ISSN (Print) : 2319-8613
ISSN (Online) : 0975-4024

Multi-View Data Visualization Based Big Data Analysis Using Clustering Similarity Measure

Srinivasa Rao Madala[#1], V N Rajavarman[#2], T.Venkata Satya Vivek [#3]
[1,3]Research scholors, 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.
[1]mr.srinu13@gmail.com, [2]nrajavarman2003@gmail.com, [3]tvsvivek1990@gmail.com

***Abstract* ---In big data, data visualization is an impressive concept to represent data for efficient data analysis to handle high dimensional data. In data visualization, there are three main properties i) to represent without loss of data patterns ii) without any changes in data pattern change the attributes iii) data visualization with structure and unstructured data attributes for data analysis. There are many types of data visualization are presented practically to define data analysis (i.e. topic based data visualization, attribute based data visualization, audio based data visualization and text based data visualization in different data sets). Parallel co-ordinate is an efficient 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. Similarity measure is a categorization property to represent data with relationship objects in data set evaluation with different pair of attributes. We need to improve parallel coordinate tool to support multi-attribute object relations, so we propose 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 achieve assessment based similarity index with data visualization. Using multi viewpoint, we present theoretical analysis based on multi attributes presentation. Our experimental results gives best data representation in data visualization with efficient similarity measure on real time document evaluation with different known collected clustering approaches.**

Key words: Data Visualization, Parallel Co-ordinate, Multivariate attributes, Similarity Measure and Multi – viewpoint and Clustering Methods.

I. INTRODUCTION

Structured and unstructured data in big data visualization contains different forms of data like image, audio and video and collected this data from different multiple data sets based on the size time and space complexity evaluation. For example Face book generates 25 GB of data which contains following user's personal details and their sharing data with mutual and personal friends. Thousands, hundreds of different dimensional attributes by monthly providing different data to analyze multiple attribute dimensions to handle data visualization. Because of increasing rapid usage of big data in various applications, different authors proposed different association and classification and clustering to analyze high dimensional data. Parallel coordinate data visualization is one of the promising approaches to represent data without change their data patterns from overall data. Sample data visualization with different dimensions as shown fig 1.

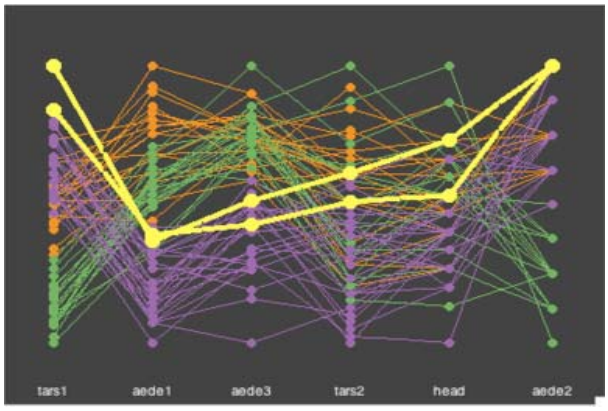

Fig .1. Parallel coordinate plot of the data visualization for flea data.

Neighbor attribute partitioning is as shown in fig 1 with different node axes values in dimensional set. Some of the researchers and introduced to define effective data visualization with different node data presentation with same attributes. Mainly data visualization consists three data representations in real time data presentations, topic based data visualization, which consist about particular topic with algorithm process like network traffic visualization 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 particular data sets like social and network, oriented data sets with different data patterns. To represent data in these three ways, traditionally develop Parallel coordinate $5W_s$ density model, in that analyze data attributes and represent those attributes in parallel axes presentation for multiple data set with data types and topics evaluation in real time data set presentation. To measure similarity between attributes in data visualization with relationships then parallel coordinate 5W density presentation not satisfied in data evaluation. So in this paper, we propose and develop novel Similarity Measure Centered with Multi Viewpoint (SMCMV)) approach and related clustering approaches to represent data. This approach follows multi-view data representation with dimensions in attribute relationship. In that clustering is an aggressive concept and topic in data retrieval based on attributes, in that we intrinsic the structure data formation and formulate them into required and meaningful data presentation. So our proposed approach follows clustering properties to read and present data in different dimensions with attribute relations. And also our approach exclusively follows multi-view data presentation with respect attribute presentation. We also calculate similarity measure in attribute partitioning in data set exploration; similarity measure plays important aspect in success and failure of data representation in clustering procedure. Main objectives of our proposed approach as follows:

1. We present and propose an approach to find the similarity between data objects with different relations in high dimensional data evaluation.
2. Proposed similarity measure with different clustering calculations with provable quality and performance consistent.
3. Display multi-view data visualization with different data patterns.
4. Give efficient data visualization with multiple attributes with clustering calculations.

Remaining of this section organized as follows: Section 2 relates 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.

II. RELATED WORK

Comparative compose plots, as a standout amongst the most prominent strategies, was first suggested by Inselberg [3] and Wegman recommended it as something for extraordinary viewpoint data inquire about [4]. Directions of n-dimensional data can be authorized upon in parallel tomahawks in a 2-dimensional airplane and connected by straight sections. As demonstrated out in the scholarly works [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 symbolizes N-dimensional data in a 2-dimensional zone with factual meticulousness. Visual grouping, pivot reordering 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. With the go for averting over-plotting and ensuring thickness data, Kai Lun Chung and Wei Zhuo [2] outlined two visual explanatory apparatuses: decision outlines and respects diagrams, to lessen visual chaos running in parallel blends. The decision outline is a brushing gadget which helps clients underlines the regions picked. 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] organized the data point of view into anomalies, 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] spread components running in parallel blends to union parallel facilitates and scatterplot climbing, which diminished information swarming. Clients can reorder the blends by pulling the tomahawks in their interface for better obvious watcher. No past work, to the best of our insight, has made two

extra tomahawks by utilizing the densities for the parallel sort out creation. We analyzed the data design to begin with keeping in mind the end goal to procure the of SD and RD, and after that pictured the data hubs. These data styles diminished information over-lapping and populating. 5Ws strength parallel 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

2.2.1. *Dimensional Model*: As the name (5Ws Dimensions) suggest that When data occur, Where data from, What data contains, Why data occur and Who receive data. 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.

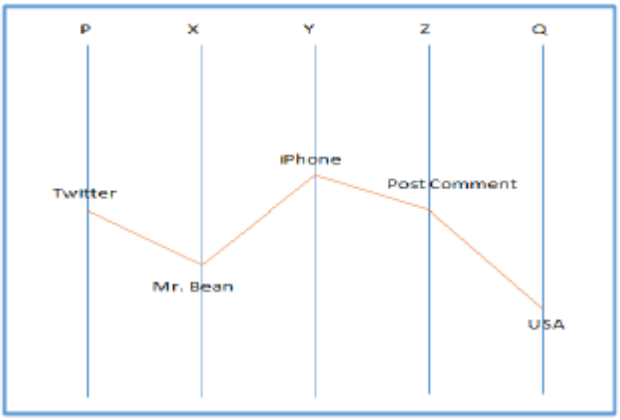

Fig .2. Parallel coordinate 5Ws density model presentation.

This is example presentation of density levels with data patterns p=α, x=β, y=χ, z=δ and q=ε. Then the mapping functions to represent these parameters with following function i.e. $f(t, p, x, y, z, q)$. Where t | T { } is sufficient time seal for each information occurrence. p |P{ } symbolizes where the information came from, such as “Twitter”, “Face book” or “Sender”. x | X{ } symbolizes what the data content was, such as “like”, “dislike” or “attack”. y | Y{ } represents how the information was moved, such as “by the Internet”, “by phone” or “by email”. z | Z{ } symbolizes why the data occurred, such as “sharing photos”, “finding friends” or “spreading a virus”. q | Q{ } symbolizes who obtained the information, such as “friend”, “bank account” or “receiver”.

2.2. *Sending & Receiving:* Sender density (SD) used to calculate and measure of sender with data patterns for accurate or particular attribute p=α, x=β, y=χ, z=δ in time t, then sending density and receiving density (RD) as follows concurrently:

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

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.

2.3. *Parallel Data Visualization Axes:* make two extra tomahawks by proposing two densities SD ( ) and RD ( ), which each have an incentive for every information design, for parallel 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 techniques for Big Data representation. No information designs have been lost amid the investigation and perception handle.

2.4. *Re-order with Clustering:* The 5Ws density parallel directions with re-requesting and grouped perspectives give visual structures and examples to outline the cozy connection between the tomahawks in a realistic format. It obviously exhibits Big Data designs for various datasets, diverse themes and distinctive information sorts in perception.

III. 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 shown in table 1.

TABLE 1. Different parameter description.

Parameter	Description
n,m,c,k,d	Number of documents, terms, classes, clusters, and document factor ||d||=1
S = {d1, . . . , dn} ,Sr	Set of documents in cluster r
$D=\sum_{d_i \in S} d_i$	Composite vector of documents
$D_r=\sum_{d_i \in S_r} d_i$	Composite documents for cluster r
$C = D / n$	Centroid vector documents
$C_r = D_r / n_r$	Centroid vector documents 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:

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

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:

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

Cosine similarity for 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.

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

Where o and 0 represents vector with origin point in different data point evaluation, the evaluate requires 0 as one and only blueprint. The likeness between two documents di and dj is established w.r.t. the angle between the two factors when looking from the original source. To build a new idea of likeness, it is possible to use more than just one referrals factor. We may have a more precise evaluation of how near or

remote a couple of factors are, if we look at them from many different viewpoints. A assumption of group subscriptions has been created before to the evaluate. The two things to be calculated must be in the same group, while the points from where to determine this statistic must be outside of the group. We refer to this as offer the Multi-Viewpoint based Similarity. Similarity measure for different documents presentation with attributes as follows:

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

The likeness between two factors di and dj inside cluster Sr, considered from a factor dh outside this group, is equal to the item of the cosine of the position between di and dj looking from dh and the Euclidean ranges from dh to these two points.

3.3. *Implementation:* In this section, we present the implementation procedure of our 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, MVS (d_i, d_j) and MVS (d_i, d_l), papers dj is more similar to papers di than the other papers dl is, if and only if. Implementation procedure of the MVS with similar attributes as show in following figure 3.

1: **procedure** BUILDMVSMATRIX(A)
2: **for** $r \leftarrow 1 : c$ **do**
3: $D_{S \setminus S_r} \leftarrow \sum_{d_i \notin S_r} d_i$
4: $n_{S \setminus S_r} \leftarrow |S \setminus S_r|$
5: **end for**
6: **for** $i \leftarrow 1 : n$ **do**
7: $r \leftarrow$ class of d_i
8: **for** $j \leftarrow 1 : n$ **do**
9: **if** $d_j \in S_r$ **then**
10: $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S \setminus S_r}}{n_{S \setminus S_r}} - d_j^t \frac{D_{S \setminus S_r}}{n_{S \setminus S_r}} + 1$
11: **else**
12: $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S \setminus S_r} - d_j}{n_{S \setminus S_r} - 1} - d_j^t \frac{D_{S \setminus S_r} - d_j}{n_{S \setminus S_r} - 1} + 1$
13: **end if**
14: **end for**
15: **end for**
16: **return** $A = \{a_{ij}\}_{n \times n}$
17: **end procedure**

Fig .3. Procedure MVS (multi view Similarity) in similarity matrix.

Fig. 3. First of all, the external blend w.r.t. each category is determined. Then, for each row ai of A, i = 1, . . , n, if the happy couple of records di and dj, j = 1, . . . , n are in 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 measured as in range 12. This is the similarity matrix procedure to define different attributes in data sets.

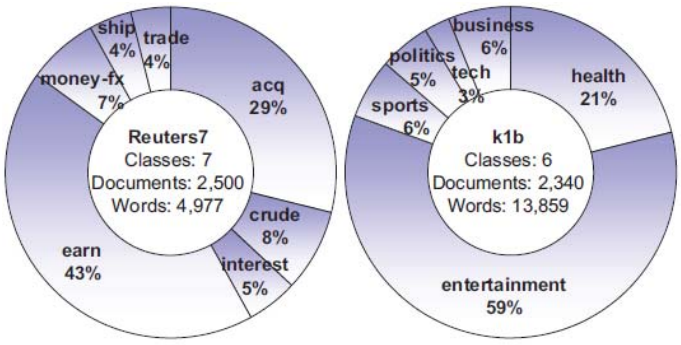

Fig .4. Multi-view data visualization for different real time data sets with different characteristics.

3.4. *Cluster Label Data Presentation:* Two genuine report datasets are utilized as cases in this legitimacy test. The first is reuters7, a subset of the renowned gathering, Reuters-21578 Distribution 1.0, of Reuter's newswire articles1. Reuters-21578 is one of the most broadly utilized test gathering for content order. In our legitimacy test, we chose 2,500 archives from the biggest 7 classifications: "acq", "rough", "intrigue", "acquire", "cash fx", "ship" and "exchange" to shape reuters7. A portion of the archives may show up in more than one classification. The second dataset is k1b, an accumulation of 2,340 website pages from the Yahoo! subject progression, including 6 points: "wellbeing", amusement", "brandish", "legislative issues","tech" and "business". It was made from a past review in data recovery called WebAce [6], and is presently accessible with the CLUTO toolbox [9]. The two datasets were preprocessed by stop-word expulsion and stemming. In addition, we evacuated words that show up in under two reports or over 99.5% of the aggregate number of archives. At last, the reports were weighted by TF-IDF and standardized to unit vectors. The full attributes of reuters7 and k1b are displayed in Fig. 4. The validity test has shown the prospective benefits of the new multi-viewpoint centered likeness evaluate in comparison to the cosine evaluate.

IV. 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:* The information corpora that we used for tests include of 20 conventional papers datasets. Besides reuters7 and k1b, which have been described in information previously, we involved another 18 written text selections so that the study of the clustering techniques is more thorough and comprehensive. Just like k1b, these datasets are offered together with CLUTO by the toolkit's writers [19]. They had been used for trial examining over the documents, and their resource and resource had also been described in information. Table 2 summarizes their features. The corpora existing a variety of dimension, variety of sessions and category stability. They were all preprocessed by conventional techniques, such as stop-word removal, arising, elimination of too unusual as well as too 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 compare them with five other clustering techniques on the 20 datasets in Desk 2. To sum up, the seven clustering techniques are:

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

Our MVSC-IR and MVSC-IV applications are implemented in Coffee. The controlling aspect α in IR is always 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 every strategy, we performed grouping a couple times with haphazardly instated values, and 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. Figure 6 show 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.

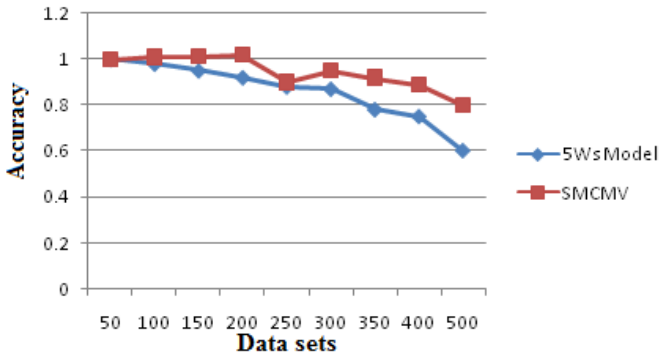

Fig .5. Accuracy of different data sets in different data visualization.

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 figure 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

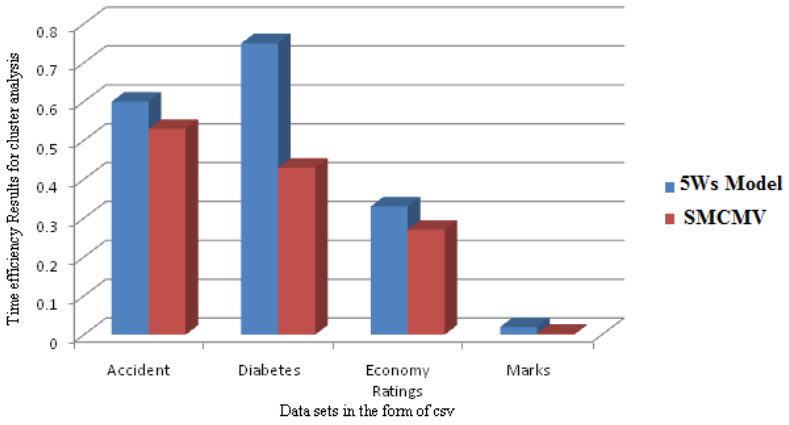

Fig .6. Time efficiency values of both proposed and traditional approaches with different data sets.

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.

V. 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 improvement 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.
[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.
[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
[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.
[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
[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
[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
[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
[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
[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
[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, DOI: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.

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