ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Implementing Clustering using CSI by K-means

U.Vignesh, P.Valarmathi, S.Arun

Abstract*—This paper presents a new clustering method called Correlation Similarity Indexing (CSI). It deals with the aspect of instances, which are not been pre-classified in early and do not have a defined attribute to be associated with them. CSI use to reveal the unnoticed class of regarded items. It forms patterns into different group of finding similarity between them with the similarity measure space calculated on the given documents. CSI concentrates in dividing the identified datasets into a number of several clusters that should not have an overlapping aspect with each other. CSI performs proximity analysis to identify the dissimilarity measure between the neighborhoods assigned to the document. It deals with the similarity measure space result between the objects. A cluster level process is proposed to perform the object extraction with similarities. If the given object is in the form of an image then they can also to be clustered by finding the similarity between them by using CSI. CSI involves the K-means clustering algorithm perception build over to it and their action are also to be performed in the aspect of clustering done on documents and images or even in the regional aspect. CSI finds the number of clusters and then groups the similar item sets into respective clusters using K-means have been done on in this system. CSI can be used for large class of applications such as GIS system, Image database exploration, Medical imaging etc.

Index Terms— **Similarity Measure Space, Clustering, Indexing, Proximity.**

I. INTRODUCTION

With the rapid advancement in clustering concepts, new proposal regarding the algorithm have achieved tremendous availability for efficient clustering involved in the concept of mixing. Here, we propose a new algorithm based on clustering concepts known as Correlation Similarity Indexing (CSI). With the indexing at first, the varieties of similarities involved in the documents are to be identified then with the use of K-means algorithm, clustering the similarities have been grouped and the results have been noted for further references of CSI.

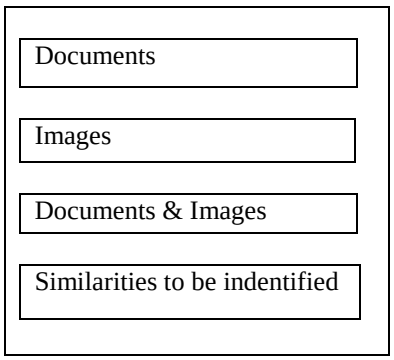


Fig.1 Sketch of the system components

Fig.1 shows the correlation similarity indexing process, at first from the given input, it notes whether it belongs to the category of document, image or both document and images, then it assigns a weight calculation for the similar object and completes its work of identification. The weight calculations are referred to as similarity measure space. It measures the distance by which the neighbor document how close to them. In case of images it first calculates the similarity of region then goes to the similarity of image. In case of both document and image, the distance and regions are to be first noted then the comparative similarities have to be calculated. Thus, CSI paves the way for time consumption compared to latent preserving indexing in previous cases.

CSI can be applied to any of the large and small class of applications, but it can be highly preferable to these applications ass mentioned in table I. CSI can also individually do clustering in case of some clustering without the interruption of an K-means clustering (Eg.: GIS system, Image database exploration). CSI with K-means clustering proves the efficient group in case of patterns mining and medical imaging aspect, where the frequent item set has to be identified and grouped into clusters. Basically, for a correlation similarity indexing, the input can be given for clustering are of four types as shown in fig.2 viz. text, image, audio and video. Text are similar as documents, where the files can be uploaded and maintained whereas image, which deals with the color, size and regional similarities for the purpose of clustering.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Table I Applications to apply on CSI

Application	**Clustering**
DNA pattern mining	CSI + K-means
GIS	CSI
Image Database Recognition	CSI
Medical Imaging	CSI + K-means

.

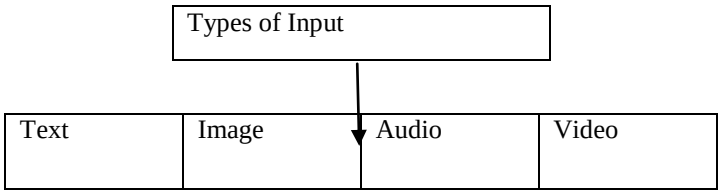


Fig.2 Input Recognition

II. RELATED WORK

S.C.Deerwester et.al proposed the concept of Latent Semantic Indexing (LSI) for the purpose of clustering involved in the document. It has the process of identifying the better subspace in approximate category to adapt given document space by localizing the international reconstruction fault known as the Euclidean distance. With this Euclidean distance their aspect of finding the relevant document from the bunch of document and grouping of those similar documents has been done over here. Whereas, D.R.Hardoon et. al proposed the CCA method. CCA stands for canonical correlation analysis. In this CCA method, they find a projection view for the datasets of pair and their correlations between low dimensional representations in projections are to be maximized. It can be expressed as

$$\frac{\sum_{wx,wy}(Xwx,\ Ywy)}{\|Xwx\| \cdot \|Ywy\|} \quad (1)$$

Taping Zhang et. al proposed a Correlation Preserving Indexing (CPI), where they estimates a process, if two considered documents are nearer to each other in the given document space, then they are to be considered as similar and grouped into clusters. In case of vice versa, they are to be considered as dissimilar and grouped into different clusters as shown in fig. 3, the projections are to be noted for similar and dissimilar distances. D.M.Blie et. al have found better clustering in the Latent Dirichlet Allocation (LDA). LDA designed to note significant aspect of intra document statistical structure via the distributional model.

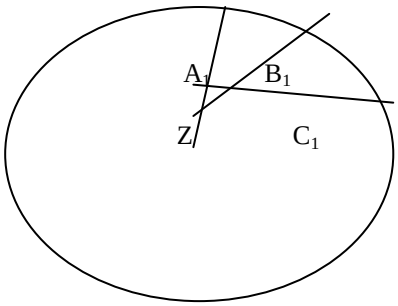


Fig.3 CPI projection

CPI projection have been clearly understood from the fig. 3, the terms Z which includes the combination of probability that they occurs are expressed as

$Z=\alpha+\beta+\gamma,$ (2)

$\alpha+\beta+\gamma=A_1(p)+B_1(p)+C_1(p)$ (3)

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

III. OUR PROPOSAL

A. Architecture

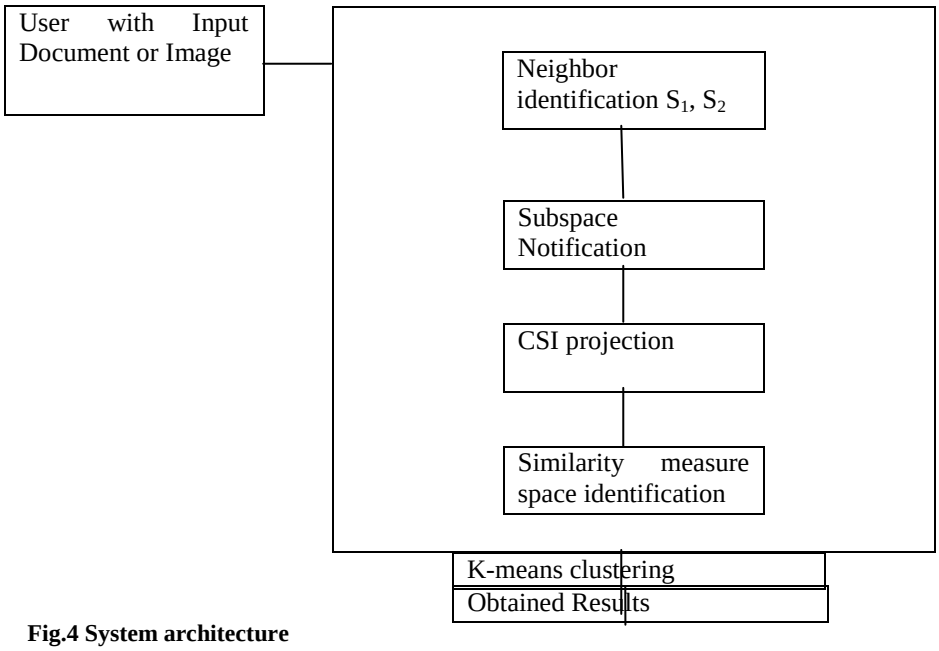


Fig.4 System architecture

Fig. 4 shows the clear overview of the system that it processes. Here the process starts from the user with the given input. The input refers to a document, image or both document and image, which contains various types of variations built on towards it. As soon as the user deals with the input, CSI action has been started or the input directly flows towards CSI. Here, at first the neighbor is to be identified in the matrix manner with the representation of S_1 and S_2. Then with regards to S_1 and S_2, the subspace notification has to be calculated. After these subspace calculation, CSI projection has to be noted or projected towards 2D or 3D representations corresponding to the given input. Then the CSI calculates the similarity measure space by mentioning the nearest document identification and the regional identification has to be done in case of images for finding the similarities between them. After finding those similarities cluster have to be formed by using the allotment of K in the K-means clustering algorithm. The K-means clustering algorithm results the formed clusters with their similarity degrees.

B. CSI by K-means

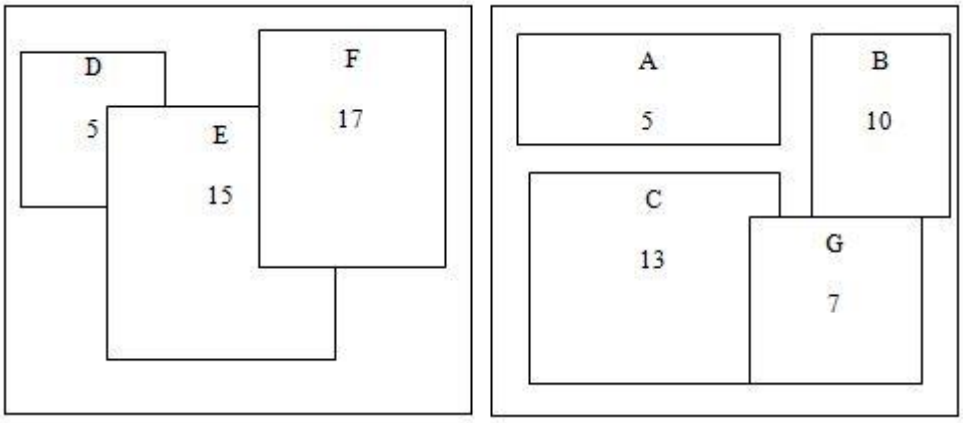
Correlation Similarity Indexing (CSI), which deals with the calculation of correlations distance between the documents that have been given for the aspect of clustering, then with the findings of similarities, K-means plays its role to cluster the similar documents based on the equations shown in (4) and (5).

$$Sim_{size}(S_1, S_2) = 1.0 - \frac{|\, size(S_1) - size(S_2) \,|}{Max(size(S_1), size(S_2))} \quad (4)$$

$$Sim_{image}(I, J) = \frac{\sum_{m=1}^{n} Sim_{color}(I_m, J_m)}{K} \quad \text{if both of } I_m \text{ \& } J_m. \quad (5)$$

CSI first constructs the nearest neighbor identification with the matrix of S_1 and S_2 as in the format of matrix. Then they are to be projected towards the document vectors by finding the subspace allocation rather than comparison with the zero singular values. Then the CSI projection has to be notified with corresponding to input. Then the documents are to be clustered by means of K-means clustering using the similarities identified by CSI.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013



(a) (b)

Fig. 5 (a) Document A (b) Document B

Fig. 5 (a) and (b) shows the documents of two types that they have been in the database. By using CSI, with this identified documents we have to calculate the similarities of degrees that they have already allotted and produced as input. Here with these documents the similarity ratio which obtains the positive result are to be considered as (D, G) with 0.60; (E, B) & (E, C) with 0.89 & 0.80; (F, B) & (F, G) with 0.70 & 0.60 are the obtained result to be clustered by K-means clustering algorithm as shown in table II.

Table II Similarities degrees corresponding to Fig. 5 (a) and (b)

X / Y	A	B	C	G
D	-	-	-	0.60
E	-	0.89	0.80	-
F	-	0.70	-	0.65

IV. CSI AND CONTROL FLOW

Fig. 6 shows the control flow diagram of our system, by which the flow of an aspect that CSI and K-means works can be easily identified by using it. Flow aspect starts with the user to get an input files as an document or the document as bunches of documents or the images, whatever is necessary to be clustered. Then in the concept of CSI, CSI projection has been projected with the parameters identified using the subspace notification. Followed by CSI projection, correlation similarity measure space has to be calculated and to be given as input to the K-means clustering algorithm. K-means clustering concept has been carried out with degrees of similarities, which gives us an output clusters.

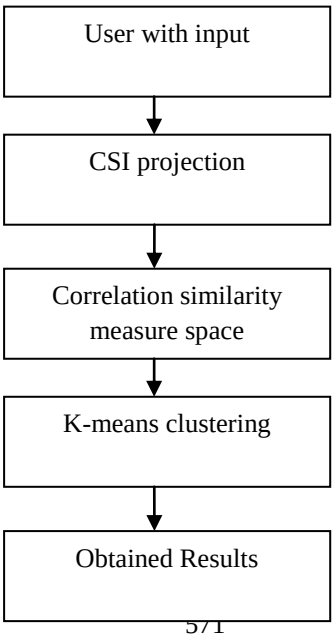


Fig. 6 Control flow

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

V. PERFORMANCE ANALYSIS

In the analysis of performance metrics of CSI has been carried out and proved to be better by comparing it with CLI, LPI, etc. of existing clustering methods. When comparison results have been noted as shown in fig. 7 based on the aspect of speed, accuracy and efficiency CSI has an advantage of better report in the clustering process.

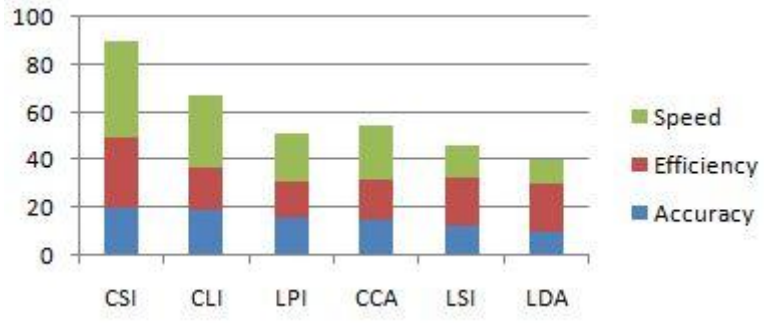


Fig. 7 Comparison graph in metrics

VI. CONCLUSION

This paper has introduced new clustering method called Correlation Similarity Indexing (CSI). By using correlation similarity measure space distance between the given documents or images are to be calculated. Thus, CSI produces a similarity of documents. Then K-means clustering has been applied to form a clustering of similarities achieved from the documents. Possible extensions and improvements of our model includes meta features and CSI to act independently without the support of K-means clustering in the large class of applications.

ACKNOWLEDGMENT

The authors would like to thank Mr. M. Prabakaran, RVS College of Engineering and Technology/Anna University, for his valuable advice, support and guidance during the preparation of this paper.

REFERENCES

[1] S. Kotsiantis and P. Pintelas, “Recent Advances in Clustering: A Brief Survey,” WSEAS Trans. Information Science and Applications, vol. 1, no. 1, pp. 73-81, 2004.

[2] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, “Indexing by Latent Semantic Analysis,” J. Am.Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[3] D. Cai, X. He, and J. Han, “Document Clustering Using Locality Preserving Indexing,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 12, pp. 1624-1637, Dec. 2005.

[4] D.R. Hardoon, S.R. Szedmak, and J.R. Shawe-taylor, “Canonical Correlation Analysis: An Overview with Application to Learning Methods,” J. Neural Computation, vol. 16, no. 12, pp. 2639-2664, 2004.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[6] X. Zhu, “Semi-Supervised Learning Literature Survey,” technical report, Computer Sciences, Univ. of Wisconsin-Madison, 2005.

[7] G. Lebanon, “Metric Learning for Text Documents,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 4, pp. 497-507, Apr. 2006.

[8] F.Y.M. Wan, Introduction to Calculus of Variations and Its Applications. Chapman Hall, 1995.

[9] Encyclopedia of Mathematics. M. Hazewinkel, ed., Springer-Verlag, http://eom.springer.de/L/l057190.htm, 2002.

[10] I.S. Dhillon and D.M. Modha, “Concept Decompositions for Large Sparse Text Data Using Clustering,” Machine Learning, vol. 42, no. 1, pp. 143-175, 2001.

[11] P. Strobach, “Bi-Iteration SVD Subspace Tracking Algorithms,” IEEE Trans. Signal Processing, vol. 45, no. 5, pp. 1222-1240, May 1997.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

[12] D. Zeimpekis and E. Gallopoulos, "Design of a Matlab Toolbox for Term-Document Matrix Generation," Proc. Workshop Clustering High Dimensional Data and Its Applications at the Fifth SIAM Int'l Conf. Data Mining (SDM '05), pp. 38-48, 2005.

[13] L. Lovasz and M. Plummer, Matching Theory. Elsevier, 1986.

[14] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral Relaxation for k-Means," Neural Information Processing Systems, vol. 14 (NIPS 2001), pp. 1057-1064, MIT Press, 2001.

[15] D. Cheng, R. Kannan, S. Vempala, and G. Wang, "A Divide-and- Merge Methodology for Clustering," ACM Trans. Database Systems, vol. 31, no. 4, pp. 1499-1525, 2006.

AUTHOR BIOGRAPHY

U.Vignesh completed his Master of Technology in Information Technology from Veltech Multitech Dr.Rangarajan Dr.Sakunthala Engineering College/Anna University – Chennai in 2012. He has published papers in the IOSR-JCE and ESIRJ. His research interests lie in the field of data mining, cloud computing and networking. He is working as an AP/IT in Mookambigai College of Engineering/Anna University – Pudukkottai, Tamilnadu.

P.Valarmathi completed her Master of Engineering in Information Technology from Vinayaka Mission University in 2007. She is doing research in the field of image processing and data mining. She is working as an Professor (HOD)/CSE in Mookambigai College of Engineering/Anna University – Pudukkottai, Tamilnadu.

S.Arun completed his Master of Technology in Information Technology from Veltech Multitech Dr.Rangarajan Dr.Sakunthala Engineering College/Anna University – Chennai in 2012. His research interests lie in the field of web services, image processing and networking. He is working as an AP/CSE in Madha Institute of Engineering and Technology/Anna University – Chennai, Tamilnadu.