Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 11, November 2018, pg.195 – 203

An Different Similarity Measures with N-Grams For Text Documents Comparison

R.Anushya[1], A.Finny Belwin[2], A.Linda Sherin[3], Dr. Antony Selvadoss Thanamani[4]

[1]Research Scholar Department of Computer Science & Bharathiar University, India
[2]Research Scholar Department of Computer Science & Bharathiar University, India
[3]Research Scholar Department of Computer Science & Bharathiar University, India
[4]Professor and Head Department of Computer Science NGM College, Pollachi, India
[1] anushya7373@gmail.com; [2] belwin35@gmail.com; [3] linz15sherin@gmail.com; [4] selvdoss@gmail.com

Abstract— Data analysis is a new, emerging field in research area and business. The huge numbers of documents are available in form of unstructured, semi-structured and structured data. Estimating similitude between writings is a critical errand for a few applications. In the existing many similarity algorithms has been proposed for text similarity calculation based on distance between documents in the text processing field. The increased attention has led to many of techniques for measuring semantic based document similarity algorithms. The document similarity application teachers or other users can easily search documents containing some specific terminology. In this paper propose a different type of document similarity calculation based on cosine similarity, Jaccard similarity and Euclidean distance with n-grams algorithms. The similarity metrics between documents can be defined in several ways depending on the representation of the documents. The experimental result compares that three similarity algorithms and finally evaluate which is best similarity measure.

Keywords— Document Similarity, N-Grams set, Cosine Similarity, Jaccard Similarity, Euclidean distance.

I. INTRODUCTION

Discovering similitude between composed records in substantial assemblages of content is helpful; one may need to, for example, consequently recognize spam messages from genuine ones, or gathering news articles by their subject [1]. Bunching, a machine learning strategy used to discover groupings inside datasets dependent on shared attributes is a successful method for achieving this objective [2]. Be that as it may, applying bunching on inalienably loud information, for example, content information, regularly has disadvantages; since it is almost constantly done in high-dimensional space, it is fantastically hard to check whether

one's grouping calculation is suitable or if one's bunches bode well. In this manner, picturing high-dimensional information in fewer measurements is a useful initial step when performing group examination on information [3].

Math is utilized to ascertain closeness where it commands the field. After the beginning of two new fields in a century ago, Information Theory and Computer Science, the point of comparability has not turned out to be littler by any stretch of the imagination. Rather by utilizing the PC it has been simpler to discover how comparative at least two things are to one another. This possibility caused by these fields where science can be connected and the ease to make the estimations quick, have influenced people to create calculations to make better approaches to compute similitude less demanding, quicker and as right as conceivable [4]. The instinctive ideas of likeness ought to be about the equivalent for basically everybody.

One of the predetermined classifications is printed closeness; taking at least two strings and contrasting them with one another with discover how comparable they are. This classification is so intriguing and loaded up with advantage that numerous college educators and understudies have considered far into this classification and composed numerous papers on the literary comparability. Purpose behind this is people are extraordinary; they got diverse thoughts and distinctive limits with regards to how comparative something depends on a scale [5]. The many have examined which calculations to utilize and when to utilize them, with regards to printed similitude. In the wake of separating from the majority of calculations to indicate likeness, at that point to data recovery and last to printed closeness, there are as yet numerous lefts. Still they can be put into each their containers where three of those zones for content comparability stand out more than the rest;

- Vector space model
- Edit distance
- Ontology based

The issue does not just comprise in the amount they stand out yet additionally how great they are for a given errand. Some are sufficient to run an inquiry on closeness on shorts strings, while the long ones give a poor outcome, others the turnaround [6]. Some run quickly on short strings while others the invert. Closeness score is the measure to indicate how comparative two arrangement of information are to one another. The arrangement of information can be about as for this situation around two unique writings [7]. To discover the similitude is to discover the correlation between the two messages and grade it after a score framework. For Vector Space Model there is no truly need to create a scoring framework since it is an outcome by means of the cosine work gives a range from 0 to 1, where 0 implies that the vectors are symmetrical and subsequently absolutely inverse significance the two writings pieces are altogether unique while 1 implies that the vectors are pointing at precisely the same bearing and in this way a similar importance both content pieces are completely comparable otherwise known as they are the equivalent. The numbers in the middle of 0 and 1 demonstrates how comparable the two writings are relying upon the two vectors blessed messenger to one another. Taking these similitudes and increasing with 100 gives the level of how comparative the two writings are.

Information Science is another, developing field in research and industry [8]. The meaning of an information researcher ranges from having the capacity to apply a couple of factual calculations to having the capacity to perform information mining, information examination and machine learning. Archive closeness is the metric that characterizes how comparative two given writings are. The writings being referred to are written in human dialect however should be examined by projects [9]. This is the final result of this

examination and furthermore the simple means by which we assess the consequences of various methodologies.

Rest of the paper is organized as follows, section I contain overview of cloud computing and different type of cloud deployment. Section II contain review of exiting cloud scheduling and power ware resource algorithms, Section III contain proposes system and module implementations, Section IV contain result and discussion , performance analysis , Section V concludes.

II. RELATED WORK

In existing factual measures are should have been incorporated to upgrade the likeness between two records. Similitude measures have as of late turned into a rising theme of enthusiasm among information mining and Big Data inquire about networks. Closeness measures have been generally utilized in different spaces of research, for example, for remaking phylogenetic trees. Comparability measure is likewise used to ascertain the similitude between two arranged trees. The goal is to discover transformative relationship among different natural species or different elements dependent on the comparability and contrasts among their physical or hereditary highlights. Similitude measure has likewise been utilized in separation based ordering for string closeness to enhance database look. Further, comparability measurements have additionally been adjusted for looking at diagrams and ascribed trees. Closeness measures have additionally been utilized for assessing the significance of highlights in information mining and looking at data content.

The element choice process makes this disparity measure particularly appropriate in huge, high dimensional archive accumulations. Its execution is approved on a few test sets starting from institutionalized datasets. The uniqueness measure is contrasted with the notable cosine divergence measure utilizing the normal F-proportions of the various levelled agglomerative bunching result [10]. This new divergence measure results in an enhanced grouping result acquired with a lower required computational time.

The current similitude measure for a specific application field dependent on foundation learning and highlight parameters explicit to that field; rather we build up a general numerical hypothesis of closeness that utilizes no foundation information or highlights explicit to an application region. Thus it is, without changes, relevant to various regions and even to accumulations of items taken from various zones. The strategy naturally zooms in on the predominant likeness angle between each two items. To understand this objective, we initially characterize a wide class of closeness separations. At that point, we demonstrate that this class contains a specific separation that is all inclusive in the accompanying sense: for each combine of articles the specific separation is not exactly any "powerful" remove in the class between those two items [11]. This all inclusive separation is known as the "standardized data remove" (NID), it is appeared to be a measurement, and, naturally, it reveals all likenesses all the while that powerful separations in the class reveal a solitary comparability each [12].

The semantic measures to assess the semantic relatedness between two ideas. This measure makes utilization of the examples for semantically right ways and the data theoretic worldview presented. The SNOMED-CT cosmology of medicinal ideas [13]. The measures incorporate two way based measures, and three estimates that expand way based measures with data content insights from corpora. The separation based ordering methods for string nearness look. We first demonstrate that string separation proportions of intrigue, for example, the pressure remove and weighted character alter separate give nearly measurements. We at

that point demonstrate to adjust vantage point trees and other separation based ordering strategies to oblige relatively metric separations. First methodology for evaluating the selectivity of TF-IDF based cosine similitude predicates [14]. We assess our methodology on three distinctive genuine datasets and demonstrate that our technique regularly delivers evaluates that are inside 40% of the real selectivity. In this paper, we talk about a system for evaluating the selectivity of TF-IDF based cosine similitude predicates. We make utilization of a factual synopsis of the conveyance of various tokens in the database.

The current Euclidean separation measure, alternate measures have practically identical viability for the Partitional content archive bunching assignment. Pearson connection coefficient and the found the middle value of KLD uniqueness measures are marginally better in that their subsequent bunching arrangements are more adjusted and have a closer match with the physically made class structure [15]. In the interim, the Jaccard and Pearson coefficient estimates discover more rational bunches. The EEG signals are portrayed by bends on the complex of intensity ghostly thickness grids [16]. By enriching the complex with a Riemannian measurement, we acquire the Riemanian separate between two on the complex [17]. In light of this, the proportion of difference is then characterized. To best encourage the order of comparable and unique EEG flag sets, we acquire the ideally weighted Riemannian separation intending to render motions in various classes more detachable while those in a similar class smaller.

III.PROPOSED METHODOLOGY

In this proposed framework the closeness measurements between records can be characterized in a few different ways relying upon the portrayal of the reports, if the archives are spoken to as vectors where every component is a word at that point approaches dependent on coordinating coefficient can be utilized. This dataset contained all reports made by a whole class bunch amid one semester.

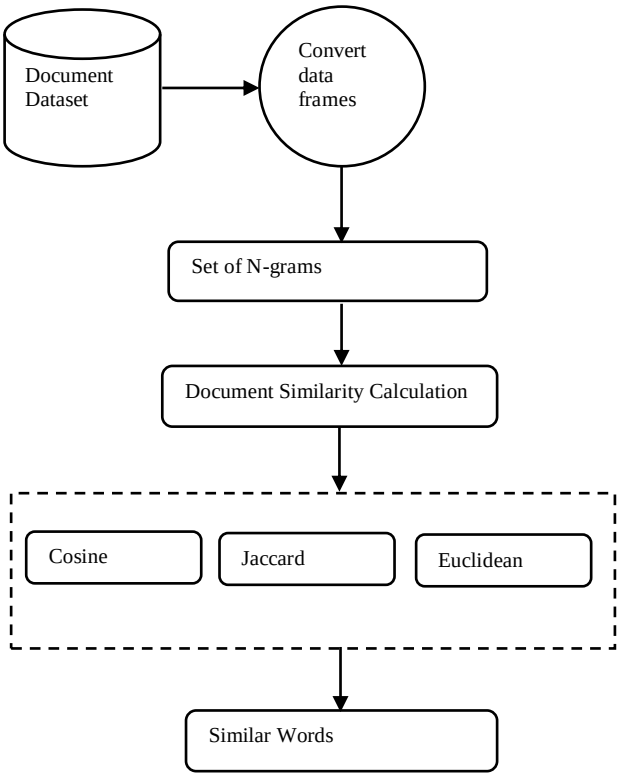

Figure 1 Overview of Proposed Flow Diagram

In this figure 1 demonstrate the proposed stream outline of report likeness computation, first dataset changes over into information outline dependent on information outline transformation. After each word calculate N grams. At last, archive likeness figuring utilizing three closeness calculation and think about.

A. Set of N-Grams

In this first steps utilizing N-gram filtration Scheme as the foundation method. In this procedure, first play out a basic pre-preparing assignment, which remove stop-words, straightforward boisterous sections and fixes the phrase(s) and sentence(s) boundaries. It at that point, readies a lexicon of N-grams, utilizing word reference arrangement. At last, it applies an example filtration calculation, to play out an early gathering of higher length N-grams that fulfils the base recurrence [18]. All the while, all events of these N-grams in records are supplanted with one of a kind capitalized alphanumeric mixes.

This substitution keeps the odds of pointless describing of N-grams (that have a higher estimation of N) by other halfway organized N-grams having a lower estimation of N. From each qualified term, it gathers the accompanying highlights:

F = frequency of N-gram in the document.

P = total weighted frequency in the document,

Referring to the situation in these sentences, either the underlying half or the last three-fourths of a sentences'. The weight an incentive because of the situation of K-gram, D, can be represented as:

$$D = (S - S_f) \qquad (1)$$

S = total number of sentences in the document, S_f= sentence number in which the given K-gram occurs first, in the document.

B. Jaccard with K-Grams.

Take into account two sets A = {0, 1, 2, 5, 6} and B = {0, 2, 3, 5, 7, 9}. How similar are A and B? The Jaccard similarity is defined

$$JS(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (2)$$

$$= \frac{|\{0,2,5\}|}{|\{0,1,2,3,5,6,7,9\}|} = \frac{3}{8} = 0.375$$

More documentations, given a set A, the cardinality of A signified $|A|$ include what number of components are A. The crossing point between two sets A and B is signified $A \cap B$and uncovers all things which are in the two sets. The relationship between two sets A and B is indicated $A \cup B$ and uncovers all things which are in either set.

Affirm that JS fulfills the properties of a likeness.

To completely sum up set likenesses (in any event those that are amiable to extensive scale systems) we present a third set activity. The symmetric distinction between two sets A and B is denoted$A \triangle B = (A \cup B) \backslash (A \cap B)$. Note that n is called set minus and A\B is the majority of the components in A, aside from those likewise in B. In this way, the symmetric contrast of A and B portrays all components in A or B, however not in both.

Here consider the follow class of similarities. It uses$\overline{A \cup B} = [n] \backslash (A \cup B)$, where [n] is a superset that all sets A and B we consider a subset from.

$$S_{x,y,z,z'}(A,B) = \frac{x|A \cap B| + y|A \cup B| + z|A \Delta B|}{x|A \cap B| + y|\overline{A \cup B}| + z'|A \Delta B|} \qquad (3)$$

It can define several concrete instances.

Jaccard Similarity defined

$$JS(A,B) = S_{1,0,0,1}(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \Delta B|} = \frac{|A \cap B|}{|A \cup B|} \qquad (4)$$

So how do we put this together? Consider the (N = 2)-grams for each D1, D2, D3, and D4:

D1: [I am], [am Sam]

D2: [Sam I], [I am]

D3: [I make], [make not at all], [not at all like], [like green], [green eggs], [eggs also], [also ham]

D4: [I make], [make not at all], [not at all like], [like them], [them Sam], [Sam I], [I am]

Now the Jaccard similarity is as follows:

JS (D1, D2) = 1/3 ≈ 0.333

JS (D1, D3) = 0 = 0.0

JS (D1, D4) = 1/8 = 0.125

JS (D2, D3) = 0 = 0.0

JS (D3, D4) = 2/7 ≈ 0.286

JS (D3, D4) = 3/11 ≈ 0.273

The unique dynamic structure of sets to figure this separation proficiently and at greatly expansive scale.

C. Euclidean Distance Document Similarity

Euclidean distance is a standard measurement for geometrical issues. It is the customary separation between two and can be effortlessly estimated with a ruler in a few dimensional space. Euclidean separation is generally utilized in bunching issues, including grouping content. It fulfils all the over four conditions and in this manner is a genuine metric [19]. It is likewise the default separate measure utilized with the K-implies calculation.

Measuring distance between content archives, given two reports c and d spoken to by their term vectors and separately, the Euclidean separation of the two records is characterized as

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{1/2} \qquad (5)$$

Where the term set is $T = \{t_1, \ldots\ldots$. As mentioned previously, it uses the tf value as term weights, that is

$$w_{t,a} = tfidf(a).$$

D. Cosine Document Similarity

At the point when documents are represented to as term vectors, the comparability of two records relate to the connection between's the vectors. This is measured as the cosine of the point between vectors, that is, the alleged cosine comparability [20]. Cosine likeness is a standout amongst the most prominent closeness estimates connected to content records, for example, in various data recovery applications.

The both archives $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$, their cosine similarity is

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|} \qquad (6)$$

Where and are m-dimensional vectors by gone the term set $T = \{t_1, \ldots\ldots t_n$. Each dimension represents to a term with its weight in the report, which is non-negative. Accordingly, the cosine closeness is non-negative and limited between [0, 1].

An important property of the cosine closeness is its freedom of archive length. For instance, joining two indistinguishable duplicates of an archive d to get another pseudo record d0, the

cosine likeness among d and is 1, which implies that these two reports are respected to be indistinguishable. Then, given another record l, d and will have a similar similitude incentive to l, that is, $sim(\overrightarrow{t_d}, \overrightarrow{t_l}) = sim(\overrightarrow{t_{d'}},)$. As it were, records with a similar structure yet extraordinary sums will be dealt with indistinguishably. Entirely, this does not fulfil the second state of a measurement; in light of the fact that after all the blend of two duplicates is an alternate protest from the first archive. Be that as it may, by and by, when the term vectors are standardized to a unit length, for example, 1, and for this situation the portrayal of d and d0 is the equivalent.

IV. RESULTS AND DISCUSSION

The data for performing this experiment comprises of datasets, which are selected from standard datasets used in text mining research. These datasets are much broadened; these are from various sources, from various applications, distinctive kind of reports, and they contain diverse number of classifications. The similarity between document pairs is the most important fact to retrieve related documents in the IR system. The greater the similarity score, the higher the similarity has between documents. In the same way, the more common keyword contains in each document pair, the higher similarity score has for this document pair. There are many types of similarity functions to calculate the similarity scores. Among them, cosine and overlap are more popular than others. However, in some condition, they can't decide the similarity score correctly.

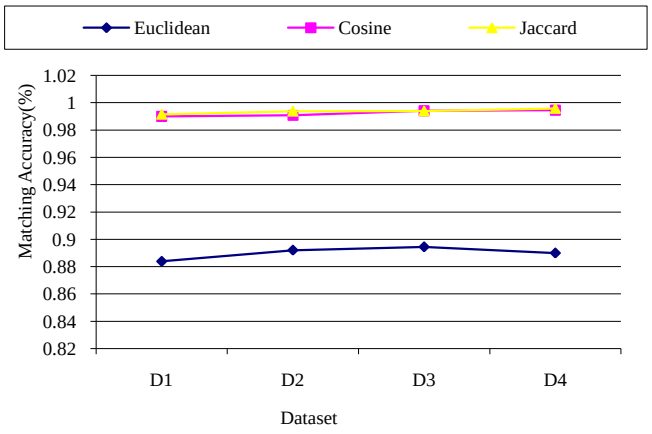

Figure 2 Comparison of matching accuracy with different documents

V. CONCLUSION

Text Mining is the excavations completed by the PC to get something new that originates from data separated consequently from information wellsprings of various contents. In this paper propose three kinds of record comparability computation calculations. Relative examination for discovering the most important document for the given arrangement of watch words by utilizing three likenesses. This dataset contained all archives made by a whole class amass amid one semester. Which implies that in a large portion of the cases an educator needs to peruse and recollect the substance of every one of their understudies work? Despite the fact that, contingent upon the situation he or she may just need to know the primary substance of the reports for class exercises, for example, doing class bunches dependent on the closeness of the records. The propose framework utilize diverse sort of archive closeness check utilizing

Cosine, Jaccard and Euclidean separation. Recent years have seen are established enthusiasm for growing new likeness methods. The exploratory near outcome indicates Jaccard likeness give high precision than existing system.

REFERENCES

[1] J. A. Aslam and M. Frost, "An information-theoretic measure for document similarity," in Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, 2003, pp. 449–450.

[2] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," IEEE Trans. Knowl. Data Eng., vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[3] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou, "Pairwise-adaptive dissimilarity measure for document clustering," Inf. Sci., vol. 180, no. 12, pp. 2341–2358, 2010.

[4] C. G. Gonz´alez,W. Bonventi, Jr., and A. L. Vieira Rodrigues, "Density of closed balls in real-valued and autometrized boolean spaces for clustering applications," in Proc. Brazilian Symp. Artif. Intell., 2008, pp. 8–22.

[5] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vit´anyi, "The similarity metric,"IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 3250–3264, Dec. 2004.

[6] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A similarity measure for text classification and clustering,"IEEE Trans. Knowl. Data Eng., vol. 26, no. 7, pp. 1575–1590, Jul. 2014.

[7] L. Mazuel and N. Sabouret, "Semantic relatedness measure using object properties in an ontology," inProc. Int. Semantic Web Conf., 2008, pp. 681–694.

[8] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," J. Biomed. Inform., vol. 40, no. 3, pp. 288–299, 2007.

[9] S. C. Sahinalp, M. Tasan, J. Macker, and Z. M. Ozsoyoglu, "Distance based indexing for string proximity search," inProc. 19th Int. Conf. Data Eng., 2003, pp. 125–136.

[10] S. Tata and J. M. Patel, "Estimating the selectivity of tf-idf based cosine similarity predicates,"ACM Sigmod Rec., vol. 36, no. 2, pp. 7–12, 2007.

[11] J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," Bioinformatics, vol. 23, no. 10, pp. 1274–1281, 2007.

[12] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, 2007.

[13] Y. Zhao and G. Karypis, "Comparison of agglomerative and partitional document clustering algorithms," Defense Tech. Inf. Center Document, Fort Belvoir, VA, CA, Tech. Rep. TR-02-014, 2002.

[14] V. K. Shrivastava, N. D. Londhe, R. S. Sonawane, and J. S. Suri, "First review on psoriasis severity risk stratification: An engineering perspective," Comput. Biol. Med., pp. 52–63, vol. 63, 2015.

[15] A. Cardoso-Cachopo, "Improving methods for single-label text categorization," Ph.D. dissertation, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, Lisboa, Portugal, 2007.

[16] A. Huang, "Similarity measures for text document clustering," in Proc. 6th New Zealand Comput. Sci. Res. Student Conf., Christchurch, New Zealand, 2008, pp. 49–56.

[17] k. Sashi, A.S. Thanamani, dynamic replication in a data grid using a modified bhr region based algorithm, future generation computer systems 27 (2), (2011), pp. 202 210.

[18] K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation Journal: International Journal for Research in Science & Advanced Technologies, Vol 1.Issue-2,2013,Ms.R.Malarvizhi and Dr. Antony Selvadoss Thanamani.

[19] R. Malathi Ravindran and Antony Selvadoss Thanamani, "K-Means Document Clustering using Vector Space Model", Bonfring International Journal of Data Mining, Volume 5, Issue 2, July 2015, Pages 10-14.

[20] Umajancy.S, Dr. Antony Selvadoss Thanamani"An Analysis on Text Mining-Text Retrieval and Text Extraction "International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013.

[21] V Chitraa, Dr.Antony Selvadoss Thanamani, "A survey on preprocessing methods for web usage data", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011.

[22] K Jothimani , Dr.Antony Selvadoss Thanamani, "An Algorithm for Mining Frequent Itemsets", IJCSET |March 2012| Vol 2, Issue 3,1012-1015

[23] Kanchana S. and Antony Selvadoss Thanamani, "BOOSTING THE ACCURACY OF WEAK LEARNER USING SEMI SUPERVISED CoGA TECHNIQUES", VOL. 11, NO. 15, AUGUST 2016 ISSN 1819-6608 ARPN Journal of Engineering and Applied Sciences ©2006-2016 Asian Research Publishing Network (ARPN). All rights reserved.

[24] Priyadharsini.C, Dr. Antony Selvadoss Thanamani, "Finding Phytochemical Component Analysis in Medicinal Plants a Data Mining Approach", International Journal of Modern Computer Science and Applications (IJMCSA) Volume No.-4, Issue No.-5, September, 2016. ISSN: 2321-2632 (Online) Page No: 63 to 66

[25] Priyadharsini.C, Dr. Antony Selvadoss Thanamani, "Prediction of Hidden Patterns and Relationship Using Ensemble Learning Algorithm", International Journal of Modern Computer Science (IJMCS) Volume 5, Issue 1, February, 2017. ISSN: 2320-7868 (Online) Page No: 17 to 19

[26] N.Raveendran1 Dr. Antony Selvadoss Thanamani, "Autonomic Cloud Services to Enhance Secure Data Sevices for Digital Library", American Research Journal of Computer Science and Information Technology (ARJCSIT) Volume 1, 10 pages

[27] Anu Mendiz. R , Dr. Antony Selvadoss Thanamani, "Data Mining Approach in Ethnopharmacolgy Based on Cloud Storage", IJSRD (International Journal for Scientific Research and Development), Volume : 4, Issue : 8 Publication, Date: 01/11/2016 Page(s): 104-106

