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. 12, December 2018, pg.239 – 248

A Similarity Measure for Documents Using Clustering Technique

R.Anushya[1], A.Linda Sherin[2], A.Finny Belwin[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— Text clustering is a critical use of information mining. It is worried about gathering comparable content archives together. Content report grouping assumes a vital job in giving natural route and perusing systems by sorting out a lot of data into few important clusters. Grouping technique needs to implant the reports in an appropriate similitude space. In this paper we look at four prominent similitude measures: cosine similarity, Jaccard similarity, Euclidean distance and Correlation Coefficient related to various sorts of vector space portrayal (Boolean, term recurrence and reverse report recurrence) of archives. Clustering of archives is performed utilizing summed up k-Means; a Partitioned constructed grouping strategy in light of high dimensional inadequate information speaking to content reports. Execution is estimated against a human-forced arrangement of Topic and Place classes. We led various tests and utilized entropy measure to guarantee factual noteworthiness of results. Cosine, Pearson relationship and Jaccard similitude rise as the best measures to catch human categorization conduct, while Euclidean measures perform poor.

Keywords- Clustering, Jaccard similarity, Cosine similarity, Euclidean measure, Correlation coefficient, K-means.

I. INTRODUCTION

Today, with the quick advancements in innovation we can amass colossal measures of information of various types. Information mining developed as a field worried about the extraction of valuable learning from information [1]. Information mining procedures have been connected to illuminate an extensive variety of true issues. Clustering is an unsupervised information mining procedure where the names of information objects are obscure. It is the activity of the clustering system to recognize the order of information protests under examination. Clustering can be connected to various types of information including content. When managing literary information, articles can be reports, sections, or words [2]. Content clustering alludes to the way toward gathering comparable content records together. The issue can be detailed as pursues: given an arrangement of reports it is required to partition them into various gatherings, with the end goal that archives in a similar gathering are more like each other than to records in different gatherings. There are numerous uses of content grouping including: archive association and perusing, corpus outline, and record arrangement clustering has been proposed for use in perusing an accumulation of reports [3] or in sorting out the outcomes returned by a web search tool because of client's question [4] or help clients

rapidly recognize and centre around the pertinent arrangement of results. Client remarks are grouped in numerous online stores, for example, Amazon.com to give cooperative proposals. In communitarian bookmarking or labelling, groups of clients that share certain characteristics are recognized by their comments. Archive clustering has likewise been utilized to consequently produce Hierarchical groups of reports [5]. This paper is composed as pursues. The section 2 manages the related work in content report grouping; section 3 depicts the record portrayal utilized in the trials. Section 4 discuss about the likeness measures and their semantics. Section 5 displays the K-means grouping calculation and Section 6 clarifies experiment settings, assessment methodologies, results and investigation and Section 7 finishes up and examines future work.

II. RELATED WORK

Text Clustering is one of the critical uses of information mining. In this section, we audit a portion of the related work in this field.

Luo et al. [3] utilized the ideas of record neighbours and connections with the end goal to improve the execution of k-means and bisecting k-means clustering. Utilizing a couple shrewd closeness work and a given similitude edge, the neighbours of a record are the reports that are viewed as like it. A connection between two records is the quantity of regular neighbours. The ideas were utilized in the choice of introductory group centroids and in report closeness estimating.

Many grouping systems have been proposed in the writing. Clustering calculations are fundamentally ordered into Hierarchical and Partitioning strategies [2, 3, 4, 5]. Various levelled clustering strategy works by gathering information objects into a tree of groups [6]. These strategies can additionally be arranged into agglomerative and disruptive Hierarchical grouping relying upon whether the Hierarchical deterioration is shaped in a base up or top-down design. K-means and its variations [7, 8, 9] are the most notable parcelling techniques [10].

Bide and Shedge proposed a clustering pipeline to enhance the execution of k-means grouping. The creators received a partition and-overcome way to deal with clusters reports in the 20 Newsgroup dataset Documents were isolated into gatherings where pre processing, highlight extraction, and k-means clustering were connected on each gathering. Report similitude was computed utilizing the cosine comparability measure. The proposed methodology accomplished better outcomes when contrasted with standard k-means as far as both clusters quality and execution time.

Progressive strategies deliver a settled arrangement of segments, with solitary, comprehensive clusters at the best and singleton groups of individual focuses at the base. Each halfway dimension can be seen as joining two clusters from the following lower level (or part a group from the following more elevated amount). The consequence of a Hierarchical clustering calculation can be graphically shown as tree, called a dendogram.

Rather than Hierarchical strategies, Partitional clustering procedures make a one-level (un-settled) apportioning of the information focuses. In the event that K is the coveted number of groups, Partitional approaches regularly discover all K clusters without a moment's delay. Balance this with customary Hierarchical plans, which cut up a group to get two clusters or consolidation two groups to get one. Obviously, a Hierarchical methodology can be utilized to produce a level parcel of K groups, and moreover, the rehashed use of a Partitional plan can infer Hierarchical clustering.

There are various Partitional methods; however we will just portray the K-means calculation which is broadly utilized in archive grouping. K-means depends on the possibility that an inside point can speak to a clusters. Specifically, for K-means we utilize the idea of centroids, which is the mean or middle purpose of a gathering of focuses. Note that centroids never relates to a real information point. The calculation is examined in detail in section 5

III.METHODOLOGY

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

A. Data Collection

This work tries different things with two bench mark datasets "Reuters 21578 distribution 1.0" and Classic dataset gathered from uci.kdd archives. The Reuters-21578 gathering is dispersed in 22 documents. Every one of the initial 21 records (reut2-000.sgm through reut2-020.sgm) contains 1000 reports, while the last (reut2-021.sgm) contains 578 archives.

B. Document Representation

In order to diminish the intricacy of the records and make them less demanding to deal with, the archive must be changed from the full content adaptation to a report vector which depicts the substance of the report. The portrayal of an arrangement of archives as vectors in a typical vector space is known as the vector space model. In the vector space model of IR, reports are spoken to as vectors of highlights speaking to the terms that happen inside the gathering. It is additionally named as pack of words, where words are expected to show up freely and the request is irrelevant. The estimation of each component is known as the term weight and is typically an element of term's recurrence (or tf-idf) in the archive, alongside different variables.

Vector Space portrayal of a report includes three steps [7]. Initial step is the archive ordering where content bearing terms are extricated from the records. The second step is to register the weights of filed terms to upgrade recovery of archives significant to the client. The last advance is recognizing the likenesses between the records.

The vector space show is a typical portrayal of content records. Give D a chance to be a gathering of reports and let T= $\{t1, t2, \ldots, tn\}$ be the arrangement of terms showing up in D. A record $x \in D$ can be represented to as an n−dimensional vector in the term space T. Let wx, ti be a chance to be the occasions a term $ti \in T$ shows up in x, at that point the vector of x is characterized as:

$$x = \{wx, t1, wx, t2 \ldots wx, tn\} \quad (1)$$

C. Extracting Index Terms

It includes pre processing content records, apply stemming, expel stop words and tokenize the content. Reports in vector space can be spoken to utilizing Boolean, Term Frequency and Term Frequency – Inverse Document Frequency.

In Boolean portrayal, in the event that a term exists in a record, the relating term esteem is set to one else it is set to zero. Boolean portrayal is utilized when each term has break even with significance and is connected when the reports are of little size.

In Term Frequency and Term Frequency Inverse Document Frequency the term weights must be set. The term weights are set as the basic recurrence include of the terms the archives. This mirrors the instinct that terms happen much of the time inside an archive may mirror its significance more emphatically than terms that happen less every now and again and should therefore have higher weights.

Each archive d is considered as a vector in the term-space and spoken to by the term recurrence (TF) vector:

$$dtf = [tf1, tf2, \ldots .. tfD]$$

Where tfi is the recurrence of term i in the archive and D is the aggregate number of exceptional terms in the content database.

The second factor is utilized to give a higher weight to words that just happen in a couple of reports. Terms that are restricted to few records are valuable for separating those archives from whatever is left of the gathering, while terms that happen much of the time over the whole accumulation aren't useful. The backwards archive recurrence term weight is one method for doling out higher weights to these more discriminative words. IDF is characterized by means of the portion N/ni, where, N is the aggregate number of archives in the accumulation and ni is the quantity of reports in which term i happens.

Subsequently, the tf– idf portrayal of the report d is:

$$dtf - idf = [tf1\ log(n/df1), tf2\ log(n/df2), \ldots .., tfD\ log(n/dfD)]$$

To represent the reports of various lengths, each record vector is standardized to a unit vector $(i.e., ||dtf - idf||| = 1)$.

IV. SIMILARITY MEASURES

There are numerous measurements for estimating archive similitude. We centre around four regular measures in this space which are: cosine similarity [7], Jaccard similarity coefficient, Euclidean measure and Correlation Coefficient.

A. *Cosine Similarity Measure*

For record clustering, there are distinctive likeness estimates accessible. The most regularly utilized is the cosine work. For two records di and dj, the similitude between them can be figured

$$CosineSimilarity(x_1 + x_2) = \frac{V(x_1) \cdot V(x_2)}{|V(x_1)||V(x_2)|}$$

Since the report vectors are of unit length, the above condition is disentangled to:

$$Cos(x1, x2) = \quad x1.x2$$

At the point when the cosine esteem is 1 the two reports are indistinguishable, and 0 if there is nothing in like manner between them (i.e., their record vectors are symmetrical to one another).

B. *Jaccard Coefficient*

The Jaccard coefficient, which is once in a while alluded to as the Tanimoto coefficient, measures closeness as the convergence isolated by the association of the articles. For content report, the Jaccard coefficient thinks about the whole weight of shared terms to the aggregate weight of terms that are available in both of the two archives however are not the common terms.

The Cosine Similarity might be stretched out to yield Jaccard Coefficient if there should arise an occurrence of Binary properties

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$JD(A,B) = 1 - J(A,B) or JD(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

C. *Euclidean Similarity*

This is the most common, "regular" and natural method for figuring a separation between two examples. It considers the contrast between two examples straightforwardly, in view of the extent of changes in the example levels. This separation type is generally utilized for informational collections that are appropriately standardized or with no unique dispersion issue.

$$\text{Euclidean Distance } |x \to y| = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$$

D. *Pearson Correlation coefficient*

This separation depends on the Pearson correlation coefficient that is computed from the example esteems and their standard deviations. The relationship coefficient 'r' takes esteems from – 1 (huge, negative connection) to +1 (expansive, positive relationship). Adequately, the Pearson separate - dp – is figured as $dp = 1 - r$ and lies between 0 (when relationship coefficient is +1, i.e., the two examples are most comparable) and 2 (when connection coefficient is - 1).

$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum xY^2 - \frac{(\sum xY^2)}{N}}}$$

V. CLUSTERING ALGORITHM

For consequent trials, the standard K-means calculation is picked as the grouping calculation. This is an iterative Partitional grouping process that means to limit the minimum squares blunder standard [6]. As made reference to already, Partitional grouping calculations have been perceived to be more qualified for dealing with extensive record datasets than Hierarchical ones, because of their moderately low computational prerequisites [17, 19, 18]. The standard K-means calculation fills in as pursues. Given an arrangement of information objects D and a pre-indicated number of group's k, k information objects are haphazardly chosen to introduce k clusters, every one being the centroids of clusters. The rest of the items are then appointed to the clusters spoken to by the closest or most comparative centroids. Next, new centroids are recomputed for each cluster and thusly all records are re-allocated dependent on the new centroids. This progression emphasizes until a joined and settled arrangement is achieved, where all information objects stay in a similar group after a refresh of centroids. The produced clustering arrangements are locally ideal for the given informational collection and the underlying seeds. Diverse decisions of beginning seed sets can result in altogether different last parcels. Techniques for discovering great beginning stages have been proposed [20]. Be that as it may, we will utilize the fundamental K-means calculation on the grounds that improving the clustering isn't the principle focal point of this paper.

K-means are outstanding and broadly pertinent clustering calculations. Here, we give a short portrayal of these calculations

K-means is an iterative grouping calculation. It depends on dividing information focuses into k groups utilizing the idea of centroids. The cluster centroids is the mean estimation of the information focuses inside a group. The created parcels highlight high intra-clusters similitude and between group variety. The quantity of clusters, k, is a pre-decided parameter of the calculation.

K-means functions as pursues:

1) K information focuses are self-assertively chose as clusters centroids
2) The similitude of every datum point to each group centroids is computed. At that point information point are re-allotted to the clusters of the nearest centroids
3) The k centroids are refreshed dependent on the recently allocated information focuses.
4) Stages 2 and 3 are rehashed until the point that combination is come to.

VI. EXPERIMENTAL RESULTS

Here, we cluster our dataset utilizing k-means [24] with each grouping procedure, we fabricate models utilizing diverse estimations of k and the four similitude estimates depicted previously. The Rapid Miner plat-shape was utilized in our analysis. This open source stage gives a well disposed GUI and backings every one of the means of Knowledge Discovery from Data, including: information pre-handling, information mining, demonstrate approval, and result perception.

A. Dataset

This work tries different things with two seat stamp datasets "Reuters 21578 dispersion 1.0" and Classic dataset gathered from uci.kdd vaults. The Reuters-21578 gathering is circulated in 22 documents. Every one of the initial 21 records (reut2-000.sgm through reut2-020.sgm) contains 1000 archives, while the last (reut2-021.sgm) contains 578 reports. Records were increased with SGML labels, and a comparing SGML DTD was delivered, so the limits of vital segments of reports are unambiguous. Every REUTERS tag contains express determinations of the estimations of traits, for example, TOPICS, LEWISSPLIT, CGISPLIT, OLDID, and NEWID. These ascribes are intended to distinguish records and gatherings of reports. Eg: <TOPICS>, </TOPICS>, <PLACES>, </PLACES>, <BODY>, </BODY>. Each will be delimited by the labels <D> and </D>. There are 5 classes Exchanges, Organizations, People, Places and Topics in the Reuters dataset and every classification has again sub classes altogether 672 sub classifications. We have gathered the TOPICS and PLACES classification sets to shape the dataset. The TOPICS class set contains 135 classifications and PLACES class set contains 175 classes. From these records we gather the legitimate content information of every class by separating the content which is in the middle of <BODY>, </BODY> and put in a content report and named it as indicated by theme and place.

Classic dataset comprises of four unique accumulations CACM, CISI, CRAN and MED. We have thought about 800 archives of the aggregate 7095 records.

In these datasets, a portion of the archives comprises single word just, so it is useless to take such reports for record dataset. For disposing of these invalid reports we apply record decrease on every classification, which restores the archives that bolsters mean length of every classification. For record decrease we build the Boolean frameworks of all reports by classification astute and compute mean length of every class and expelled the archives from the dataset which doesn't bolster mean length. By this we got legitimate records. From these substantial archives we have gathered 800 reports of four classifications each. From Reuters we have thought about 200 records of every classification (ACQ, EARN of TOPICS classification and UK, USA, of PLACES classification) totalling to 800 reports and from exemplary dataset 200 archives of every classification again totalling to 800 records.

VII. EVALUATION MEASURES

We utilize entropy as a proportion of nature of the groups (with the proviso that the best entropy is acquired when each clusters contains precisely one information point). Give CS a chance to be a grouping arrangement. For each clusters, the class circulation of the information is ascertained first, i.e., for group j we figure pij, the "likelihood" that an individual from clusters j has a place with class I. At that point utilizing this class dissemination, the entropy of each group j is figured utilizing the standard equation.

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

Where the entirety is assumed control over all classes. The aggregate entropy for an arrangement of bunches is figured as the whole of the entropies of each group weighted by the span of each cluster:

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

Where nj is the measure of group j, m is the quantity of bunches, and n is the aggregate number of information focuses

VIII. RESULTS AND DISCUSSION

In this area, we examine the nature of the gotten grouping models dependent on the estimations of the bunching assessment measures. We contrast all the gotten models with locate the best mix of bunching method and closeness measure.

In this work seed focuses are statically picked, however proficiency can be enhanced if seeds chosen are irregular or run the code more than once to check the productivity. As appeared in Tables 1a, 1b and Tables 2a 2b, Euclidean separation performs most exceedingly bad while the execution of different measures is very comparable. From our outcomes it is seen that Boolean portrayal with Pearson measure has non-zero groups. Consequently the general entropy for Boolean portrayal table shows NAN esteems for different measures as a portion of bunches is vacant. On a normal, the Jaccard and Pearson measures are somewhat better in producing more lucid bunches, which implies the groups have bring down entropy scores. Table 3a indicates one segment as created by the Boolean Pearson measure utilizing Reuter's dataset, and Table 3b demonstrates one segment as produced by the TF-IDF Jaccard Coefficient measure utilizing Classic dataset which has the least entropy esteem.

	Cosine	Jaccard	Euclidean	Pearson
Boolean	NAN	NAN	NAN	0.33
Freq. Count	0.36	0.36	0.42	0.38
TF-IDF	0.36	0.38	0.42	0.37

Table 1 a Portrayals Entropy Results of Different Vector Space Using Reuters dataset

	Cosine	Jaccard	Euclidean	Pearson
Boolean	NAN	NAN	NAN	0.06
Freq. Count	0.16	0.12	0.30	0.06
TF-IDF	0.06	0.07	0.30	0.06

Table 2b Entropy Results of Different Vector Space Representations Using Classic dataset

Comparative as over, the Euclidean separation is again turned out to be an insufficient metric for demonstrating the closeness between reports. The Jaccard and Pearson's coefficient will in general beat the cosine closeness.

	Cosine	Jaccard	Euclidean	Pearson
Cluster[0]	0.41	0.16	0.38	0.41
Cluster[1]	0.33	0.38	0.44	0.33
Cluster[2]	0.26	0.40	0.40	0.28
Cluster[3]	0.31	0.16	0.42	0.30

Table 3a: TF-IDF Entropy Results

	Cosine	Jaccard	Euclidean	Pearson
Clusters[0]	0.05	0.01	0.30	0.01
Clusters[1]	0.01	0.08	0.30	0.04
Clusters[2]	0.06	0.07	0.30	0.07
Clusters[3]	0.13	0.11	0.00	0.10

Table 4b: TF-IDF Entropy Results using Classic dataset

	ACQ	EARN	UK	USA	LABEL
Cluster[0]	173	71	64	8	ACQ
Cluster[1]	18	12	107	57	UK
Cluster[2]	8	115	15	12	EARN
Cluster[3]	1	2	14	123	USA

Table 5a: Bunching Results from Boolean Pearson Correlation Measure utilizing Reuters dataset

	CAC	CIS	CRA	MED	LABEL
Cluster[0]	0	1	0	166	MED
Cluster[1]	8	5	199	30	CRA
Cluster[2]	3	166	0	4	CIS
Cluster[3]	189	28	1	0	CAC

Table 6b: Clustering Results from TFIDF Jaccard Measure using Classic dataset

We have utilized the grouping precision as a proportion of a clustering result. Grouping precision r is characterized as

$$r = \frac{\sum_{i=1} ai}{n}$$

Where ai is the quantity of examples happening in both clusters i and it's relating class and n is the quantity of occurrences in the dataset. The clustering precision is more for TF-IDF portrayal with Pearson's and Jaccard coefficient measures. The great dataset has appeared over 94 percent precision.

IX. CONCLUSION AND FUTURE WORK

In this study we discovered that every one of the measures have huge impact on Partitional clustering of content reports with the exception of the Euclidean separation measurer. Pearson connection coefficient is marginally better as the subsequent clustering arrangements are more adjusted and is closer to the physically made classes. The Jaccard and Pearson coefficient estimates discover more sound groups. Considering the kind of group investigation engaged with this examination, we can see that there are three parts that influence the last outcomes—portrayal of the records, separation or comparability estimates considered, and the clustering calculation itself. In our future work our intension is to apply semantics learning to the archive portrayals to speak to connections among terms and concentrate the impact of these similitude measures thoroughly.

REFERENCES

[1] Han, J., Kamber, M... *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 3rd ed.; 2011. ISBN 978-0-12-381479-1.

[2] Aggarwal, C.C., Zhai, C... A Survey of Text Clustering Algorithms. In: Aggarwal, C.C., Zhai, C., editors. *Mining Text Data*. Springer US; 2012, p. 77–128.

[3] Luo, C., Li, Y., Chung, S.M... Text document clustering based on neighbours. *Data & Knowledge Engineering* 2009; 68(11):1271–1288.

[4] Hartigan, J.A... *Clustering Algorithms*. New York, NY, USA: John Wiley & Sons, Inc.; 99th ed.; 1975. ISBN 978-0-471-35645-5.

[5] Elkan, C.. Using the Triangle Inequality to Accelerate k-Means. In: Fawcett, T., Mishra, N., editors. *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*. AAAI Press; 2003, p. 147–153.

[6] Kaufman, L. and Rousseeuw, P.J., . Clustering by means of Medoids. In: Y. Dodge and North-Holland, editor. *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Springer US; 1987, p. 405–416.

[7] Blair, D.C... Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London: Butterworths; 1979: 208 pp. Price: $32.50. *Journal of the American Society for Information Science* 1979; 30(6):374–375.

[8] Bide, P., Shedge, R.. Improved Document Clustering using k-means algorithm. In: *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 2015, p. 1–5.

[9] Lang, K.. 20 newsgroups data set. 2008 (accessed 2015-12-18). "http://www.ai.mit.edu/people/jrennie/20Newsgroups/".

[10] C. J. Van Rijsbergen, (1989), Information Retrieval, Buttersworth, London, Second Edition.

[11] R. Malarvizhi, Dr. Antony Selvadoss Thanamani, "K-Nearest Neighbour in Missing Data Imputation", International Journal of Engineering Research and Development, Volume 5 Issue 1-November-2012,

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

[13] Used Mathematical Models For Finding Multiple Data Imputation In Main Stream, International Journal of Emerging Trends in Science and Technology, IJETST- Vol.||03||Issue||05||Pages 540-545||May||ISSN 2348-9480. Mrs. P.Logeshwari and Dr.Antony Selvadoss Thanamani.

[14] Chitra, V & Antony Selvadoss Thanamani 2010, 'A Survey on Pre-processing Methods for web Usage Data', (IJCSIS) International Journal of Computer Science and Information Security,vol. 7, no. 3, pp. 78-83.

[15] 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.

[16] 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.

[17] 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

[18] 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

[19] 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.

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

[21] 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.

[22] Priyadharsini.C, Dr. Antony Selvadoss Thanamani , "Imputation of Missing Data Using Ensemble Algorithms" International Journal of Modern Computer Science (IJMCS) Volume 5, Issue 1, February, 2017. ISSN: 2320-7868 (Online) Page No: 20 to 23

[23] Priyadharsini.C, Dr. Antony Selvadoss Thanamani ," An Improved Novel Index Measured Segmentation Based Imputation Algorithm for Missing Data Imputation", International Journals of Advanced Research in Computer Science and Software Engineering (UGC Approved No: 48958) Registered DOI: www.dx.doi.org / 10.23956/ IJARCSSE. ISSN: 2277-128X (Volume-7, Issue-6) June, 2017 page no: 283-286

[24] Priyadharsini.C, Dr. Antony Selvadoss Thanamani, "A Novel Index Measured Segmentation Based Imputation Algorithm (with Cross Folds) for Missing Data Imputation" International Journal of Electrical Electronics & Computer Science Engineering.*(UGC Approved No:44927),* Volume 4, Issue 3 (June, 2017) | E-ISSN : 2348-2273 P-ISSN : 2454-1222, page no: 22-24

[25] M. Ramaraj, Dr. Antony Selvadoss Thanamani " Plagiarism detection paradigm for web content using similarity analysis approach" ISSN2348 -9928IJAICTVolume1,Issue5,September2014Doi:01.0401/ijaict.2014.04.01Published on05 (10) 2014 © 2014 IJAICT(www.ijaict.com)

[26] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl Constrained K-means clustering with Background Knowledge. Proceedings of the Eighteenth International Conference on Machine Learning, pages 577 − 584, 2001.

[27] Johny Antony P, Dr. Antony Selvadoss Thanamani, "A Privacy Preservation Framework for Big Data (Using Differential Privacy and Overlapped Slicing)", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 10, October 2016

[28] R.NANDHAKUMAR1 AND ANTONY SELVADOSS THANAMANI, "A Survey on E-Health Care for Diabetes Using Cloud Framework", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 4, Issue 10, October 2017

