International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5C, May 2019

Efficient Retrieval Of Html Documents Using Hybrid Meta-Heuristic Approaches In Web Document Clustering

Manjit Singh, Anshu Bhasin, Surender

Abstract:- ***With the rapid growth of web documents on WWW, it is becoming difficult to organize, analyze and present these documents efficiently. 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. Web search is usually performed using only features extracted from the web page text. HTML tags with particular meanings have been found to improve the efficiency of the information retrieval System. However, organizing documents in a way that will improve search without additional cost or complexity is still a great challenge. Clustering can play an important role to organize such a large number of documents into several groups. However due to limitations in existing techniques of clustering, scientists have begun using Meta-heuristic algorithms for the clustering problem of documents. In this paper, we presented a document clustering method that uses HTML tags and Meta-heuristic approaches. The hybrid PSO+ACO+K-means algorithm is used for clustering the documents. In the proposed approach, results are analyzed on WEBKB dataset***

I. INTRODUCTION

Information retrieval is the science of storage, data search and finding out information within data. In a computing context data are raw objects controlled by a computer device. These include text, web documents, pictures, videos and audio clips. The primary objective of the traditional IR was to find documents which correspond to the user's needs, which were expressed by means of inquiries. After several decades of effort, research and development, the IR field have developed significantly and has now capable of retrieving in just a few seconds textual and non-textual information out of millions of documents. Hyper Text Markup Language (HTML), which is consist of a set of markup tags helped to define the content, the presentation, and the layout of the web page, is mostly used to write web pages on the Internet. However, little effort has been done to utilized HTML documents effectively in an information retrieval context by combining the interior structure of the HTML language with the semantics of the document. This has been noticed that by considering the terms in the HTML tags of a web document, the performance of document retrieval systems can be improved. However, organizing documents in a way to improved search without additional costs and complexity is a major challenge. Clustering can play a vital role to organize a large amount of web pages into groups called clusters. In each cluster, as per some similarity measure, documents share some common attributes. Still, a related web document could be rated lower in an information retrieval system in the absence of any query terms. However, if we consider the terms within the HTML tags of a web document, it could improve the relevancy of the document to the query. In this way, document clustering can improve the performance of an IR system by considering the terms in HTML Tags. It has been found that K-means are mostly used for clustering large datasets. However, K-means suffer many drawbacks due to their choice of initializations. To get rid of such problems, techniques based on optimization have been considered and that consider data clustering as an optimization problem. In the last decade, there has been an immense expansion in the field of meta-heuristics which are used as solution for various issues related to real-world optimization. Use of optimization has considerably improved the accuracy and efficiency of clustering. Particle Swarm Optimization and Ant Colony Optimization are commonly used Meta-heuristics.

II. META-HEURISTIC APPROACHES

Meta-heuristic approaches are used in web crawling process optimization. In the last decade, there has been an immense expansion in the field of meta-heuristics as a solution for various issues related to real-world optimization. In order to design an effective web search algorithm that makes use of HTML tags and can retrieve the web documents, this paper planned a retrieval algorithm which utilizes the following Meta-heuristics algorithms.

2.1 Particle Swarm Optimization (PSO)

It is a stochastic based on Kennedy and Eberhart's population search algorithm. A large number of optimization problems have been solved by utilized it widely. This algorithm has characteristic of computation and fast convergence ability. It is used in different clustering problems and image processing.

Manuscript published on 30 May 2019.
* Correspondence Author (s)
Manjit Singh, Ph.D Research Scholar, Department of Computer Applications, IKG Punjab Technical University, Kapurthala, Punjab, India(Email: manjitbehniwal@rediffmail.com)
Anshu Bhasin, Assistant Professor, Department of Computer Science & Engineering, IKG Punjab Technical University, Main Campus, Kapurthala, Punjab, India.(Email: dr.anshubhasin@ptu.ac.in)
Surender, Assistant Professor, Department of Computer Science, Guru Tegh Bahadur College, Bhawani garh, Sangrur, Punjab, India. (Email: jangra.surender@gmail.com)

Each individual particle and the swarm is composed of particles in this algorithm. In it, solution space of the problem is made as a search space. The solution to the problem is provided by each position in search space. Each particle imitates the behavior of animals in the search space which has velocities, positions, and population. In the beginning, initialization of population randomly in PSO, each particle flies in search space and best positions are stored. The best positions are communicated among the swarm members and on the basis of these good positions, velocity and positions are dynamically adjusted. The particles and their neighbor's past behavior are adjusted on the basis of the velocity. In this way, a better search area can be obtained in this process. It imitates the behavior of flocking birds for global search. PSO gives a positive outcome in complex and nonlinear problems. This algorithm used some parameters to adjust and can provide computational results which are fast and accurate, that's why it is used to find the solutions in clustering also [12, 18].

2.2 Ant Colony Optimization (ACO)

This algorithm was proposed by Dorigo and his colleagues. It is also based on population meta-heuristic algorithm that use for solving difficult problems of optimization like quadratic assignment problems. In this algorithm, every individual of the population is an artificial agent. It can find a solution incrementally and stochastically to the given problem. This algorithm is based on ants search space which lives together in colonies and can find food using shortest path with the help of pheromones. Pheromones act as a critical medium of communication between ants and help to determine the next movement. On the other hand, shortest path can be found by ants, depending on the intensity of pheromones. ACO algorithm is also used in feature selection. The feature selection is the selection process of the most important characteristics that can increase the efficiency of any clustering algorithm. The feature selection approaches are based on the supervised method in which class labels are used as a guide. But in the unsupervised feature selection, there are no class labels. ACO algorithm is used in unsupervised methods which are used to find the optimal feature using many iterations. It doesn't use any learning algorithms. It has low computational complexity and used in many datasets which have very high dimensions.

III. RELATED WORK

Salton (1971) [1] proposed vector space model to represent text documents in vectors in a feature space. The values of feature are measured by using different weighting schemes. [25] provided the details of cluster analysis theory. In their work clustering algorithms used came under the partition category. [2] and [8] observed 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.[4] used a generalized suffix-tree to obtain information of phrases. After that, they used these phrases for clustering of documents. [6][7][15] Demonstrated that algorithms of partition clustering (mainly bisecting K - means) were useful to cluster a large document collection because to low computational necessities and improved clustering results. [3][5][9][10] Has developed approaches to enhance retrieval performance using HTML tag weight. Genetic algorithms were also used to obtain optimum tag weight. They found that the terms inside HTML Tags were useful to improve the retrieval efficiency .In this paper [23] it was suggested a way of achieving a better result than any other existing algorithm by hybridizing K-means partitioning algorithm and Fuzzy C-means algorithms with Particle Swarm Optimizing (PSO).It was also noticed that FCPSO gives better results as compared to KPSO as FCPSO deals nicely with the overlapping nature of document.[18] noted that, in terms of efficiency, time and accuracy, PSO, its modification and its hybridization with other algorithms yield improved results as compared to other evolutionary algorithms.[11] showed that incorporation of the Document Index Graph(DIG) and an incremental document clustering algorithm into a phrase-based document index model creates robust and accurate document similarity calculation. The authors suggested in [21] a hybrid algorithm combining basic Ant Colony Optimization (ACO) and Tabu search. The experimental consequences showed that the proposed algorithm provides a quality clusters as compared to those produced by K-means. Experiments in [14] verify the statement of the successful application of the Ant algorithms in the processing of text documents. The results from experiments characterize the number of resulting categories by good quality, speed and flexibility. A literary survey on document clustering was presented in [17]. The report stated that if documents are clustered in a sensitive order, it can optimize the indexing and retrieval operations. [22] They proposed a novel algorithm for clustering by altering and constructing a smart form of movement for ants in the standard ant-clustering algorithm. They found that it provides time saving and increases the quality of clusters. [12] In this paper authors presented a hybrid PSO+K document clustering algorithm which performs rapid document clustering and which can prevented from being caught in a local optimal solution. [13] This paper proposed Ant Colony Optimization (ACO) to improve clustering. They select preliminary seeds for k-means clustering that are based on statistical models and proposed a novel method that is ant based refinement algorithm to increase the cluster quality.[16]They proposed a hybrid clustering algorithm (PSOKHM) based on PSO and KHM (K-harmonic means). They found that the PSOKHM algorithm helps to escape from local optima of KHM clustering and also assist PSO algorithm to get rid of the limitation of the slow convergence speed. In[19], the GSA-KM hybrid algorithm (gravitational search algorithm (GSA) and K-Means) was introduced to help to prevent the K-means algorithm from getting in local optima. The convergence rate of the GSA algorithm is increasing. [20] Ammar Sami Al-Dallas proposed a GA-based search model for the retrieval of HTML papers. By applying the genetic algorithm, they achieve high recall and precision with HTML documents.[24]

The authors discussed developments in cluster techniques based on the optimization of Particle Swarm. They systematically examined the work and presented the outcome of the growing trends in the swarm intelligence literature and particle swarm optimization and data clustering based on PSO. In [26] authors comprehensively reviewed the work and the literature survey reveals that there is a huge increase in the attractiveness of meta-heuristics, SI, and PSO based data clustering. It was found that the PSO and hybrid PSO based data clustering techniques have been superior to many existing techniques. In [27] Authors have studied various types of clustering technologies and have summarized them. They also discuss the merits and demerits of the clustering techniques.

IV. PROPOSED SYSTEM MODEL

The processing divides into two phases in this proposed model:

4.1 Phase I:

Step 1: Please take HTML document as an input document.

Step 2: Construct features vectors which represents documents and each vector have weight.

Step 3: In a features database, save weighted features vectors.

Step 4: Documents in the database are clustered using either K-means algorithm or hybrid PSO+ACO+K-means algorithm into different categories.

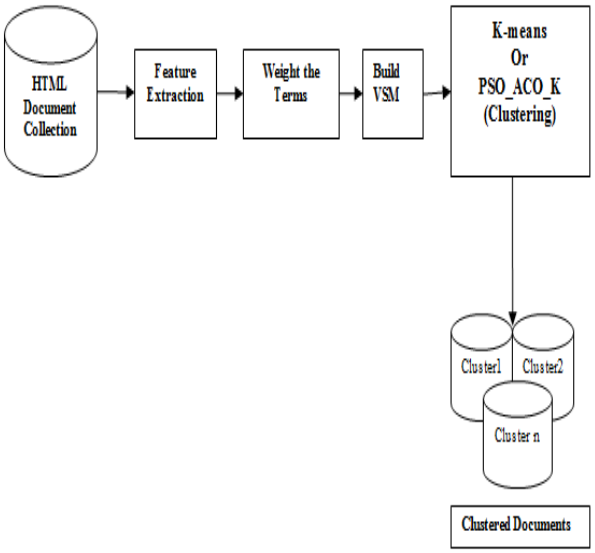

Fig 1. Clustering of documents

4.2 Phase II:

Step 1: Build the features vector for the input query by using the same approach as given in phase 1.

Step 2: For the input query, calculate the weighted feature vector.

Step 3: Find the closest cluster by calculating the distance between the input query and the centroid of each cluster. Cluster with smallest distance with input query will be considered as the closest cluster.

Step 4: Calculate the distance between the input query and the documents in the closest cluster.

Step 5: Retrieve documents that are more similar to the input query.

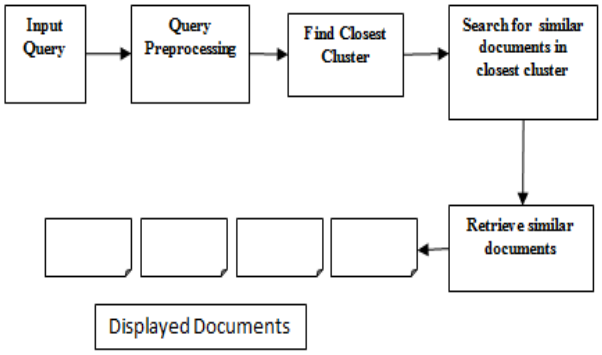

Fig 2 . Retrieval of documents

V. RESULT ANALYSIS

5.1 Data Set

The effectiveness of proposed model can be enhanced by using a WEB KB dataset. The WEB KB dataset collects web pages (HTML documents) from the departments of computer science of many universities. It has total 8,282 pages, but in this work, only 1000 pages have been used which are classified into different categories. This dataset has web pages from the Texas, Cornell, Wisconsin and Washington universities.

5.2 Evaluation Metrics

Two matrices Precision and Recall are used for the assessment of the proposed system. Generally, the precision is calculated by dividing the number of relevant documents retrieved to a total number of documents retrieved. A recall is calculated by dividing the number of relevant documents retrieved to a total number of relevant documents in the database. F-measure is a weighted harmonic mean of Precision and Recall.

Table 1. Results of WEB KB dataset

Methods	Recall	Precision	F-measure
PSO_ACO_K with weight	82.67	80.78	81.71
PSO_ACO_K without weight	79.93	78.75	79.34
K-means with weight	52.52	42.40	46.92
K-means without weight	50.87	44.59	47.52

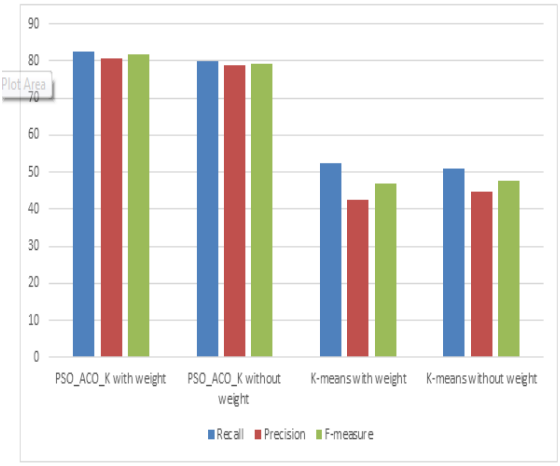

Fig.3 Precision, Recall and F-measure by using different methods

Fig.3 represents the result of PSO_ACO_K with and without weight, K-means with and without weight using parameters Precision, Recall, and F-measure. Result analysis of above figure represents that there is a significant improvement in above-mentioned parameters by using PSO _ACO_K with a weight. In K-means, Recall increased, but Precision and F-measure reduced. Thus the results of PSO_ACO_K with weight is better than K-means with weight.

Table 2. Improvement in results using a WEB KB dataset with weight

Methods	Recall	Precision	F-measure
PSO_ACO_K with weight	2.74	2.03	2.37
K-means with weight	1.65	-2.19	-0.6

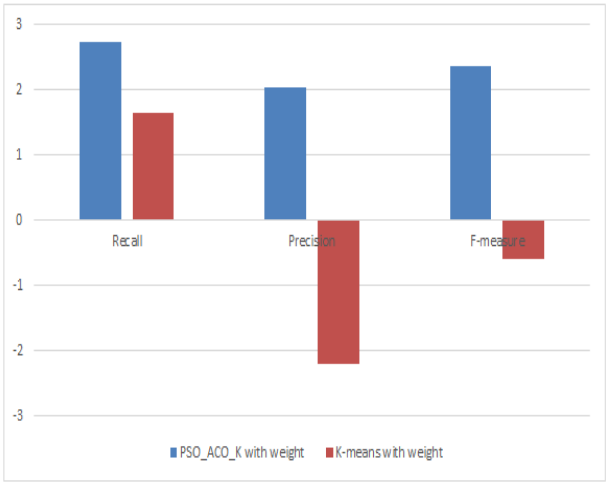

Fig.4 PSO_ACO_K and K-means with weight

Fig.4 shows the improvement in results of PSO_ACO_K as compared to K-means using weights. These results are assessed using precision and recall parameters, and F -measurement. There is a significant improvement in results of PSO_ACO_K optimization with the weighted features that is 2.37 as compared to PSO_ ACO_K without weight. On the other hand, there is an improvement in Recall using K-means with weight that is 1.65 as compared to K-means without weight. But in K-means, performance is decreased because PSO_ACO_K optimization reduces the boundary conditions and significantly increased the difference between the clustering optimization. These weights can increase the specific domain knowledge and increase the efficiency of convergence optimization.

Table 3: Improvement in results on WEB KB dataset with optimization

Methods	Recall	Precision	F-measure
PSO_ACO_K with weight	30.15	38.38	34.79
PSO_ACO_K without weight	29.06	34.16	31.82

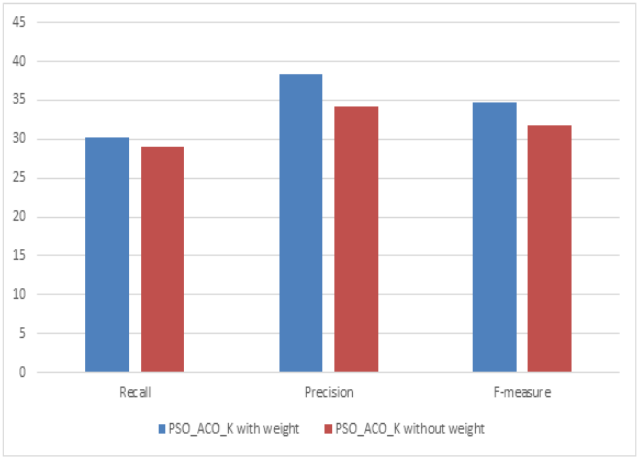

Fig.5 PSO_ACO_K optimization with and without weight

Fig.5 shows the result of PSO_ACO_K optimization with and without the weighted feature. This is actually the comparative improvement of weight and without weight by using optimization. The results show the significant improvement in using optimization in weight and without weight scenario. It improves the Recall, Precision, and F-measure parameters by 30.15, 38.38 and 34.79 by using the difference of the PSO_ACO_K with weight and K-mean with weight. Similarly 29.06, 34.16 and 31.82 by using the difference of the PSO_ACO_K without weight and K-mean without weight respectively.

VI. CONCLUSION

Clustering in web search is an effective solution for informational retrieval. Depending on information requirement, documents having the same cluster behave similarly. In this paper, an improvement in clustering is done by weighted HTML tags features and optimization in K-means clustering. Different approaches have been analyzed in this paper.

Features are used in two different ways for the purpose of determining effective document collection by adding weight and without adding weight. Precision, Recall, and F-measure are parameters used for examining the different clustering approaches. There is a substantial growth in the mentioned parameters by using hybrid PSO_ACO_K. In K-means, however, Recall improved but other parameters reduced. Our experimental results illustrate that hybrid PSO_ACO_K clustering algorithm perform better than K-means alone.

REFERENCES

1. Salton,G.,1971,"The SMART Retrieval System-Experiment in Automatic Document Processing", Prentice-Hall, Englewood Cliffs, New Jersey.
2. 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, 1996, pp. 107-112.
3. Cutler, M., Shih, Y., and Meng, W.(1997),"Using the structure of HTML documents to improve retrieval", The USENIX Symposium on Internet Technologies and Systems, pp. 241–251. Monterey, California.
4. Zamir,O.and O. Etzioni, 1998,"Web document clustering: A feasibility demonstration", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 24—28, ACM Melbourne, Australia,pp: 46-54.
5. 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.
6. Michael Steinbach , George Karypis, and Vipin Kumar, "A comparison of document clustering techniques," In KDD Workshop on Text Mining, 2002.
7. Ying Zhao and George Karypis,"Evaluation of Hierarchical Clustering Algorithms for Document Datasets", Technical Report, Jun. 2002.
8. Andrea Molinari, Gabriella Pasi,"An indexing model of HTML documents, "In proceedings of the 2003 ACM symposium on applied computing.
9. Kim, S., and Zhang, B-T. (2003),"Genetic mining of html structures for effective web document retrieval", Applied Intelligence, vol. 18, no.3, pp.243-256.
10. Byurhan Hyusein et al, "Significance of Html Tags for document indexing and retrieval" International Conference WWW/Internet 2003.
11. Khaled M. Hammouda and Mohamed S. Kamel, "Efficient Phrase-Based Document Indexing for Web Document Clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 16, No. 10, Oct. 2004.
12. Xiaohui Cui and Thomas E. Potok, "Document clustering analysis based on hybrid PSO+K-means algorithm", Special Issue, 2005.
13. 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.