International Journal of Engineering & Technology, 7 (4) (2018) 4503-4508

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 19458
Research paper

A novel approach: big data analysis based on multi-view data visualization using clustering similarity measure

Srinivasa Rao Madala [1] *, V. N. Rajavarman [2], T. Venkata Satya Vivek [1]

[1] Research scholars, Department of Computer Science & Engineering, Dr M.G.R Educational And Research Institute University, Chennai, India
[2] Professor, Department of Computer Science & Engineering, Dr M.G.R Educational And Research Institute University, Chennai, India
**Corresponding author E-mail: mr.srinu13@gmail.com*

Abstract

In big data, data visualization is annotable concept to represent data for competent data analysis to handle high dimensional data. In data visualization, there are three main properties i) to characterize without loss of data patterns ii) without any changes in data pattern change the attributes iii) data visualization among structure and unstructured data attributes for data examination. There are various types of data visualization are existing virtually to identify data analysis (i.e. topic based data revelation, attribute based data visualization, audio based data visualization and text based data visualization in different data sets). Parallel coordinate is proficient and effective data visualization tool to analyze and handle multi attribute high dimensional data. It is based 5Ws density sending and receiving data visualization, it also read data patterns and attributes with reduces the overlapping to data patterns. Parallel measure is a labeling property to characterize data with affiliation objects in data set appraisal with different pair of attributes. We need to get better parallel coordinate tool to sustain multi-attribute object relations, so we recommend and implement novel method i.e. (Similarity Measure Centered with Multi Viewpoint (SMCMV)) approach and related clustering approaches to represent data. Using multi-viewpoint, we can accomplish assessment based similarity index with data visualization. Using multi viewpoint, we present hypothetical analysis based on multi attributes presentation. Our experimental results gives best data representation in data visualization with capable similarity measure on real time document evaluation with different known collected clustering approaches.

Keywords: *Data Visualization; Parallel Co-Ordinate; Multivariate Attributes; Clustering Methods; Similarity Measure; Multi Viewpoint.*

1. Introduction

Structured and unstructured data in big data visualization contain unusual forms of data like image, audio and video and unruffled this data from different multiple data sets based on the dimension time and space complexity evaluation. For example Face book generates 25 GB of data which contain following user's individual details and their sharing data with mutual and personal friends. Thousands, hundreds of different dimensional attributes by journal providing different data to analyze multiple attribute dimensions to hold data visualization. Because of escalating rapid usage of big data in various applications, different authors projected unusual association and classification and clustering to scrutinize high dimensional data. Parallel coordinate data visualization is one of the capable approach to signify data lacking change their data patterns from overall data. Sample data visualization with different dimensions as shown fig 1.

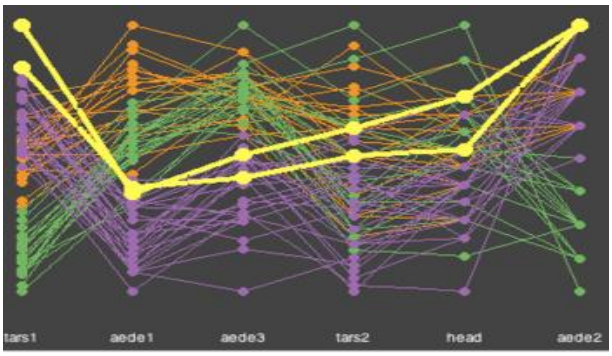

Fig. 1: Parallel Coordinate Plot of the Data Visualization for Flea Data.

Neighbor attribute partition is as shown in fig 1 with different node axes values in dimensional set. Some of the researchers and introduced to characterize efficient data visualization with different node data management with same attributes. Primarily data visualization consists three data representations in real time data presentations, topic based data visualization, which consist about exacting topic with algorithm process like network traffic revelation cloud data visualization. Data type based data visualization, which consists accurate type of data like text based data visualization, audio and video data visualization with different formations. Data set visualization, which consists meticulous data sets like social and network, slanting data sets with different data patterns. To represent data in these three ways, conventionally develop

Parallel coordinate 5Ws density model, in that scrutinize data attributes and represent those attributes in parallel axes arrangement for several data set with data types and topics evaluation in real time data set presentation. To measure resemblance between attributes in data visualization with relationships then parallel coordinate 5W density management not contented in data evaluation. Therefore in the contextual discussion we recommend and render approaches encompassing Similarity Measure Focused on Multi-Viewpoint (SMCMV)) approach and interrelated clustering approaches to characterize data. This advance follows multi-view data depiction with dimensions in attribute relationship. In that clustering is an insistent concept and topic in data repossession based on attributes, in that we essential the structure data formation and formulate them into obligatory and significant data presentation. So our proposed advance follows clustering properties to read and present data in dissimilar dimensions with attribute relations. And also our approach absolutely follows multi-view data presentation with respect attribute presentation. We also calculate similarity measure in attribute partitioning in data set exploration; There is a huge significance of Similarity Measures while considering aspects like success and failure in case of data presentation in clustering Procedure. Main objectives of our proposed approach as follows:

We are presenting and recommending a significant approach to identify the similarity between the data objects with different relations in high dimensional data evaluation.

1) Proposed similarity measure with different clustering calculations with provable quality and performance consistent.
2) Display multi-view data visualization with different data patterns.
3) Give efficient data visualization with multiple attributes with clustering calculations.

The rest of the section in the contextual discussion is arranged as mentioned: Section 2 relates the related work about visual data presentation techniques, section 3 discuss about parallel coordinate density model with data visualization. Section 4 describes proposed approach i.e. SMCMV and it's implementation procedure. Section 5 formalizes the computational and performance evaluation of proposed approach with real time data sets and plot the results, section 6 concludes overall conclusion.

2. Related work

Inselberg [3] suggested in his earlier work about the proportional compose plot which is one of the significant work amongst the prominent strategies and Wegman suggested the proportional compose strategy as something with a extraordinary viewpoint data inquire about [4]..The n-definitional data set and the corresponding direction could be authorized upon when parallel tomahawks are connected with the straight section of a two definitional airplane. As established in the academic research work [5], numerous strategies have been prescribed to offer comprehension of multivariate information utilizing engaging creation strategies. Comparable facilitate plots, as a straightforward however capable geometrical high-dimensional information creation strategy and signifies the set of N-dimensional data available within the 2-dimensional region associated with factual diligence. Graphical grouping, pivot reorganization and point of view concentrating are normal approaches to reduce jumbles running in parallel fits. Dasgupta et al. [6] suggested one in view of screen-space measurements to pick the tomahawks structure by enhancing sets up of tomahawks. Huh et al. given a related region between two nearby tomahawks as opposed to the proportional territory in regular PCP parallel tomahawks. Additionally, the shapes having a few measurable property associating data considers on close-by tomahawks are portrayed in artistic works [7] too. Zhou et al. [8] changed over the straight-line sides into shapes to moderate up the obvious chaos in grouped creation. They additionally utilized the splatting structure [2] to recognize gatherings and lessen noticeable wreckage. Kai Lun Chung and Wei Zhuo [2] contributed to the guideline to prevent over plotting of the data and safeguarding the thickness data and developed visually explanatory devices: decision outlines and respects diagrams, to lessen visual chaos running in parallel blends. The clients could understand the chosen regions through the decision outline which is a brushing gadget. The respects diagram masterminds gatherings and offer communications for clients to see more about the associations between gatherings. Julian Heinrich et al [9] planned BiCluster Audience that consolidates heat maps and parallel blends plots to find information designs. The BiCluster Audience contains numerous intelligent elements, for example, pivot obtaining, go shading, or cruising that diminish data filling in noticeable diagram. Matej Novotny and Helwig Hauser [8] have collected the irregular data set and at that point inclined and focused on the viewpoint in masterminded parallel directions to moderate up the filling issues. Xiaoru Yuan et al [7] distributed the components as it is running in equivalent combinations to combine equivalent facilitates as well as scatterplot distribution that reduced the information swarming. Clients have the facility to rearrange the combination by dragging the tomahawks in their viewing range within the boundary. As per our knowledge, no previous attempts have used two parallel tomahawks within the design parameters. We started the analysis of the data design to obtain the desired outcome consisting of SD and RD that was considered in the data center. The data styles diminished information conjunction as well as populating. The 5Ws strength equivalent directions have considerably decreased information filling for Big Data analysis and creation.

2.1. 5WS parallel coordinate model

Main suggestions of this model represent as follows

2.1.1. Dimensional model

As the name (5Ws Dimensions) suggest that when data come about, where data occur, what data holds, why data come about and who collect the data. Therefore, 5Ws dimensions illustrated with following axes to define data in various conditions.

i) $T=\{t_1,t_2,t_3,.....,t_i\}$ represents when data occurred

ii) $P=\{p_1,p_2,p_3,.....,p_i\}$ represents where data from

iii) $X=\{x_1,x_2,x_3,.....,x_i\}$ represents what data contain

iv) $Y=\{y_1,y_2,y_3,.....,y_i\}$ represents how data transfer from one to other

v) $Z=\{z_1,z_2,z_3,.....,z_i\}$ represents why data occur

vi) $Q=\{q_1,q_2,q_3,.....,q_i\}$ Represents who received.

Model of the density with attributes access as shown in fig 2.

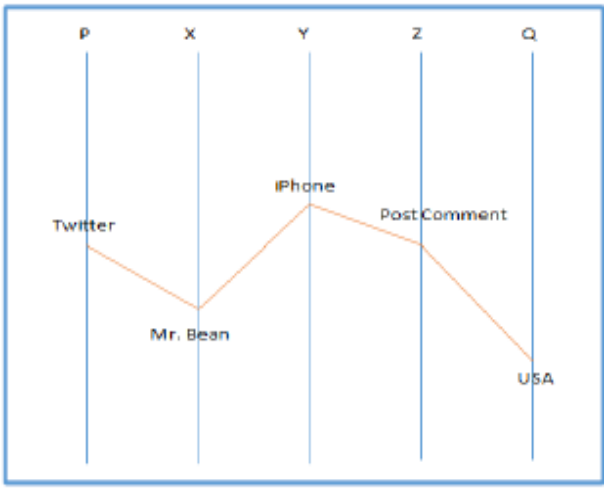

Fig. 2: Parallel Coordinate 5Ws Density Model Presentation.

This is example presentation of density levels with data patterns p=α, x=β, y=χ, z=δ and q=ε. and the mapping functions to represent

these parameters with following function such as $f(t,p,x,y,z,q)$. In the operational parameter, the t | T { } is adequate closure for respective information occurrence and p |P{ } signifies the information came forth which is "Twitter" or "Face book" or "Sender". The x | X{ } denotes about what the data is containing which comes forth from the source that is "like" or "dislike" or "attack". In the following section, y | Y{ } characterizes how the information was relaying from one point to another point such as "by phone" or "by internet" or "by email". z | Z{ } denotes specified reason of occurrence that is why the data occurred that is activities like "Photo sharing" or "finding friends" or "distribution of a virus" and the q | Q{ } denotes who got the specific information, such as "friend" or "bank account" or "receiver".

2.1.2. Sending & receiving

The Sender Density (SD) is the accurate measure and calculation of the data patterns with particular elements such as p=α, x=β, y=χ, z=δ in time duration t therefore the SD and (RD) are respectively as follows:

$$SD_{(\alpha,\beta,\gamma,\delta)} = \frac{F(\alpha,\beta,\gamma,\delta)}{|F|} \times 100\%$$

This equation presents 5Ws sender data pattern with different data attributes, then receiving density is as follows:

$$RD_{(\delta,\alpha,\beta,\gamma)} = \frac{F(\delta,\alpha,\beta,\gamma)}{|F|} \times 100\%$$

This equation presents 5Ws receiver data pattern with different data attributes.

2.1.3. Parallel data visualization axes

Arranging two extra tomahawks by suggesting two respective densities such as SD ( ) and RD ( ) and these two each have a stimulus for every information design and for similar facilitate representation so as to enhance exactness in parallel organize representation. The estimations of SD ( ) and RD ( ) in both tomahawks speak to the information stream designs appeared as poly-lines among the 5Ws measurements. This decreases information jumbling in the chart since one subset has just a single poly-line. 5Ws thickness parallel tomahawks, consolidated with the in sequential order tomahawks and numerical tomahawks, have given more scientific tools for the Big Data illustration. Therefore, there is no information designs is lost within the search and observation button.

2.1.4. Re-order with clustering

The 5Ws density corresponding to the directions as it is re - requesting and assembled to give visual structures and examples to the framework within the convenient connection between the tomahawks in a realistic format. Therefore, it obviously exhibits Big Data designs for various datasets, diverse themes and distinctive information sorts in perception.

3. System design & implementation

In this section, we discuss about our proposed approach similarity measure procedure with different attributes and relations and indexes in practical examples. Consider the procedure discussed in section 3, to represent data with multi-view cluster based on similarity measure. To design this implementation then following modules are required to define efficient attribute relations.

3.1. Related work

Based on term and document frequency in uploaded data sets, we calculate Euclidian distance between words and similarity between documents with attribute relations. Description of different parameters used in our approach is mentioned below in Table 1:
Table1: Different parameter elaboration

Parameter	Description
n,m,c,k,d	Count of documents, terms or classes, or clusters and document factor ‖d‖=1
$S = \{d1, \ldots, dn\}$,Sr	The cluster r contains set of documents ranging from 1 to n
$D=\sum_{d_i \in S} d_i$	Complex vector structure of documents in the set
$D_r=\sum_{d_i \in S_r} d_i$	Compound documents within the set for cluster namely r
$C = D / n$	Centroid vector document
$C_r = D_r / n_r$	Centroid vector documents specified for cluster r

This table summarizes basic used notations used in this paper to calculate different data representations. Euclidian distance evaluation for different documents as follows:

$$\text{Dist}(di, dj) = \|di - dj\|$$

Distance with cluster formation in different attributes in relationships as follows:

$$\min \sum_{r=1}^{k} \sum_{d_i \in S_r} \| d_i - C_r \|^2$$

Based on vector presentation from overall data sets with similar data objects as follows:

$$\text{Sim}(di, dj) = \cos(di, dj) = d_i^t \; d_j$$

The Cosine similarity of different attributes shown in above equation presentation for k-means with Euclidian distance, similarity magnitudes are main difference between Euclidian distance and k-means distance from overall data sets. Some of the researchers define more sequential clustering data presentation to access different attributes in cosine similarity attribute presentation.

3.2. Similarity measure

Cosine similarity for different attributes considers sim equation in above section without changing their meaning in different attributes.

$$\text{Sim}(di, dj) = \cos(di-0, dj-0) = (di-0)^t \; (dj-0)$$

The contextual research process requires zero as the single design framework as 0 signifies the origin point at various data point. The resemblance of the documents di and dj is proven in context with the approach among two significant factors searching for the source or origin. To develop a general idea about the similarity between the two documents, more than one referral factors could be used in the evaluation process. The farthest and nearest outcome of the evaluation enables the researcher to identify precise results which could be possible if different viewpoints would be taken into the consideration. The researcher pre-assumed the group subscription before the beginning of the evaluation process. The two factors need to be within the same group before considering for the evaluation process whereas the statistical outcome of the evaluation must be outside of the group. Researcher considers the outcome as the multi-viewpoint Similarity. Therefore, Similarity size for different documents presentation with attributes as follows:

$$\text{MVS}(d_i, d_j \mid d_i, d_j \in S_r = \frac{1}{n - n_r} \sum_{d_i \in S \setminus S_r} (d_i - d_h)^t (d_j - d_h)$$

The resemblance of two factors that is di and dj which resides in cluster Sr would be measured considering the factor dh placed outside this group. Therefore the factor is equivalent to cosine of the location between di and dj which is viewed from dh. Therefore the Euclidean arrays are ranging from dh to these two points.

3.3. Implementation

The following section presents the implementation proceedings of the proposed approach to define efficient data presentation in different dimensions with effective similarity measures between data objects. Multi view point similarity measure for structure documents as follows:

$$\mathrm{MVS}(d_i, d_j \mid d_i, d_j \in S_r) = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} (d_i^t d_j - d_i^t d_h - d_j^t d_h + d_h^t d_h)$$

$$= d_i^t d_j - \frac{1}{n - n_r} d_i^t \sum_{d_h} d_h - \frac{1}{n - n_r} d_i^t \sum_{d_h} d_h + 1, \| d_h \| = 1$$

Compare two similar documents with attributes relations for all documents such as MVS (di, dj) and MVS (di, dl) where the papers dj is more alike to papers di compared to the other papers dl. Implementation procedure of the MVS with similar attributes as show in following figure 3.

```
1:  procedure BUILDMVSMATRIX(A)
2:    for r ← 1 : c do
3:      D_{S\S_r} ← Σ_{d_i ∉ S_r} d_i
4:      n_{S\S_r} ← |S \ S_r|
5:    end for
6:    for i ← 1 : n do
7:      r ← class of d_i
8:      for j ← 1 : n do
9:        if d_j ∈ S_r then
10:         a_ij ← d_i^t d_j − d_i^t D_{S\S_r}/n_{S\S_r} − d_j^t D_{S\S_r}/n_{S\S_r} + 1
11:       else
12:         a_ij ← d_i^t d_j − d_i^t (D_{S\S_r} − d_j)/(n_{S\S_r} − 1) − d_j^t (D_{S\S_r} − d_j)/(n_{S\S_r} − 1) + 1
13:       end if
14:     end for
15:   end for
16:   return A = {a_ij}_{n×n}
17: end procedure
```

Fig .3: Procedure MVS (Multi View Similarity) in Similarity Matrix.

Fig. 3. First of all, the external combination in context with each category would be determined. Whereas, for every individual row ai of A and i = 1, n, couple of records di and dj where j = 1, n are resides within the same category, aij is measured as in range 10, Fig. 3. Otherwise, dj is believed to be in di's category, and aij is counted in range of 12. This is the similarity matrix procedure to define different attributes in data sets.

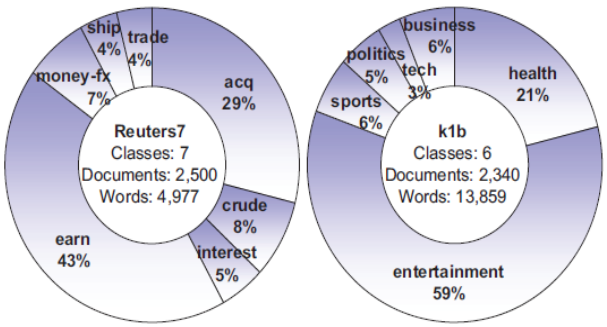

Fig .4: Multi-View Data Visualization for Different Real Time Data Sets With Different Characteristics.

3.4. Cluster label data presentation

The validity test utilizes two distinctive and genuine dataset in the research process. First of all, the Rauter-21578 has been taken into consideration and a subset of the same is used in this case. The test assembly namely Reuters-21578 is one of the most widely utilized subset in research arena.. In the validity testing procedure, researchers picked 2,500 files among the biggest 7 classifications such as "acq", "rough", "intrigue", "acquire", "cash fx", "ship" and "exchange" which enables us to shape reuters7. A fragment of the files could be visible in more than one classification. The secondary choice of dataset is k1b which is an buildup of 2,340 website pages belongs to the Yahoo! subject development considering 6 points such as "wellbeing", amusement", "brandish", "legislative issues", "tech" and lastly "business". The points are made from a past evaluation in data recovery namely Web Ace [6] which are presently available within the CLUTO toolbox [9].The stop-word expulsion and stemming are used to pre-process the dataset prior to the evaluation . Additionally, we reportedly eliminates the words that comes appears in two reports or archives over 99.5% of the total number of files. Lastly TF-IDF is used to weight the reports and standardizing the unit vectors hence complete characteristics set of reuters7 and k1b are stated in Fig. 4. Therefore, the validity test significantly exhibited the possible benefits of the new multi-viewpoint centered likeness evaluate in comparison to the cosine evaluate.

4. Computational evaluation

In this section, we discuss performance evaluation procedure regarding data visualization for both parallel coordinate density model and our propose approach Similarity Measure Centered with Multi View Point for different data objectives. For that we are taking different software parameters like JDK 1.8 and Net Beans 8.0 for user interface construction to upload data sets and process data sets using different parameters in reliable data stream evaluation with respect to data presentation in different formats.

4.1. Data sets collection

There are 20 conventional papers dataset have been used in the test as the part of the information set. Apart from reuters7 and k1b, we eventually entangled more 18 written text collections for detailed elaboration and comprehensiveness of the cluster technique. These datasets are accessible within the CLUTO by the toolkit's writers [19] just like k1b. These datasets are utilized for trial examining over the documents and the resource also had been described in information set. The features of the dataset have been summarized in the Table 2. The data set is visible as variety of dimension, variable sessions and category stability. Conventional techniques have been used for the pre-process like stop-word removal, arising, elimination of unusual and regular terms with normalization.

Table 2: Sample Document Datasets in Different Formats Collected from Various Data Available Links

Data	Source	c	n	m	Balance
fbis	TREC	17	2,463	2,000	0.075
hitech	TREC	6	2,301	13,170	0.192
k1a	WebACE	20	2,340	13,859	0.018
k1b	WebACE	6	2,340	13,859	0.043
la1	TREC	6	3,204	17,273	0.290
la2	TREC	6	3,075	15,211	0.274
re0	Reuters	13	1,504	2,886	0.018
re1	Reuters	25	1,657	3,758	0.027
tr31	TREC	7	927	10,127	0.006
reviews	TREC	5	4,069	23,220	0.099
wap	WebACE	20	1,560	8,440	0.015
classic	CACM/CISI/ CRAN/MED	4	7,089	12,009	0.323
la12	TREC	6	6,279	21,604	0.282
new3	TREC	44	9,558	36,306	0.149
sports	TREC	7	8,580	18,324	0.036
tr11	TREC	9	414	6,424	0.045
tr12	TREC	8	313	5,799	0.097
tr23	TREC	6	204	5,831	0.066
tr45	TREC	10	690	8,260	0.088
reuters7	Reuters	7	2,500	4,977	0.082

4.2. Experimental results

To illustrate how well MVSCs is capable of doing, we associate the data set with five significant clustering techniques on the 20 datasets in Desk 2. To sum up, seven clustering techniques are mentioned below:

- MVSC-IR: MVSC using requirements operate IR
- 5Ws Density Model : MVSC using requirements operate IV
- K-means: conventional k-means with Euclidean distance
- Skeins: rounded k-means with CS
- graphics: CLUTO's chart technique with CS
- graph EJ: CLUTO's chart with prolonged Jaccard
- MMC: Spectral Min-Max Cut criteria [13]

Our MVSC-IR and MVSC-IV applications are implemented in Coffee. Therefore the controlling aspect that is α in IR is always within the set at 0.3 during the tests. Nothing unless there are other options calculations are ensured to discover worldwide ideal, and every one of them are introduction subordinate. Henceforth, for evSSery strategy, we accomplished grouping twice with haphazardly instated values which picked the best trial as far as the relating target work esteem. In every one of the analyses, each trial comprised of 10 trials. In addition, the outcome detailed here on each dataset by a specific bunching technique is the normal of 10 trials. Fig 6 shows the accuracy of our proposed approach with different data sets evaluation procedure on text oriented documents with feasible parameters with values shown in Table 3.

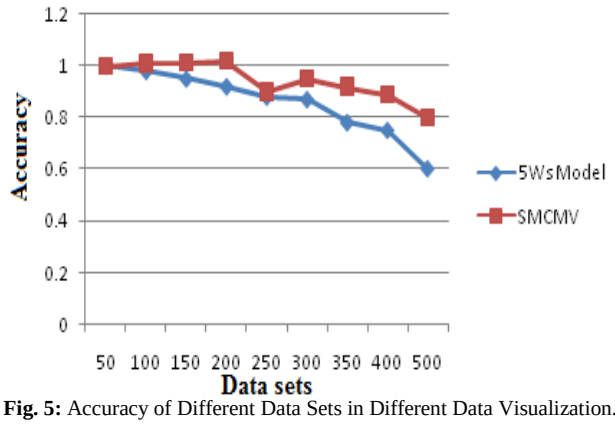

Fig. 5: Accuracy of Different Data Sets in Different Data Visualization.

Table 3: Accuracy Values

Documents	5Ws Model	SMCMV
50	1	1
100	0.98	1.01
150	0.95	1.015
200	0.92	1.02
250	0.88	0.9
300	0.87	0.95
350	0.78	0.92
400	0.75	0.89
500	0.6	0.8

Time efficiency results are plotted with following values show in Table 4. The presented of performance evaluation of our proposed approach with traditional approach shown in fig 6 with respect to time efficiency in real time data set processing.

Table 4: Time Efficiency Values

Documents	SMCMV	5Ws Model
15	0.015	0.04
30	0.014	0.03
45	0.012	0.035
60	0.011	0.02
75	0.009	0.025
90	0.008	0.015

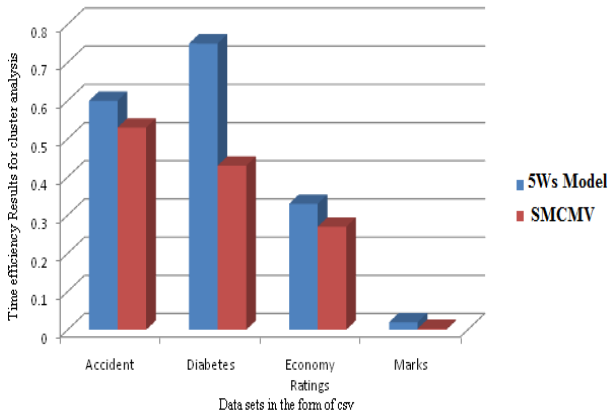

Fig. 6: Time Efficiency Values of both Proposed and Traditional Approaches with Different Data Sets.

Finally, we describe and conclude SMCMV approach gives better and efficiency results than 5Ws density model for different types of documents related to different types of documents.

5. Conclusion

In this paper, we present to discuss about data visualization with different data sets, and also discuss about parallel coordinates data visualization in data representation based on topic, type of data and data sets. For similarity measure of different data objects in data sets, for that we propose to develop novel method i.e. Similarity Measure Centered with Multi Viewpoint (SMCMV) with cosine similarity for different text, image, video documents. We also compare data visualization difference between parallel coordinate density model presentation and our proposed approach in both theoretical and practical for large data documents. The main key point of our proposed approach to define data sets in multi view data representation. Further enhancement of our proposed approach is to define data documents in parallel processing using advanced machine learning approaches with real time data sets.

References

[1] Jinson Zhang, Wen Bo Wang," Big Data Density Analytics using Parallel Coordinate Visualization", 2014 IEEE 17th International Conference on Computational Science and Engineering.

[2] Pingdom, "Internet 2012 in numbers", posted on Jan 16, 2013, http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/.

[3] J. Sanyal, S. zhang, J. Dyer, A. Mercer, P. Amburn, and R.J. Moorhead, "Noodles: A Tool for Visualization on Numerical Weather Model Ensemble Uncertainty", IEEE Transactions on Visualization and Computer Graphics, vol. 16, no 6, pp 1421-1430, Nov/Dec 2010. https://doi.org/10.1109/TVCG.2010.181.

[4] S. Hadiak, H.J Schulz, and H. Schumann, "In Situ Exploration of Large Dynamic Networks", IEEE Transactions on Visualization and Computer Graphics, vol. 17, no 12, pp 2334-2343, Dec 2011. https://doi.org/10.1109/TVCG.2011.213.

[5] Y.S. Wang, C. Wang, T.Y. Lee, and K.L. Ma, "Feature-Preserving Volume Data Reduction and Focus+Context Visualization", IEEE Transactions on Visualization and Computer Graphics, vol. 17, no 2, pp 171-181, Feb 2011 https://doi.org/10.1109/TVCG.2010.34.

[6] S. Afzal, R. Maciejewski, Y. Jang, N. Elmqvist, and D.S. Ebert, "Spatial Text Visualization Using Automatic Typographic Maps", IEEE Transactions on Visualization and Computer Graphics, vol. 18, no 12, pp 2556-2564, Dec 2012. https://doi.org/10.1109/TVCG.2012.264.

[7] A.H. Meghdadi, and P. Irani, "Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization", IEEE Transactions on Visualization and Computer Graphics, vol. 19, no 12, pp 2119-2128, Dec 2013 https://doi.org/10.1109/TVCG.2013.168.

[8] E. Lamboray, S. Wurmlin, and M. Gross, "Data Streaming in Telepresence Environments", IEEE Transactions on Visualization and Computer Graphics, vol. 11, no 6, pp 637-648, Nov/Dec 2005 https://doi.org/10.1109/TVCG.2005.98.

[9] L. Shi, Q. Liao, X. Sun, Y. Chen and C. Lin, "Scalable Network Traffic Visualization Using Compressed Graphs", In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 606-612, Oct 2013

[10] W. Cui, Y. Wu, S. Liu, F. Wei, M.X. Zhou, and H. QU, "Context-Preserving, Dynamic Word Cloud Visualization", IEEE Computer Graphics and Applications, vol. 30, no 6, pp. 42-53, Nov/Dec 2010 https://doi.org/10.1109/MCG.2010.102.

[11] J. Zhang and M.L Huang, "5Ws Model for Big Data Analysis and Visualization", In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 1021-1028, Dec 2013 https://doi.org/10.1109/CSE.2013.149.

[12] A. Shiravi, H. Shiravi, M. Tavallaee, and A.A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," Computers & Security, vol. 31, no. 3, pp 357-374, May 2012 https://doi.org/10.1016/j.cose.2011.12.012.

[13] W.S. Seol, H.W. Jeong, B. Lee and H.Y. Youn, "Reduction of Association Rules for Big Data Sets in Socially-Aware Computing", In Proc. 2013 16th IEEE International Conference on Computational Science and Engineering (CSE), pp. 949-956, Dec 2013 https://doi.org/10.1109/CSE.2013.140.

[14] Z. Wang, W. Xiao, B. Ge, and H. Xu, "ADraw: A novel social network visualization tool with attribute-based layout and coloring", In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 25-32, Oct 2013

[15] J. Zhang and M.L. Huang, "Density approach: a new model for BigData analysis and visualization", Concurrency and Computation: Practice and Experience. Publish online July 2014, https://doi.org/10.1002/cpe.3337.

[16] Z. Wang, J. Zhou, W. Chen, C. Chen, J. Liao and R. Maciejewski, "A Novel Visual analytics Approach for Clustering Large-Scale Social Data", In Proc. 2013 IEEE International Conference on Big Data (IEEE BigData 2013), pp. 79-86, Oct 2013.

[17] Duc Thang Nguyen, Lihui Chen," Clustering with Multi-Viewpoint based Similarity Measure", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. XX, NO. YY, 2011.

[18] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," Mach. Learn., vol. 55, no. 3, pp. 311–331, Jun 2004. https://doi.org/10.1023/B:MACH.0000027785.44527.d6.

[19] G. Karypis, "CLUTO a clustering toolkit," Dept. of Computer Science, Uni. of Minnesota, Tech. Rep., 2003, http://glaros.dtc.umn.edu/gkhome/views/cluto.

[20] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in Proc. of the 17th National Conf. on Artif. Intell. Workshop of Artif. Intell. For Web Search. AAAI, Jul. 2000, pp. 58–64.

[21] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," Pattern Recognit. Lett. vol. 28, no. 1, pp. 110 – 118, 2007. https://doi.org/10.1016/j.patrec.2006.06.006.

[22] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in Proc. of the 8th Int. Symp. IDA, 2009, pp. 83–94. https://doi.org/10.1007/978-3-642-03915-7_8.

[23] P. Lakkaraju, S. Gauch, and M. Speretta, "Document similarity based on concept tree distance," in Proc. of the 19th ACM conf. on Hypertext and hypermedia, 2008, pp. 127–132. https://doi.org/10.1145/1379092.1379118.

[24] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," IEEE Trans. on Knowl. In addition, Data Eng., vol. 20, no. 9, pp. 1217–1229, 2008.

[25] Madala S.R., Rajavarman V.N., Venkata Satya Vivek T. (2018) Analysis of Different Pattern Evaluation Procedures for Big Data Visualization in Data Analysis. In: Satapathy S., Bhateja V., Raju K., Janakiramaiah B. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 542. Springer, Singapore. https://doi.org/10.1007/978-981-10-3223-3_44.