# Web Document Clustering using Metaheuristic Approaches

Manjit Singh[1], Anshu Bhasin[2], Surender Jangra[3]
[1]*Ph.D Scholar, Department of Computer Applications, IKG Punjab Technical University Kapurthala, Punjab, India*
[2]*Department of Computer Science & Engineering, IKG Punjab Technical University Main Campus, Kapurthala, Punjab, India*
[3]*Department of Computer Application, Guru Teg Bahadur College, Bhawanigarh, Sangrur, Punjab, India*

***Abstract-*** Internet is a gigantic information resource, which is rapidly growing day by day as more and more data are being added to the World Wide Web. It is now becoming increasingly difficult to locate useful information from this environment. With the increase in use of mark-up languages, a new scenario has arisen into the information retrieval field. However, in the recent times, the volume of data on the World Wide Web is still rapidly growing day by day. This leads to the need for the development of new approach that may aid users in navigating, summarizing and organizing the required information. One of the techniques that could be useful to achieve this goal is web document clustering. However, existing techniques suffer from local optima problems. Various efforts have been made to address such drawbacks. This includes the utilization of various Meta-heuristic approaches as well. This paper provides a review of the currently use of HTML Tags in information retrieval and Meta-heuristics approaches used in web document clustering.

***Keywords-*** *World Wide Web, clustering, retrieval meta-heuristic, HTML.*

## I. INTRODUCTION

With rapid grow of web documents on WWW, it is becoming difficult to organize, analyze and present these documents efficiently. Web search engines can help the web user to browse and locate the documents in quick fashion. Normally web search engines return many documents to the web user, out of which some are relevant and some irrelevant documents to the topic, for the given query. Usually web search is a currently being done using features that are extracted from the web page-text only. The web documents present on the web are written using Hyper Text Mark-up Language mostly. These web pages(HTML) are consist of a set of mark-up tags that describe the layout, the presentation, and the content of the web page. It has been observed that by considering the terms within HTML tags of a web document, the performance of documents retrieval system can be improved [3,10]. Still, it is a big challenge to organize the documents in a manner that results in better search without extra cost and complexity. Clustering can play an important role in organizing such a large amounts of documents into groups called clusters. In each cluster, as per some similarity measure, documents share some common attributes. A related web document could be rated low in a information retrieval system in the absence of some query terms. However, if we consider the terms within HTML tags of a web document, it could improve the relevancy of the document to the query. Thus by considering terms within HTML Tags, document clustering can enhance the performance of an IR system [23]. Web document clustering works on the base of inter document resemlance. The resemlance between webs documents is determined by the set of terms (including terms within HTML Tags) shared between web documents. It is found that for clustering large datasets, K-means is mostly used [34]. But due to its choice of initializations, K-means suffers from many drawbacks. To get rid of these problems, optimization based techniques have been considered and they treat data clustering as an optimization problem [31]. Use of optimization has significantly improved the accuracy and efficiency of clustering. Stochastic optimization approaches are good at preventing convergence to a locally optimal solution. Hence, these approaches could be helpful to find global near-optimal solution. Meta-heuristics being commonly used are Ant Colony Optimization, Particle Swarm Optimization and Genetic Algorithms etc.

### *A. Models for information Retrieval*

Several models are in use for information retrieval. Some of these are

*1) Boolean retrieval model:* The model is based on set theory and Boolean operators where documents and the query are treated as a set of terms. During retrieval the documents containing the query terms are returned to the user, the retrieval strategy is binary, either the document contain the term or not.

*2) Probabilistic retrieval model:* It is a formalism of information retrieval and helpful to obtain ranking functions for ranking similar documents as per their relevancy to a given search query. It makes an estimation of the probability of finding whether a document dj is similar to a query q.

*3) Vector space retrieval model:* The documents and the query are represented as vectors of features in this model

[1]. For instance, a document $D_i$ and a query $Q_j$with N total number of terms are represented as

$$D_i = [T_{i1}, T_{i2}, ..., T_{ik}, ..., T_{iN}] \quad (1)$$

$$Q_j = [Q_{j1}, Q_{j2}, ..., Q_{jk}, ..., Q_{jN}] \quad (2)$$

*B. Similarity Measure*

It is necessary to determined document similarity before performing a clustering of documents in a cluster analysis. A similarity measure computes the degree of relevance between a pair of vectors. As documents and queries are both vectors, cosine similarity between a document $D_i$ and a query $Q_j$ will be:

$$\text{Sim}(D_i, Q_j) = \frac{\sum_{k=1}^{N} D_{ik} \cdot Q_{jk}}{\sqrt{\sum_{k=1}^{N} D_{ik}^2 \cdot \sum_{k=1}^{N} Q_{jk}^2}} \quad (3)$$

## II. WEB DOCUMENT CLUSTERING

Web document clustering is clustering of web documents. Document clustering algorithms aids to discover groups of common patterns in documents. There are different ways to cluster documents. Commonly used clustering methods are: Partitional and Hierarchical clustering.

*A. Partitional Clustering*

A partitional clustering algorithm finds all the non-overlapping clusters at once by dividing the set of documents based on an objective function. These algorithms try to minimize or maximize an objective function.

*1) K-means Clustering:*

This technique is a most commonly applied partitional clustering technique. The K-means clustering algorithm tries to minimize the objective function O(N, K)

$$O(N, K) = \sum_{j=1}^{K} \sum_{d_i \in c_j} \left\| d_i - c_j \right\|^2 \quad (4)$$

Where $\left\| d_i - c_j \right\|^2$ represents the distance between the document($d_i$)and the centroid($c_j$).

The centroid of the documents in a cluster $C_j$ can be computed as

$$c_j = \frac{1}{|c_j|} \sum_{d_i \in c_j} d_i \quad (5)$$

*2) Hierarchical Clustering:*

The objective of Hierarchical Clustering techniques is to build a diagram that represents the hierarchy of the clusters. They often build a hierarchical two dimensional structure called a dendogram. These are of two types that depend on the way the dendogram is built.

The first Hierarchical Clustering paradigm is the agglomerative strategy that starts with each pattern belonging to one cluster. Then, iteratively, clusters are joined until one unique cluster contains all the data. Examples of agglomerative techniques are the Single Linkage and the Complete Linkage algorithms.

The other Hierarchical Clustering paradigm is the divisive strategy that starts with a unique cluster containing all the patterns and repeatedly subdivides the clusters until each pattern belongs to a different cluster.

## III. METAHEURISTIC APPROACHES IN WEB DOCUMENT CLUSTERING

Meta-heuristics algorithms are now been extensively used to solved a variety of NP hard optimization problem including web document clustering. In this paper, we discussed in some details two Meta-heuristics algorithms Particle swarm optimization (PSO) and Ant colony optimization (ACO) and their usage in web document clustering.

*A. Particle Swarm Optimization*

PSO learns from a scenario where birds group searches for food, randomly, in a specific area and applies it to solve optimization problems. Each solution in the search space, in PSO, is alike a "bird", known as "particle". The characteristic of each particle in the group is represented by current position and current velocity of the particle. Data clustering has been viewed as an optimization problem and algorithms such as PSO have been used to solve it.

*B. Ant Colony Optimization*

Ant algorithms are a family of algorithms inspired by the behaviour of real ant colonies. When in search of food, ants deposit a chemical substance called pheromone on their path. If they succeed in finding a food source they again deposit pheromone on their path back to the colony. Also, if the multiple ants find the same food source, the one which followed the shorter path will deposit the pheromone first. Thus, the shorter paths and the paths which lead to better food sources will be followed by more and more ants and the pheromone on them will be further enriched. Also, because pheromone evaporates in atmosphere with the passage of time, paths which are no longer useful or followed will lose the pheromone deposited on them. However, ants are always free to choose their own paths ensuring the explore for new food sources [17].This behaviour of ants can be utilized to design algorithms to solve optimization problems including web document clustering.

## IV. COMPARATIVE ANALYSIS OF CLUSTERING ALGORITHMS

Comparison of various clustering algorithms is provided in Table1.

**TABLE I. COMPARISON OF VARIOUS CLUSTERING ALGORITHMS**

| Author/s | Algorithm | Similarity Function | Objective / Fitness Function | F-Measure | Data Sets | Initial Seeds | Results |
|---|---|---|---|---|---|---|---|
| Douglass R. Cutting, David R.Karger,Jan O. Pedersen and John W. Tukey,1992 [2] | Scatter/ Gather | $\frac{\langle g(c(\alpha)),g(c(\beta))\rangle}{\|g(c(\alpha))\|\|g(c(\beta))\|}$ | ..... ..... | ..... | New York Times News Service | ..... | 1.Dynamic Clustering 2. Clusters presented with summaries 3. Fast |
| Merwe V. D. And Engelbrecht,2003[13] | gbest PSO, K-Means and Hybrid PSO | Euclidean distance | $\frac{\sum_{j=1}^{N_c}\left[\sum_{\forall z_p \epsilon c_{ij}} d(z_p,m_j)/\lvert c_{ij}\rvert\right]}{N_c}$ | ..... | Artificial 1,Artificial 2, Iris, Wine,Breast-cancer and Automotive | Pso provide initial seeds | PSO approaches have better convergence to lower quantization errors and provide clusters of good quality. |
| Xiaohui Cui, Thomas E. Potok, Paul Palathingal, 2005[15] | K-Mean,PSO,Hybrid PSO | Cosine, Euclidian | $\frac{\sum_{i=1}^{N_c}\frac{\sum_{j=1}^{p_i} d(o_i,m_{ij})}{p_i}}{N_c}$ | ---------- | TREC | PSO provide initial seeds | Higher compact clustering is generated by Hybrid PSO algorithm than using PSO or the K-means. |
| Xiaohui Cui, Thomas E. Potok,2005 [14] | K-mean,pso,Hybrid pso | Euclidian and Cosine | $\frac{\sum_{i=1}^{N_c}\frac{\sum_{j=1}^{p_i} d(o_i,m_{ij})}{p_i}}{N_c}$ | ..... | Text REtrieval Conference (TREC) | Pso provide initial seeds | Higher compact clustering is generated by Hybrid PSO algorithm than using PSO or the K-means. |
| C.IMMACULATE MARY, DR. S.V. KASMIR RAJA(2005-09)[16] | K-Means with Mode, K-Means with ACO | Using some distance measure | ------- | K-Means with ACO improve the F-measure | Wisconsin Breast Cancer Dataset,Dermatology Dataset | Initial cluster centers are selected based on statistical mode | Ant Colony Optimization algorithm improve the cluster quality than k-means |
| Yulan He, Siu Cheung Hui, and Yongxiang Sim,2006 [17] | AHC, K-means, aiNet, and ant-based | Euclidian | $\frac{\sum_{i=1}^{N_c}\frac{\sum_{j=1}^{N_i} \mathrm{dist}(c_i,d_j)}{N_i}}{N_c}$ | Ant-based clustering method improve the F-measure | 20 Newsgroup data set | ..... | The performance of Ant-based clustering method is proved to be better than K-means and AHC and aiNet |
| K.Premalatha, Dr. A.M. Natarajan, 2009[19] | DPSO with GA operators algorithm | Cosine | $\frac{\sum_{i=1}^{N_c}\frac{\sum_{j=1}^{p_i}\frac{m_{ij}.o_i}{\|m_{ij}\|.\|o_i\|}}{p_i}}{N_c}$ | ..... | CISI, Cranfield and ADI | ..... | DPSO with GA operators algorithm avoid the stagnation behavior of the particles. |
| Fengqin Yang, Tieli Sun,and Changhai Zhang,2009 [20] | KHM ,PSO, and PSO-KHM | ..... | $\sum_{i=1}^{n}\frac{k}{\sum_{j=1}^{k}\frac{1}{\|x_i-c_j\|^p}}$ | F-Measure | ArtSet1, ArtSet2, Iris, Glass, Cancer, CMC and Wine | PSO provide initial seeds in hybrid psokhm | The performance of PSO-KHM is compatible to KHM and better than PSO in terms of F-measure. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Kayvan Azaryuon, Babak Fakhar ,2013 [24] | ACO based Novel Document Clustering Algorithm, Standard clustering k-means | $j^{\Sigma sim(o,oj)/N(o)}$ | ..... | Proposed Novel ant-colony based clustering algorithm improve the Presioion and Recall | 21578 Reuters Information Bank | ..... | The proposed algorithm is useful in quality of produced clusters and algorithm run time compared to the standard ant clustering algorithm and the K-means algorithm |
| Sunita Sarkar, Arindam Roy and B.S.Purkayastha,2014 [25] | K-Mean,PSO,Hybrid PSO | Cosine | $f = \frac{\sum_{i=1}^{N_c}\{\frac{\sum_{j=1}^{P_i} d(o_i m_{ij})}{P_i}\}}{N_c}$ | ..... | N EPALI W ORD NET | PSO provide initial seeds in hybrid pso+k-mean | Authors found that Hybrid PSO+K-means perform better than PSO and K-means algorithms |
| R. Malathi Ravindran and Dr. Antony Selvadoss Thanamani, 2015[27] | K-Mean | Cosine | $\text{argmin}_{c,n}\sum_{x=1}^{n}\left\|e_x - c_{n(x)}\right\|^2$ | ..... | ..... | ..... | In this paper it has been proved that when even the dimension is high ,documents can be clustered efficiently |
| Oi-Mean Foong and Suet-Peng Yong , 2016[32] | LSA, PSO, Fuzzy-PSO and LSA-PSO | Euclidian | $f = \frac{\sum_{i=1}^{N_c}\{\frac{\sum_{j=1}^{P_i} d(o_i m_{ij})}{P_i}\}}{N_c}$ | Precision, Recall and F1 measure | DUC2002 | ..... | LSA-PSO algorithm achieved satisfactory results with average F1 measure |
| Neha Garg, R.K.Gupta, 2016[28] | K-means, Hybrid GA-KM | Cosine | Objective function through ranking method | ---------- | Cnae9 Dataset ,Reuter 21578 Dataset | GA provide initial seeds for hybrid GA-KM | Hybrid GA-KM algorithm achieves better results when applied to real datasets than the k-means algorithm. |
| Mr. Hardeep Singh,2016 [29] | K-Mean and Spherical K-Mean | Cosine | ..... | ..... | 20 Newsgroups data set,Faults in an urban waste water treatment plant' | ..... | SKM is partially sensitive to noise,provides better results for large databases, low average running time, average density of clusters |
| Irwan Bastian, Rozaliyana, Metty Mustikasari,2016[30] | K-Mean | Cosine and Winnowing algorithm | ..... ..... | K-Means improve precision average | Indonesian archive online news portal consisting of kompas.com, tempo.com, and detik.com. | Initial cluster centers randomly from given data | The execution time not influenced by the number of clusters and precision values show that algorithm is quite accurate. |
| Monika Raghuvanshi , Rahul Pate.2017 [33] | Modified K-mean Alogrithm,K-Mean | Euclidian | avg_dist from min_obj and max_obj. | ..... | ..... | Select initial cluster centers randomly from given data | Modified K-mean algorithm provides high accuracy and purity of clustering. |

## V. RESULTS AND DISCUSSIONS

In given section, we present an overview of web document clustering algorithms in some detail.

Salton (1971) proposed vector space model to represent text documents in vectors in a feature space. Terms in a document collection were taken as features. The feature values are computed by using various weighting schemes. Term-Frequency and Inverse Document Frequency (tf*idf) is the most commonly used term weighting scheme [1]. In [5] authors provides the details of cluster analysis theory and they divides the methodologies mainly into partional and hierarchical techniques. In their work clustering algorithms used comes under the partitonal category. It is found that a lightweight document clustering method can process tens of thousands of documents and group them into several thousand clusters [6]. In [34] author discussed the process of text document clustering for different types of clustering techniques. He observed that for huge dataset k-means performs better than the others, DBSCAN performs superior to other techniques for outlier detection. Partitional clustering algorithms (mainly bisecting K-means) due to low computational requirements and better clustering results are useful for clustering big documents collection [8,9, 18]. It is noticed that if term weighting is given as per the importance of tags in which they appear, the significance degree of an index term can be computed [3, 10]. The approaches that use the HTML tag weights to improve retrieval performance have been developed and also used genetic algorithm to find the optimal tag weights. It is observed that the terms within HTML Tags are useful for improving the retrieval effectiveness [4,7, 11, 12]. Ammar Sami Al-Dallal proposed a searching model that is based on GA to retrieve HTML documents. He achieved a high recall and precision with HTML Documents by applying Genetic Algorithm [23]. It is observed that PSO, its modification and hybridization with other algorithms gives better results in terms of execution time, efficiency and accuracy as compared with other evolutionary algorithms [22]. A literature survey on document clustering is presented mentioning that indexing and retrieval operation can be optimised, if documents are clustered in a sensible order [21]. Particle Swarm optimization based clustering techniques have been discussed and provides systematically surveyed the work and presented the results of increasing trends in the literature of swarm intelligence, Particle Swarm Optimization and PSO-based data clustering [26]. A comprehensive reviewed the work and the literature survey reveals that PSO and hybrid PSO based data clustering techniques have outperformed many existing techniques [31].

## VI. CONCLUDING OBSERVATIONS

In this paper we studied and summarized different kind of Meta-heuristic approaches used in clustering. Document clustering helps to improve the precision and recall of the IR system. We have also discussed the use of HTML Tags in information retrieval. It has also been observed that inclusion of various HTML Tags can further improve retrieval of information in web search. Keeping in view these facts and available literature on document clustering and use of HTML Tags in web search it should be possible to develop more effective web search algorithms for HTML documents. We proposed to develop some such types of search algorithms.

## VII. ACKNOWLEDGEMENTS

We are grateful to authorities of Ambala College of Engineering & Applied Research (ACE), Ambala for providing necessary facilities. We also thankful to Dr. Chander Mohan, Professor, Department of CSE for their help and valuable suggestions in writing of this paper.

## VIII. REFERENCES

[1]. G.Salton,"The SMART retrieval system-experiment in automatic document processing," Prentice-Hall, Englewood Cliffs, New Jersey,1971.

[2]. D.R.Cutting,D.R.Karger,J.PedersenandJ.Tukey,"Scatter/Gather: A cluster-based approach to browsing large document collections,"Proceeding of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval June 21-24, Copenhagen Denmark,pp:318-329,1992.

[3]. Molinari, Andrea and Gabriella Pasi,"A fuzzy representation of HTML documents for information retrieval systems,"In: Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, vol. 1, pp. 107-112,1996.

[4]. M.Cutler,Y.Shih, and W.Meng,"Using the structure of HTML documents to improve retrieval,"The USENIX Symposium on Internet Technologies and Systems, pp. 241–251. Monterey, California ,1997.

[5]. A.K. Jain, M.N. Murty and P.J. Flynn," Data clustering: a review," ACM Computing Surveys, pp. 31, 3, 264-323,1999.

[6]. Sholom Weiss, Brian White and Chid Apte,"Lightweight document clustering," IBM Research Report RC-21684, 2000.

[7]. Sun Kim and Byoung-Tak Zhang,"Web-document retrieval by genetic learning of importance factors for HTML tags,"PRICAI 2000 Workshop and Text and Web Mining,Melbourne,pp. 13-23 , August 2000.

[8]. Michael Steinbach , George Karypis and Vipin Kumar, "A comparison of document clustering techniques," In KDD Workshop on Text Mining, 2002.

[9]. Ying Zhao and George Karypis,"Evaluation of hierarchical clustering algorithms for document datasets," Technical Report, Jun. 2002.

[10]. Andrea Molinari and Gabriella Pasi,"An indexing model of HTML documents,"In proceedings of the 2003 ACM symposium on Applied computing.

[11]. S.Kim and B.T.Zhang,"Genetic mining of html structures for effective web document retrieval," Applied Intelligence, vol. 18, no.3, pp.243-256,2003.

[12]. Byurhan Hyusein and Ahmed Patel, "Significance of HTML tags for document indexing and retrieval," International Conference WWW/Internet 2003.

[13]. V. D. Merwe and A. P. Engelbrecht,"Data clustering using particle swarm optimization," Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia. pp. 215-220.

[14]. Xiaohui Cui and Thomas E. Potok, "Document clustering analysis based on hybrid PSO+K-means algorithm", Special Issue, 2005.

[15]. X.Cui,T.E.Potok,P.Palathingal,"Document clustering using particle swarm optimization,"In: Proceedings in SIS. Pp 185–191,2005.
[16]. C.Immaculate Mary, DR. S.V. Kasmir Raja,"Refinement of clusters from k-means with ant colony optimization," Journal of Theoretical and Applied Information Technology,JATIT 2005-2009.
[17]. Yulan He, Siu Cheung Hui and Yongxiang Sim,"A novel ant-based clustering approach for document clustering," Springer-Verlag Berlin Heidelberg 2006,LNCS 4182, pp. 537–544, 2006.
[18]. R. Kashef and M.S.Kamel,"Enhanced bisecting k-means clustering using intermediate cooperation,"Journal of Pattern Recognition, vol. 42, issue 11, pp. 2557-2569, Nov. 2009.
[19]. K.Premalatha and A.M. Natarajan,"Discrete PSO with GA operators for document clustering,"Int. J. of Recent Trends in Engineering and Technology, Vol. 1, No. 1, Nov 2009.
[20]. Fengqin Yang,Tieli Sun and Changhai Zhang," An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization,"Expert Systems with Applications 36 9847–9852,2009 .
[21]. K. Premalatha and A.M. Natarajan,"A literarure review on document clustering,"Information Technology Journal,ISSN 1812-5638,2010.
[22]. S. Rana, S. Jasola and R. Kumar, "A review on particle swarm optimization algorithms and their applications to data clustering," Artificial Intelligence Review, vol. 35, pp. 211–222, 2011.
[23]. Ammar Sami Al-Dallal,"Enhancing recall and precision of web search using genetic algorithm," A thesis submitted for the degree of Doctor of Philosophy, School of Information Systems Computing and Mathematics, Brunel University, August 2012.
[24]. Kayvan Azaryuon and Babak Fakhar,"A novel document clustering algorithm based on Ant Colony Optimization algorithm,"Journal of mathematics and computer Science 7 , 171-180,2013.
[25]. Sunita Sarkar, Arindam Roy and B. S. Purkayastha,"A comparative analysis of Particle Swarm Optimization and K-means algorithm for text clustering using Nepali Wordnet,"International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.
[26]. Shafiq Alam, Gillian Dobbie ,Yun Sing Koh, Patricia Riddle and Saeed Ur Rehman," Research on particle swarm optimization based clustering: A systematic review of literature and techniques,"Swarm and Evolutionary Computation 17 1–13,2014.
[27]. R. Malathi Ravindran and Antony Selvadoss Thanamani," K-means document clustering using Vector Space Model,"Bonfring International Journal of Data Mining, ISSN 2277 – 5048,Vol. 5, No. 2, July 2015.
[28]. Neha Garg and R.K. Gupta." Document clustering analysis based on hybrid clustering algorithm,"IJARCCE,ISSN (Print) 2319 5940,Vol. 5, Issue 4, April 2016.
[29]. Hardeep Singh," Clustering of text documents by implementation of K-means algorithms,"Streamed Info-Ocean,Volume 1, Issue 1, January-June 2016.
[30]. Irwan Bastian, Rozaliyana and Metty Mustikasari," Web document clustering system using K-Means algorithm",IJARCSSE,ISSN: 2277 128X,Volume 6, Issue 8, August 2016.
[31]. K. Nafees Ahmed and T. Abdul Razak,"A study on Metaheuristic Optimization approach for data clustering,"IJETST- Vol.03,Issue 08,Pages 94-101,August,ISSN 2348-9480,2016.
[32]. Oi-Mean Foong and Suet-Peng Yong,"Swarm LSA-PSO clustering model in text summarization,"Int. J. Advance Soft Compu. Appl, Vol. 8, No. 3,ISSN 2074-8523., December 2016.
[33]. Monika Raghuvanshi and Rahul Patel ,"An improved document clustering with multiview point similarity /dissimilarity measures," IJECS Volume 6 Issue 2, Page No.20285-20288,Feb. 2017.
[34]. Mohit,"Text clustering techniques: A survey ", IJESRT, ISSN: 2277-9655, May, 2017.