Journal of Theoretical and Applied Information Technology
15th June 2021. Vol.99. No 11
© 2021 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID APPROACH FOR WEB SEARCH RESULT CLUSTERING BASED ON GENETIC ALGORITHM WITH K-MEANS

[1*]BOURAIR AL-ATTAR, [1]AHMED J. ALLAMI, [1]ALI THOULFIKAR A. IMEER, , [1]YUSOR FADHIL ALASADI, [2]NORITA MD. [2]NORWAWI, [1]HAWRAA M. KADHIM

[1]University of Al-Ameed, Karbala, Iraq

[2]Faculty of Science and Technology, University Sains Islam Malaysia

Corresponding author: *

ABSTRACT

Nowadays, search engines tend to use the latest technologies in enhancing the personalization of web searches, which leads to a better understanding of user needs. One of these technologies is web search results clustering which returns meaningful labeled clusters from a set of Web snippets retrieved from any Web search engine for a given user's query. Search result clustering aims to improve searching for information from the potentially huge amount of search results. These search results consist of URLs, titles, and snippets (descriptions or summaries) of web pages. Dealing with search results is considered as treating large-scale data, which indeed has a significant impact on effectiveness and efficiency. However, unlike traditional text mining, queries and snippets tend to be shorter which leads to more ambiguity. K-means tend to converge to local optima and depend on the initial value of cluster centers. In the past, many heuristic algorithms have been introduced to overcome this local optima problem. Nevertheless, these algorithms suffer several shortcomings. In this paper, we present an efficient hybrid web search results clustering algorithm referred to as G-K-M, whereby, we combine K-means with a modified genetic algorithm. The AOL standard dataset is used for evaluating web data log clustering. ODP-239 and MORESQUE are used as the main gold standards for the evaluation of search results clustering algorithms. The experimental results show that the proposed approach demonstrates its significant advantages over traditional clustering. Besides, results show that proposed methods are promising approaches that can make search results more understandable to the users and yield promising benefits in terms of personalization.

Keywords: *Personalized Search Engine, Search Results Clustering, G-K-M Clustering Algorithm, K-Means*

1. INTRODUCTION

With the huge growth of information on the Internet, it has become very difficult for users to find relevant web pages. In response to the user's query, currently available search engines return a ranked list of web pages along with their snippets. If the query is general, it is extremely difficult to identify the specific web page which the user is interested in. Search results clustering (SRC) is a challenging algorithmic problem that requires clustering together the results returned by one or more search engines in topically coherent clusters [1]. SRC systems return meaningful clusters from a set of Web snippets retrieved from any Web search engine for a given user's query. The web search results tend to be large-scale repository which can significantly influence the effectiveness and efficiency of the search system. The huge and continually increasing amount of information on the web creates many challenges for the researchers of web search [1-6]. Web search results clustering not only attracts commercial interest, it is also an active research area, with a large number of published researches discussing specific issues and systems. Web search results clustering is connected with document clustering [2]. With the growth in the number of Web users, the problem of personalization of web search engines has become very critical and popular [7]. It is highly needed to personalized Web Search effectively as it is an open problem in the information retrieval community [8]. Clustering search result for personalization of Web search brings several interesting challenges[3, 4]: first, since fast retrieval is one of the primary concerns in a web search, clustering methods are desired to have a quick response time. Second, limited data is available for clustering i.e. the data consists of a URL, a title, and a small description or snippet of a web page. The limited amount of data makes a clustering task more difficult. Third, in search clustering methods, different sizes and numbers of clusters can result from different user queries. So efficient clustering algorithm is desired to enhance the effectiveness and efficiency of the personalization of Web search [5, 7]. Thus it brings interesting clustering challenges in the personalized search framework. Clustering results should change dynamically to detect the changes in the user's interest and to reflect the personalized ranking of search results. However, traditional clustering algorithms are regarded as "static" since the clustering result cannot reflect the changes in the user's interest and reflect the personalized ranking of search results[8,9]. Besides, the high dimensionality and sparsity of the text feature space and phenomena such as polysemy and synonymy can only be handled in a way that is provided to measure term similarity [10-12]. Traditional clustering algorithms do not consider the semantic relationships among words[10]. The sensitivity to initial values and cluster centers of the traditional clustering algorithms reduces its best [13-15]. These methods are still sensitive to the selection of the initial cluster prototypes [16], and require the number of clusters to be specified in advance. However, in search clustering methods, different sizes and numbers of clusters can result from different user queries. The number of clusters in web search clustering is generally unknown [4]. Therefore, search clustering methods need to have a mechanism that can determine the number of clusters in the data [4]. In order to alleviate the shortcomings of traditional clustering methods. This research used in this article presents a hybrid clustering method G-K-M which combines a novel genetic algorithm with K-means clustering method for personalized web search engine. This paper is split into four main sections: In Section Two, we discussed related works on personalized web search engine documents [17]. Then in Section three, we described how do we implementing out our review. And Section four will be on the experimental findings, and finally, Section five and will be the conclusion of our work.

2. RELATED WORK

Many methods and approaches have been proposed in terms of web search personalization. For instance, In the area of Personalization of web search, [11] uses a combination of heuristics and k-means technique using cosine similarity. Their heuristic approach detects the initial value of k for creating initial centroids. This eliminates the problem of external specification of the value k, which may lead to unwanted results if wrongly specified. The centroids created in this way are more specific and meaningful in the context of web search results. Another advantage of the proposed method is the removal of the objective means function of k-means which makes clusters' sizes the same. The result of the proposed approach consists of different clusters of documents having different sizes [18] propose a method of search result clustering based on a heuristic search on the graph induced by the hyperlinks among the documents of search result [19] uses a genetic algorithm to improve the quality of clusters and produce better results. In this paper, the genetic algorithm is used for improving the cluster quality for effective Personalization of web search based on clustered query sessions [20] presents an approach to provide personalized query suggestions based on a genetic algorithm-based clustering technique. This improves retrieval effectiveness and relevancy by expanding the query with additional words. The main objective of this work is to improve the retrieval of information by expanding the user's query based on the user's domain of interest [21]. The Genetic Algorithm(GA) is used for cluster optimization to improve the quality of clusters for effective personalized web search. The processing involved in applying the genetic algorithm for clusters optimization is done offline and has no impact on the performance of online processing of personalized Web Search using these optimal sets of clusters. The latest work investigates the limitations of existing text clustering methods and addresses these limitations by providing five new text clustering methods–Query [22]. Sense Clustering (QSC), Dirichlet Weighted K-means (DWKM), Multi-View Multi-Objective Evolutionary Algorithm (MMOEA), Multi-objective Document Clustering (MDC), and Multi-Objective Multi-View Ensemble Clustering (MOMVEC). These five new clustering methods showed that the use of rich features in text clustering methods could outperform the existing state-of-the-art text clustering methods Multi-Objective Document Clustering (MDC) and Multi-Objective Multi-View Ensemble Clustering (MOMVEC). These five new clustering methods showed that the use of rich features in text clustering methods could outperform the existing state-of-the-art text clustering methods. These new text clustering methods demonstrated that the rich features are very useful in clustering to determine the similarity of documents and can play an important role in deriving high-quality clusters. Their result proves clustering method based on semantic feature weighting provides better quality clusters with meaningful labels as compared to traditional algorithms. With the growth in the number of Web users, the problem of personalization of web search engines has become very critical and popular [7]. It is highly needed to personalized Web Search effectively as it is an open problem in the information retrieval community [8]. Search result clustering is considered to be a special case of document clustering because of its following unique challenges [23,24]: first, since fast retrieval is one of the primary concerns in a web search, search result clustering methods are desired to have a quick response time. Second, unlike document clustering, limited data is available for search result clustering methods. Generally, the data consists of a URL, a title, and a small description or snippet of a web page. The limited amount of data makes a clustering task more difficult. Third, in search clustering methods, different sizes and numbers of clusters can result from different user queries. As stated by [25], future pointers for enhancing search results clustering problem is to introduce new approaches for estimating the number of clusters instead of a set number of clusters to a fixed size. Therefore, search clustering methods need to have a mechanism such meta-heuristic search algorithm which can determine the number of clusters in the data [26].

3. PROPOSED METHOD

This study framework includes all the stages of implementing the proposed method. Such stages determine the process of collecting the dataset, preprocessing tasks, feature extraction, and the clustering method. The comparative analysis and evaluation of clustering are carried out by using precision, recall, and overall F-measure. Figure 1 shows the framework of data logs and web search results clustering.

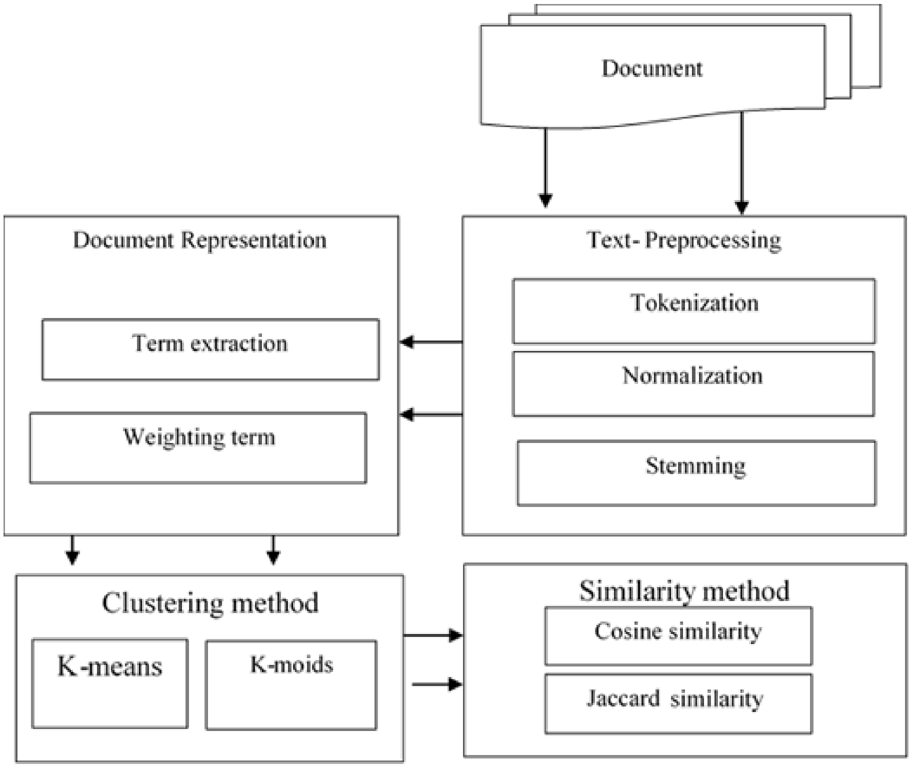

Figure 1: Research Methodology.

3.1 Preprocessing

The preprocessing phase aims to clean the input documents from all characters and terms that can affect the quality of cluster descriptions. To become a full-featured clustering algorithm, the process of finding cluster labels and contents must be preceded by some preprocessing of the input collection. This stage should encompass text filtering, document language recognition, stemming, and stop words identification. steps. There are three main steps in preprocessing phase; text filtering to remove HTML tags, entities, and non-letter characters, language identification, and finally stemming and stop word removal.

3.2 Text representation

Directly applying most learning algorithms to text information, in a direct way without representing, it has been proved to be impossible, due to the complex nature of the text information. Therefore, before applying the text using a machine learning method, it is essential to convert the content of a textual document to a compact representation is necessary. They are Document representation is efficiently used as a language-independent method, since they are it is independent of the meaning of the language and performs well in case of noisy text. Term Frequency × Inverse Document Frequency (TF×IDF) weighting is also recognized as a simple method for term weighting.

$$W_i = tf_i . \log\left(\frac{N}{n_i}\right) \qquad 1$$

Term Frequency × Inverse Document Frequency (TF×IDF) weighting is seen as the most popular method used for term weighting since it considers this property. By using this approach, assigning the weight of term *i* in document d to the number of times the term appears in the document is proportional, and it is in inverse proportion to the number of documents in the corpus, in which the term appears.

3.3 New hybrid similarity measure

Document clustering is the process in which similar documents are grouped to form a coherent cluster. The accuracy of clustering depends on a precise definition of the closeness between a pair of objects, in terms of either the pairwise similarity or distance. define a new similarity measure that combines several measures including cosine similarity measure, WordNet-based similarity, and corpus-based similarity.

Cosine

Cosine similarity is one of the most well-known similarity measures which is applied to text documents such as in numerous information retrieval applications and clustering. In measuring the given two documents. $\overrightarrow{\mathbf{t_a}}$ and $\overrightarrow{\mathbf{t_b}}$, their cosine similarity is:

$$SIM_C(\overrightarrow{\mathrm{t_a}}, \overrightarrow{\mathrm{t_b}}). \frac{\overrightarrow{\mathrm{t_a}} . \overrightarrow{\mathrm{t_b}}}{|\overrightarrow{\mathrm{t_a}}| * |\overrightarrow{\mathrm{t_b}}|} \qquad 2$$

WordNet-based similarity

First of all, documents p and q are analyzed to extract all the included WordNet synsets [27]. For each WordNet synset, we keep synsets and put them into the set of synsets associated with the sentence, Cp and Cq, respectively. Given Cp and Cq as the sets of concepts contained in sentences p and q, respectively, with |Cp| ≥ |Cq|, the similarity between p and q is calculated as:

$$sim_{wn}(p,q) = \frac{\sum_{c1 \in Cp} \max_{c2 \in Cq} s(c1,c2)}{|Cp|} \qquad 3$$

Where s(c1, c2) is calculated using the ProxiGenea3 measure [28] is defined as:

$$s(c1,c2) = \frac{1}{1 + d(c1) + d(c2) - 2 \cdot d(c0)} \qquad 4$$

where c0 is the most specific concept that is present both in the synset path of c1 and c2. The function returning the depth of a concept is noted with d.

Corpus-based similarity

The similarity between two documents p and q is determined as:

$$sim_c(p,q) = \frac{1}{2}\left(\left(\frac{\sum_{w \in p} \max_{w2 \in q} ws(w,w2) \cdot idf(w)}{\sum_{w \in p} idf(w)}\right) + \left(\frac{\sum_{w \in q} \max_{w1 \in p} ws(w,w1) \cdot idf(w)}{\sum_{w1 \in q} idf(w)}\right)\right) \qquad 5$$

where idf(w) is calculated as the inverse document frequency of word w, the semantic similarity between words is calculated as:

$$ws(wi\, , wj\,) = \max_{ci \in wi, cj \in wi} s_{cb}(c1, c2) \qquad 6$$

$$s_{cb}(c1,c2) = \frac{1}{IC(c1) + IC(c2) - 2 \times IC(LCS(c1,c2))} \qquad 7$$

Hybrid similarity measure

We believe that clustering documents using the clustering approach require a special similarity measure that not only considers the syntactic information (words frequencies) of the documents but also takes the semantic information into account. Consequently, this research design a new hybrid similarity method, we take the advantages of both traditional cosine similarity measure and semantic similarity measure and corpus-based similarity measure in order to improve the quality of similarity result. The similarity between two documents p and q will be computed by applying the following Equation.

$$\mathbf{sim_{overall}}(\mathbf{p},\mathbf{q}) = \boldsymbol{\lambda_1} \times \boldsymbol{sim_{wn}}(\boldsymbol{p},\boldsymbol{q}) + \boldsymbol{\lambda_2} \times \boldsymbol{sim_c}(\boldsymbol{p},\boldsymbol{q}) + \boldsymbol{\lambda_3} \times \boldsymbol{sim_{cosine}}(\boldsymbol{p},\boldsymbol{q}) \qquad \mathbf{8}$$

Where $0 < \lambda < 1$. Moreover, $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

3.4 G-K-M clustering technique

G-K-M algorithm used to find optimal clusters' seeds & their number according to the following. Steps

a). Population initialization (selects n chromosomes randomly)

1). Set a radius value r randomly that can produce several clusters where $r \leq 0.20$, the right r value can vary from data set to the data set and is generally unknown to a data miner, the use of different r values to create different chromosomes for a GA can be useful.

2). selects n chromosomes randomly to select a chromosome we calculate the density of each document d of the data set as follows:

$$Density(di) = |\{dj: dist(di,dj) \leq r\,; \forall j\}. \qquad 9$$

The document di having the highest density (i.e. $Density(\mathrm{d}i) > Density(\mathrm{d}j\,); \forall j$) is then chosen as the first seed Sj1, and all documents $(\{dj: dist(di,dj) \leq r\,; \forall j\})$ within the r distance of di are removed from the data set.

3). continue the selection of the subsequent seeds as long as we get a record with a density greater than a user-defined threshold T. Therefore, for an r value we get several seeds to form a chromosome Cdj.

4). repeat step 1 to 3 to form n chromosome (each chromosome contain the different size of seeds) The number of seeds of a chromosome is randomly chosen between $[2,\sqrt{\mathrm{n}}]$, where n is the number of documents.

b). Selection operation:

Sort the n chromosomes in the descending order of their fitness values. We then choose the n/2 number of best chromosomes from the initial population of n chromosomes using the fitness function

c). Fitness computation:

first identifies the seed Sj of a cluster Cj; ∀j and calculates the distance between two seeds Si and Sj to estimate the separation between the clusters, instead of calculating the average distance between all pairs of documents d_a∈Ci an d_b∈Cj. We also calculate the distance between the documents of a cluster and its seed dist(da,Si); ∀d_a∈Ci in order to calculate the compactness of a cluster Ci.

$$\boldsymbol{comp_j} = \frac{\sum_{d_a \in Cj} \mathbf{dist(da,Sj),}}{|\boldsymbol{cj}|} \qquad \mathbf{10}$$

$$\boldsymbol{sep_j} = \min_{\forall i \neq j} \mathbf{d(Si,Sj)} \qquad \mathbf{11}$$

$$\boldsymbol{Fitness_j} = \sum_{\forall j}(\boldsymbol{comp_j} - \boldsymbol{sep_j}) \qquad \mathbf{12}$$

d). Crossover operation:

We first sort the |r| chromosomes in descending order according to their fitness values. All chromosomes participate in the crossover operation pair by pair since for a crossover operation we need a pair of chromosomes. Then apply the twin removal operation on the new population made of the offspring chromosomes.

e). Elitism operation:

Elitism keeps track of the best chromosome throughout the generations and also keeps improving the quality of the population in each generation. If the fitness of the worst chromosome (i.e. the chromosome having the worst fitness among all chromosomes of the new generation) is less than the fitness of the best chromosome Cdb then the worst chromosome is replaced by Cdb.

f). Mutation operation:

The basic idea of the mutation operation is to randomly change some of the chromosomes in order to explore different solutions. While adding random changes to the chromosomes we use a probabilistic approach where a chromosome with a low fitness has a high probability of getting a random change and vice versa.

g). K-Means:

We use the seeds of the best chromosome Cdb as the initial seeds of the K-Means clustering algorithm.

4. EVALUATION METRICS

In this work, we evaluate the clustering method for clustering search results. Different gold standards have been used for the evaluation of search result clustering algorithms among which the most cited are: ODP-239, and MORESQUE. MORESQUE Dataset: MORESQUE (MORE Sense-tagged Queries) is a dataset designed for the evaluation of subtopic information retrieval. The dataset consists of 114 topics (i.e., queries), each with a set of subtopics and a list of 100 top-ranking documents. MORESQUE was developed as a complement to AMBIENT following the guidelines provided by its authors. The aim is to study the behavior of Web search algorithms on queries of different lengths, ranging from 1 to 4 words. MORESQUE provides dozens of queries of length 2, 3, and 4, together with the 100 top results from Yahoo! for each query annotated as in the AMBIENT dataset (overall, we tagged 11,400 snippets). We decided to carry on using Yahoo! mainly for homogeneity reasons. ODP-239 Dataset is generated from Open Directory Project with a total of 23900 documents i.e. 100 web documents for each of 239 ambiguous queries. This dataset has more ambiguous queries than AMBIENT and MORESQUE. Clusters of this dataset are very hard to distinguish as they often have similar documents, which makes this dataset more complex compared to AMBIENT and MORESQUE.

Now, in order to evaluate clustering methods, the common information retrieval metrics precision, recall, and f-measure will be used. Precision aims to evaluate the cluster based on the number of correct retrieved candidates out of the total number of retrieved candidates. Hence, we can calculate precision and recall. Precision aims to evaluate the cluster based on the number of correct retrieved candidates out of the total number of retrieved candidates. It can be calculated as follows:

$$Precision\ (cluster\ i) = \frac{\#\ of\ correct\ instance\ in\ cluster\ i}{total\ \#\ of\ cluster\ i\ instances} \qquad 13$$

Whereas, recall aims to evaluate the cluster based on the number of correct retrieved candidates out of the total number of correct instances in the dataset. It can be computed as follows:

$$recall\ (cluster\ i) = \frac{\#\ of\ correct\ instance\ cluster\ i}{total\ \#\ of\ correct\ instances\ in\ the\ dataset} \qquad 14$$

Now it can be possible to calculate the f-measure as follows:

$$f - mesaure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
15

5. EXPERIMENTAL RESULTS

5.1 Evaluation of enhanced K-means clustering on clustering web data log

This phase aims to evaluate the enhanced k-Means clustering method with hybrid similarity measure in terms of the ability to identify categories of data log's queries. This has been made with multiple numbers of clusters which are 3, 4 and 5 clusters. Therefore, the results of precision, recall and f-measure will be stated with the three-cluster size which are 3, 4 and 5 clusters. Table 7.1 shows the results of this phase.

Table 1: Evaluation results of enhanced K-means model over AQL datasets

# of cluster	Precision	Recall	F-measure
3	88.44	90.91	89.66
4	94.45	90.59	92.48
5	89.67	91.76	90.7

As shown in Table 1, the precision, recall, and f-measure have been calculated for each cluster size. The greatest values of precision, recall, and f-measure have been obtained when the number of cluster k = 4 respectively.

To compare between the performance of k-means when it uses traditional similarity measures Cosine, Dice, and Jaccard and when it uses the new hybrid similarity measure in terms of the performance. In order to establish the comparison, Table 2 shows the best performance for all traditional similarity measures and new similarity measures.

Table 2: Comparison between the similarity measures

Similarity Measure	Precision	Recall	F-measure
Cosine	94%	87%	90%
Dice	93%	85%	89%
Jaccard	87%	85%	86%
New hybrid Similarity Measure	94.4	90.5	92.5

As shown in Table 2, k-means has achieved the best performance with the new hybrid similarity measure by achieving 94%, 90%, and 92.5% of precision, recall, and f-measure respectively.

5.2 Evaluation of Enhanced K-medoids clustering on clustering web data log

This phase aims to evaluate the enhanced k-Medoids clustering method with hybrid similarity measure in terms of the ability to identify categories of data log's queries. Basically, this has been made with multiple numbers of clusters which are 3, 4, and 5 clusters. Therefore, the results of precision, recall, and f-measure will be stated with the three-cluster size which are 3, 4, and 5 clusters. Table 3 shows the results of this phase.

Table 3: Evaluation results of enhanced K-medoids model over AQL datasets

# of cluster	Precision	Recall	F-measure
3	88.86	88	88.43
4	86.19	94.29	90.06
5	83.96	91.1	87.38

As shown in Table 3, the precision, recall, and f-measure have been calculated for each cluster size. In fact, the greatest values of precision, recall, and f-measure have been obtained when the number of cluster k = 4 respectively.
To compare between the performance of k-medoids when it uses traditional similarity measures Cosine, Dice, and Jaccard and when it uses the new hybrid similarity measure in terms of the performance. In order to establish the comparison, Table 4 shows the best performance for all traditional similarity measures and new similarity measures.

Table 4: Comparison between the similarity measures

Similarity Measure	Precision	Recall	F-measure
Cosine	87%	83%	84%
Dice	89%	84%	86%
Jaccard	89%	85%	87%
New hybrid Similarity Measure	86.19	94.29	90.06

As shown in Table 4, k-medoids has achieved the best performance with the new hybrid similarity measure by achieving 86%, 94%, and 90 % of precision, recall, and f-measure respectively.

5.3 Evaluation of enhanced K-means and K-medoids clustering on clustering web search results

This subsection aims to evaluate the enhanced k-Means and k-medoids clustering methods which use a new hybrid semantic similarity measure in terms of the ability to identify categories of web search results. Search result clustering differs from classical text clustering as the partitioning shape, more precisely the distribution of the Web snippets into clusters shows evidence of some particularity. Indeed, it is well-known that subtopics on the Web are not equally distributed [29]. In this evaluation, after we tokenize and remove the stop word of the web snippet, we executed k-medoids clustering models over ODP-239 and MORESQUE datasets. Experiments are performed on two datasets. The values of precision, recall, and f-measure were computed. The results of the two algorithms over the dataset of MORESQUE and ODP-239 are depicted in Table 5. The results show our new methods k-Means and k-medoids clustering methods which use a new hybrid semantic similarity measure are found to have a statistically significant improvement over their results with traditional similarity methods on ODP-239 and MORESQUE datasets. Overall, enhanced k-medoids performed well and generally outperformed the enhanced k-means method on both ODP-239 and MORESQUE datasets.

Table 5: Evaluation results of enhanced k-Means and k-medoids methods over MORESQUE and ODP239 datasets

Model	MORESQUE			ODP-239		
	P	R	F	P	R	F
K_means method (hybrid semantic similarity measure)	88.48	89.25	88.86	85.6	87.99	86.78
K_medoids (hybrid semantic similarity measure)	87.73	92.23	89.92	87.39	90.59	88.96

5.4 Evaluation of the proposed Gen-K Clustering Algorithms

As discussed earlier, we propose Gen-k clustering Algorithms that are capable of automatically finding the right number of clusters and identifying the right seeds through a novel initial population selection approach. The advantages of the proposed algorithms are that they can find the actual number of clusters present in the search result. The main reason is that in search clustering methods, different sizes and numbers of clusters can result from different user queries.

Traditional centroid Based clustering methods i.e. k-Means and k- Medoids avoid having to determine the number of clusters by predefining the number of clusters. These centroid Based clustering methods are still sensitive to the selection of the initial cluster prototypes and require the number of clusters to be specified in advance. However, in search clustering methods, different sizes and numbers of clusters can result from different user queries. The number of clusters in web search clustering is generally unknown. As stated by [21]. future pointers for enhancing search results clustering problem is to introduce new approaches for estimating the number of clusters instead of a set number of clusters to a fixed size. Therefore, it is important for search clustering methods to have a mechanism that can determine the number of clusters in the data. adapt the K-means algorithm to a third-order similarity measure and propose a stopping criterion to automatically determine the “optimal” number of clusters. Experiments are run over two gold standard data sets, ODP-239 *and MORESQUE* [30], and show improved results overall state-of-the-art text-based SRC techniques so far. This section aims to evaluate the Gen-k clustering algorithms in terms of clustering web data log and web search results clustering. The following sub-sections illustrate each experiment based on the mentioned parameters.

5.5 Evaluation of Gen-k clustering algorithms on clustering web data log

In this section, the genetic algorithm is combined with both k-Means and k-medoids clustering to improve the cluster quality for effective personalization of web search based on clustered query sessions.

This section evaluates the Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) in terms of the ability to identify categories of data log’s queries. As described earlier, the Gen_k algorithm automatically determines the “optimal” number of clusters (K). we run Gen_k five times each since it can give a different number of clusters in different runs. The optimal number of clusters (K) fed into K-Means or k-medoids to produce clustering results Basically, the experiments are made with a different number of clusters which are 3, 4, and 5 clusters. Therefore, the results of precision, recall, and f-measure will be stated with the three cluster sizes which are 3, 4 and 5 clusters. Table 7, 6 shows the estimated number of clusters and shows present the precision, recall, and f-measure for both the methods. The result of our proposed method is remarkable as it exactly found the actual number of clusters present in the datasets.

Table 6: Evaluation results of the Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) over AQL datasets

	Gen-k-Means (K = 18)			**Gen-k-medoids (K = 18)**		
# of cluster	**P**	**R**	**F**	**P**	**R**	**F**
3	89.42	88.67	89.04	89.34	90.3	89.82
4	86.82	89.79	88.28	87.43	93.4	90.32
5	84.41	91.55	87.84	85.29	92.43	88.72

As shown in Table 6, the Genetic algorithm has identified the near-optimal number of the cluster as K = 18 for the AOL dataset using both k-means and k-medoids.

5.6 Evaluation of Gen-k clustering algorithms on clustering web search results

This subsection aims to evaluate the Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) in terms of the ability to identify categories of web search results.

Search result clustering differs from classical text clustering as the partitioning shape, more precisely the distribution of the Web snippets into clusters shows evidence of some particularity. Indeed, it is well-known that subtopics on the Web are not equally distributed.

In this evaluation, after we tokenize and remove stop words of web snippets, we executed Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) over ODP-239 and MORESQUE datasets. Experiments are performed on two datasets. The values of precision, recall, and f-measure were computed. The results of the two algorithms over the dataset of MORESQUE and ODP-239 are depicted in Table 7.6 and Table 7.

Compare Gen-k clustering algorithms and traditional clustering algorithms in the previous paragraph shows that new Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) methods which combine a novel genetic method and enhanced k-Means and k-medoids clustering are found to have a statistically significant improvement over traditional clustering methods results on ODP-239 and MORESQUE datasets. Overall, enhanced k-medoids performed well and generally outperformed the enhanced k-means method on both ODP-239 and MORESQUE datasets.

Table 7: Evaluation results of Gen-k-Means and Gen-k-medoids over MORESQUE and ODP239 datasets

Methods	**MORESQUE (K = 15)**			**ODP-239 (K = 12)**		
	P	**R**	**F**	**P**	**R**	**F**
Gen-K_means method (hybrid semantic similarity measure)	88.48	89.25	88.86	85.6	87.99	86.78
Gen-K_means method (hybrid semantic similarity measure)	87.73	92.23	89.92	87.39	90.59	88.96

As shown in Table 7, GA has identified the near-optimal number of clusters as K = 15 for MORESQUE dataset and K = 12 for ODP-239 dataset.

6. CONCLUSION

This paper has provided the evaluation process of the enhanced and optimized web data clustering models for web search personalization in terms of mining data log and web search clustering. First, this chapter evaluates the enhanced k-Means and k-Medoids clustering method with hybrid similarity measure that combines several measures including cosine similarity measure, semantic similarity, and corpus-based similarity measures. Finally, this chapter evaluates Gen-k clustering algorithms (Gen-k-Means and Gen-k-medoids) in terms in terms of mining data log and web search clustering. Results show that enhanced and optimized clustering methods significantly outperformed traditional methods in both mining data log and web search clustering tasks.

REFERENCES

[1] M. I. Jalgaon, "An efficient and novel approach for web search personalization using web usage mining," *Journal of Theoretical and Applied Information Technology,* vol. 73, no. 2, 2015.

[2] M. Sah and V. Wade, "Evaluation of Personalized Concept-Based Search and Ranked Lists over Linked Open Data," *in UMAP Workshops*, 2014., Malaysia

[3] M. Sah and V. Wade, "Personalized concept-based search and exploration on the web of data using results categorization" in *Extended Semantic Web Conference*, May 2013, pp. 532-547: Springer. Ireland

[4] A. Kumar and M. Ashraf, "Personalized web search engine using dynamic user profile and clustering techniques," in *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, 1-13 March 2015, pp. 2105-2108: IEEE. India

[5] K. W. T. Leung, D. L. Lee, and W. C. Lee, "Pmse: A personalized mobile search engine," *IEEE transactions on knowledge and data engineering,* vol. 25, no. 4, April 2013, pp. 820-834, doi: 10.1109/TKDE.2012.23.

[6] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck, "Probabilistic models for personalizing web search," WSDM '12: Proceedings of the fifth ACM international conference on Web search and data mining February 2012, pp. 433-442: ACM, doi.org/10.1145/2124295.2124348.

[7] P. Kumar, R. S. Bapi, and P. R. Krishna, "SeqPAM: a sequence clustering algorithm for Web personalization," *International Journal of Data Warehousing and Mining,* vol. 3, no. 1, 2007, p. 29-53, doi: 10.4018/978-1-59904-645-7.ch002.

[8] A. B. Rajmane, P. M. Patil, and P. J. Kulkarni, "Personalization of Web Search Using Web Page Clustering Technique," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 9, no. 10, January 2015, pp. 2171-2176, doi.org/10.5281/zenodo.1124957.

[9]. Kenneth Wai-Ting Leung, Wee Keong Ng, Dik Lee, "Personalized Concept-Based Clustering of Search Engine Queries", *Transactions on Knowledge and Data Engineering* Vol 20, No.11, 2008, pp 1505-1518, doi: 10.1109/TKDE.2008.84.

[10]. Montserrat Batet, David Sánchez, "Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic

information content" *Artificial Intelligence Review,* Vol 53, No.3 June 2019, pp 2023–2041, doi: 10.1007/s10462-019-09725-4

[11]. Gaël Dias Rumen Moraliyski, "Relieving Polysemy Problem for Synonymy Detection", October Conference: Proceedings of the 14th Portuguese Conference on Artificial Intelligence: *Progress in Artificial Intelligence* 2009, doi:10.1007/978-3-642-04686-5_50,

[12]. Chen, C.s., Wang, T., ZHENG, W. , CHEN, J.S.. "Design and realization of topic search based on transferring of searching engine" *Computer Engineering and Design,* Vol 21, 2008, pp 66.

[13]. Hatamlou, A., Abdullah, S. and Nezamabadi-Pour, H.. "A combined approach for clustering based on K-means and gravitational search algorithms". *Swarm and Evolutionary Computation* Vol6, 2012, pp47-52.

[14].Chen Zhang Shixiong Xia, K-means Clustering Algorithm with Improved Initial Center, "Conference: Knowledge Discovery and Data Mining, WKDD 2009", *Second International Workshop on Research Interest*, DOI: 10.1109/WKDD.2009.210

[15]. M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm", *Expert Systems with Applications* Vol.40No.1, 2013, pp 200–210DOI: 10.1016/j.eswa.2012.07.021

[16]. R.J. Kadhim, Bourair Sadik Mohamad Taqi, B. Shuaib, "Development a prototype of academic performance among university students", 2012, *International Journal of Independent Research and Studies*, 1(1), 39-49, January 2012, 11 Pages.

[17]. Bourair Alattar, Norita Md Norwawi, "A personalized research engine based on correlation clustering method", *Journal of Theoretical and Applied Information Technology* 30th, . Vol.93. No.2, November 2016, pp 345-352.

[18]. Mansaf Alam, Kishwar Sadaf, "Web Search Result Clustering based on Heuristic Search and k-means", *Project: Big Data Analytics,* August 2015

[19]. Luís Filipe da Cruz Nassif Eduardo R Hruschka Eduardo R Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" 2013 *Transactions on Information Forensics and Security* 8(1):46-54 Follow journal DOI: 10.1109/TIFS.2012.2223679

[20]. Indumathi, D., Chitra, A. and Girthana, K.. "Search Query Expansion using Genetic Algorithm–based Clustering". *Smart CR,* Vol 3 No.1, 2013, pp14-23.

[21]. Chawla, S. 2015. "Optimization of Clusters of Web Query Sessions using Genetic Algorithm for Effective Personalized Web Search". *International Journal of Computer Applications* 122(9).

[22]. Mohiuddin Ahmed , Raihan Seraj, Syed Mohammed Shamsul Islam , "The k-means Algorithm: A Comprehensive Survey and Performance" *Evaluation, Electronics* 2020, 9(8), 1295; https://doi.org/10.3390/electronics9081295.

[23]. Wahid, Abdul, "Improving Clustering Methods By Exploiting Richness Of Text Data", URI: http://hdl.handle.net/10063/5336", 2016, *Victoria University of Wellington.*

[24]. Yordan P. Raykov , Alexis Boukouvalas , Fahd Baig, Max A. Little, "What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm" *PLOS ONE* 26, 2016, doi.org/10.1371/journal.pone.0162259.

[25]. Youcef Djenouri, Asma Belhadi, Djamel Djenouri and Jerry Chun-Wei Lin, "Cluster-based information retrieval using pattern mining", *Applied Intelligence* 2020, doi.org/10.1007/s10489-020-01922-x, 1016.

[26]. Laith Abualigah , Amir H. Gandomi, Mohamed Abd Elaziz , Husam Al Hamad , Mahmoud Omari, Mohammad Alshinwan, and Ahmad M. Khasawneh, "Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering", *Electronics,* 2021, 10, 101. doi.org/10.3390/10,101, 1-29.

[27]. Davide Buscaldi, Joseph Le Roux, Jorge J. Garc´ıa Flores, Adrian Popescu, "LIPN-CORE: Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features", Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 1: Proceedings of the

Main Conferenceand the Shared Task, May 2013, pages 162–168.

[28]. Davide Buscaldi, Jorge J. Garc´ıa Flores, Ivan V. Meza and Isaac Rodr´ıguez "SOPA: Random Forests Regression for the Semantic Textual Similarity task", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 132–137,

[29]. Antonio Di Marco, Roberto Navigli, "Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction", *Computational Linguistics*, 2012, 39(3), 709-754. doi:10.1162/COLI a 00148.

[30[. Sudipta Acharya, Sriparna Saha, Jose Moreno, Gael Dias. "Multi-Objective Search Results, Clustering".25th International Conference on Computational Linguistics (COLING 2014), Aug, 2014, Dublin, Ireland. pp.99 - 108, 2014. <hal-01077207>