International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Novel Big Data Based Security Analytics Approach for Concept-Based Mining Model in Cloud Computing

M. Bhargavi Krishna[1] , Dr.K.Sarvanan[2]

[1]Student, Department of Computer Science And Engineering, Madanapalle Institute Of Technology And Science, Madanapalle, India)

[2]Assistant Professor, Department of Computer Science And Engineering, Madanapalle Institute Of Technology And Science, Madanapalle, India

ABSTRACT

The mining model can catch terms that present the ideas of the sentence, which prompts disclosure of the theme of the archive. Another concept based mining model that breaks down terms on the sentence, report, and corpus levels is presented. The concept based mining model can viably separate between non important terms as for sentence semantics and terms which hold the concepts that speak to the sentence meaning. The proposed mining model comprises of sentence-based concept analysis, document based concept analysis, corpus-based similarity measure, and concept based similarity measure. The term which adds to the sentence semantics is dissected on the sentence, record, and corpus levels as opposed to the customary investigation of the report as it were. The proposed model can effectively discover noteworthy coordinating ideas between reports, as indicated by the semantics of their sentences. The likeness between documents is figured in light of another concept based similarity measure. The proposed similarity measure takes full favorable position of utilizing the concept analysis measures on the sentence, report, and corpus levels in figuring the closeness between records.

Keywords : Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, conceptual term frequency, concept-base similarity.

I. INTRODUCTION

Text mining attempts to find out a new, previously unfamiliar information by applying techniques from natural language processing and data mining. Clustering, one of the traditional data mining techniques, is an unverified learning paradigm where clustering methods try to recognize inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intracluster similarity and low intercluster similarity.

The existing system proposed a novel big data based security examination way to deal with recognizing propelled assaults in virtualized foundations. System logs and in addition client application logs gathered occasionally from the visitor virtual machines (VMs) are put away in the Hadoop Distributed File System (HDFS). At that point, extraction of assault highlights is performed through graph based event correlation and MapReduce parser based distinguishing proof of potential assault ways. Next, assurance of assault nearness is performed through two-advance machine learning, namley strategic relapse is connected to

figure assault's restrictive probabilities regarding the properties, and conviction proliferation is connected to ascertain the faith in presence of an assault in light of them.

The proposed mining model comprises of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which adds to the sentence semantics is broke down on the sentence, document, and corpus levels as opposed to the traditional examination of the report as it were. The proposed model can proficiently discover critical coordinating ideas between documents, as per the semantics of their sentences. The likeness between documents is figured in view of another concept based similarity measure. The proposed similarity measure takes full favorable position of utilizing the concept analysis measures on the sentence,document, and corpus levels in computing the likeness between documents. Extensive arrangements of examinations utilizing the proposed idea construct mining model in light of various informational collections in content grouping are led.

II. RELATED WORK

Text mining concerns applying information mining systems to unstructured text. Information extraction (IE) is a type of shallow text understanding that finds particular bits of information in characteristic dialect records, changing unstructured content into an organized database.[1]

Measurable research in clustering has all around concentrated on informational collections depicted by ceaseless highlights [2] and its techniques are hard to apply to errands including emblematic highlights. Furthermore, these techniques are at times worried about helping the client in translating the outcomes acquired. Machine learning scientists have created theoretical clustering methods went for taking care of these issues. Following a long haul convention in AI, early theoretical clustering executions utilized rationale as the system of idea portrayal. In any case, coherent portrayals have been censured for obliging the subsequent group structures to be depicted by fundamental and adequate conditions. An option are probabilistic concepts which connect a likelihood or weight with every property of the idea definition[3]. In this paper, we propose an emblematic various leveled clustering model that makes utilization of probabilistic portrayals and broadens the customary thoughts of specificity-all inclusive statement ordinarily found in machine learning. We propose a parameterized measure that enables clients to indicate both the quantity of levels and the level of all inclusive statement of each level [5]. By giving some criticism to the client about the adjust of the all inclusive statement of the ideas made at each level and given the instinctive conduct of the client parameter, the framework[6]. It enhances client connection in the grouping procedure.

III. PROPOSED WORK.

Concept-Based Mining Model

The proposed concept- based mining model contains sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. A crude content record is the contribution to the proposed display. Each document has very much characterized sentence limits. Each sentence in the document is marked naturally in view of the PropBank documentations . In the wake of running the semantic part labeler, each sentence in the archive may have at least one marked verb contention structures. The quantity of produced marked verb contention structures is absolutely subject to the measure of data in the sentence. The sentence that has many named verb contention structures incorporates numerous verbs related with their contentions. The named verb contention structures, the yield of the part naming assignment, are caught

and broke down by the idea construct mining model with respect to sentence, report, and corpus levels.

In this model, both the verb and the contention are considered as terms. One term can be a contention to more than one verb in a similar sentence. This implies this term can have more than one semantic part in a similar sentence. In such cases, this term assumes essential semantic parts that add to the significance of the sentence. In the idea based mining model, a marked term either word or expression is considered as idea.

The target behind the concept based analysis assignment is to accomplish a precise examination of ideas on the sentence, record, and corpus levels as opposed to a solitary term examination on the archive as it were.

Document-Based Concept Analysis

To dissect every idea at the document level, the conceptbased term frequency tf , the quantity of events of an concept (word or expression) c in the first document, is figured. The tf is a nearby measure on the document level.

Corpus-Based Concept Analysis

To remove concepts that can separate between documents, the concept based document frequency df, the quantity of archives containing concept c, is ascertained. The df is a worldwide measure on the corpus level. This measure is utilized to remunerate the ideas that lone show up in few reports as these concepts can separate their documents among others. The way toward figuring ctf, tf , and df measures in a corpus is accomplished by the proposed calculation which is called Concept-based Analysis Algorithm.

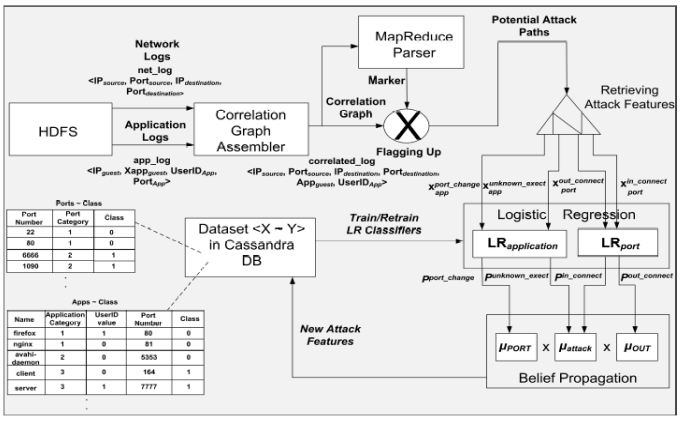

Figure 1. Information flow of BDSA in text mining.

Concept-Based Analysis Algorithm

1. ddoci is a new Document
2. L is an empty List (L is a matched concept list)
3. sdoci is a new sentence in ddoci
4. Build concepts list Cdoci from sdoci
5. for each concept ci 2 Ci do
6. compute ctfi of ci in ddoci
7. compute tfi of ci in ddoci
8. compute dfi of ci in ddoci
9. dk is seen document, where k ¼ f0; 1; . . . ; doci _ 1g
10. sk is a sentence in dk
11. Build concepts list Ck from sk
12. for each concept cj 2 Ck do
13. if (ci == cj) then
14. update dfi of ci
15. compute ctfweight ¼ avgðctfi; ctfjÞ
16. add new concept matches to L
17. end if
18. end for
19. end for
20. output the matched concepts list L

The concept-based analysis algorithm portrays the way toward ascertaining the ctf, tf , and df of the coordinated ideas in the documents. The method starts with preparing another document (at line 1) which has welldefined sentence limits. Each sentence is semantically marked by . The lengths of the coordinated concepts and their verb contention structures are put away for the concept-based similarity calculations in .Every concept (in the for circle, at line 5) in the verb contention structures,

which speaks to the semantic structures of the sentence, is handled successively. Every concept in the present document is coordinated with alternate concepts in the already handled documents. To coordinate the concepts in past documents is refined by keeping an concept list L, which holds the section for each of the past archives that offers an concept with the present document. After the document is handled, L contains all the coordinating concepts between the present concept and any past document that offers no less than one concept with the new document. At last, L is yield as the rundown of documents with the coordinating concepts and the vital data about them. The concept-based analysis algorithm is fit for coordinating every idea in another record ðdþ with all the already handled reports in Oðmþ time, where m is the quantity of ideas in d.

IV. EXPERIMENTAL EVALUATION

The BDSA approach is evaluated by creating a guest VM running CentOS 6.5 as well as another two guest VMs running Ubuntu 14.04 (64-bit) on one of the aforementioned HPC server nodes.

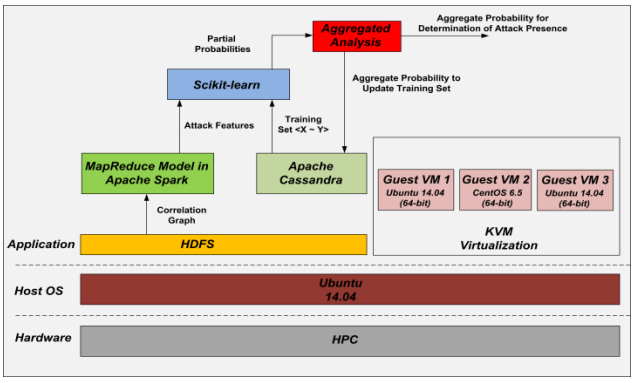

Figure 2. soft ware stack on text mining node

V. RESULT ANALYSIS

.

The use of data mining to mine attack patterns in network logs is used in Beehive security approach. Basically, Beehive detects attacks through correlating logs obtained from different points within the enterprise network. First it collects logs from different points.Example web server logs,user logs, IDS logs, etc within the enterprise network are collected over a two-week period. The logs are then parsed using known network configuration details to extract 15 features for each host based on the IP addresses of websites accessed, details of the application used, violation of
network policy, and changes in network flow characteristics caused by the accesses.

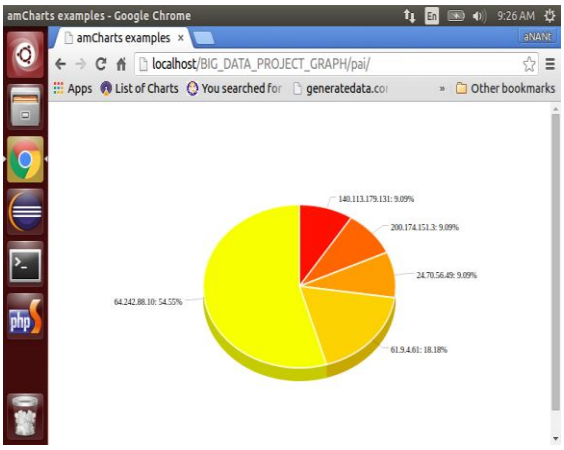

Figure 3. The rates of er ror logs in an data Set that are predicted in future

However, it is limited in providing prompt threat quarantine and elimination due to its post-factum nature.

VI. CONCLUSION

This work crosses over any barrier between regular dialect preparing and text mining disciplines. Another concept-based mining model made out of four segments, is proposed to enhance the content clustering quality. By abusing the semantic structure of the sentences in documents, a superior content clustering result is accomplished. The main segment is the sentence-based concept analysis which examines the semantic structure of each sentence to catch the sentence ideas utilizing the proposed theoretical term recurrence ctf measure. At that point, the second part, document based concept analysis, dissects every idea at the document level

utilizing the concept-based term frequency tf. The third part breaks down ideas on the corpus level utilizing the document frequency df worldwide measure. The fourth part is the concept-based similarity which permits measuring the significance of every idea regarding the semantics of the sentence, the subject of the document, and the segregation among archives in a corpus. By joining the variables influencing the weights of ideas on the sentence, archive, and corpus levels, an concept-based similarity measure that is fit for the precise estimation of pairwise documents is conceived. This permits performing idea coordinating and idea based closeness figurings among records in an extremely strong and precise way. The nature of content grouping accomplished by this model essentially outperforms the customary singleterm-based methodologies.

VII. REFERENCES

1. KJ. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.
2. B Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.
3. K Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.
4. G Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.
5. G Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
6. UY. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI '00), pp. 627-632, 2000.
7. MF. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, July 1980.
8. H Jin, M.-L. Wong, and K.S. Leung, "Scalable Model-Based Clustering for Large Databases Based on Data Summarization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.
9. M Junker, M. Sintek, and M. Rinck, "Learning for Text Categorization and Information Extraction with ILP," Proc. First Workshop Learning Language in Logic, 1999.
10. S. Soderland, "Learning Information Extraction Rules for Semi- Structured and Free Text," Machine Learning, vol. 34, nos. 1-3, pp. 233-272, Feb. 1999.