 6: SEMANTIC CLUSTERING WHEN CLUSTER CENTRE = 9

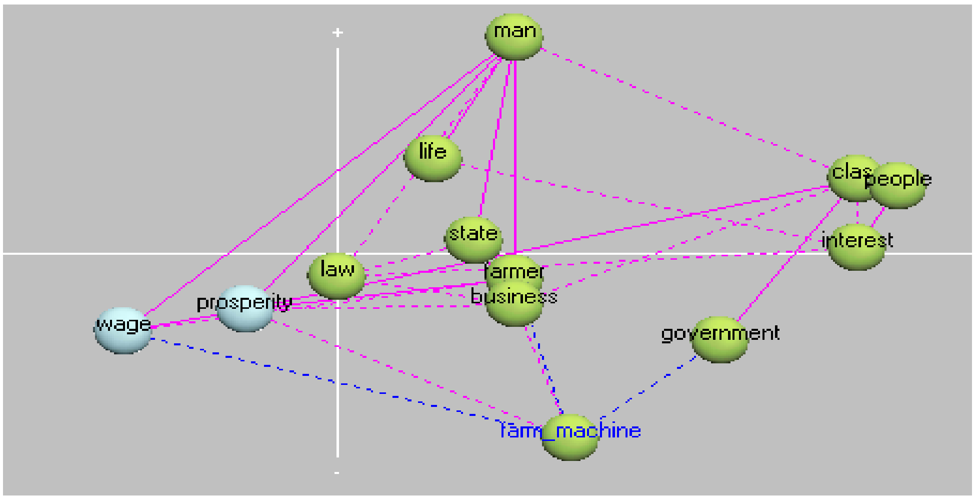


FIG 7 : SEMANTIC CLUSTERING WHEN CLUSTER CENTRE = 13

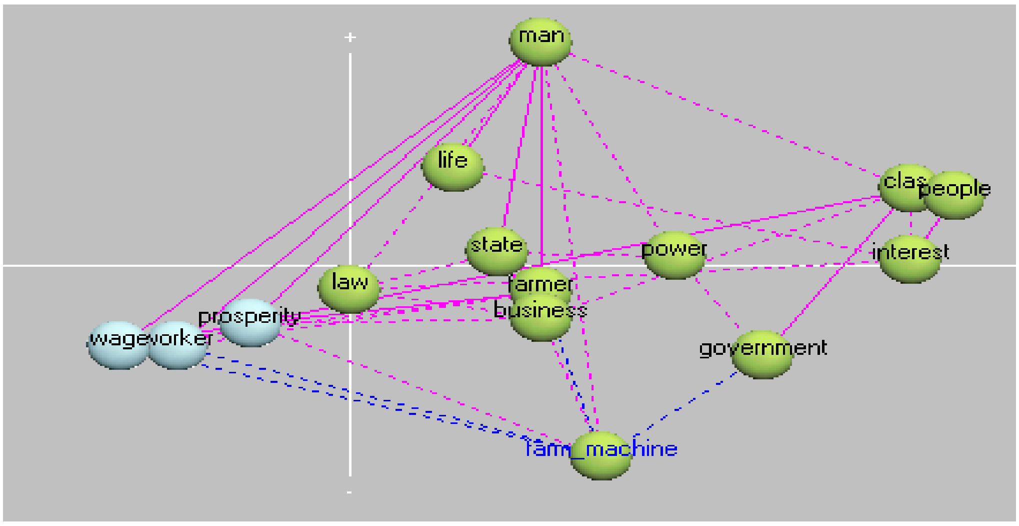


FIG 8: SEMANTIC CLUSTERING WHEN CLUSTER CENTRE = 15

TABLE 2

COMPARISON BASED ON THE MICRO F1 MEASURE

| Number of labeled data in each class | Co-training | TSVM | CBC | Proposed Approach |
|---|---|---|---|---|
| 1 | 0.23 | 0.41 | 0.48 | 0.56 |
| 2 | 0.46 | 0.483 | 0.58 | 0.62 |
| 4 | 0.556 | 0.54 | 0.62 | 0.73 |
| 8 | 0.62 | 0.6 | 0.66 | 0.76 |
| 16 | 0.67 | 0.66 | 0.69 | 0.81 |

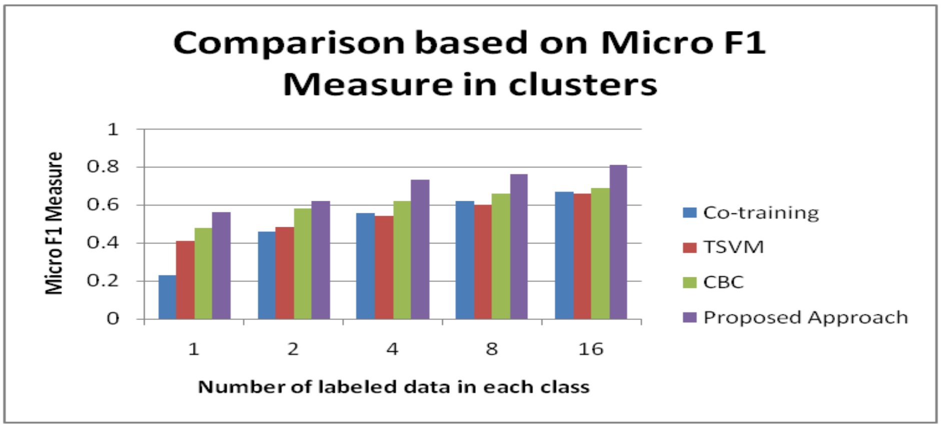


FIG 9: COMPARISON BASED ON THE MICRO F1 MEASURE

TABLE 3

COMPARISONS OF CLASSIFICATION ERROR IN PERCENTAGE

| Error Criterion | REL- ML | ABS SIG | Proposed method |
|---|---|---|---|
| Omission error | 3.04 | 4.64 | 2.98 |
| Commission error | 1.15 | 2.79 | 1.04 |
| Class Averaged Error | 2.095 | 3.715 | 2.01 |
| Total error | 1.95 | 3.58 | 1.82 |

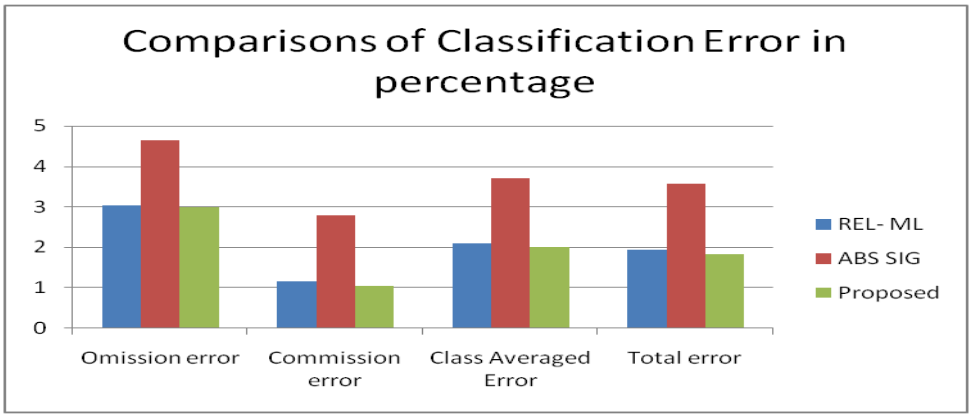


FIG 10: COMPARISONS OF CLASSIFICATION ERROR IN PERCENTAGE

## VI. CONCLUSION

In this research the Data leakage in the cloud computing transactions are highlighted. A framework based on the weighted indexing for the prevention of the data leakage is devised and discussed. The weight estimation is done through the back propagation network. It is followed by the relatedness estimation based on the various properties of the text. The sample results are presented. The results are compared with the existing methods and the graph is plotted. Semantic based clustering for the anomaly detection to find the data leak is discussed. The Clustering is further used for the classification to add up for the semi supervised classification. After classification the threat patterns are stored in the database for further preventive actions in the data transmission. The necessary theory is discussed and the proposed approach is promising with the results obtained. The future enhancement could be concentrated on extending the work to the network intrusion detection. The semantic nature of the proposed approach doesn't limit to the detection alone but paves a way for the origin and the context of the intrusion.

## REFERENCES

[1] Robert P. Mahowald, Connor G Sullivan, "Worldwide SaaS and Cloud Software 2012–2016 Forecast and 2011 Vendor Shares", Doc # 236184, International Data Corporation, August 2012.

[2] http://www.infosectoday.com/Articles/Securing_SaaS_Applications.htm

[3] Kevin Hamlen, Murat Kantarcioglu, Latifur Khan, BhavaniThuraisingham, Security Issues for Cloud Computing, International Journal of Information Security and Privacy, 4(2), 39-51, April-June 2010.

[4] Anthony Bisong, Syed M. Rahman, An Overview of the Security Concerns In Enterprise Cloud Computing, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011

[5] http://www.ironport.com/pdf/ironport_dlp_booklet.pdf

[6] Victor Delgado, Exploring the limits of cloud computing, Master's Thesis, Ocber 4, 2010

[7] Ashish Kumar, World of Cloud Computing & Security, International Journal of Cloud Computing and Services Science (IJ-CLOSER),Vol.1, No.2, June 2012, pp. 53-58

[8] Preeti Raman, HilmiGuneşKayacık, Anil Somayaji, Understanding Data Leak Prevention, Annual Symposium on Information Assurance (Asia), June 7-8, 2011, Albany, NY

[9] Vachharajani.N, M. J. Bridges, J. Chang, R. Rangan, G. Ottoni, J. A.Blome, G. A. Reis, M. Vachharajani, and D. I. August, "Rifle: An architectural framework for user-centric information-flow security," MICRO 37: Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture. Washington, DC, USA: IEEE Computer Society, 2004, pp. 243–254.

[10] White.J and D. Thompson, "Using synthetic decoys to digitally watermark personally-identifying data and to promote data security," 2006 International Conference on Security and Management, SAM 2006, June 26-29 2006, pp. 91–99.

[11] Lee.S, K. Lee, A. Savoldi, and S. Lee, "Data leak analysis in a corporate environment," in ICICIC '09: Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control. Washington, DC, USA: IEEE Computer Society, 2009, pp. 38–43.

[12] http://www.watchguard.com/tips-resources/grc/wp-data-loss-prevention.asp

[13] LI, Y., MCLEAN, D., BANDAR, Z., O'SHEA, J., AND CROCKETT, K. 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. 18, 8, 1138–1149.

[14] Islam, A. and Inkpen, D. 2008. Semantic text similarity using corpus-based word similarity and string similarity.ACM Trans. Knowl.Discov.Data. 2, 2, Article 10 (July 2008)