Step 2: Clustering testing data vectors
a-Ntest **y**
While (Iter < Max_Iteration)
Load (Aw-m, m = 1,…, mem-max , ya)
for all ant-Worker Aw, (w = 1,……., Aw-max) **do**
Find Minimum Average Distance (ya, Ci)
Compute Neighbour Function *f* (ya, Ci) Equation 3
if Pd (ya) >= t **then** Equation 2
Drop(ya, Ci);
Update Threshold (ti, tnew) **end-for** Equation 6&7
end-while

Step 3: Merging the most similar and neighboring Clusters
for all Ci, (i = 1,……., C-max) **do**
Compute Clusters Minimum Equation5
Average Distance MAD (Ci, Cj)
Compute Cluster Neighbour Function *f* (Ci, Cj) Equation 3
if f(ya, Cj) ³ t || f(yb, Ci) ³ t **then**
Merge Clusters (Ci, Cj); **end-for**

Step 4: Cluster Refinement: Detecting Small Clusters and Checking Boundaries Rule 1-3
Find Clusters mean points (Cm) R. 1
for all Cmi& Cmj , (i &j = 1,……,Cm-max) **do**
Compute Neighbour Function Equation 3
for Close Clusters *f* (Cmi , Cmj)
if Pp (Cmi) >= t **then** Equation 1
Merge Clusters (Ci, Cj); **end-for**
Find Intersection Vectors for Close Clusters (Ci, Cj) R. 2
for all ya, (a = 1,……., y-max) **do**
Compute Neighbour Function *f* (ya ,Ci) &*f* (ya, Cj) Equation 3
if Pp (ya , Ci) > Pp (ya , Cj) && R. 3
Pd (ya , Cj) >= t **then**
Add(ya , Cj); **end-for**
Hybrid Agglomerative Clustering and Ant Colony Optimization (ACO) provide high efficiency and high quality data clustering.

5. Results

This method is used to achieve high efficiency and high-quality data clustering. The method is also beneficial to be used in textual document clustering algorithms for many text domain applications. To enhance this work, we are proposing a new hybrid clustering algorithm using agglomerative clustering with ACO algorithm. In this we have done pre-processing from new document by document clustering and using agglomerative & ACO algorithms then after we evaluate the performance. By using these algorithms we can achieve high efficiency and high quality.

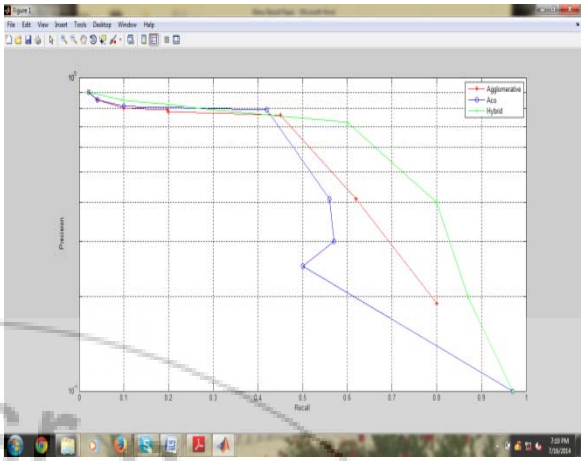


Figure 4.1: Showing Agglomerative Clustering Hybrid with ACO

Our algorithm allow for two input parameters: Precision and Recall. As in Text Mining in Clustering, Hybrid algorithms Agglomerative and ACO for high efficiency and high quality. Recall set for Hybrid is 0.9 and precision 0.2. Hybrid algorithm accurate from another is value around 20%.

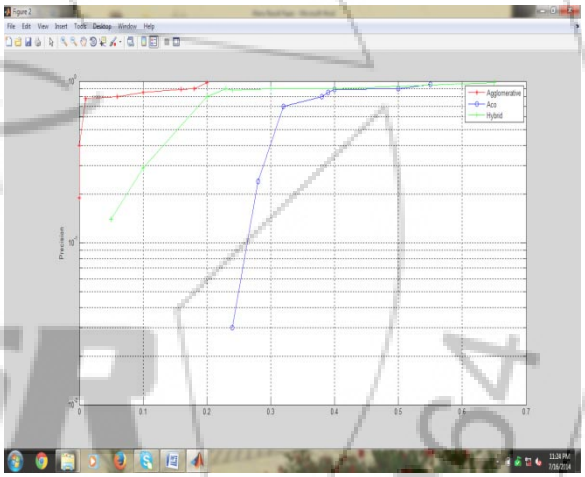


Figure 4.2: Showing Agglomerative Clustering Hybrid with ACO.

Our algorithm allow for two input parameters: Precision and Recall. As in Text Mining in Clustering, Hybrid algorithms Agglomerative and ACO for high efficiency and high quality. Recall set for Hybrid is 0.7 and precision 0.2. Hybrid algorithm accurate from another is value around 25%.

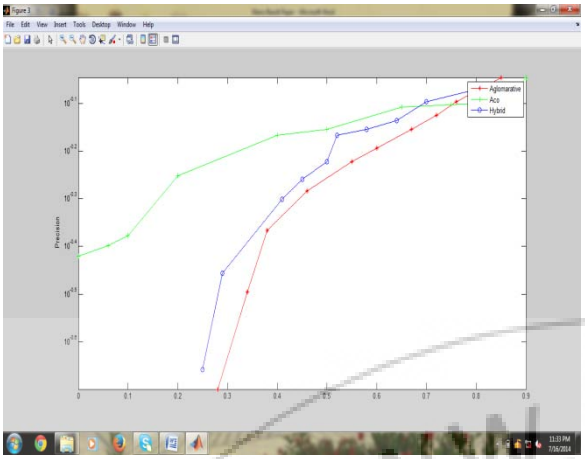


Figure 4.3: Showing Agglomerative Clustering Hybrid with ACO.

Our algorithm allow for two input parameters: Precision and Recall. As in Text Mining in Clustering, Hybrid algorithms Agglomerative and ACO for high efficiency and high quality. Recall set for Hybrid is 0.7 and precision 0.2. Hybrid algorithm accurate from another is value around 32%.

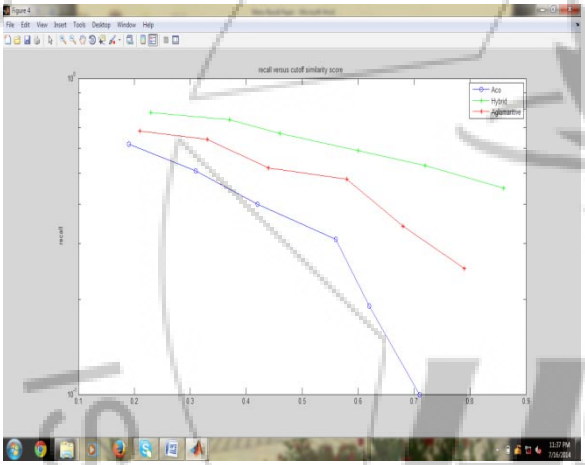


Figure 4.4: Showing Agglomerative Clustering Hybrid with ACO for Recall and Recall cut off Similarity Score.

Our algorithm allow for two input parameters: Recall and Recall cut off Similarity Score. As in Text Mining in Clustering, Hybrid algorithms Agglomerative and ACO for high efficiency and high quality. Recall set for Hybrid is 0.2 and Recall cut off Similarity Score 0.9. Hybrid algorithm accurate from another is value around 20%.

Table 1: Showing Clustering Performance Evaluation of Precision and Recall.

Partition	Cluster	Precision	Recall	F-measure
P1	C1	0.82	0.566	0.649
	C2	0.32	0.5865	0.4049
P2	C3	1	0.4255	0.6086
	C4	0.94	0.2876	0.3533
P3	C5	1	0.5	0.6543
	C6	0.4987	1	0.9765
P4	C7	1	0.3233	0.456
	C8	1	0.3211	0.52345
P5	C9	0.775	0.2356	0.28764
P6	C10	1	0.3567	0.4876
	C11	0.8	0.45	0.396
P7	C12	1	0.3567	0.4676
	C13	0.65	0.0975	0.1834
P8	C14	1	0.3032	0.453
	C15	0.3876	0.0975	0.17543
P9	C16	1	0.06756	0.13654
P10	C17	0.82	0.3	0.35678
P11	C18	0.55	0.07532	0.18765
P12	C19	1	0.27	0.45643
	C20	0.7	0.27	0.37875
P13	C21	1	0.6	0.687643
	C22	0.56	0.3	0.33333
P14	C23	1	0.71	0.16543

We have used the metrics precision, recall and F- measure shown in table 1 for evaluating the performance of the proposed approach. The evaluation metrics used in the proposed approach is given below:

F(i, j)= 2 * Recall (i, j) * Precision (i, j) / Precision (i, j) + Recall (i, j)

For each cluster, we compute the Precision, Recall and F – measure with the help of the above mentioned equation. The obtained results are shown in table 1.

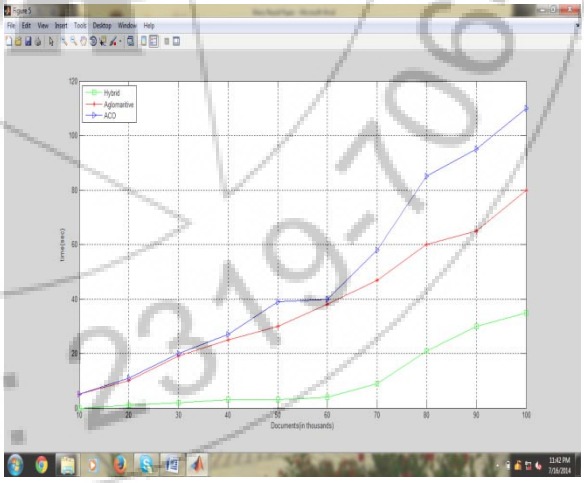


Figure 4.5: Showing Agglomerative Clustering Hybrid with ACO For Time and Document.

Our algorithm allow for two input parameters: Time and Document. As in Text Mining in Clustering, Hybrid algorithms Agglomerative and ACO for high efficiency and high quality. Time set for Hybrid is 30 and Document 100. Hybrid algorithm accurate from another is value around 30%.

6. Conclusion and Future Work

Data mining process is to extract information from a data set and transform it into an understandable structure for further use. In our work, we load a dataset on which preprocessing will be performed. Document clustering will be performed on the dataset using hybrid agglomerative clustering algorithm and ACO (Ant colony optimization) to get higher efficiency and higher quality rate. Performance will be evaluated on the basis of four parameters named as recall, precision, time and document. For future, we can enhance the performance of our work by using better optimization algorithm to get more efficiency in results.

References

[1] Wael M.S. Yafooz, Siti Z.Z. Abidin, Nasiroh Omar, Rosenah A. Halim, " Dynamic Semantic Textual DocumentClustering Using Frequent Terms and Named Entity" , IEEE 3rd International Conference on System Engineering and Technology, Vol.11, Issue.5, pp.1345-1351, Shah Alam , Malaysia, 2013.

[2] S.Murali Krishna, S.Durga Bhavani, "An Efficient Approach for Text Clustering Based on Frequent Itemsets" ,European Journal of Scientific Research, Vol.42, Issue.3, pp.385-396, 2010.

[3] Abdelmalek Amine1,2, Zakaria Elberrichi1, and Michel Simonet, "Evaluation of Text Clustering Methods Using WorldNet" ,The International Arab Journal of Information Technology, Vol.7, Issue.4, pp.590-780, October 2010.

[4] Benjamin C.M. Fung, Ke Wang, Martin Ester, "Hierarchical Document Clustering Using Frequent Itemsets" , Journal of the American Statistical Association, Vol.8, Issue.4, pp.590-780, 2010.

[5] Tru H. Cao, Vuong M. Ngo, Dung T. Hong, Tho T. Quan, "A Named – Entity - Based Multi - Vector Space Model for Semantic Document Clustering" , Faculty of Computer Science and Engineering, Vol.13, Issue.4, pp.68-75, 2012.

[6] Julien Ah-Pine, Guillaume Jacquet, "Clique-Based Clustering for improving Named Entity Recognition systems" , Proceedings of the 3rd NIST TREC Conference, Vol. 10, Issue.4, pp.105-109,2009.

[7] Florian Beil, Martin Ester, Xiaowei Xu, "Frequent Term-Based Text Clustering" , The International Arab Journal of Information Technology, Vol.8, Issue.5, pp.121-145, 2006.

[8] Hiroyuki Toda, Ryoji Kataoka, "A Search Result Clustering Method using Informatively Named Entities" ,Canadian Journal of Information, Vol.5, Issue.4, pp.133-143, 2005.

[9] Ye - Hang Zhu1, Guan-Zhong Dai1, Benjamin C. M. Fung2, De-Jun Mu," Document Clustering Method Based on Frequent Co-occurring Words" , International Journal of Computer Science & Engg., Vol. 7, Issue.5, pp.60-78, 2006.

[10] Bader Aljaber, Nicola Stokes, James Bailey, Jian Pei, "Document Clustering of Scientific Texts Using Citation Contexts" , International Journal of Computer Science & Engg., Vol. 8, Issue.4, pp.110-121, 2006.

[11] M.J.Zaki and C.J.Hsiao, "An Efficient Algorithm For Closed Itemset Mining" ,In Proc.2002 SIAM Int. Conf. Data Mining, Vol.10, Issue.5, pp.457-473, Apr.2002.

[12] F.Rosenblatt, "A Probabilistic Model For Information Storage And Organization in the Brain" Rev." , Journal of Emerging Technologies in Web Intelligence, Vol.4, Issue.2, pp.386-498, 1958.

[13] "Data Mining Knowledge Discovery Handbook" , 2nd ed., Vol.10, Issue.3, pp.190-260, 2010.

[14] Dunham, M.H., " Data Mining: Introductory and Advanced Concepts" , Pearson Education, Vol.6, Issue.5, pp.35-99, 2006.

[15] Han, J. and Kamber, M., "Data Mining:Concept & Technologies" , Morgan Kaufman Publisher, Vol.6, Issue.4, pp.1-1000, 2006.

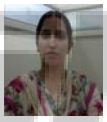
[16] Han, J. and Kamber, M., "Data Mining:Concept & Technologies" , Morgan Kaufman Publisher, Vol.11, Issue.4, pp.1-1000, 2011.

[17] N. J. Nilsson. Artificial Intelligence: A New Synthesis. Morgan Kauffmann, Vol. 1437, Issue.8, pp. 27-55, 1998.

[18] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. Science, Vol.220, Issue.4598, pp.671–680, 1983.

Author Profile

Manpreet Kaur is Student of M.Tech in the department of Computer Science and engineering at Shri Guru Granth Sahib World University (SGGSWU) , Fatehgarh Sahib,Punjab. She has done B.Tech from Yadvindra College of engineering, Talwandi Sabo under Punjabi University, Patiala.



Sukhpreet Kaur received the B. Tech degree in Computer Science and Engineering from Punjab Technical University, Punjab, India in 2007 and M.Tech Degree in Computer Engineering from Punjabi University, Patiala, India in 2010. She worked as an Assistant Professor at Department of Computer Science and Engineering in Baba Farid College of Engineering and Technology,Punjab, India till 2012. She is working as Assistant Professor at Department of Computer Science and Engineering, Shri Guru Granth Sahib World University Fatehgarh Sahib, Punjab, India since July 2012. She has published more than 30 research papers in different international journals and conferences. Her research areas include Software Engineering and Data mining.