# WEB PAGE CLUSTERING USING HARMONY SEARCH OPTIMIZATION

*Rana Forsati*[1] , *Mehrdad Mahdavi*[2] , *Mohammadreza Kangavari*[3], *Banafsheh Safarkhani*[4]

[1]Department of computer Engineering, Islamic Azad University, Karaj Branch, Karaj, Iran
[2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[3]Department of Computer &IT, Iran University of Science&Technology, Tehran, Iran
[4]Department of Computer Engineering, Tehran Azad University, Tehran, Iran

**ABSTRACT**

Clustering has become an increasingly important task in modern application domains. Targeting useful and relevant information on the World Wide Web is a topical and highly complicated research area. Clustering techniques have been applied to categorize documents on web and extracting knowledge from the web. In this paper we propose novel clustering algorithms based on Harmony Search (HS) optimization method that deals with web document clustering. By modeling clustering as an optimization problem, first, we propose a pure HS based clustering algorithm that finds near global optimal clusters within a reasonable time. Then we hybridize *K*-means and harmony clustering to achieve better clustering. Experimental results on five different data sets reveal that the proposed algorithms can find better clusters when compared to similar methods and the quality of clusters is comparable. Also proposed algorithms converge to the best known optimum faster than other methods.

***Index Terms***— clustering web pages, harmony search, global optimization

## 1. INTRODUCTION

Recently, as the web developed rapidly, a large collection of full-text documents in electronic form is available and opportunities to get a useful piece of information are increased. Clustering is the unsupervised classification of data set to reduce the amount of data by categorizing or grouping similar data items together. To apply the clustering techniques, each document is usually represented as a vector of weighted term frequencies such as Text Frequency (TF) and Inverse Document Frequency (IDF) [1]. For these vectors, it is necessary to calculate a similarity or distance measure that clustering algorithm defines between two vectors.

Some of the most conventional clustering methods can be broadly classified into two categories: The first one is hierarchical clustering. This type is used to build tree structure from data set [2-4].The another one is partitioning clustering that, cluster the data in a single level. Partitioning methods try to partition a collection of documents into a set of groups, so as to maximize a pre-defined fitness value. In recent years the partitioning clustering methods are well suited for clustering a large document dataset due to their relatively low computational requirements [5, 6]. The time complexity of the partitioning techniques is almost linear, which makes them widely used. The best known method in partitioning clustering is *K*-means algorithm [7].

Although *K*-means algorithm is simple, straightforward, and easy to be implemented and works fast in most situations, it suffers from two major drawbacks that make it inappropriate for many applications. One is sensitivity to initialization and the other is convergence to local optima. To deal with the limitations that exist in *K*-means, recently, new concepts and techniques have been entered into web data mining. One of these techniques is optimization methods that try to optimize a pre-defined objective function.

One of the advantages of partitioning-based clustering algorithms is that they use information about the collection of documents when they partition the dataset into a certain number of clusters. So, the optimization methods can be employed for partitional partitioning. Regarding to this definition, *K*-means can be considered as an optimization method. In addition to the *K*-means algorithm, several algorithms, such as Genetic Algorithm [8], Particle Swarm Optimization (PSO) [9] and Ant Clustering [10] have been used for web page clustering.

A meta-heuristic algorithm, mimicking the improvisation process of music players, has been recently developed and named Harmony Search (HS) [11]. Harmony search algorithm had been very successful in a wide variety of optimization problems, presenting several advantages with respect to traditional optimization techniques [12].

In this paper, we propose algorithms based on HS to cluster web documents. The first algorithm, namely Harmony Search Clustering (HSCLUST), which is good at finding promising areas of the search space but not as good as *K*-means at fine-tuning within those areas. Second algorithm, Harmony *K*-means Clustering (HKCLUST) combines power of the HSCLUST with the speed of a *K*-means. To demonstrate the effectiveness and speed of HSCLUST and

978-1-4244-1643-1/08/$25.00 ©2008 IEEE