NAME OF THE SCHOLAR	**:**	**KISHWAR SADAF**
NAME OF THE SUPERVISOR	**:**	**DR. MANSAF ALAM**
DEPARTMENT	**:**	**DEPARTMENT OF COMPUTER SCIENCE,** **FACULTY OF NATURAL SCIENCES**
TITLE	**:**	**INFORMATION RETRIEVAL USING CLUSTERING TECHNIQUES**

Information Retrieval (IR) systems are the tools to access the information available on the web. This information is unstructured, disorganized, dynamic and heterogeneous in nature and enormously large. Moreover, the process of retrieval is highly affected by vague query put up by the average users. Search engines are the best examples of IR systems on the web. Although today's search engines are smarter than earlier IR systems, ambiguous queries are still a major problem. To answer all the possible meaning of an ambiguous query, search engines return too many results which are not necessarily relevant to the user's need. Usually, a user has to traverse several search result pages to get to the desired result. A way of assisting users in finding what they are looking for quickly, is to group the search results by topic. The user does not have to reformulate the query, but can merely click on the topic most accurately describing his or her specific information need. This grouping of result is called Clustering. More specifically, it is a process of grouping similar documents into clusters so that documents of one cluster are different from the documents of other clusters.

The main objective of clustering the search result is not to improve the ranking of documents in the search result but to give a user a prevue of the retrieved search result. Hence, it can be seen as a complimentary process to retrieval rather than as an alternative. Clustering when applied on search engine's result documents, is referred to as Web Search Result Clustering (WSRC). This technique not only helps users to find their desired information but also organizes search results into thematic groups, letting users narrow down their search. There are many search engines, but not mainstream ones, that provide clustered view of search result. Kartoo, Carrot^2, Yippee! etc. are some of the clustering engines available on the web.

To cluster search result, we proposed an approach that is based on Heuristic Search and k-means methods. The advantage of the proposed method is the removal of external specification of k for k-means. Our heuristic obtains the initial center points that are

subsequently used in initializing k-means. We exploit the hyperlink feature of retrieved documents to establish the relatedness between them. Our heuristic is based on the notion that related pages tend to have connections between them. We evaluated our method on three different datasets. All the documents are taken from Google search engine’s result. Experimental results show that our method achieved superior precision, recall and f-score values as compared to simple k-means and spherical k-means.

We proposed another WSRC technique, called WSRC-CSCC (**W**eb **S**earch **R**esult **C**lustering**-** **C**uckoo **S**earch and **C**onsensus **C**lustering), which is based on meta-heuristic Cuckoo search and Consensus Clustering. Cuckoo search method is based on the parasitic breeding behaviour shown by cuckoos. A cuckoo lays its eggs into the nests of other birds. To do so, it searches for the best nest and impersonates its eggs like the eggs of the host bird. If the host bird finds alien eggs in the nest, it either destroys the cuckoo’s eggs or abandons the nest. We adopted this behaviour of cuckoos for clustering search result. Instead of calculating the quality of a nest, we apply consensus clustering on all of the nests. To validate our method, we applied our approach on five datasets generated using DMOZ web directory. The experimental results show good result as compared to another method based on cuckoo search.

A cluster is not effective if its label cannot define its content. A user will never select it even if it contains the relevant documents. To address the problem of labeling, we proposed a method to create labels for clusters of search result documents. In this method, we applied our heuristic which finds the linked documents of a cluster. The titles of these linked web documents are then searched for frequent itemsets using famous Apriori algorithm. To assess our method, we applied it on the “jaguar” dataset. The method produces appropriate cluster labels.